NASA Conference Publication 3189, Volume 2

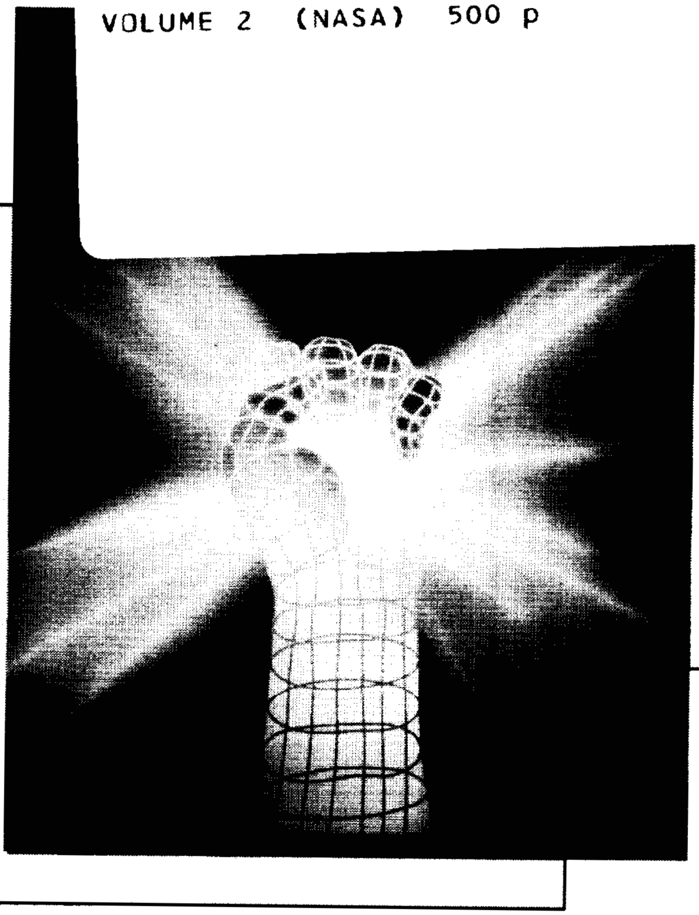
The Third National Technology Transfer Conference & Exposition

December 1-3, 1992 • Baltimore, MD

(NASA-CP-3189-Vol-2) TECHNOLOGY 2002: THE THIRD NATIONAL TECHNOLOGY TRANSFER CONFERENCE AND EXPOSITION, VOLUME 2 (NASA) 500 p

N93-22149
--THRU--
N93-22203
Unclas

H1/99 0150644



Conference Proceedings

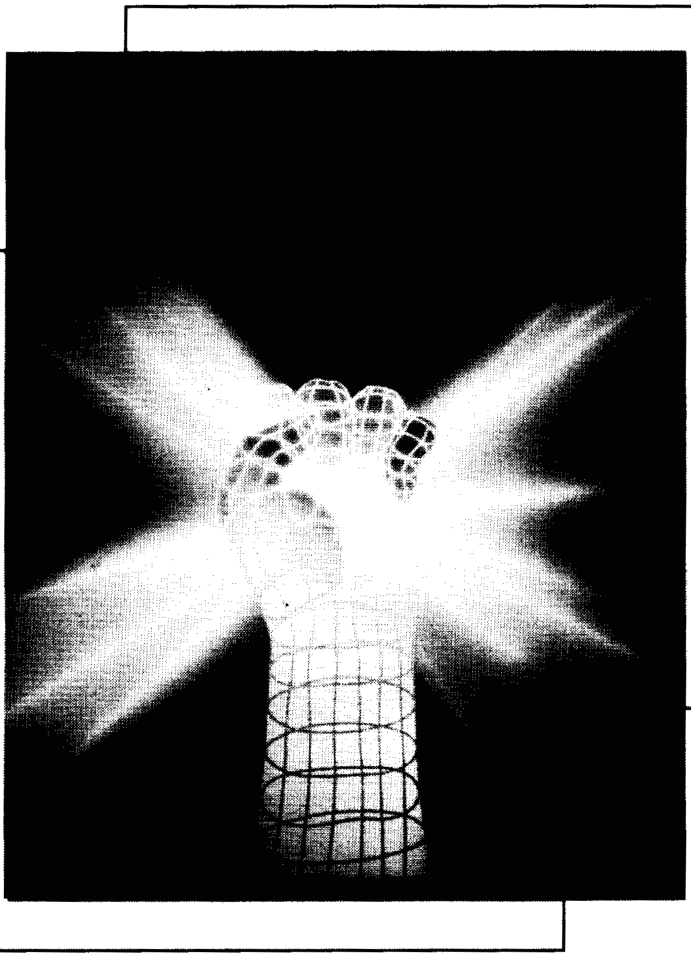
Sponsored by NASA, the Technology Utilization Foundation, and NASA Tech Briefs Magazine

NASA
STI PROGRAM
SCIENTIFIC &
TECHNICAL
INFORMATION

NASA Conference Publication 3189, Volume 2

The Third National Technology Transfer Conference & Exposition

December 1-3, 1992 • Baltimore, MD



Conference Proceedings

Sponsored by NASA, the Technology Utilization Foundation, and NASA Tech Briefs Magazine

NASA
STI PROGRAM
SCIENTIFIC &
TECHNICAL
INFORMATION

TABLE OF CONTENTS

Symposia:

Information and Communications Part 3: Computer Graphics and Display Technologies

Data Management, Storage, and Processing

Manufacturing Technology Part 3: Robotics

Biotechnology and Life Sciences Part 3

Energy and Environment Part 4: Environmental Technologies

Information and Communications Part 4: Computer Software

Information and Communications Part 5: Computer Simulation, Video, and Imaging Technology

Manufacturing Technology Part 4: Intelligent Tools

Microelectronics/Optoelectronics Part 2

Advanced Materials Part 4

Artificial Intelligence Part 3

Biotechnology and Life Science Part 4: Computers in Medicine

Data Management, Storage, and Processing Part 2

Manufacturing Technology Part 5

Sensors and Signal Processing

omit

INFORMATION AND COMMUNICATIONS PART 3: COMPUTER GRAPHICS AND DISPLAY TECHNOLOGIES

TRANSPORTABLE APPLICATIONS ENVIRONMENT (TAE) PLUS

a NASA tool for Building and Managing Graphical User Interfaces

Martha R. Szczur
NASA/Goddard Space Flight Center
Greenbelt, MD 20771 USA
mszczur@postman.gsfc.nasa.gov
301 286-8609

ABSTRACT

The Transportable Applications Environment (TAE) Plus, developed at NASA's Goddard Space Flight Center, is an advanced portable user interface development environment which simplifies the process of creating and managing complex application graphical user interfaces (GUIs). TAE Plus supports the rapid prototyping of GUIs and allows applications to be ported easily between different platforms. This paper will discuss the capabilities of the TAE Plus tool, and how it makes the job of designing and developing GUIs easier for application developers. TAE Plus is being applied to many types of applications, and this paper discusses how it has been used both within and outside NASA.

BACKGROUND

With the emergence of low-cost graphic workstations and the subsequent demands for highly interactive systems, the design and develop of the user interface software has become more complex and difficult. With high resolution workstations, the user interface designer has to be cognizant of multiple window displays and asynchronous events from users and windowing systems, the use of color, graphical interaction objects and icons, and various user selection techniques (e.g., mouse, trackball, tablets).

To make user interfaces easier to create, many different types of tools have been developed. The X Window System™[2] has had a major impact on the user interface software in the UNIX and VMS-based workstation environments, and "X" has become the standard windowing system across these platforms. To make the task of programming the user interface using the X Window System easier, *toolkits* have evolve, which provide higher level services to the programmer along with a set of interaction objects (e.g., menus, buttons, scroll bars). Using the toolkit services, programmers can configure the objects to their specification. The most common toolkit in the UNIX and VMS workstation environment is the Open Software Foundation's Motif™ toolkit. Although X and Motif provide programmatic tools to aid the programmer, they are very complex to learn and do not offer any productivity advantages.

The tools that hold the most promise of dramatic productivity gains for user interface developers are What-You-See-Is-What-You-Get (WYSIWYG) user interface design and management tools. These tools allow you to directly layout your user interface, rehearse the graphical user interface (GUI), and even generate the source code, which will manage your application's user interface during operation.

During the evolution of the various GUI technologies, the Data Systems Technology Division at Goddard Space Flight Center built a user interface development environment, called the Transportable Applications Environment (TAE) Plus. This software tool has been built with the objective of providing a stable environment to support our on-going and diverse development efforts. Our two primary objectives were (1) to improve productivity of application user interface development and (2) to provide a buffer from technology changes.

To improve productivity we defined the following goals:

- support WYSIWYG design of the GUI objects
- support evolution from rapid prototype to baseline operational system
- provide reusable software components
- provide less complex set of application services
- support user interface experts, who may not be programmers

To protect our investments in the development of large-scale space applications that have a long "lifespan", we wanted to provide a mechanism that would allow GUI technology changes to be integrated into the systems with minimal impact on the application-specific software. To provide this "change" buffer, we defined the following goals:

- separate the GUI definition from the application
- provide application programs with toolkit-independent runtime services
- support portability of applications across workstations (e.g., UNIX, VMS)

Elements of these goals were addressed in the early 1980's when GSFC recognized that most large-scale space applications, regardless of function, required software to support human-computer interactions and application management. This lead to the design and implementation of the Transportable Applications Executive (now, referred to as TAE Classic), which abstracts a common core of system service routines and user dialog techniques used by all applications [1]. Over the years, TAE Classic matured into a powerful tool for quickly and easily building and managing consistent, portable user interfaces, but only for the standard alphanumeric terminal.

We took advantage of the lessons learned in the TAE Classic development when we decided to support the GUI environment. By utilizing some of the internal data structures and features of the original TAE software, we developed a set of tools which support the building and management of graphical user interfaces. This advanced version of TAE is called TAE Plus (i.e., TAE *plus* graphics support).

WHAT DOES TAE PLUS PROVIDE?

To meet the defined goals, services and tools were developed for creating and managing window-oriented user interfaces. It became apparent, due to the flexibility and complexity of graphical user interfaces, that the design of the user interface should be considered a separate activity from the application program design. The interface designer can then incorporate human factors and graphic art techniques into the user interface design. The application programmer only needs to be concerned about what results are returned by the user interaction and not the look of the user interface.

In support of the user interface designer, an interactive *WorkBench* application was implemented for manipulating interaction objects ranging from simple buttons to complex multi-object panels. As illustrated in Figure 1, after designing the screen display, the WorkBench saves the specification of the user interface in resource files, which can then be accessed by application programmers through a set of runtime services, Window Programming Tools (WPTs). Guided by the information in the resource files, the routines handle all user interactions. The WPTs utilize Open Software Foundation's Motif™ and the standard MIT X Window System™ to communicate with the graphic workstations.[2] As a further aid to the UI developer, the WorkBench provides an option to generate the source code (C, C++ or Ada) which will display and manage the designed user interface during runtime. This gives the programmer a working template into which application-specific code can be added.

INTERACTION OBJECTS AS BUILDING BLOCKS

The basic building blocks for developing an application's GUI are a set of interaction objects. All visually distinct elements of a display that are created and managed using TAE Plus are considered to be interaction objects and they fall into three categories: selection objects, text objects, and data-driven objects. Selection objects are mechanisms by which an application can acquire directives from the end user. They include menu bar with cascading menus, radio buttons, check boxes, scrolling selection list, icon button, option menu, scale (slider) pulldown menus and push buttons. Text objects are used by an application to request text information or to instruct or to notify the user. They include keyin, optimized dynamic text object that is updated dynamically by the application, label, multi-line edit and text displays (e.g., message, status, help). Data-driven objects are vector-drawn graphic objects which are linked to an application data variable; elements of their view change as the data values change. Examples are dials, thermometers, and strip charts. When creating user dialogues, any of these objects can be grouped and arranged within panels (i.e., windows) in the WorkBench. There is also support for a X Workspace into which applications can write directly using X Window services. Refer to Figure 2 for a sample of the TAE Plus interaction objects.

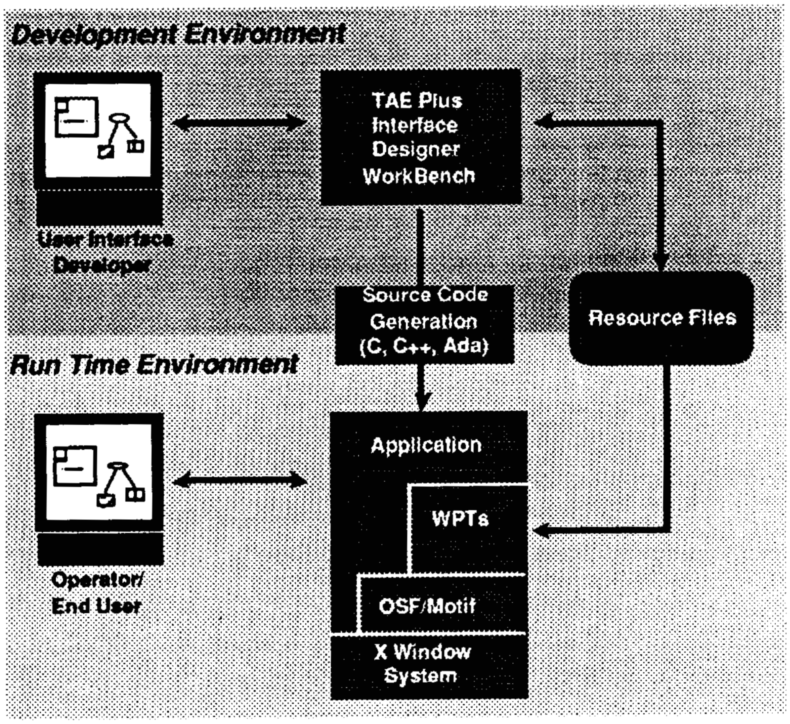
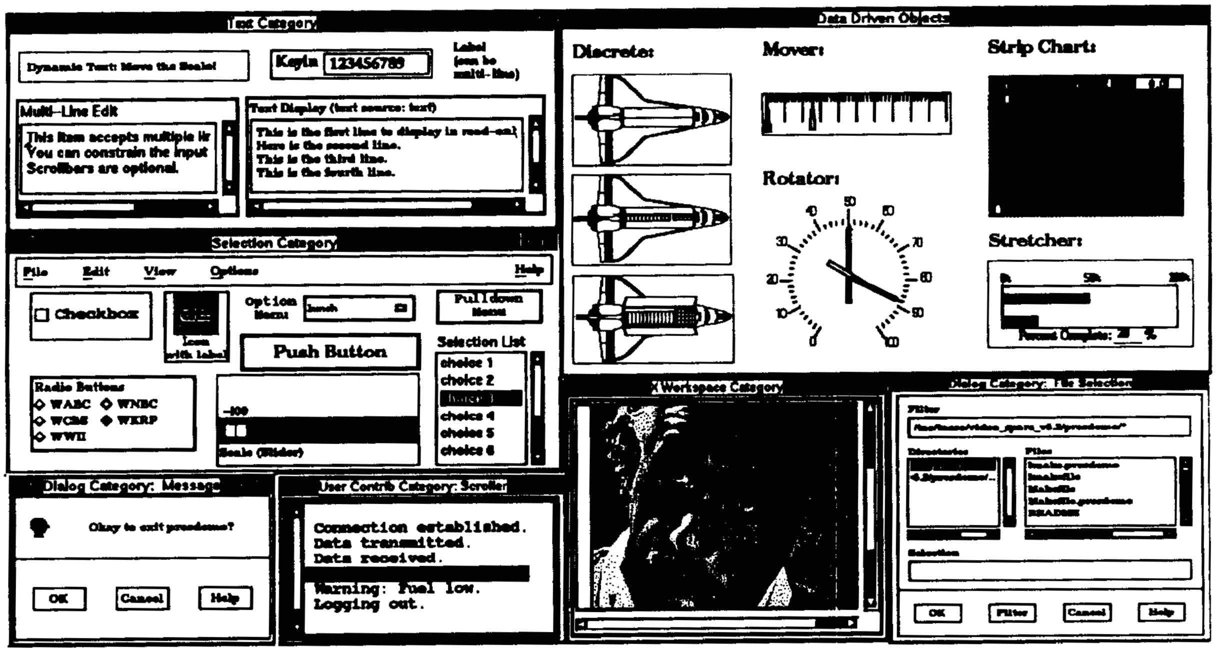


Figure 1: TAE Plus Structure

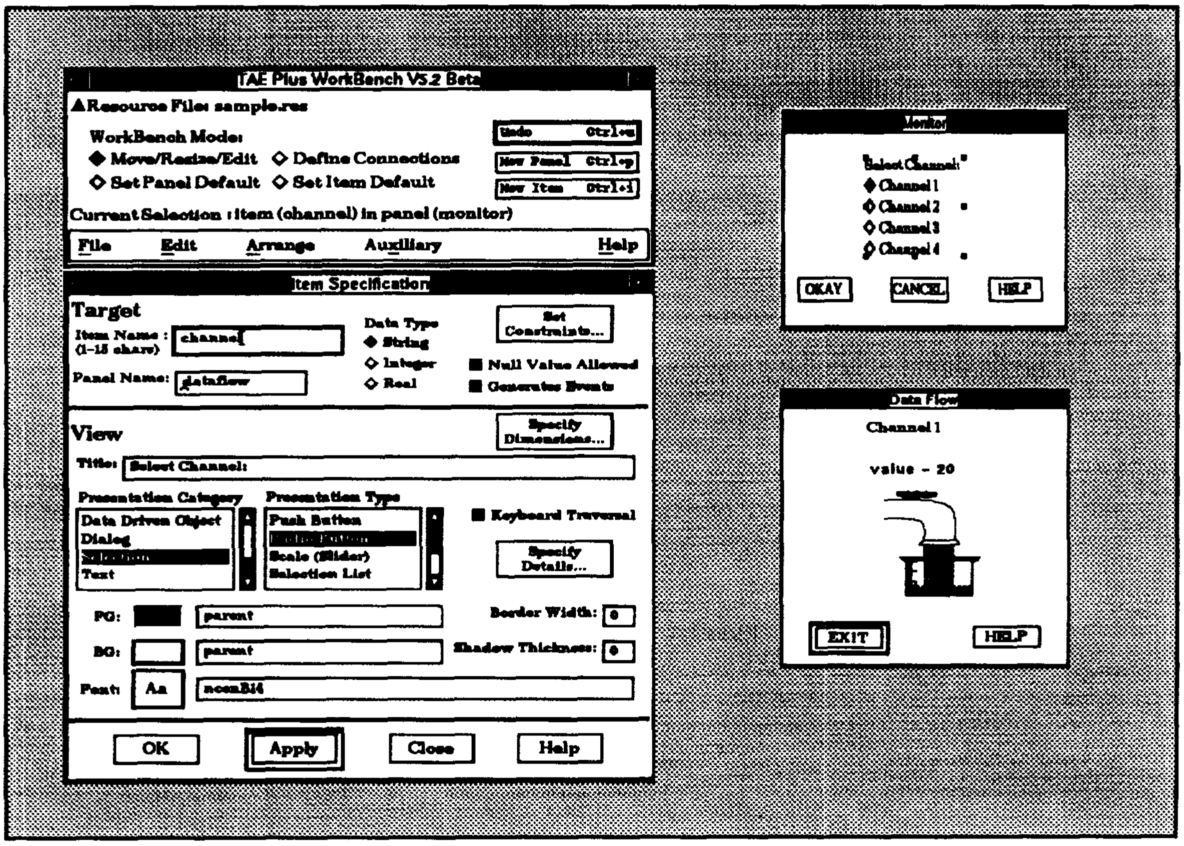


```
Connection established.
Data transmitted.
Data received.

Warning: Fuel low.
Logging out.
```

Figure 2: TAE Plus Interaction Objects

TAE PLUS WORKBENCH

The WorkBench provides an intuitive environment for defining, testing, and communicating the look and feel of an application system. Functionally, the WorkBench allows an application designer to dynamically lay out an application screen, defining its static and dynamic areas. The tool provides the designer with a choice of pre-designed interaction objects and allows for tailoring, combining and rearranging of the objects. To begin the session, the designer needs to create the base panel (i.e., window) into which interaction objects will be specified. The designer specifies presentation information, such as the title, font, color, and optional on-line help for the panel being created. The designer defines both the presentation information and the context information of all interaction items to reside in the panel by using the item specification window (refer to Figure 3). For icon support, the WorkBench has an icon editor, within which an icon can be drawn, edited and saved. As the UI designer moves, resizes, and alters any of the item's attributes, the changes are dynamically reflected on the display screen.


TAE Plus WorkBench V5.2 Beta

▲Resource File: sample.res

WorkBench Mode:
◆ Move/Resize/Edit ◇ Define Connections
◇ Set Panel Default ◇ Set Item Default

Current Selection : item (channel) in panel (monitor)

Item Specification

Target
Item Name : (1-15 chars) channel
Panel Name: dataflow

Data Type
◆ String
◇ Integer
◇ Real

■ Null Value Allowed
■ Generates Events

View
Title: Select Channel:

Presentation Category
Data Driven Object
Dialog
Selection
Text

Presentation Type
Push Button
Radio Button
Scale (Slider)
Selection List

■ Keyboard Traversal

FG: parent
BG: parent
Font: Aa ncenBI4
Border Width: 0
Shadow Thickness: 0

Monitor
Select Channel:
◆ Channel 1
◇ Channel 2
◇ Channel 3
◇ Channel 4

Data Flow
Channel 1
value - 20

The designer also has the option of retrieving palettes of previously created items. The ability to reuse interaction objects saves programming time, facilitates experimenting with different combinations of items in the prototyping process, and contributes to standardization of the application's look and feel. If an application system manager wants to ensure consistency and uniformity across an entire application's UI, all developers could be instructed to use only items from the application's palette of common items.

When creating a data-driven object, the designer goes through a similar process by setting the associated attributes (e.g., color thresholds, maximum, minimum, update delta) in the specification panels. To create the associated graphics drawing, the WorkBench provides a drawing tool within which the static background and dynamic foregrounds of a data-driven object can be drawn, edited, and saved.

Most often an application's UI will be made up of a number of related panels, sequenced in a meaningful fashion. Through the WorkBench, the designer defines the interface connections. These links determine what happens when the user selects a button or a menu entry. The designer attaches events to interaction items and thereby designates what panel appears and/or what action executes when an event is triggered. Events are triggered by user-controlled I/O peripherals (e.g., point and click devices or keyboard input).

Having designed the layout of panels and their attendant items and having threaded the panel and items according to their interaction scenario, the designer is able to preview (i.e., rehearse) the interface's operation from the WorkBench. With this potential to test drive an interface, to make changes, and to test again, iterative design becomes part of the prototyping process. With the rehearsal feature, the designer can evaluate and refine both the functionality and the aesthetics of a proposed interface. After the rehearsal, control is returned to wherever the designer left off in the WorkBench and the designer can either continue with the design process or save the defined UI in a resource file.

As a further aid to the application developer, the WorkBench has a "generate" feature, which produces a fully annotated and operational body of code which will display and manage the entire WorkBench-designed UI. Currently, source code generation of C, C++, Ada and the TAE Command Language (TCL) (an interpreted prototyping language) are supported. Providing this code template helps in establishing uniform programming method and style across large applications or within a family of interrelated software applications.

WINDOW PROGRAMMING TOOLS (WPTS)

The Window Programming Tools (WPTs) are a package of application program callable subroutines used to control an application's user interface. Using these routines, applications can define, display, receive information from, update and/or delete TAE Plus panels and interaction objects. The WPT package utilizes the the MIT X Window System, as its standard windowing system and the Motif toolkit and window manager.

The WPTs provide a buffer between the application program and the Motif toolkit. For instance, to display a WorkBench-designed panel, an application makes a single call to Wpt_NewPanel (using the panel name specified in the WorkBench). This single call translates into a function that can make as many as 50 calls to Motif library routines. For the majority of applications, the WPT services and objects supported by the WorkBench provide the necessary user interface tools and save the programmer from having to learn the complexities of programming directly with Motif and X. This can be a significant advantage, especially when considering the learning curve differential between 40 WPT routines versus over 400 X Toolkit intrinsics and over 200 Xlib services. Refer to Figure 4 for a sample list of the WPTs.

Wpt_AddEvent	Add other sources for input/output/exception
Wpt_BeginWait	Display busy indicator cursor
Wpt_CloseItems	Close Items on a Panel
Wpt_ConvertName	Get the X Id of a named window
Wpt_Endwait	Stop displaying busy indicator cursor
Wpt_Init	Initializes interface to X Window System
Wpt_ItemWindow	Gets the window Id of the window containing a parameter
Wpt_MissingVal	Indicates if any values are missing
Wpt_New Panel	Displays a user interface panel
Wpt_NextEvent	Gets next panel-related event
Wpt_PanelErase	Erases the displayed panel from the screen
Wpt_PanelMessage	Displays message in "Bother Box"
Wpt_PanelReset	Resets object values to initial values
Wpt_PanelTopWindow	Gets panel's parent shell window Id
Wpt_PanelWidgetId	Return the Widget Id of a Wpt Panel Widget
Wpt_PanelWindow	Returns the X Id of a panel
Wpt_ParmReject	Generates a rejection message for a given value
Wpt_ParmUpdate	Updates the displayed values of an object
Wpt_Pending	Check if a WptEvent is pending from X, Parm or file.
Wpt_RemoveEvent	Remove a previously registered event
Wpt_SetTimeOut	Set/Cancel timeout for gathering Wpt events.
Wpt_ViewUpdate	Updates the view of a parameter on a displayed panel

Figure 4: Sample List of Window Programming Tools (WPTs)

IMPLEMENTATION

The TAE Plus architecture is based on a separation of the user interaction management from the application-specific software. The current implementation is a result of having gone through several prototyped and beta versions of a WorkBench and user interface support services during the 1986-89 period, as well as building on the TAE Classic structure.

The "Classic" portion of the TAE Plus code is implemented in the C programming language. In selecting a language for the WorkBench and the WPT runtime services, we felt a "true" object-oriented language would provide us with the optimum environment for implementing the TAE Plus graphical user interface capabilities. (See Chapter 9 of Cox [4] for a discussion on the suitability of object-oriented languages for graphical user interfaces.) We selected C++ [5] as our implementation language for several reasons [6]. One of the reasons was the availability of existing, public domain C++ object class libraries. Delivered with the X Window System is the InterViews C++ class library and a drawing utility, idraw, both of which were developed at Stanford University [7]. The idraw utility is a drawing editor which we integrated into the WorkBench to support creating, editing and saving the graphical data-driven interaction objects. This reuse of existing software enabled the addition of a major new function without the significant cost and time of implementing a drawing editor from scratch.

TAE PLUS AS A PRODUCTIVITY TOOL

There are several ways that TAE Plus can contribute to improving software productivity. It provides a development tool that aids in prototyping; gets the best from people; makes steps more efficient; and supports the reuse of software components.

Prototyping

Most organizations now recognize the importance of prototyping and getting the end-user involved in the design process. However, prototyping is not usually thought of as a way to improve productivity. In fact, the prototyping step is frequently avoided or only carried out in a half-hearted manner because of the fear that the end-user will want numerous changes and thereby slow down the design process. This "ostrich head in the ground" syndrome frequently ends in an unpleasant confrontation when the application is delivered to the end-user and the UI fails to meet user expectations. The resultant retrocoding and correcting is often difficult and has to be absorbed as a maintenance cost. Creating a prototype, which allows easy changes and iterative rehearsing of the UI, improves the efficiency of the design and development phase and reduces the likelihood of serious UI changes in the delivered system.

Prototyping fosters a dialog between the developers and the user that can solidify the real system requirements and specifications. As a tool that enables rapid prototypes to be built quickly and easily, TAE Plus can be used to design more effective and user-accepted applications.

Getting the best from people

To get the maximum productivity from each member of a development team individuals should be utilized in the areas that they have an expertise. Too often the people designing application user interfaces are the programmers, who frequently do not have any training in human factors or graphic art techniques. This tends to be an ineffective use of the programmer's expertise, and often results in a less than optimum user interface. The WorkBench was designed to eliminate this problem by giving the user interface design experts a tool that is easy to use (i.e., does not require programming skills), while freeing up the programmer to concentrate on the application specific code.

Making steps more efficient

Another productivity option [8] is to automate a previous manual step, thus eliminating the step entirely. In several of the existing user interface development tools (e.g., Telesoft's TeleUSE™, Visual Edge's UIMX™) including TAE Plus, there is the capability to automatically generate the application code that manages the designed UI. This eliminates the process of the application programmer having to manually generate and key in this code, thus reducing the likelihood of keyboard errors or incorrect function calls. Particularly in cases where the application is heavily interactive, this automatic code generation can account for the majority of the application code and significantly improve productivity of the development process.

Reusing Components

Another way to reduce the amount of source code written for an application, thereby reducing the development cost, is to reuse existing software. In TAE Plus, the WPT runtime services offload all of the display and management of the UI from the application code. This approach enables the application programmer to concentrate fully on the application-specific functions, and not be concerned with the UI code. Also, TAE Plus itself reuses existing windowing software (e.g., MIT's X Window System, OSF/Motif, Stanford's Interview object classes), thus improving the productivity of its own development.

TAE PLUS USERS' EXPERIENCES

One way to measure how effective TAE Plus is as a productivity tool is to develop the same application twice, one time using TAE Plus and another time not using TAE Plus. While most users feel certain that TAE Plus is saving them development time, they are on tight development schedules and do not have the interest in building parallel UIs. However, a few case studies in which the same user interface was developed with and without TAE Plus give evidence that the productivity gain can be impressive.

In Case 1, a programmer from General Electric developed a simple screen copy utility which gathers information through radio buttons, action icons, and text input. Then, it sends the information to an HP printer, as well as updating a text widget on the screen. When he did not use TAE Plus and wrote the UI code directly within the application code, it took him 80 hours to develop an operational application. When he used the TAE Plus WorkBench to develop the same operational application, it took him 4 hours. This productivity gain of 95% is illustrated in Figure 5. However, it should be noted that the gain does not take into account the unmeasured factor that "it is always easier the second time around."

Figure 6 illustrates Case 2. A programmer at NASA with no TAE Plus experience, but with X Window System experience, was tasked to write a simple application and account for the time spent on developing it with and without TAE Plus. The application has two panels, a few action icons, a radio button bank, and a dynamic mover object that moves along a static background when the associated data value changes. Including the time it took to learn how to use the WorkBench to the completion of the operational application, it took him 9 hours. (Note: an experienced TAE Plus user did the same application in 1.5 hours.) The application developed without TAE Plus (thus, making direct calls to the X Window System) took him 52 hours, and this implementation was still a "bit buggy." Even as a beginner TAE Plus user, it took him over four times longer to develop the application without TAE Plus. In the case of the experienced TAE Plus user, the productivity gain was even more dramatic, with a 96% increase in development of the application. A year later we had another programmer write this application with and without using TAE Plus. He was experienced with using the Motif toolkit and he developed the application in 17 hours making direct calls to the Motif toolkit. Using the WorkBench (which he had never used before) it took him 5 hours. Even with an experienced Motif programmer there was a 70% improvement in development time when using TAE Plus for the first time.

Although these case studies certainly do not provide enough statistical data to allow any grandiose conclusions to be made, they do demonstrate real cases in which using a GUI development tool, in this case TAE Plus, has significantly decreased the time it takes to develop the application. In general, TAE Plus reduces the time it takes a developer to create, test and deliver a software system.

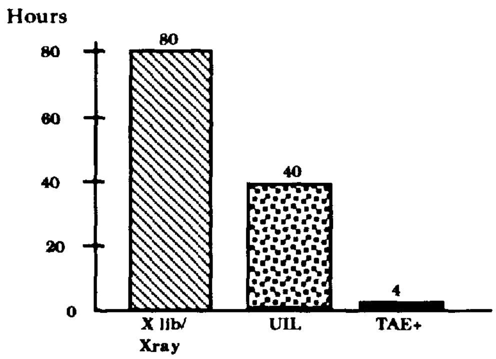


Figure 5: Case Study 1

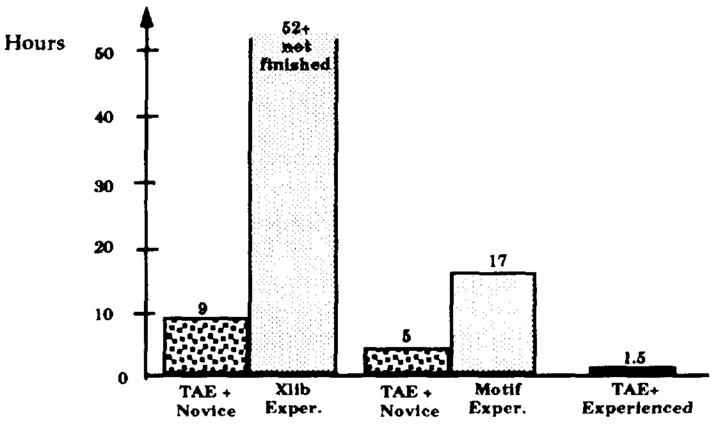


Figure 6: Case Study 2

AVAILABILITY AND MAINTENANCE

In December 1992 the latest version of TAE Plus (V5.2) became available from COSMIC, the NASA's software distribution center located at the University of Georgia. TAE Plus may be licensed by the public for a nominal fee and it is available on a variety of platforms: Sun workstations, Vaxstation II , Decstation 3100, HP9000, Masscomp, Silicon Graphics Iris and IBM RISC 6000. It is also available on the Vaxstation II under VMS and the NEC company has ported it onto their NEC EWS 4800/220 workstation for use by their customers.

Maintenance of a software system is a key factor in its success, and while every system is maintainable, how easy it is to maintain is the real issue. We knew when we began development that TAE Plus was targeted for wide application utilization and for different machines, so ease of maintenance has always been important. By providing the application-callable WPTs, applications are isolated from the windowing system. Thus, when the latest release or next generation windowing system shows up, only the WPTs will require updating or rewriting; the application code will not be affected.

User support is another facet of maintainability. Since the first release of TAE Classic in 1981, we have provided user support through a fully staffed Support Office. Users receive answers to technical questions, report problems, and make suggestions for improvements. In turn, the Support Office keeps users up-to-date on new releases, provides a newsletter, and sponsors user workshops and conferences. This exchange of information enables the Project Office to keep the TAE software and documentation "in working order" and, perhaps most importantly, take advantage of user feedback to help direct our future development.

APPLICATIONS USING TAE PLUS

Since 1982 over 900 installation sites have received TAE Classic and/or TAE Plus. Just over the past year, COSMIC has issued licenses to over 300 customer sites. The applications built or being built with TAE perform a variety of different functions. TAE Classic usage was primarily used for building and managing large scientific data analysis and data base systems (e.g., NASA's Land Analysis System (LAS), Atmospheric and Oceanographic Information Processing System (AOIPS), and JPL's Multimission Image Processing Laboratory (MIPL) system.) Within the NASA community, TAE Plus is also used for scientific analysis applications, but the heaviest concentration of user applications has shifted to support of realtime control and processing applications. This includes supporting satellite data capture and processing, monitor and control of spacecraft and science instruments, prototyping user interface of the Space Station Freedom crew workstations and supporting diagnostic display windows for realtime control systems in ground operations. For these types of applications, TAE Plus is principally used to design and manage the user interface, which is made up of a combination of user entry and data-driven interaction objects. TAE Plus becomes a part of the development life cycle as projects use TAE Plus to prototype the initial user interface design and have this designed user interface evolve into the operational UI.

Outside the NASA community, TAE Plus is being used by an assortment of other government agencies (13%), universities (15%), and private industries (40%). Within the government sector, users range from the National Center for Atmospheric Research, National Oceanographic and Atmospheric Administration, U.S. Geological and EROS Data Center, who are developing scientific analysis, image mapping and data distribution systems, to numerous Department of Defense laboratories, who are building command-and-control systems. Universities represented among the TAE community include Cornell, Georgia Tech, MIT, Stanford, University of Maryland and University of Colorado. Applications being developed by University of Colorado include the Operations and Science Instrument Support System(OASIS), which monitors and controls spacecraft and science instruments and a robotics testbed for research into the problems of construction and assembly in space. [9] Private industry has been a large consumer of the TAE technology and a sample of the companies that have received TAE Plus include Loral Aerospace, Martin Marietta, Computer Sciences Corp., TRW, Lockheed, IBM, Northern Telecom, Mitre Corp., General Dynamics and GTE Government Systems. These companies are using TAE Plus for an assortment of applications, ranging from a front-end for a corporate database to advanced network control center. Northern Telecom, used TAE Plus to develop a technical assistance service application which enables users to easily access a variety of applications residing on a network of heterogeneous host computers.[10] General Software Corporation uses TAE in their commercial product, METPRO, a meteorological information processing system, which has been distributed in seven countries. Another company, Global Imaging, Inc. has embedded TAE into their commercial image processing system. Because of the high cost associated with programming and software-development, more and more software development groups are looking for easy-to-use productivity tools, and TAE Plus has become recognized as a viable tool for developing an application's user interface.

NEXT STEPS

The current TAE Plus provides a useful tool within the user interface development environment -- from the initial design phases of a highly interactive prototype to the fully operational application package. However, there are many enhancements and new capabilities that will be added to TAE Plus in future releases.

In the near term, the emphasis will be on enhancements and extensions to the WorkBench. All the requested enhancements are user-driven, based on actual experience using TAE Plus, or requirement-driven based on an application's design. For example, on the enhancements list are extensions to the interaction objects, (e.g., graph data-driven object, form fill-in), support for importing foreign graphics, and extensions to the dialog connections feature (e.g., graphic representation of the connection mapping, item-to-item connections).

Future advancements include expanding the scope of TAE Plus to include new tools and technologies. For instance, the introduction of hypermedia technology and the integration of expert system technology to aid in making user interface design decisions are targeted for investigation and prototyping.

CONCLUSION

With the emergence of sophisticated graphic workstations and the subsequent demands for highly interactive systems, the user interface becomes more complex and includes multiple window displays, the use of color, graphical objects and icons, and various selection techniques. Software tools, such as TAE Plus, are providing ways to make user interface developer's tasks easier and improve the overall productivity of the development process. This includes supporting prototyping of different user interface designs, as well as development and management of the operational application's user interface.

TAE Plus is an evolving system, and its development will continue to be guided by user-defined requirements. To date, each phase of TAE Plus's evolution has taken into account advances in windowing systems, human factors research, standardization efforts and software portability. With TAE Plus's flexibility and functionality, it is providing a useful productivity tool for building and managing graphical user interfaces.

ACKNOWLEDGEMENTS

TAE Plus is a NASA software product being developed by the NASA/Goddard Space Flight Center with prime contract support by Century Computing, Inc. The work is sponsored by the NASA Office of Space Operations.

TAE is a registered trademark of National Aeronautics and Space Administration (NASA). It is distributed through NASA's distribution center, COSMIC, (706) 542-3265. For further information, contact COSMIC and/or the TAE Support Office at GSFC, (301) 286-6034.

REFERENCES

1. Perkins, D.C., Howell, D.R., Szczur, M.R., "The Transportable Applications Executive -- an interactive design-to-production development system," Digital Image Processing In Remote Sensing, edited by J-P Muller, Taylor & Francis Publishers, London, 1988.

2. Scheifler, Robert W., Gettys, Jim., "The X Window System," MIT Laboratory for Computer Science, Cambridge, MA, October 1986.

3. Open Software Foundation, Inc., OSF/Motif™ Programmer's Reference Manual, Revision 1.1, 1990

4. Cox, Brad J., Object Oriented Programming, An Evolutionary Approach, Addison-Wesley Publishing Company, Reading, MA, 1986.

5. Stroustrup, Bjarne, The C++ Programming Language, Addison-Wesley Publishing Company, Reading, MA, 1987.

6. Szczur, Martha R., Miller, Philip, "Transportable Applications Environment (TAE) Plus: Experiences in 'Object'ively Modernizing a User Interface Environment," Proceedings of the OOPSLA Conference, September 1988.

7. Linton, Mark A., Vlissides, John M., Calder, Paul R., "Composing User Interfaces with Interviews, "IEEE Computer, February, 1989.

8. Boehm, Barry, "Improving Software Productivity", IEEE Computer, September, 1987, pp. 43-57

9. Klemp, Marjorie, "TAE Plus in a Command and Control Environment", Proceedings of the TAE Eighth Users' Conference, June, 1990

10. Sharma, Alok, et al., "The TAS Workcenter: An Application Created with TAE", Proceedings of the TAE Eighth Users' Conference, June, 1990

ADVANCED DISPLAY OBJECT SELECTION METHODS FOR ENHANCING USER-COMPUTER PRODUCTIVITY

Dr. Glenn A. Osga
Naval Command Control & Ocean Surveillance Center,
RDT&E Division
San Diego, Ca. 92152-5000

ABSTRACT

The User-Interface Technology Branch at NCCOSC RDT&E Division has been conducting a series of studies to address the suitability of commercial off-the-shelf (COTS) graphic user-interface (GUI) methods for efficiency and performance in critical naval combat systems. This paper presents an advanced selection algorithm and method developed to increase user performance when making selections on tactical displays. The method has also been applied with considerable success to a variety of cursor and pointing tasks. Typical GUIs allow user selection by: 1) moving a cursor with a pointing device such as a mouse, trackball, joystick, touchscreen, and 2) placing the cursor on the object. Examples of GUI objects are the buttons, icons, folders, scroll bars, etc. used in many personal computer and workstation applications. This paper presents an improved method of selection and the theoretical basis for the significant performance gains achieved with various input devices tested. The method is applicable to all GUI styles and display sizes, and is particularly useful for selections on small screens such as notebook computers. Considering the amount of work-hours spent pointing and clicking across all styles of available graphic user-interfaces, the cost/benefit in applying this method to graphic user-interfaces is substantial, with the potential for increasing productivity across thousands of users and applications.

INTRODUCTION

Many varieties of cursor selection tasks exist across graphic user-interface (GUI) applications. Selectable objects include: text on menus, radio buttons, check boxes, icons, file folders, buttons, and labels. They may also include graphic objects in computer-aided design, architectural, design and graphics layout programs. The selection of objects by cursor placement can be described in human performance terms by Welford's variation of Fitts's Law, (refs. 1-2). This law states that, within certain limits, cursor distance moved to the object and object size affect user performance time. When positioning time is plotted by Fitts's index of difficulty ($\log^2$ Distance/Size + .5) performance speed is affected by design factors such as control/display ratio, cursor velocity, and quality of visual feedback (refs. 2-4). These critical design parameters related to task difficulty (e.g. distance and object size), and quality of visual feedback were manipulated to yield the advanced selection algorithm (ASA) described in this paper.

The first parameter manipulated by the ASA is cursor-object distance. Current user-interface designs typically require the user to place the cursor on the desired object. Thus, distance traveled by the cursor is dependent on the distribution of objects, size of the display, and efficiency of the user-interface design with respect to cursor travel. In contrast, the ASA reduces the cursor travel distance required by determining the distance from the cursor to the closest selectable object, not requiring cursor placement on the object to be selected.

The second parameter is directly related to cursor-object distance calculation. For selection purposes, objects are treated as larger than they visually appear on the display due to the calculation of relative object-cursor distance. Thus, an object's selection area is directly related to the spacing and distribution of objects, vs. their apparent size.

The third parameter provides the user with necessary constant visual feedback to allow variable cursor-object distances. Methods of visual feedback have become "standardized" across many software applications. With the exception of pull-down or pop-up menus, visual feedback (highlighting) is typically presented <u>after</u> the user performs an action selecting an object. The ASA displays visual highlighting for a "selectable" object constantly as the cursor is moved on the display. Thus, object selectability is shown to the user before a selection action is made, and is independent of object size or absolute location.

METHOD

Complementary software methods were developed to implement the Advanced Selection Aid (ASA). The "selectable" cursor target, defined as the object closest to the current cursor position, is constantly determined. Methods used are dependent on the type of task performed. The first method applies to the selection of objects which are spaced or arranged irregularly on the display. As shown in Figure 1, the computer determines x and y coordinates of object and cursor "hotspot" location. The minimum distance from the cursor to the closest object can be computed using the Pythagorean Theorem. If the cursor is on the same x or y axis as the closest object (e.g. either x or y is less than the object height or width) then the absolute value of x or y is the calculated distance. This distance to all displayed objects is computed in real-time as the cursor is moved. If the cursor was equidistant from two or more objects, the last closest object is used. In Figure 1, object 2 is closer to the cursor and the expanded outline around it indicates that it is currently "selectable" and would be selected if an appropriate selection action is taken.

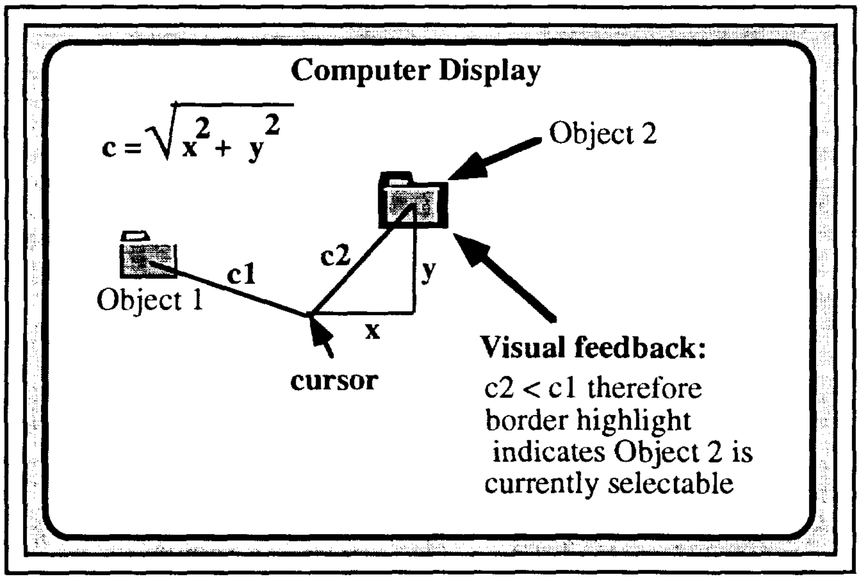
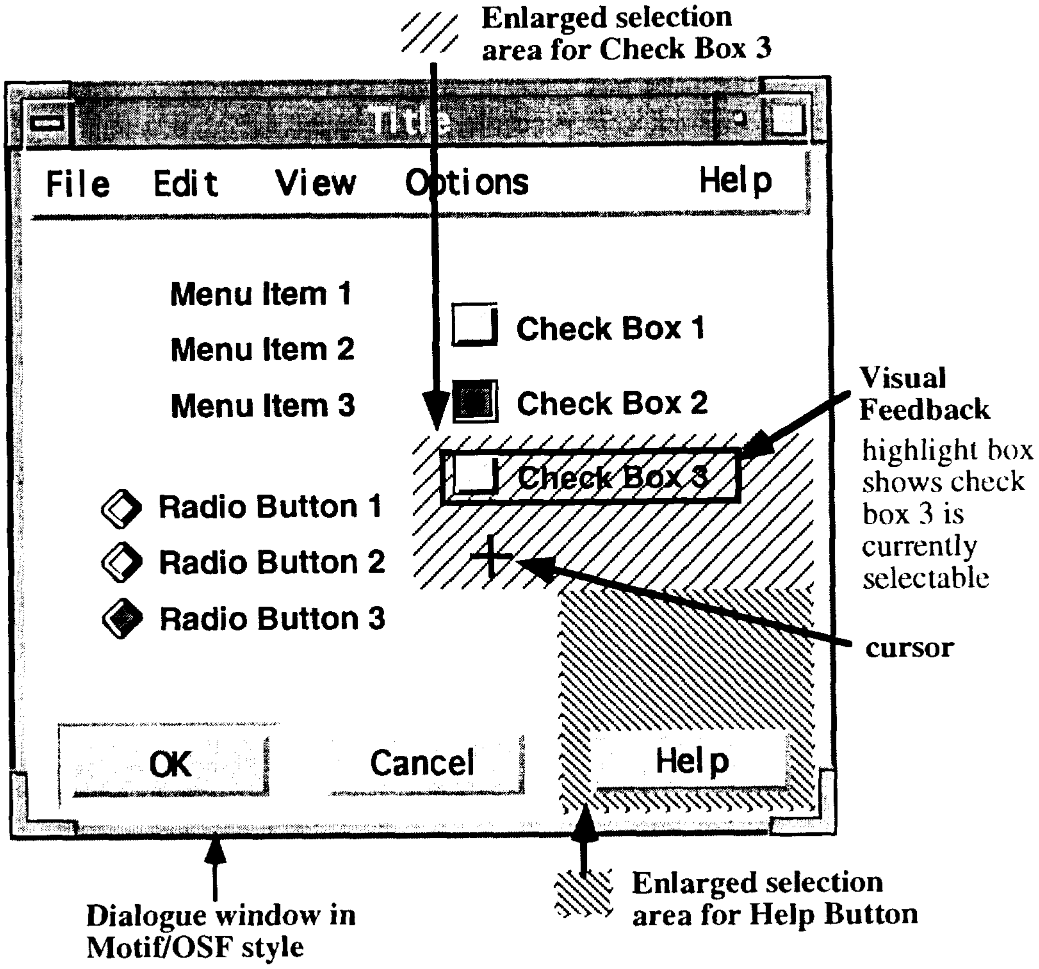


Figure 1: Advanced Selection Algorithm cursor-object distance calculation and visual feedback example

For display items which are adjacent and fixed in relative location to each other, such as dialogue window items, the selection area can be determined by defining selection areas which surround each selectable item. A simpler method defines the distance between the cursor and the edge of the object. With collections of irregular shaped objects, the edge must be used to calculate cursor-object distance in lieu of the object center, or results are not accurate. In Figure 2, all objects have larger selection areas than their physical appearance alone would indicate, as illustrated by the selection areas for "Check Box 3" and the "HELP" button. The ASA calculates the distance from the cursor to the nearest object edge, which is "Check Box 3" in this example. The object is highlighted by appropriate methods such as the outlined box shown. Other highlighting methods can be used to indicate the "selectable" object such as color, or outline. We used inverse video and the typical GUI check box or radio button indicators to indicate final object selection.

The ASA impact on machine performance using available PC technology has been negligible. We have tried hundreds of objects simultaneously displayed on a 19" color monitor in 8-bit color mode without any performance decrement. Being computationally simple, the effect of visual highlighting and cursor-object distance computation does not appear to slow other machine calculations. The net effect of these computations and visual highlighting is a closed-loop human-computer control system which does not require the user to make precise cursor movements to accomplish selection tasks. The next section describes the impact of the ASA on human performance.



Enlarged selection area for Check Box 3

Visual Feedback highlight box shows check box 3 is currently selectable

cursor

Dialogue window in Motif/OSF style

Enlarged selection area for Help Button

Figure 2: Application of Advanced Selection Algorithm to a typical graphical user-interface dialogue window

EVALUATIONS

The ASA was applied to the selection of symbols on tactical displays and to dialogue window selection tasks. Details of test results and methods are described elsewhere (ref. 5). Additional data and studies were conducted during 1992 yielding similar results. Input devices tested have included touchscreens, touchtablets, trackballs, and mouse. Figure 3 presents results for a task involving the selection of graphics objects which were included both dispersed and closely placed configurations. At least six users were tested for each input method during controlled studies, with significant increases in user performance for speed and error reduction observed. With the addition of the ASA, some input technologies, such as touchscreens and small touchtablets, transition from being either non-practical or tedious devices to being practical and efficient methods. Other input technologies, such as the mouse or trackball become considerably easier to use. Results for the ASA for menu selection tasks indicate a significant reduction in selection errors for adjacent menu items, with no increase in speed. These results would be expected since object selection size cannot be enhanced if objects are of equally sized and placed adjacent. For selection of adjacent objects, error reduction indicates increased user satisfaction would likely result, however.

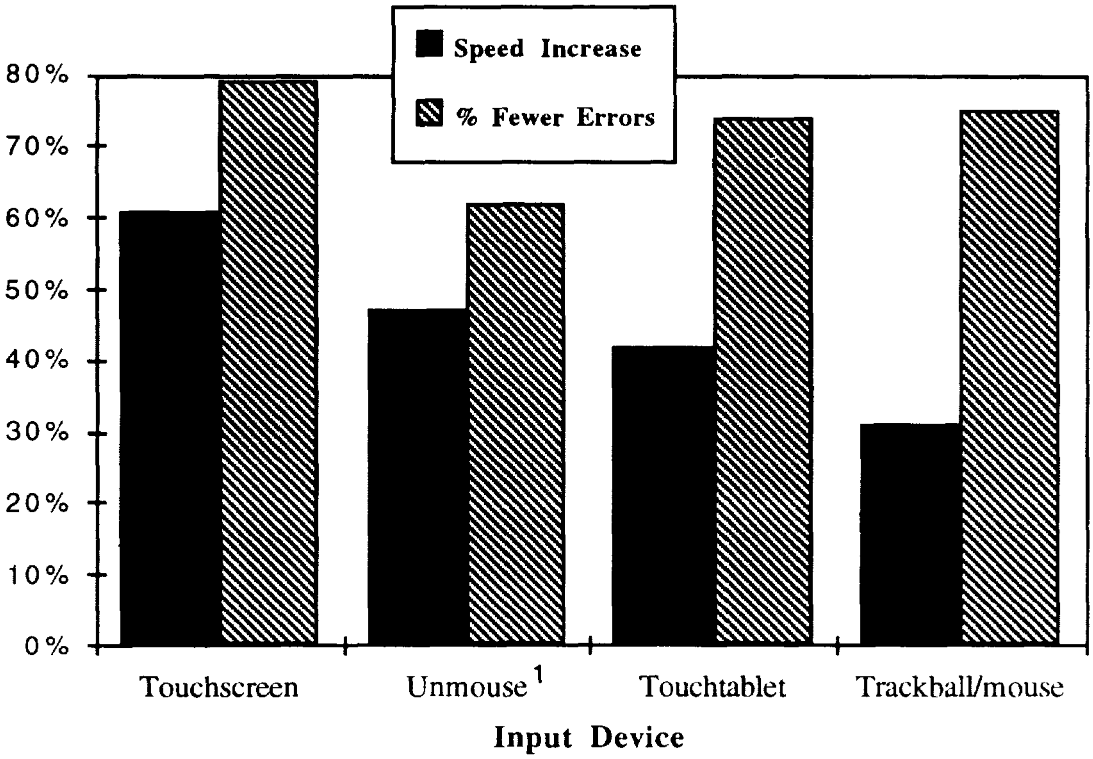


Figure 3: Average performance increase in user speed and accuracy during selection of graphic objects using the advanced selection algorithm[1]

APPLICATIONS

Current GUI users are faced with a myriad of small objects which are constant targets for cursor placement throughout the working day. Software designers include small squares, icons and objects depicting window close boxes, "sizing icons", etc. across all popular GUIs. These selection tasks become more difficult on smaller screens such as portable or notebook computers. The ASA can be applied to computer desktop GUI applications to support tasks such as the manipulation of file folders or window icons. Icons can be displayed with enhanced borders or inverse colors. We have successfully modified dialogue windows in the Apple Macintosh[2] GUI to incorporate the ASA.

Cursor travel distances become larger as two-page size displays are used in workstations and desktop publishing applications. The ASA is particularly useful for these configurations. Touchtablet devices which are designed as small replacements for a mouse or trackball, become difficult to use with large screens due to small finger movements resulting in large cursor movements. The ASA improves performance with small tablets considerably. Several negative design aspects typically associated with touchscreen use are virtually eliminated with the ASA. Cursor offset relative to the finger placement and the impact of obscuring of display objects is reduced. Display objects do not require reformatting with larger buttons or selection areas. Other selection techniques such as eye or head cursor tracking and selection would also benefit from ASA use.

Many GUI programs and applications are impractical for use without pointing and clicking. A segment of the user population with marginal cursor manipulation and pointing skills who are unable to

[1] Unmouse is a trademark of MicroTouch Inc.

[2] Macintosh is a trademark of Apple Computer Inc.

comfortably use GUI methods may find point-and-click computing to be practical with ASA implemented. ASA sensitivity can be varied. For example, we are also implementing other selectability criteria to be combined with cursor-object distance, such as the type of object. When selecting from a particular class of objects on a cluttered display containing many object types, a function could be applied which selectively allows the user to apply ASA to specific object classes while excluding others.

CONCLUSIONS

Significant user performance enhancement has been shown in user experiments using ASA for object selection. These effects have been demonstrated for a variety of pointing devices, and should apply to all cursor pointing methods. We have observed no negative effects in machine speed or processing delays in a typical PC configuration. Application of this method is possible in any graphic interface application where cursor-object distance can be computed. Considering the labor-hours spent by thousands of graphic user-interface users moving and placing a cursor for pointing and selection tasks, the application of this technique by software developers will substantially improve productivity across a broad base of end users and applications.

ACKNOWLEDGMENTS

The advanced selection algorithm is currently U.S. patent pending. Development and evaluation was sponsored by the Office of Naval Technology Human Factors Project, and the Aegis Shipbuilding Program. The author thanks Mr. Adam Richardson and Mr. Roger Keating for programming and data collection efforts. Opinions expressed within are those of the author, and do not represent official U.S. Navy or D.O.D policy, position, or decision.

REFERENCES

1. Fitts, P.M. (1954) The information Capacity of the Human Motor System in Controlling Amplitude of Movement, *Journal of Experimental Psychology*, 47, 381-391.

2. Card, S.K, Moran, T.P. and Newell, A. (1983) *The Psychology of Human Computer Interaction*, Lawrence Erlbaum Associates, Hillsdale New Jersey, p 55.

3. Parng, A.K. (1988) *An Automated Test of Fitts' Law and Effects of Target Width and Control-Display Gain Using a Digitizer Tablet*, Technical Document 1214, Naval Ocean Systems Center, San Diego.

4. Greenstein, J.S. & Arnaut, L. Y. (1988) Input Devices in M. Helander, (Ed.), *Handbook of Human-Computer Interaction*, Elsevier, New York.

5. Osga, G.A. (1991) Using Enlarged Target Area and Constant Visual Feedback to Aid Cursor Pointing Tasks *In Proceedings of the Human Factors Society 35th Annual Meeting*, San Fransisco.

53-82

Universal Index System

N93-22152

Steve Kelley
Nick Roussopoulos
Timos Sellis
Sarah Wallace

Advanced Communication Technology Inc.
1209 Goth Lane
Silver Spring, Maryland 20905

ABSTRACT

The Universal Index System (UIS) is a index management system that uses a uniform interface to solve the heterogeneity problem among database management systems. UIS provides an easy-to-use common interface to access all underlying data, but also allows different underlying database management systems, storage representations, and access methods.

1. Introduction

Today, there is a great diversity of computers, operating systems, database management systems, and communication protocols. As a result of this heterogeneity, computer users are required to learn many different data access methods in order to obtain the information they need. This causes an attitude of "it's too much trouble to learn all these different systems," which leads to a significant amount of software and data duplication.

There are several approaches that can be taken to solve the heterogeneity problem: two of which are *standardization* and *uniformization*. *Standardization* is the concept of choosing one specific system to use, and expecting or requiring everyone to follow this standard. This, however, does not provide an adequate solution because it could be extremely costly to change to the standard if a different system was being used. *Uniformization* is the concept of creating a layer on top of current systems that provides uniform access to all data, regardless of the underlying system. This allows the underlying systems to remain unchanged, yet also provides a single common access method for users to access data.

This paper presents the Universal Index System (UIS), an index management system that uses a uniform interface to solve the heterogeneity problem among database management systems. UIS provides an easy-to-use common interface to access all underlying data, but also allows different underlying database management systems, storage representations, and access methods.

2. UIS Components

UIS is a system that manages and maintains indexes, sets, indexsets, and indexkits. An *index* is an object that associates terms with pointers. A simple example of an index is the index of a book. It associates a term used in the book with the page number(s) on which that term appears. Another example of an index is a subject index in a library catalog, which associates library books with different subjects.

A *set* is an object that contains only pointers. Usually sets are created by extracting the pointer field from an index. Using the example of a book's index, a set could be created from the index by the definition all the page numbers that contain the words 'database', 'data model', 'data definition language', or 'data manipulation language'.

An *indexset* is a catalogued collection of indexes and sets. Every index and set must be associated with exactly one indexset. In addition to the indexes and sets belonging to an indexset, an indexset also contains an index catalog to maintain all the information for managing indexes, and a set catalog to maintain all the information for managing sets.

An *indexkit* is a logical grouping of an introduction, index, dictionary and thesaurus. The introduction component of an indexkit is an object which contains a textual description of the index. The dictionary component of an indexkit is an object that associates terms given in the index with their definition. It is used to assist the user in accessing the index. The thesaurus component of an indexkit is an object that associates terms given in the index with other terms. The thesaurus supports both generalization and specialization of terms in the index. The thesaurus is also used to assist the user in accessing the index. The introduction, dictionary and thesaurus components are neither managed nor maintained by UIS. Figure 1 shows the relationships among the different objects managed by UIS.

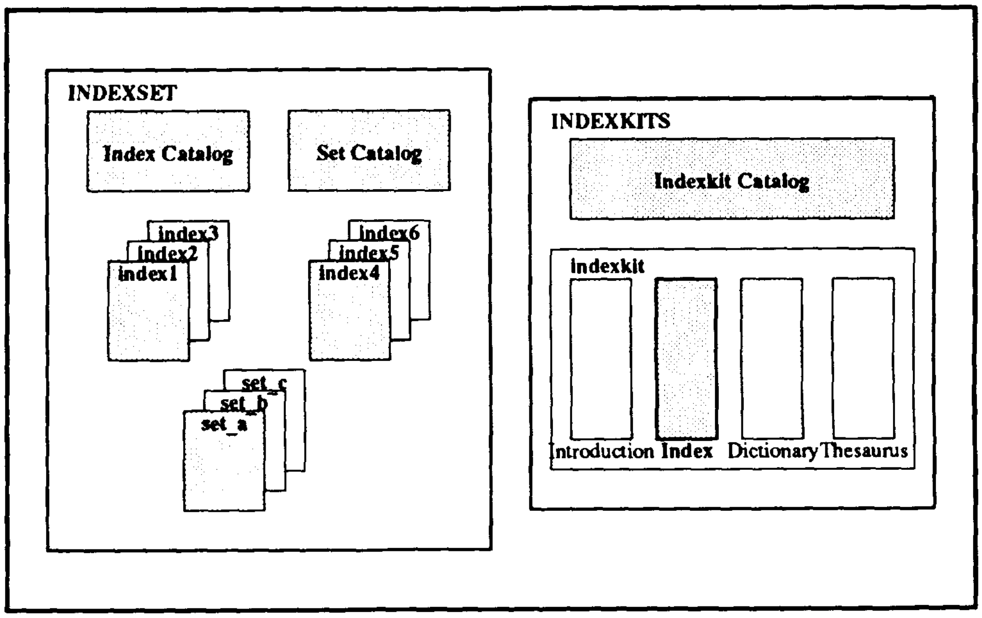


Figure 1 - Relationships among Indexes, Indexsets, Indexkits and Kitsets.

3. UIS Capabilities

UIS provides commands that allow the user to create and manipulate indexes, sets, indexsets, and indexkits.

3.1. Index Commands

UIS uses the notion of *current* objects to simplify the index commands. The user specifies which instance of an object is to be *current* i.e. to be worked on, and then subsequent commands are performed on the *current* object. The index commands rely on the existence of a *current index*, *current index row*, and *current index boolean*.

The *current index row* is set to be the tuple in the *current index* that was most recently accessed by the navigation routines (see below for a description of the navigation routines). The *current index boolean* is a boolean condition chosen by the user to assist in navigation.

UIS provides a relationally complete set of commands for indexes. In addition to commands that allow the user to create, insert into, delete from, save and destroy indexes, there are routines that allow the user to retrieve a previously created index for either modification or read only, return an index (the opposite of retrieve) and pick an index to be the *current index*.

There are commands to allow the user to navigate both forward and backward through an index, accessing a single tuple at a time. UIS provides the user with *index booleans* and *index selects* to assist in this navigation. An *index boolean* is a boolean condition defined by the user to restrict the search to a subset of the index. For example, the user could define an index boolean, camseq = LFP1010" to restrict the search on an index to only those tuples of an IUE index having LFP1010 as camera sequence number. The user can create index booleans during a user session, but they do not persist beyond the end of that session. UIS provides commands to create, modify, list (display), pick (make as current), and delete index booleans. There are also commands to allow the user to reproduce indexes. These include copying and moving an index to an indexset.

To support interfaces to programming languages, there are commands to allow the user to bind attribute values to program variables, i.e. embedding UIS commands in an application written in C. There are two commands for binding to program variables, one for binding a single attribute (column) from an index, and one that allows for binding a whole row from an index. These commands cannot be used during an interactive session.

3.2. Indexset Commands

UIS provides a few commands to manipulate indexsets. At this point a user can only create and destroy indexsets. In the future, we plan to add commands such as include copy, subset, intersect, subtract and union, and commands to copy and move indexsets.

3.3. Indexkit Commands

Although not implemented in the current prototype, several commands to manipulate indexkits have been designed for UIS. In addition to commands that allow a user to create and destroy indexkits, there are commands to allow the user to reproduce indexkits. These include copy, subset, intersect, subtract and union. Subsetting an indexkit is defined to be a new indexkit, whose components are the result of subsetting each of the components in the original indexkit. Intersecting two indexkits is defined to be a new indexkit, whose components are the result of intersecting corresponding components of the two original indexkits. Similar definitions hold for union and subtraction.

3.4. Command Summary

Tables A, B, C, and D and the end of this paper provide a summary of the index, set, indexset and indexkit commands, respectively.

4. The Design of UIS

The development of the UIS prototype was divided into several phases: the requirements phase, the design phase, the implementation phase and the testing and integration phase. This approach was taken in an attempt to resolve any conflicts in the proposed system as early as possible.

The requirements document contains a functional description of what the system should do. The purpose of the design phase is to convert the functional description of **what** the system should do into an algorithmic description of **how** the system should do it.

The design phase primarily concentrated on two tasks. First, we had to determine what information needed to be available to the system during execution and what information needed to be available from one execution to the next (persistent information). Second, we needed to translate the functional requirements of the user commands into design specifications. These two tasks were performed in a stepwise fashion to yield a cohesive and consistent design.

4.1. System Information

UIS manages and maintains four different types of objects: indexes, sets, indexsets, and indexkits. In order to do so properly and efficiently, the system needs to have available certain information about each object. As an example, consider a library: how useful or efficient would a library be if it did not have a catalog that listed what books were contained in the library, or where they were located? Probably not very useful, definitely not very efficient. In the same way that a library catalogs all the **objects** that it manages, so must UIS. This section describes which information UIS needs to efficiently manage its objects.

4.2. System Catalog — Indexes

In Section 2 we defined conceptually what an index is. To determine what persistent information we need for indexes, we need to know what an index is structurally. Structurally, an index is a table in which some of the columns are the items indexed and the last column is the pointer. An index is of type k if it has k-item tuples (columns). The format of an index depends on the internal representation of the index. Examples of formats are B-trees, R-trees, and heaps.

Given this structural definition, we see that some of the information that needs to be stored include the name of the index, its type, and its format. Other information that is necessary are the attribute or column names, their types, lengths and their location within the tuple (offset). This information is necessary when checking whether or not a user's command is valid, and to assist the system in locating and extracting attribute values. Another piece of information used to assist the system in index manipulation and validation is the index's tuple width (the total size of the tuple). In addition, we decided it would be helpful to store whether or not a given index had an indexkit associated with it. This would allow us to remain consistent with the indexkit system information (discussed later).

Because an index can have any number of attributes, we decided it would be easier to have two system catalogs. The first one contains all the information about the index except for the attribute information. A second catalog contains the attribute information. This approach was taken to simplify the catalog access routines (if a single catalog were used, the access routines would have to support variable length entries). Figure 2 describes pictorially the system catalog

Index Catalog

Index Name	Index Typ	ndex Forma	Indexkit Name	Indexkit set	Indexspace Name	Tuple Widt
FOLLET_EOTN	2	*BTREE*	*FOLLET_EOTN*	*FOLLET*	*EARLY_WORKS*	44
SUBJECTS	5	*RTREE*	*SUBJECT_LIB*	*LIBRARY*	*PG_COUNTY*	204

Attribute Catalog

Index Name	Attribute Name	Attr Type	Attr Length	Attr Offset
FOLLET_EOTN	*TERM*	*STRING*	40	0
FOLLET_EOTN	*PAGE_NUM*	*INT*	4	40
SUBJECTS	*SUBJECT_TERM*	*STRING*	40	0
SUBJECTS	*AUTHOR*	*STRING*	40	40
SUBJECTS	*TITLE*	*STRING*	100	80
SUBJECTS	*ISBN_NUMBER*	*STRING*	12	180
SUBJECTS	*LC_NUMBER*	*STRING*	12	192

Index Format =
{ BTREE, RTREE, HEAP }

Attr Type =
{ INT, FLOAT, CHAR, STRING }

Figure 2 - System Catalog Information for Indexes.

information for indexes. It contains two example indexes: FOLLET_EOTN (a book index for Ken Follet's *The Eye Of The Needle* and SUBJECTS (a library catalog of subjects which references books).

4.3. System Catalog — Indexsets

An indexset has several components (see Figure 3). It contains an index catalog discussed in the previous section, a set catalog, a transaction log, and then the indexes and sets themselves that belong in the indexset. The transaction log contains information about updates to the indexes and sets in the indexset. It is used in transaction management (currently unimplemented). UIS allows the user to explicitly specify all the buffer management constants needed for the management of indexset components. As a result, the system catalog information for indexsets must store all this information.

Before explaining the system catalog information for indexsets, we need to clarify what is meant by *databook* and *indexspace*. When defining an indexset, the user creates a logical space in which indexes and sets will belong at some point in the future. The *databook* objects are these logical spaces. An *indexspace* is the physical storage space on the disk that corresponds to the logical space defined by the databooks. Indexspaces can contain several databooks, and databooks can span more than one indexspace. Having the user be able to specify both logical space and physical space allows the user to place indexes physically near each other or logically near each other.

Given these new objects, an indexset is composed of the following components: index catalog, set catalog, transaction log, any number of databooks, and any number of indexspaces. For each of these components, the system needs to have information about the names of each of these components, the initial physical size of these components, their maximum size, and the rate at which these components can increase (when an insertion needs to be made and there is no space, an increase is requested and as long as the maximum size has not been reached, the increase is allowed).

Storing all this information creates a complicated system catalog structure. The databook and indexspace information for indexsets is stored in its own catalog. This is due to the fact that there can be any number of these objects in an indexset (similar to the attribute information for indexes). Since the directory, index catalog, set catalog and transaction log components are required for each indexset, and an indexset can contain at most one of each component, all of this information can be stored in a single catalog along with the indexset name. In addition, it was decided to have entries in this catalog for the total number of databooks and indexspaces in the indexset, to assist in retrieval from the other catalogs.

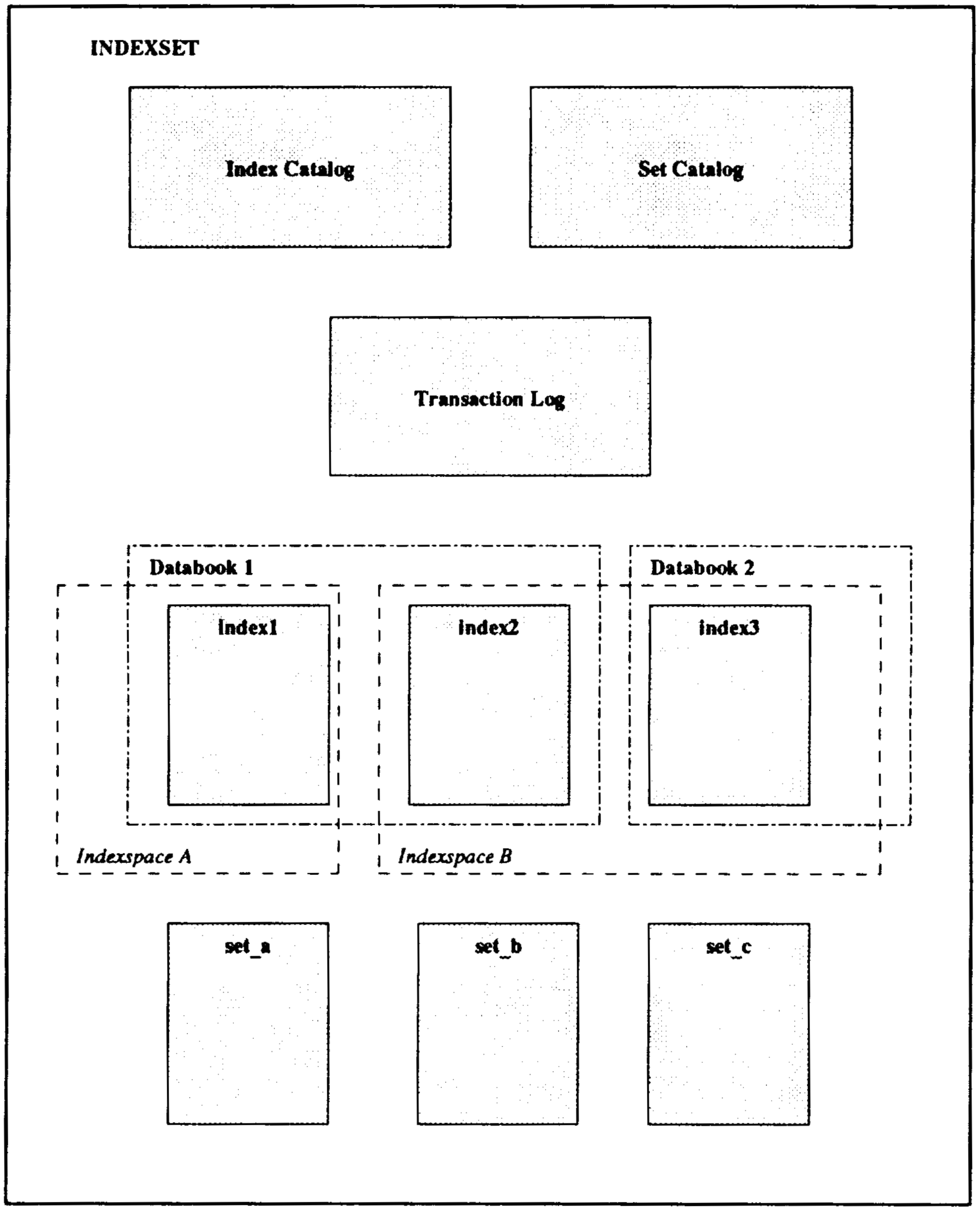
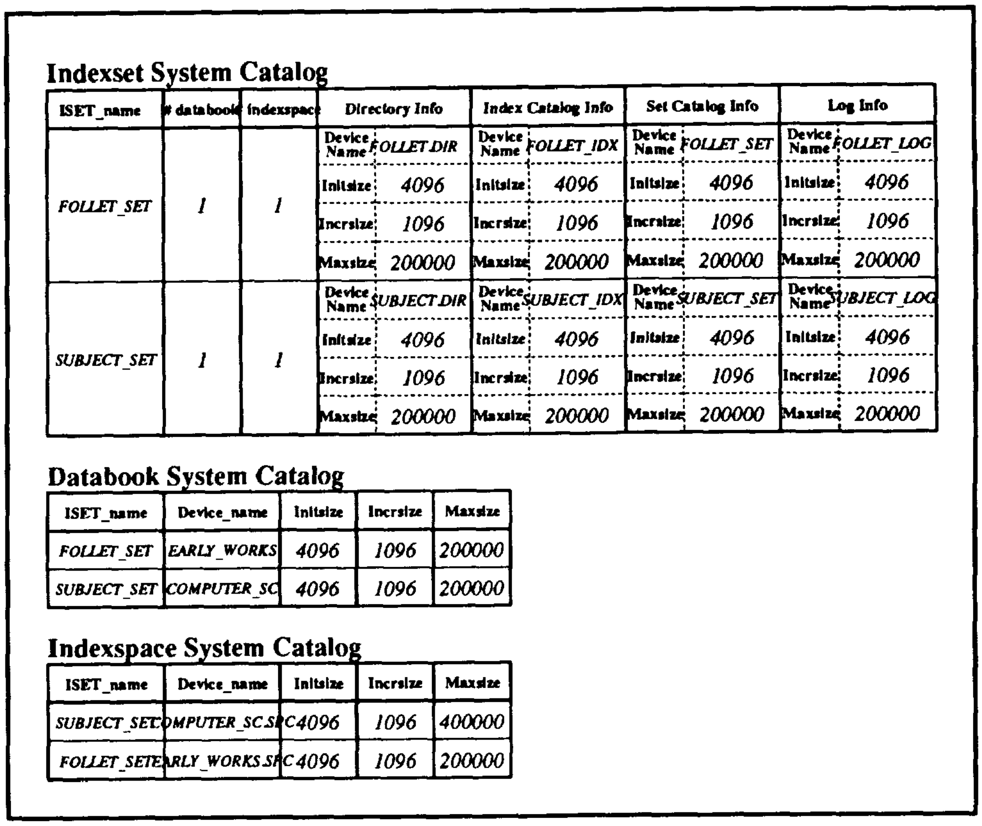


Figure 3 - Physical Structure of an Indexset

Figure 4 describes pictorially the system catalog information for indexsets. It contains two examples of indexsets: FOLLET_SET (an indexset that contains all the index information about Ken Follet's books) and SUBJECT_SET (an indexset that contains all the subject information at a specific library). For example, the index catalog component for FOLLET_SET says that the index catalog is located in the file *FOLLET_IDX*. Its initial size is 4096 bytes, and when the system needs more space for the index catalog, space is allocated in blocks of 1096 bytes. If the size of the index catalog reaches 200000, no more space will be allocated to the index catalog. The **Databook System Catalog** and the **Indexspace System Catalog** contain similar information about the databooks and indexspaces in the indexset.

4.4. System Catalog — Indexkits

As defined an Section 2, an indexkit is a logical grouping of an introduction, index, dictionary and thesaurus. In order for the system to understand this logical grouping, it needs to keep track of which instances of each component



Indexset System Catalog

ISET_name	# databook	indexspace	Directory Info		Index Catalog Info		Set Catalog Info		Log Info	
FOLLET_SET	*1*	*1*	Device Name	*FOLLET DIR*	Device Name	*FOLLET_IDX*	Device Name	*FOLLET_SET*	Device Name	*FOLLET_LOG*
			Initsize	*4096*	Initsize	*4096*	Initsize	*4096*	Initsize	*4096*
			Incrsize	*1096*	Incrsize	*1096*	Incrsize	*1096*	Incrsize	*1096*
			Maxsize	*200000*	Maxsize	*200000*	Maxsize	*200000*	Maxsize	*200000*
SUBJECT_SET	*1*	*1*	Device Name	*SUBJECT DIR*	Device Name	*SUBJECT_IDX*	Device Name	*SUBJECT_SET*	Device Name	*SUBJECT_LOG*
			Initsize	*4096*	Initsize	*4096*	Initsize	*4096*	Initsize	*4096*
			Incrsize	*1096*	Incrsize	*1096*	Incrsize	*1096*	Incrsize	*1096*
			Maxsize	*200000*	Maxsize	*200000*	Maxsize	*200000*	Maxsize	*200000*

Databook System Catalog

ISET_name	Device_name	Initsize	Incrsize	Maxsize
FOLLET_SET	*EARLY_WORKS*	*4096*	*1096*	*200000*
SUBJECT_SET	*COMPUTER_SC*	*4096*	*1096*	*200000*

Indexspace System Catalog

ISET_name	Device_name	Initsize	Incrsize	Maxsize
SUBJECT_SET	*COMPUTER_SC.SPC*	*4096*	*1096*	*400000*
FOLLET_SET	*EARLY_WORKS.SPC*	*4096*	*1096*	*200000*

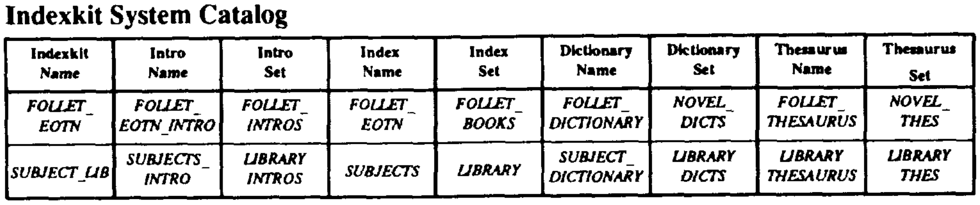
Figure 4 - System Catalog Information for Indexsets.

make up this logical grouping. As a result, the system information needed for each indexkit is the name of the indexkit, the introduction name and its location (intro_set), the index name and its location (indexset), the dictionary name and its location (dict_set), and the thesaurus name and its location (thes_set) (Remember that the introduction, dictionary and thesaurus components are not managed by UIS). With this information, the system can efficiently execute all the indexkit commands.

Figure 5 illustrates the system catalog information for indexkits. It contains two example indexkits (they correspond to the two index examples of Figure 2: FOLLET_EOTN (an indexkit corresponding to the index, having the same name), and SUBJECT_LIB (an indexkit corresponding to the subject index of a library catalog). Indexkits are not implemented in the current prototype.

5. Run-Time Information

In addition to persistent information about each object in the system, during execution, there is a need to track additional information about the state of objects currently being manipulated or accessed by the system. Tracking such



Indexkit System Catalog

Indexkit Name	Intro Name	Intro Set	Index Name	Index Set	Dictionary Name	Dictionary Set	Thesaurus Name	Thesaurus Set
FOLLET_ EOTN	*FOLLET_ EOTN_INTRO*	*FOLLET_ INTROS*	*FOLLET_ EOTN*	*FOLLET_ BOOKS*	*FOLLET_ DICTIONARY*	*NOVEL_ DICTS*	*FOLLET_ THESAURUS*	*NOVEL_ THES*
SUBJECT_LIB	*SUBJECTS_ INTRO*	*LIBRARY INTROS*	*SUBJECTS*	*LIBRARY*	*SUBJECT_ DICTIONARY*	*LIBRARY DICTS*	*LIBRARY THESAURUS*	*LIBRARY THES*

Figure 5 - System Catalog Information for Indexkits.

information is essential to maintaining a consistent system. This information will be particularly crucial in a multi-user environment, when it is possible for different users to try to update the same data at exactly the same time. If the system were keeping no information about objects currently in the system, then it would have no way of preventing different users from updating the same data at the same time; there would be no way to guarantee a consistent system. This section describes what information UIS needs during execution to maintain consistency of the objects.

5.1. Run-Time Information — Indexes

As described in Section 3, the index routines support the notion of a *current* index. What this means in terms of execution, is that a user can have any number of indexes retrieved at a time (i.e. open and accessible), of which at most one may be the *current* index. We adopted the notion of using a tag (unique identifier) to identify indexes that have been retrieved to allow us to quickly access the indexes. As a result, anytime an index is retrieved, an index tag is assigned to it. For each index that is retrieved by the system, the tag must be readily available in order to manipulate the index. This run-time variable is designated by **Index Tag**.

A pointer into the index file must also be readily available to the system if the index is to be accessed at all. Clearly, if the index weren't going to be accessed at all, there would be little reason for the user to retrieve it. Therefore, a file descriptor for each index must also be kept as run-time information while the system is being used. This run-time variable is designated by **F_ptr**.

An index can be retrieved for either modification or read only. There are two pieces of run-time information that need to be kept related to the retrieval mode of indexes. The first is the actual retrieval mode. The system needs to know whether an index has been retrieved for modification or read only in order to prevent the user from trying to modify an index that was retrieved for read only. This is especially crucial in a mult-iuser environment, when more than one user may want to access the same index. This run-time variable is designated by **Mode**. Secondly, the system needs to keep track of whether the index has actually been modified (in the case of retrieval for modification). This information is used in the "save index" command. An index that has been retrieved for modification, but not actually modified does not need to be saved even if the user issues the save index command. Having this information available permits the system to detect these occurrences and not waste its time saving an index that has not actually changed. This run-time variable is designated by **Dirty**. **Dirty** is set to TRUE if the index has been modified, but not saved. **Dirty** is set to FALSE if the index has not been modified since the last time it was saved.

Finally, the system needs to know which indexes that are currently in the system have been created, but not saved. The reason for this is as follows. We cannot guarantee that a newly created index will be small enough to be completely contained in main memory. Therefore, when the user creates a new index, all persistent information is entered into the system catalog and the index files are created. The system needs to be able to distinguish these "created but not saved" indexes from those that either have been recently created but saved, or those that were retrieved. This distinction is necessary because if the user quits the system without saving these indexes, the system needs to know that they are to be deleted. This run-time variable is designated by **Saved**. **Saved** is set to TRUE if the index was retrieved during this user session (i.e. created sometime in the past) or if the index was created during this user session and has already been saved. **Saved** is set to FALSE if the index was created during this user session but has not yet been saved.

The remaining run-time information that needs to be available is the information found in the system catalog. Therefore, a pointer to the system catalog information is also needed at run-time. This run-time variable is designated by **SC_info**. Figure 6 shows the information that UIS needs to manage and manipulate indexes correctly.

Index Tag	Saved?	Mode	Dirty?	SC_Info	F_ptr
i1	*FALSE*	*MODIFY*	*TRUE*	••	*3*
i2	*TRUE*	*READ_ONLY*	*FALSE*	••	*4*

Figure 6 - Run-Time Catalog Information for Indexes.

5.2. Run-Time Information — Indexsets

As defined in Section 2, an indexset is a catalogued group of indexes and sets. Therefore, when an index or set is to be retrieved from an indexset, its system catalog information is found in the catalog components of the indexset (refer to Figure 1). At execution time, the system needs to maintain file descriptors to the catalog components of the indexset in order to be able to retrieve indexes and sets. These run-time variables are designated by **Fd_I_cat, Fd_I_attr_cat**, and **Fd_S_cat**. They correspond to the index catalog, the index attribute catalog and the set catalog components of the indexset, respectively.

If multiple indexes or sets are retrieved from a single indexset, we need to be very careful in making sure that only one set of catalog file descriptors are used for that indexset. If every retrieved index and set has its own file descriptor information for the indexsets catalog, then it would be very easy for the system to encounter read/write conflicts in the indexsets catalog components. Therefore, we need to have a way to maintain a single copy of the indexset information, and still know exactly how many indexes and sets from that indexset are currently retrieved. This suggests a need for run-time variables to count the number of retrieved indexes and sets for each indexset. This has two advantages. First, it prevents having multiple file descriptors to the indexset catalog components and prevents read/write conflicts. Second, it allows us to have the indexset retrieved for as small an amount of time as necessary. By keeping track of how many indexes and sets are currently retrieved, the system is able to return the indexset as soon as those numbers are zero. The run-time variables that designate these counts are **I_count** for indexes, and **S_count** for sets.

The remaining run-time information that needs to be available is the information found in the system catalog. Therefore, pointers to the system catalog information are also needed at run-time. These run-time variables are designated by **SC_info**, **Databook**, and **Indexspace**, which point to the different system catalog entries for the indexset. Figure 7 shows the information that UIS needs to manage and manipulate indexsets correctly.

5.3. Run-Time Information — Indexkits

There is no run-time information needed for indexkits. Because an indexkit is nothing more than a collection of system catalog information, all commands involving indexkits update only this system catalog information. As a result, the catalog is only accessed at the exact moment a request is made. There is no notion of **retrieving** an indexkit, and at some later time making some modification to it.

Fd_I_cat	Fd_I_attr_cat	I_count	Fd_S_cat	S_count	SC_Info	Databook	Indexspace
3	*4*	*2*	*5*	*0*	**		##
6	*7*	*1*	*8*	*1*	**		##

Figure 7 - Run-Time Catalog Information for Indexsets.

Table A: Index Commands

Index Management Commands		
create index	drop index	insert index
update index	move index	delete index

Index Reproduction Commands		
copy index	intersect index	subset index
subtract index	union index	

Index Searching Commands	
find term in index	build set with term
build set with list	build set with range

Index Browsing Commands		
retrieve index	pick index	save index
return index	list indexes	

Index Navigation Commands		
first in index	next in index	fetch using index
last in index	previous in index	
build index boolean	list index booleans	pick index boolean
modify index boolean	drop index boolean	
build index select	list index selects	pick index select
modify index select	drop index select	

Index Run-Time Environment Commands	
bind index column	bind index table

Table B: Set Commands

Set Management Commands		
build empty set	drop disk set	insert set
delete set	update set	

Set Reproduction Commands		
combine sets	restrict sets	sort sets

Set Browsing Commands			
retrieve set	pick set	build empty memory set	save set
return set	list sets	drop set	

Set Navigation Commands		
first in set	next in set	fetch using set
last in set	previous in set	
build set boolean	list set booleans	pick set boolean
modify set boolean	drop set boolean	
build set select	list set selects	pick set select
modify set select	drop set select	

Set Run-Time Environment Commands
bind set column

Table C: Indexset Commands

Indexset Management Commands	
create indexset	drop indexset
alter indexset	move indexset

Indexset Reproduction Commands		
copy indexset	intersect indexset	subset indexset
subtract indexset	union indexset	

Indexspace Commands	
create indexspace	alter indexspace

Table D: Indexkit Commands

Indexkit Management Commands	
create indexkit	drop indexkit
update indexkit	move indexkit

Indexkit Reproduction Commands		
copy indexkit	intersect indexkit	subset indexkit
subtract indexkit	union indexkit	

THE DEVELOPMENT OF A NATURAL LANGUAGE INTERFACE

TO A GEOGRAPHICAL INFORMATION SYSTEM

Sue Walker Toledo, Ph.D.
Senior Scientist, Netrologic
5080 Shoreham Place, Suite 201
San Diego, Ca. 92122

Bruce Davis, Ph.D.
Project Manager
NASA at the John C. Stennis Space Center
Stennis Space Center, Ms. 39529-6000

November 27, 1992

Introduction

This paper will discuss a two and a half year long project undertaken to develop an English-language interface for the geographical information system GRASS. The work was carried out for NASA by a small business, Netrologic, based in San Diego, California, under Phase I and II Small Business Innovative Research contracts. We consider here the potential value of this system whose current functionality addresses numerical, categorical and boolean raster layers and includes the display of points sets defined by constraints on one or more layers, answers yes/no and numerical questions, and creates statistical reports. It also handles complex queries and lexical ambiguities, and allows temporarily switching to UNIX or GRASS.

The Need for More Natural Computer Interfaces

Let us first review some of the more obvious reasons for why one might want to undertake developing a natural language interface to a geographical information system (GIS). More subtle, but very important, reasons for developing simpler interfaces than we have today will come up later.

If we could find a way to enable all computer interfaces to handle our language as we do, there would be a greatly reduced time spent in learning the computer systems we employ, there would be easier communication between separate groups using the same or similar data, and we would be likely to experience less frustration in our work, and generally be more efficient in our use of our computers. The computer would suddenly become accessible to many more people, especially if good speech recognition and machine translation were standard parts of such systems. All we have to do to realize these things, if we are using computers very intensely in our work, is to think of the tomes of manuals we have had to read, and count the number of different programming languages, word processors, network protocols, etc. we have had to learn in our lifetime. It helps to imagine how long it would take to teach a high school student all we know about the computer to do our jobs ...say a sub-Saharan African high school student. Or we can think of the frustration an ARC/INFO user finds when he comes to visit a colleague who uses INTERGRAPH or GRASS or ERDAS. If we use a GIS under UNIX, the manuals for those two pieces of software alone can sometimes consume two long book shelves. In addition, we probably also have hardware, and editor or word processor (and probably C) documentation as well in the same bookcase. It seems no one any more even aspires to knowing every capability of the software systems they use, as many did twenty years ago when systems were less complex. Undoubtedly many errors are made because of the difficulty busy people have in finding the time to learn the complicated systems they must use.

Current Solutions Often Proposed to Solve the Problem: GUIs and Natural Language Interfaces

Today many say that the solution to the problem of the complexity of command line interfaces to the computer (such as DOS, UNIX, C, and the basic programming language of all GISs) are the GUIs - graphical user interfaces. It is important to make a distinction here, however. What a GUI is actually required to be is a multitasking windows environment. It is clear that such GUIs, whose windows are generally now opened and controlled through an interface that is standard across all systems, are very valuable additions to the interface scene. But what goes on in the individual windows are separate processes, each of which may have quite different interfaces.

When people say that GUIs are a solution to the interface problem, they usually have in mind not only that there is a standardized multitasking windows interface being used, but also that there is a menu interface (also following industry-wide standards in its construction) inside the control windows, involving icons and/or simple, easily understood English phrases that the user can choose between by pointing and clicking with a mouse. An excellent menu-driven GUI is also unquestionably a great improvement over a command line interface to computer software of any complexity.

Another solution to this problem that is sometimes proposed is a natural language interface. Of course one would usually want it embedded in a GUI, as ours in fact is: it includes several windows, for instance a GIS display window, and a system operation log window in which the user can watch what is going on behind the scenes if he wishes, and the basic control window, which allows the user to give English language commands or queries, or GRASS or UNIX commands, or to make other choices from a GUI-type menu. The question we want to address in this paper is how valuable the addition of the natural language interface can be.

Thus we will discuss here the relationship between three specific kinds of interfaces that can occur in any environment, 1) menu-driven interfaces, involving choices through icons, slide bars, numbers or natural language phrases, 2) natural language interfaces, receiving questions or commands in full natural language sentences, and 3) interfaces focussing on graphical interactions with the user, e.g. allowing interactions such as the user outlining a region for the system to analyze, or for a display to zoom to, or the user dragging parts of a drawing into other positions within the same drawing in a CAD system. (Note that in our terminology we are not calling any GUI a graphical interface, in spite of every GUI having simple graphical capabilities of dragging things from window to window, allowing the user to point to choices, etc. We will be using the term "graphical" only for interactions that go beyond the ones embedded in the generic GUI.) The three kinds of interfaces have overlapping, but sometimes different, advantages and disadvantages. Since a graphical interface is indispensable whenever the task is primarily graphical, and since graphical interfaces can be integrated with both menu and natural language interfaces, we will focus our attention on comparing menu-driven interfaces and natural language interfaces.

Menu-driven interfaces have the following advantages over natural language interfaces:

a) Their very structures teach the capabilities and the limits of the software to which they interface (very important). Good menu interfaces for systems of limited functionality are easy to learn, and give users a sense of confidence that they understand the system they are employing.

b) The "state-of-the-art" is more advanced, in that standards for menu construction have been established and helpful tools have been developed to make menu interface construction and maintenance easier and less time-consuming than in the past.

c) People have worked out well how to integrate menu interfaces with graphical interactions critical in applications.

d) Sometimes a menu can be structured in such a way as to force users to use the functionality efficiently.

e) Today's menu interfaces generally take much less space and work much faster than natural language interfaces.

f) Menu-driven interfaces embedded in GUIs are sometimes preferred by people who cannot type well, but can nevertheless manipulate a mouse.

However, a good natural language interface would have certain advantages over a good menu-driven interface for the same system. Here are some advantages of natural language interfaces, including some that arise in situations in which a menu-driven interface is not possible.

1) Systems of the future will be larger, and integrate many functionalities, which will mean wide and deep menu structures (thousands of icons are not an easy thing to remember, for instance; and a menu interface built up with hundreds of thousands of words or phrases is bewildering to try to grasp as a whole). In a situation of this kind of complexity, natural language interfaces will also save the user from the tedium of being forced to go down through many levels of menus, or to choose from long lists.

2) If the menu choices involve many submenus and a long succession of different requests, the user can easily loose track of what he is doing, as menus disappear from his screen and his mind begins to forget the sequence of choices he has made. Compare the situation to having a natural language query, or even a number of them, before one's eyes. The idea is that our minds seem to grasp an integrated, "streamed" task request better than one broken into many parts that have to be carried out in sequence.

3) Natural language interfaces allow single-sentence requests which cannot be directly specified through the menu-structure. To accomplish them in a menu-driven system would involve manually working out the equivalents of conditionals and loops, for example.

4) The natural language user doesn't have to think in terms of any specific structure on the overall functionality of the system. He just asks for what he wants in one of the natural ways to request some functionality. (The particular structure the menu builder has put on the functionality may be only one of several choices, and might be in conflict with the user's natural way of organizing the same material in his head.)

5) The natural language user does not suffer the same kind of disturbance when a great deal of functionality is added to the system, while the menu user can see his whole set of menus reorganized (ways of requesting things that have become automatic are now changed).

6) Ways of responding to people with different levels of education (the beat patrolman, the police detective, the chief of police) would not be evident, as in menu interfaces, where the standard procedure is to use completely different interfaces for different groups of people (so that each user can completely understand his interface).

7) Machine translation is one kind of natural language interface to a free-text (or structured) database. The need for this kind of natural language interface becomes increasingly critical as the world comes rapidly together economically and politically.

8) Only a highly developed natural language capacity in the computer will permit realistic virtual reality scenarios involving simulated human speech and the interpretation of real human speech.

9) Without a robust natural language capacity it will be hard for us to easily extract all the information we will need from the large free text databases that will become commonplace soon.

10) Natural language interfaces would very likely allow people to think more freely in their work, with the result that they might be significantly more creative and more efficient.

Strange as it may seem, more people initially resist the idea of natural language interfaces because they say they can't type, or because they can't spell, than for any other reason. (Menu enthusiasts should remember that other users claim more difficulty with mousing than with typing.) Spelling- (and grammar-) correctors and speech recognition can now eliminate such issues. The speech-recognition technology is just coming on the scene, of course, and has quite a few problems of its own, many of which are tied in with the linguistic problems that have a lot to do with what make the state of the art of constructing natural language computer interfaces less "advanced" than that of building menu interfaces.

The great advantage of the menu-driven interface is that it lays out the capabilities of the system for the user (assuming there are clear explanations behind every menu choice, accessible through the help facility of the system, which is not always the case, of course). The best that the most developed natural language systems generally do now is give the user access to a short discussion of system functionality if the user asks a question like "what can you do?" As a result the user of a natural language interface will too often ask the system to do things it has not been designed to handle. Which points out a second important advantage of a menu-driven interfaces - that presumably the system will always be able to do what you ask of it (although if you do not quite understand what you are doing nothing can prevent you from asking for the wrong thing). It is also the case that new users to natural language systems often employ natural language constructions that the linguistic side of the system cannot yet handle, i.e. the limitations on what the system can accomplish through a natural language interface go beyond the limitations of the software behind the interface.

Regarding the teaching capabilities of the two kinds of system, it is obvious that the natural language system builder can put answers to a large variety of questions about the system into his interface, but the teaching side of natural language systems is likely to be very limited for some time, while developers are spending their time working out some of the more critical linguistic problems. This true disadvantage of the natural language interface is closely connected with one of its greatest advantages, however, as I will try to explain in a moment. It would seem (for many reasons) that the ideal interface would be one that integrated all the things we do "naturally,"so that all three kinds of interfaces we are discussing would be included, and generally within the GUI multitasking windows environment. In such a system the user would have the benefit of both kinds of teaching styles just mentioned, the more discursive one that would be natural in response to natural language questions, and the structured one that the menu interface laced with good "help" automatically provides (of course there is discursive help behind each menu item as well). He could also have visual instruction, which would be necessary to teach him about many graphical interactions.

I would first like to make somewhat graphic the first two advantages claimed for a natural language interface over a menu interface. The computer systems of the future are likely to be very large - e.g. the database containing all national hospital information (one for all military personnel now exists). Consider the national spatial data infrastructure database about to come on line. The whole GIS world is attending to the data standards that will be used to make all this data simultaneously available to all the federal agencies that can make use of it. Now consider all the sciences behind the queries that will be asked of this database. The menu structure that would allow all the scientists to *easily* query such a database would be broad and deep indeed, and take a very long time to construct, moreover.

For instance the soil scientist that wishes to get information out of one of the soil layers in the database will not want to first consult a book or computer file giving the names of the layers for the different counties in the United States, then look at the layer containing his data to see the names of the soils in his region (generally on the order of fifty to a hundred different soils), then locate the soil manual for the county to find out which soils have the characteristics that he is concerned with, and finally request a map of the soils in question. He will want a menu structure that will help him do this.

For instance, if he wants to see the wetlands soils in Hancock County, Mississippi, and there is no natural language interface that permits him to just type in that request, then he will want a menu system that will allow him to specify the county he is concerned with (presumably he would first choose the state out of a list of fifty, then choose the county in that state out of a list of similar length), and then that he wants soil information, and then that he wants to see the wetlands areas. Note that if the menu structure is to allow him the choice of requesting the wetlands soils, it will also probably allow him to specify other kinds of soils of interest to soil scientists, e.g. soils with different drainage, erosion, runoff, acidity etc. properties. So this choice part of a menu can get quite complicated. Also the menu must somehow allow him to construct at least boolean constraints involving a number of different layers simultaneously, which adds more complexity.

Thus what points 1 and 2 above are getting at is that to build a menu structure that allows a user to easily query the basic data of a science means to build most of the concepts of the science into the menu structure. And since our personal screen of vision can only scan a limited number of items at once without getting lost, we must have many submenus (or allow typed input in our menu system).

The natural language query "what county in the state has the smallest number of property owners?" would presumably illustrate point 3.

I would like to add a little to the content of points 4 and 10, for they may concern factors of much greater importance than anyone now realizes. Recall that I claimed that one of the main disadvantages of a natural language interface, that a user will often ask for something the system cannot do, is connected with one of its most interesting advantages. The natural language interface user will be likely to spend more time thinking about what he wants the system to give him, about setting his goal, (for that is what one wants to specify in a natural language command or query), while the menu interface user will be more likely to think more about what he can do with the system before him, or about how the system can be used to do the particular thing he wants today. He can actually often leave his goal only partially formed and depend on the path he will be forced to take down through the menu structure to make it more concrete (since he knows he can only do what the menu allows anyway). The menu interface user generally lives intellectually within the confines of his system, and doesn't tend to waste much time questioning its limitations.

The natural language interface user is not so obviously confined, and when he asks for something the system can't yet do, he will tend to immediately think of asking developers to add new functionality. In this sense the natural language interface is an "open" interface in terms of the users' mindset, while the menu interface is a "closed" interface. Except in situations where what is expressed is most naturally expressed in images, drawings, non-language sound, or gestures, it would seem that natural language could express anything expressible in a menu interface. Certainly the expressive powers of natural language are limitless, which the expressiveness of even a command language is limited by the capacities of the hardware it runs on. In any case, it would seem that being accustomed to focusing on one's goals, the object one requires, rather than on "how to" issues, would tend to lead to more creative thought in one's work.

Related to this is that in a situation of a very complex menu structure, where doing anything of much content requires many choices being made, a robust natural language interface that would often allow the user to just simply specify what he wants done in the language that comes first to mind would be greatly appreciated by the scientists using the interface, due to the savings in time involved. Thus a high quality natural language interface might be employed a very great part of the time by highly educated users. (This requires, however, that natural language interfaces be developed in such a way that the scientist can ask for all the information he needs to know to make sure the work he is doing meets the highest standards of his science. For instance, he must be able to ask about the accuracy and precision of the data he is accessing, the dates it was gathered and by whom, precisely how it is being manipulated by the programs responding to his request.)

I would like to give a little more evidence for believing that the freedom of thought that might be more encouraged by an excellent natural language interface could significantly increase creativity in one's work. But first I want to distinguish two different kinds of GIS menu interfaces for you. One is a generic interface, designed to just let you access the basic underlying functionality of the GIS in a clearer way than through short command line function calls (for instance, for GRASS, the excellent interface put out by Osiris). The second is one that has been designed from the point of view of the tasks that the user wishes to carry out. It uses icons or the ordinary vocabulary of the user, and allows the user to specify that tasks be carried out by choosing icons or phrases that are natural for him to use. This would usually be called an application interface. The kind of natural language interface that we developed, and which I am concerned with discussing here, is also an application natural language interface.

(Note that being an "application interface" does not mean that the interface is useful in only one application, but that it is designed to be used by people in applied work. For instance, an interface developed to do environmental monitoring will have to have the vocabulary of all the many sciences that are brought in during the analysis that must be carried out. As we try to use computers to integrate more and more of the data of our society, we will want systems that have built into them more and more of the vocabulary we think in.)

Consider doing something fairly simple through one of the application menu interfaces. The soil science example a few paragraphs back already shows how the menu interface is likely to become tedious when one wants to make a few constraints involving a number of layers (and we started the example several choices down). Having to go to so much trouble to ask for something relatively simple slows the user down, interrupts his normal thought processes (introduces frustration into the situation, which tends to focus attention on the cause of the frustration and take it off the goal of the endeavor).

The situation is worse for the generic menu interface. Say you want to do something that in your GIS takes a hundred lines of commands. The situation is far better than when you didn't have the menu interface, since you now don't have to remember the precise syntax of the many commands in the underlying GIS language. But you are still going to have to execute those same lines of GIS program, just now by choosing more intuitive things from the menu rather than writing cryptic command line statements.

The same problem will hold for the application natural language interface, of course, IF the complex program in question is not using any of the application concepts. Command line programmers, menu users, and natural language users will always try to decompose any long and complex program they have to write to see if there are pieces that will be employed over and over again. These they build into macros in the first two cases (this is often not possible in menu-driven systems, however), while the natural language user will define a new concept (the equivalent of a function or macro name). And as we have more and more macros and functions in the system that a person will want to use, the choice lists for them in a menu will keep getting longer, i.e. the choice is harder to get at. As the only true language manipulating animal, however, we seem to have the capacity to be able to simultaneously remember the meanings of millions of different words/phrases, so that we shouldn't really need to be prompted. We can just say/write what we want. Menus can be built in such a way to help us get to our choice more easily, of course, even prompt us when we can only think of something related to our choice, or let us type in individual phrases. Once this is permitted, then the question again becomes which is faster and more accurate for the user, the menu, or natural language sentences.

An example I like to think about in this command level and generic GIS interface context is the following. Writing a mathematical proof in a formal logical system is very similar to writing a command line GIS program, and proofs so written out are checkable by computers for correctness. But no mathematician (not even mathematical logicians) would consider doing such a thing, because they believe that not only would it increase the time required to write a paper by a very great amount, but would also reduce their creativity to zero. They want to write their papers at the higher level they think in. (Most published papers contain correct proofs, incidentally.)

Most of what I have said so far seems to be weighted towards natural language interfaces. Why are there so few people developing or using them as database interfaces? Well, the problem is that even after thirty years of work, the current parsers, the programs that take an English language query or command and attempt to interpret it for the computer, still do not do very well much of the time. The last thirty years have involved a lot of attention to what most people call syntax, determining the part of speech of the different words in a sentence, finding the subject, the predicate, and the prepositional phrases, and then what those phrases modify. This is not an easy task, incidentally, because many words in English may be of different types of speech in different contexts.

While solving this problem (at least to a great extent; most systems will stumble over some constructions, however), linguists also learned a lot about the varied efficiencies of different methods to do this. But there is not yet any consensus about how to approach the deeper problem of resolving all the subtleties of meaning that different words and phrases can have. Here are some examples of the kinds of ambiguous words of phrases that occur constantly in natural language, making difficult the task of grasping the meaning of a sentence for a computer:

Show the soils *there.*

How many soils are *there*?

Calculate the *area* covered by ocilla soils.

Show me the *area* covered by ocilla soils.

What are the poorly drained soils?

what: list the names or display a map?

poorly drained: all of "excessivelypoorly drained", "moderately poorly drained" and "poorly drained", or just "poorly drained"?

if a display wanted: what colors (default colors of the SCS soil layer, white for the points in question and black elsewhere, different colors for the three kinds of poorly drained soils and black elsewhere, or some other color scheme)?

Show me the area *near* the test stand where acid deposition was less than twice the average acid deposition at the test stand today.

Show me the area *near* Chicago where acid deposition is less than twice the average acid deposition in Chicago today.

The better quality natural language interpreters today do have means built in to engage the user in dialogue about ambiguities, which is invaluable to the natural language interface user (and which is a technique we humans use to resolve ambiguities in conversation far more than we are aware of), but they do not have enough broadly applicable techniques for resolving the masses of ambiguity we resolve (or realize it is not necessary to resolve) without dialogue.

In addition to getting in trouble with the many ambiguities involved in natural language, there is another major stumbling block in the way of natural language interface developers. A natural language interpreter requires constant maintenance of its own vocabulary/database-connections-to-the-vocabulary database, unless it is dealing with a completely static database. It is this, especially when added to the fact that the systems do not always perform linguistically the way one would like them to, which has probably prevented a wider use of natural language interfaces.

The Place of Natural Language Interfaces Today

The question then is, in these early stages of the development of natural language interfaces, should someone with limited funds attempt to use this technology. Also, is the technology sufficiently important that the government and industry should finance any significant amount of research and development in this area? I tend to think the answer is yes, to both questions (in the first case, though, only for the relatively affluent user who can afford a talented person to maintain his natural language interpreter's database), and for the following reasons.

A good natural language interface would be a very valuable thing, for the many reasons discussed above, and the only way we can obtain such systems is by serious research, for which realistic natural language commands/queries is essential. It is very difficult to gather a realistic corpus of such linguistic input data without putting a system out in the field with a reasonable amount of functionality (which is possible now) and then obtaining feedback from users about what is lacking.

Far better, though, would be to get users, before directly interacting with the computer, to give to a tape recorder the commands/queries they would really like to ask the system, even though they may know already that with their current system that would not work; for in this way linguists would not only get more detailed linguistic input, but also have both natural language and speech recognition input data for the database interface research. I believe that it is only in the context of such a functional natural language interface that the kind of detailed feedback can be obtained that is needed to set priorities on the linguistic research that must be done.

Moreover, in many situations a well-developed system based on the current technology will be easier for the user to use than a menu interface devised to address the same tasks and the same data using the same set of concepts. An attempt should be made to develop fully integrated interfaces from the beginning (menu, natural language, and graphical combined), so that we can get a better feeling for the strengths and weakness of the different aspects of the integrated interface. Moreover, if natural language is to be used, it is essential to provide such interfaces now, while the state of the art in the natural language area is in the process of active development, so that users can have something easier to use than command line programming when the natural language interface is incapable of doing what they would like to ask of it. I would predict, moreover, that we would learn to construct even more flexible menu interfaces if we created some within integrated systems where the goal was to always come as close as possible to the ease of the natural language interface it is paired with. An integrated interface such as this will be essential whenever truly graphical functions are required by the system. Moreover there will always be subsets of functionality that will be much easier to do through menu interfaces than by natural language, no matter how robust natural language systems become (for instance, situations in which a small number of finite choices are involved; examples are the ATM machine, or choosing formatting fonts and margins, or specifying map colors, the latter of which can be done very elegantly with sliding bars that change the color until it is the one you want ... but don't you really wish those phone answering machines that take you down through many submenus could be replaced by a machine that could just understand when you say what you want?!).

The Netrologic Natural Language Interface to GRASS

We hope that the system we are in the process of developing, an English-language interface to the GIS GRASS, will have sufficient functionality so that it can soon be a working prototype out in the field gathering this kind of input data for further developments. Let me now tell you a little more about it. The current and near-term functionality of the interface system is the following:

Displays points sets defined by a set of constraints

Answers yes/no and numerical questions, and prints reports

Handles categorical, numerical, and boolean data

Handles ambiguity in words, e.g. "area"

Changes map windows to fixed regions, e.g. "quadrants"

Handles complex queries, including commonly used functions

Permits UNIX and GRASS commands instead of an English query

Integration of Graphical Features of GRASS, e.g. finding of coordinates and specific data values

The developers have thought about many long-term extensions of the system that would be valuable to users, for instance the inclusion of a library of functions commonly used within the GIS community, and building various expert-system features into the system. The first extension of this kind should of course be to extend the system to cover all of GRASS's basic functionality. The basic system design also endeavored to make the system amenable to modification to enable it to be serve as an interface to GISs other than GRASS.

The Basic System Design

The system was constructed by modifying PARLANCE, a commercial natural language interface to relational databases developed by BBN, so that it no longer generates SQL (the standard command language for relational databases), and then building a bridge between the meaning representation language (MRL) output of its linguistic interpreter to programs in the GRASS command language.

The basic system design can be described by the following components, through which the data successively flows:

- The PARLANCE user interface (modified)
- The PARLANCE English query interpreter (truncated)
- A transformation module, producing "SpatialFlattenedMRL"
- The SFMRL interpreter, which translates SFMRL into calls on the GRASS drivers
- GRASS drivers
- GRASS (calculations, displays, etc.)
- BBN's output handling (in the case of a textual response)

Goal of the Natural Language Interface Community

In conclusion, let me mention that the following goals for current and future developments of a natural language capacity in the computer appear to have been taken by workers in the NL/AI community. A large number of papers and even systems exist illustrating directions proposed for addressing one or another of these points. It is possible that more attention to the first one as the appropriate starting point might be helpful.

- Gather a large representative sample set of the queries and commands people would really like to be able to make, for each discipline and task, to enable linguists to be able to do the appropriate linguistic research. Spend a large amount of time on a "preanalysis" of this data, deriving from it the most critical problems to be addressed, and a prioritization of these problems.

- Develop an appropriate meaning representation language (MRL) into which queries and commands can be translated; if possible this should be a universal "interlingual" that would work for all the computer natural language applications, and be independent of the natural language in question.

- Expand the linguistic community's understanding of the use of language so that the queries and commands a user would like to employ can be "parsed" into the meaning representation language in a way adequate to the user's needs (possibly only after some dialogue with the user).

- Work out how to use common and scientific knowledge, as well as computer-generated models of individual users, for reasoning and other "expert system" purposes, as well as to aid in the parsing of queries and commands, and determine where to embed this knowledge in the NL systems that are used.

- Develop systems that are extremely flexible, can easily change as the language changes (a very large amount of new scientific and technical vocabulary enters each natural language every year). Develop systems where the different natural language components - speech recognition, machine translation, free text database retrieval, GIS, relational, and object-oriented database interfaces, etc. - can be changed simultaneously as the state of the art in natural language interpretation develops.

- Find a way to develop systems whose underlying structure is sufficiently easily understandable that they can be maintained without great difficulty as the language of users and the underlaying functionality of the systems change.

- Natural language capabilities should be integrated into easily usable systems along with many other access techniques that are being developed.

- Learn how to develop systems that can teach users their capabilities and how to use them.

You Can Say the Same Thing Many Ways in Natural Language

(And spell-correcters and speech-recognition help with the typing problem.)

display the soils
show the soil types, please
will you please show the soil layer
give me the SCS layer
display the soil data
display the data on soils
show the soils in the region
paint the soils here
display the scs layer **ETC.!**
show the soils data
give me a soil map
map the soils
paint the soils for me
give me a map of the soils
what the hell do the soils look like
show the soils in the area
show the soils of the area

show the terrain between 5 feet and 10 feet in altitude
show me elevations ranging between 5 and 10 feet
show me where the elevation ranges between 5 feet and 10 feet
display the land which is between 5 feet and 10 feet in altitude
display places of elevation between 5 feet and 10 feet
display points where the elevation is larger than 5 feet and smaller than 10 feet

show any soil that is atmore or ocilla
display all soils that are atmore or ocilla
show the atmore soils and the ocilla soils

vc6

Parsing: **SHOW POORLY DRAINED SOILS WHOSE ELEVATION IS NOT GREATER THAN THE AVERAGE ELEVATION**

MRL:

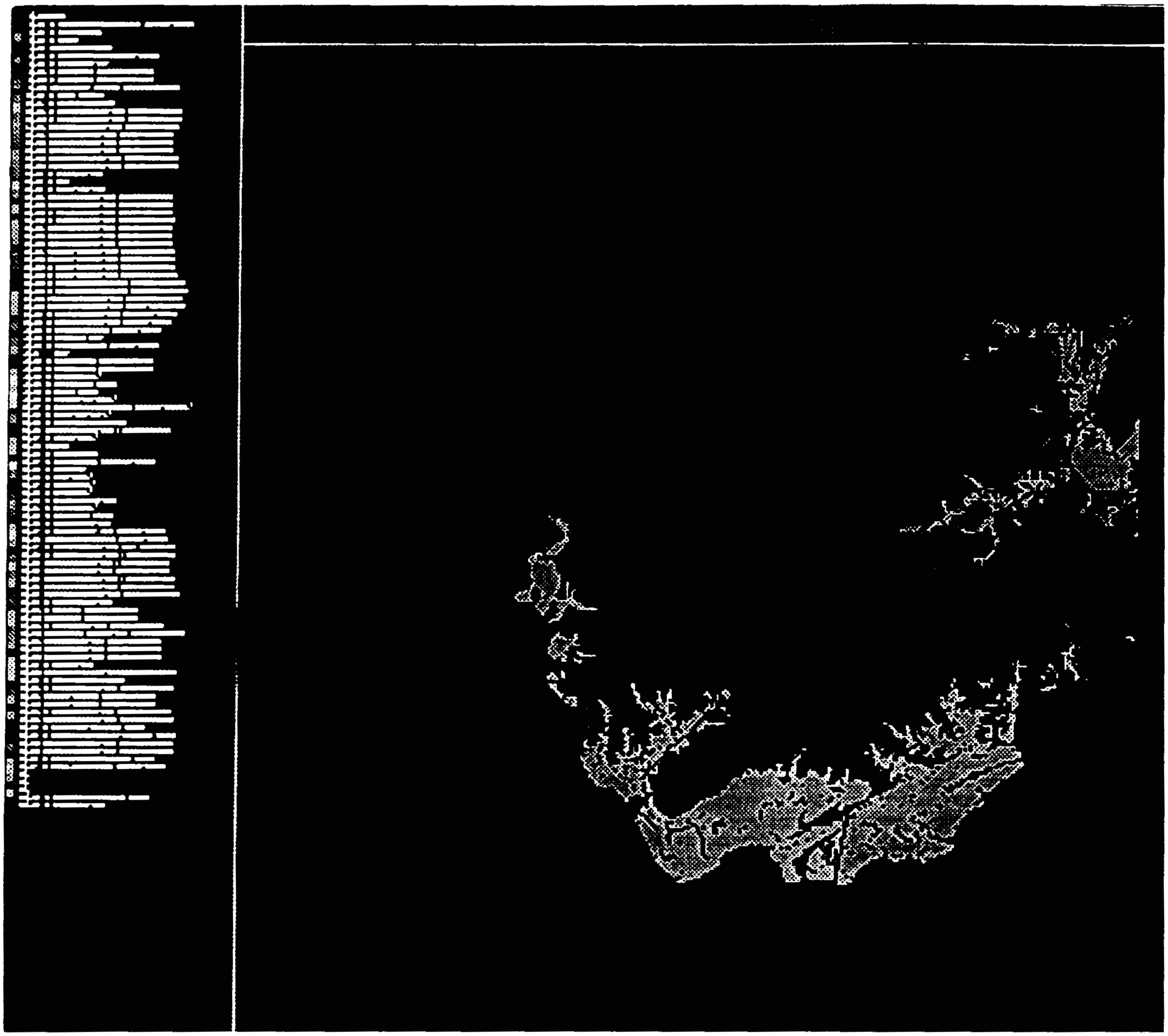
```
(FOR SOME
   X.125
   /
   PNT
   I:I
   (AND
    (FOR SOME
       X.157
       /
       SOIL-CLASS
       I:I
       T
       I;I
       (AND ( I POORLY DRAINED I  X.157)
          (EQ (PNT-SOIL-OF X.125) X.157)))
   (NOT
    (FOR THE
       X.153
       /
       LENGTH
       I:I
       (EQ (PNT-ELEVATION-OF X.125)
         X.153)
       I;I
       (FOR THE
         X.149
         /
         (GENERATE AVG
                X.150
                /
                PNT
                I:I
                NIL

                 I;I
                (PNT-ELEVATION-OF X.150))
          I:I
          NIL
          I;I
          (GT X.153 X.149)))))
 I;I
 (DISPLAY (PNT-SOIL-OF X.125)))
```

Parsing: **SHOW POORLY DRAINED SOILS WHOSE ELEVATION IS NOT GREATER THAN THE AVERAGE ELEVATION**

GRASS COMMANDS and RESPONSES

```
g.region      -d
g.remove      MASK > /dev/null
r.stats       -cm input=elev15q.20m, output=/netro/n.sum
r.stats:  complete ...  100%

Waiting for completion...


g.region      -d
g.remove      MASK > /dev/null
r.mapcalc     < /netro/n.tbl
echo          1=1 | r.reclass in=n_result out=MASK title=MASK
d.mon         select=x0
d.erase
d.colormode   fixed
d.frame       -s frame=image
d.rast        -o map=han.sol
d.frame       -s frame=legend
d.erase
d.legend      map=han.sol
d.frame       -s frame=image
d.mon         unlock=x0

PARSING EXPRESSION ...
EXECUTING n_result = ...  100%
CREATING SUPPORT FILES FOR n_result
minimum value 0, maximum value 1
expression stack size 13, execute stack size 3
```

"Show Poorly Drained Soils whose Elevation is Not Greater than the Average Elevation"



J.L. Star Aug 92

omit

DATA MANAGEMENT, STORAGE, AND PROCESSING
PART 1

AN APPLICATION PROTOCOL FOR CAD TO CAD TRANSFER OF ELECTRONIC INFORMATION

Charles C. Azu Jr.
Engineer
Naval Command Control, and Ocean Surveillance Center
Research, Development, Test, and Evaluation Division
San Diego, CA 92152

ABSTRACT

The exchange of Computer Aided Design (CAD) information between dissimilar CAD systems is a problem. This is especially true for transferring electronics CAD information such as multi-chip module (MCM), hybrid microcircuit assembly (HMA), and printed circuit board (PCB) designs. Currently, there exists several neutral data formats for transferring electronics CAD information. These include IGES, EDIF, and DXF formats. All these formats have limitations for use in exchanging electronic data. In an attempt to overcome these limitations, the Navy's MicroCIM program implemented a project to transfer hybrid microcircuit design information between dissimilar CAD systems. The IGES (Initial Graphics Exchange Specification) format is used since it is well established within the CAD industry. The goal of the project is to have a complete transfer of microelectronic CAD information, using IGES, without any data loss. An Application Protocol (AP) is being developed to specify how hybrid microcircuit CAD information will be represented by IGES entity constructs. The AP defines which IGES data items are appropriate for describing HMA geometry, connectivity, and processing as well as HMA material characteristics.

INTRODUCTION

There exists today within the Microelectronics industry a variety of established ECAD (Electronic Computer Aided Design) systems. These systems all have their own proprietary formats for representing ECAD information. To communicate with another ECAD system, design information must be converted to a neutral format. The data is then transferred to the other system which in turn translates the information from the neutral format to its own proprietary format, figure 1. This process is executed everyday within an engineering company, a company's engineering department, between a design organization and a manufacturing organization, and between a customer and a fabricator. Unfortunately, this process is not robust, numerous errors occur during the translation portion of the processes. Errors are often in the category of missing, incomplete, or extraneous information, see Table 1. As a result, the design file received into the receiving CAD system must often be edited or updated. The update process consist of returning the file into a robust state. The goal is to have the transferred file be equal (functionally and informationally) to the original file. This can often be a very tedious, expensive and time consuming process for larger CAD files depending upon the extent of repair to be done.

Table 1
Typical Transfer Problems Using IGES

1. Loss of information on different layers.
2. Loss of dimensional intelligence.
3. Alteration of text and line fonts.
4. Loss of non-geographical information.
5. Loss of connectivity information
6. Loss of components configuration info.
7. Loss of routing information.

The U.S. Navy must often bear the final cost of the problems its manufacturers/suppliers have in transferring CAD information. For this reason, the U.S. Navy, through its MicroCIM project office at NCCOSC RDT & E Division, decided to investigate this problem. The MicroCIM program was charged with working with the military hybrid microcircuit assembly (HMA) industry to implement/develop new technology. One such technology is the errorless transfer of hybrid microcircuit ECAD information between dissimilar CAD systems. A method for achieving this exchange using an established neutral format has been developed. The neutral format chosen is IGES (Initial Graphic Exchange Specification) for reasons which will be discussed later. The method was put in the form of an Application Protocol (AP), so called because the method is a protocol for applying IGES in the successful transfer of CAD information. The AP is intended to be used by manufacturers of ECAD systems and software when building their next generation systems[1]. The AP details to the manufacturer how to represent hybrid design constructs in the IGES format. It standardizes the IGES representation of a hybrid microcircuit assembly CAD file. This standardized method of representing HMA design file entities will allow the errorless transfer of HMA ECAD files. Referring to Table 1 it is seen that the majority of errors are rooted in the lack of standardization in the representation of HMA ECAD file constructs when using neutral formats.

The remainder of this paper will present some background information and then explain the AP, how it was developed, and how it can be used.

BACKGROUND

HMAs

The focus of the AP is on the electronic information necessary to fully represent hybrid microcircuit assemblies. Generally, HMAs are non-monolithic integrated circuits, made up of two or more different technologies, and may consist of semiconductor chips and capacitors attached to a ceramic substrate with printed resistors and interconnections[1]. This is the basic definition which is used in the AP. This definition is meant to be inclusive of Multi-chip Modules, thick film HMAs, thin film HMAs, and low temperature co-fired ceramic (LTCC) HMAs.

IGES

As stated earlier, IGES is a neutral format specification for describing electronic information such as CAD files. IGES is an acronym for Initial Graphics Exchange Specification. It is a specification which had it first release in the early 80s. The purpose of the standard is to provide a means by which to represent and communicate product definition data in a digital format. IGES has grown to be inclusive of almost all types of production definition data, especially CAD/CAM information. This data can be in the form of engineering drawings, documentation, 2D & 3D designs, and solid models.

In the ECAD world there are several existing neutral file specifications for various areas of electronic information[1]. Two such specifications used in the analysis and hardware areas respectively are EDIF (Electronic Design Interchange Format) and VHDL (VHSIC Hardware Design Language). IGES was chosen over EDIF and VHDL for implementation in the AP for several reasons; 1)It is a standard format available in the majority of CAD systems, ECAD, drafting, or other, 2)It is widely used in industries for transferring design file between machines, 3)IGES is a very flexible language with multiple ways to define entities, and 4)It can readily represent information within the scope of the AP.

To put ECAD information into an IGES format a translator is required. The translator operates by mapping information contained in a proprietary ECAD database into the IGES format[2]. The mapping can be in either binary or ASCII where the ASCII generates a readable IGES file. The IGES file structure contains five distinct sections. The Start Section contains 72 columns of human readable comments which are not processed by the program. The second section is the Global Section which is a free format area specifying the information needed by the pre-processor and information needed by post processor to manage a file. The Directory Entry (DE) section and Parameter Data (PD) sections are usually the largest sections in the IGES file. The DE section contains the descriptive attribute data for each entity used in the original file. The Parameter Data (PD) section follows, and it contains entity definition and actual parameters for each of the entities in the DE section. The last section in an IGES file is the terminate section which contains a single record that has the count of the records in each previous section. The IGES version 5.0 manual has more detailed information about this as well as detailed information on current entities supported by IGES.

APPLICATION PROTOCOL (AP)

An AP, in its most generic form, is a protocol for applying some type of information or technology[1]. In our case, we describe how to apply the IGES neutral data format for representing HMA ECAD information. This AP develops a standard representation for HMAs so as to minimize cost, maximize efficiency in the design process, and provide a means for handling the increasing complexity of HMAs[2]. The procedure used in this AP (and similar APs) involves identifying the information required to fully describe an application area (HMAs) and representing that information in the form of a conceptual model. This model is then used to select the appropriate IGES constructs for representing the information.

Our AP is centered around three models: AAM, AIM, ARM. The AAM, Application Activity Model, presents the generic activities needed to design and fabricate HMAs. The ARM, Application Reference Model, represents the information needed to support the AAM activities or the information generated from those activities. The physical location of the information contained in the ARM can be found in the AAM. The AIM, Application Interpreted Model, specifies the constructs of a standard, such as IGES, for use in transferring some to all the information described in the ARM. Together, these models define the appropriateness of IGES constructs for describing the geometry of the various parts of a hybrid microcircuit, its inner connectivity, and processing and material characteristics.

The scope of the AP is to support design, fabrication, and final assembly information for an HMA[1]. The AP does not support all information required for electrical testing of HMAs. The information contained in the ARM limits the AP scope to layered electrical products information which is currently contained in ECAD systems. Other sections of the AP describe a) definition of the terms used in the AAM, ARM, and AIM, b) implementation and conformance test guide lines, and c) AP relationship to Units of Functionality.

Modeling Methodology

The AAM and ARM were developed using IDEF methodology in order to represent the information being conveyed to the reader. IDEF was developed through the Air Force's Integrated Computer Aided Manufacturing Definition Program. The AAM is built using IDEF0 which is an activity modeling method. The ARM is built using IDEF1X modeling method which is an information modeling method. The AIM modeling method was created specifically for this AP and is based upon various modeling techniques. The component parts of IDEF0 and IDEF1X models are shown in figures 2a and 2b respectively. IDEF0 models are composed of ICOMs, arrows, and boxes. Each activity or function is represented by a box which takes in any combination of Inputs, Controls, Mechanisms, and Outputs through arrows. Each activity can be decomposed into further activities. An entire IDEF0 model is a hierarchal representation of a process composed of activities and functions. In each sub-level are the activities making up an upper level function. Arrows pass information, data, and product between levels as necessary.

In the IDEF1X method a piece of information is represented as an entity, a relationship, an attribute to an entity, or some type of assertion[3]. The IDEF1X structure is top-down where top entities (objects) are composed of bottom entities. Entities are represented by rectangles as shown in figure 2b. The syntax for describing relationships between entities is also shown. Entities which are beyond the scope of the model have a dashed rectangular outline. These entities are in the model to complete an open relationship or clarify a relationship.

Application Activity Model

An organization intending to implement this AP would look at the AAM to see if their information is within scope and within the context needed for planning the necessary automation changes[1]. The viewpoint of the model is from that of designers and manufacturers of hybrid microcircuit assemblies. The model is meant to be generic, i.e. it is not specific to a particular manufacturers operations. Unfortunately, the generality of the model leaves many open issues. For example, the model as it stands, applies to MCMs, thick film hybrids, thin film hybrids, etc. The fundamental differences between these technologies is not represented in the AAM. The other AP models, especially the ARM has facilities for differentiating between various hybrid technology types. The AAM shows where the information in the ARM is used.

Figure 3a and 3b show model diagrams page A-0 and A0. These are the first and second level diagrams which present the major activities necessary to produce an HMA. The A-0 shows the basic inputs, controls, and mechanisms required to produce the various outputs from a manufacture hybrid devices activity. The inputs are physical things such as Supplies & Materials and Industry Technology as well as information from Customer Requirements. Controls on the activities are documentation like military, industry, and company standards. Controls are usually those things which are not changed in any form by the activity they enter into. The outputs are not only Shipped Hybrids but also Scrap generated in production process, Prototypes built before production and required to be delivered to the customer, and response to the customers request for price quotes.

The A-0 activity is decomposed into four activities which are the core of an HMA manufacturer's operations. The first activity is the Management Of Customer Orders which uses the Customer Requirements from diagram A-0 to generate a Quote Response. The second activity is Performs Engineering which uses Industry Technology and Supplies & Materials to produce Prototypes in accordance with Standards and Customer Requirements. Data generated from prototype fabrication as well as Scrap Information is used to produce various engineering documents. This activity also produces drawings, schematics, layouts, released design, etc. The third activity Assure Product Quality, takes in drawings and other documents from Perform Engineering and Customer Requirements information to produce a quality plan. Production data from Produce Hybrids is analyzed using statistical methods and results are fed into Produce Hybrids and Perform Engineering. The final activity is the actual production of hybrids. Supplies and Materials are taken in and the hybrids along with documentation are produced according to the released design drawings and in keeping with standards. Scrap and Production data are also generated. The remainder of the AAM in the AP is composed of decompositions of A0 activities to various levels.

The AAM was arrived at by consulting previous AAM models built under Navy contract by various HMA manufacturers. Active participants in the building of the AAM were the Navy and two major military HMA manufacturers. Agreement of the AAM was received from the US Navy's MicroCIM program Ad-Hoc Advisory Panel, a group composed of government, industry, and academia interested in HMAs.

Application Reference Model

The ARM describes the hybrid product information. The model presents an enterprise-view of information of the hybrid as a product[1]. The ARM is a reference point for implementation of the AIM. It shows how various types of product information relate to one another and how a particular piece of information fits into the concept of an HMA. The documented information as presented in the ARM supports the activities of the AAM. It also provides the baseline for the development of the AIM.

Figure 4 presents the top most diagram of the ARM for HMAs. This page in the model can be read as follows (refer to figure 2b):

> The highest level entity in the model is the Hybrid CAD Presentation. This entity has one key attribute. The key attribute uniquely identifies every instance of the entity. The other attributes are characteristics of a Hybrid CAD Presentation such as; layers of an HMA are built on separate CAD Layers. The connection between the entities Hybrid CAD Presentation and Hybrid Version can be read: Hybrid CAD Presentation is a CAD design of zero, one, or more Hybrid Versions. A Hybrid Version is uniquely identified by an attribute called Hybrid ID. The Hybrid Version was designed using zero, one, or many Design Rules and a Design Rule is involved during the design of zero, one, or many Hybrid Versions. The dotted line between these two entities indicates that they are not dependant

upon one another. A Hybrid Version is zero or one Assembly Occurrence and contains zero, one, or many Assembly Occurrences. An Assembly Occurrence is uniquely identified by an Assembly Occurrence ID, it also has an attribute representing various types. An Assemble Occurrence is dependant upon its relationship with Hybrid Version. An Assembly Occurrence involves zero, one, or more Process Steps. A Process Step instance is uniquely identified by a Process Step No. and has Station, Process Description, and Log Requirements as attributes. The Process Step is dependant upon the relationship it has with Assembly Occurrence. The remaining entity to entity relationships for Process Step can be read as follows. A Process Step is produced using zero, one, or many Tools. A Process Step is used in one or more Assembly Consumables. A Process Step utilizes zero, one, or more Patterns. A Process Step is followed by zero, one, or many Process Steps. A Process Step has attached zero, one, or many Hybrid Assembly Components. A Process Step achieves an assembly using zero, one, or more Process Operations. The entities Hybrid Assembly Component, Pattern, and Process Step are dependant upon their relationship with Process Step.

The remainder of the diagram can be read as above. As stated previously, the ARM is the baseline from which the AIM is developed. The AIM shows how the information contained in the ARM is to be expressed by subsets of IGES entities.

Application Interpreted Model

The scope of the AIM is limited to LEP (layered electrical products) information which most ECAD systems contain. HMAs are a subset of the wide range of LEP types (ie. MCMs, Printed Circuit Boards, etc.). The IGES entities selected for implementation in the AIM were selected so as to minimize the total file size. The selected IGES entities have restrictions placed upon their use either through the Global, Direct Entry, or Parameter Data sections. This is done so as to restrict the number of different ways a particular entity is used within an HMA CAD file. Other IGES entities can be used within a file but they should not be used for purposes stated in the AIM[1]. Table 2 is a subset of the selected IGES entities. The Type and Form headings are IGES numbers set by the standard itself. They are listed so that an implementer of the AIM can refer to the standard for specific information on the entity. The Status field describes the entities current status. Standard means that the entity exists and does not need to be modified to be used in the AIM. Gray means that the entity is located in the Gray pages of the current IGES version document. RFC (Request For Change) means that the entity is either new or needs to be modified and an RFC exists and is in the ballot process. New means that the entity does not exist and an RFC needs to submitted. Modified means that an existing entity needs to modified to be used in the AIM. The AIM individual object definition entity models contain usage restrictions appropriate to the application. These restrictions are described in detail in the AP with the object models. Figures 5a and 5b are two sample object models from the AP.

Table 2
A Sampling of IGES Entities used in AIM

Status	Type	Form	Description
Standard	100	0	Circular Arc
Standard	102	0	Composite Curve
Standard	106	63	Copious Data
Standard	124	0-1	Transformation Matrix
Modified	125	All	Predefined Planar Shape
Standard	312	1	Text Display Template
Modified	402	18	Flow Associativity
New	402	5xxx	Net Connectivity Assoc.
Grey	406	27	Property- Generic Data
New	406	5xxx	Property- Region Fill
RFC	406	5xxx	Property- Definition Extent

The graphic notation developed for the AIM object models is meant to ease the development of unambiguous translators conforming to the AIM. The notation is composed of several principle elements; Object Definition Block, Object Instance Block, Object Value Block, and Cardinality code. The latter three elements are related and derived from the Object Definition Block which designates an IGES entity type, form, directory entry value, parameter data values, and relationships to other IGES entities. Definitions for the other graphic notations in the AIM can be found in the AP.

For readability, the diagrams in the AIM are divided into six subsections. Section one contains the AIM interface object models. These represent a perspective of an LEP in which one can exchange data. The interface object models describe the set of independent entities in a IGES file which are part of an LEP. Currently, there exist three interface objects in the AIM; Part Library, Physical Layout, and Technical Illustration.

Section 2 defines objects specific to LEPs. Display Geometry is section 3 and defines objects that are common to CAD/CAM systems that use 2D geometry. A miscellaneous section contains subordinate objects which are used in combination to form an LEP specific object. There is also a section that defines objects referenced from the Direct Entry sections of other objects. In the final section are objects that represent pre-defined Direct Entry values. Figure 5a is an example of a model from the Interface Section, figure 5b comes form the Display Geometry section.

INDUSTRY IMPLEMENTATION

As stated in the introduction it will be up to private industry to implement the AP. Specifically it is expected that ECAD system manufacturers such as Mentor Graphics, Intergraph, Cadence, Harris, and Computer Vision will implement the AIM in their next generation of translators for ECAD systems. To successfully conform to the AP, these vendors must design their ECAD system translators to be capable of reading and writing CAD/CAM files that conform to the AIM. The designers and manufacturers of HMAs can then use these systems without having to worry about the cost and loss in efficiency currently inherent when transferring CAD files between dissimilar ECAD systems and sometimes between the same type of system. The U.S. Navy ,by building this AP, has served as a catalyst for a solution to the file transfer problem. It is now up to the HMA industry to demand the implementation of this solution from ECAD system manufacturers.

FUTURE DEVELOPMENT

Conformance requirements and testing applications have not yet been fully developed for the AP. It is hoped that industry will take on these tasks as part of a continuing effort to improve this Application Protocol for hybrid microcircuit assemblies.

REFERENCES

1. Parks, C., McCollough R., et al., "IGES Hybrid Microcircuit Application Protocol (AP) version 1.0", NIST TN 1295 Draft, July 1, 1992

2. Parks, C., McCollough R., et al., "IGES Hybrid Microcircuit Application Protocol (AP) version 0.1", NIST TN 1295 Draft, April 1, 1991, NOT AVAILABLE

3. Yuhwei, Yang "IDEF1X Style Guide For PDES Users", Draft, August 31, 1989

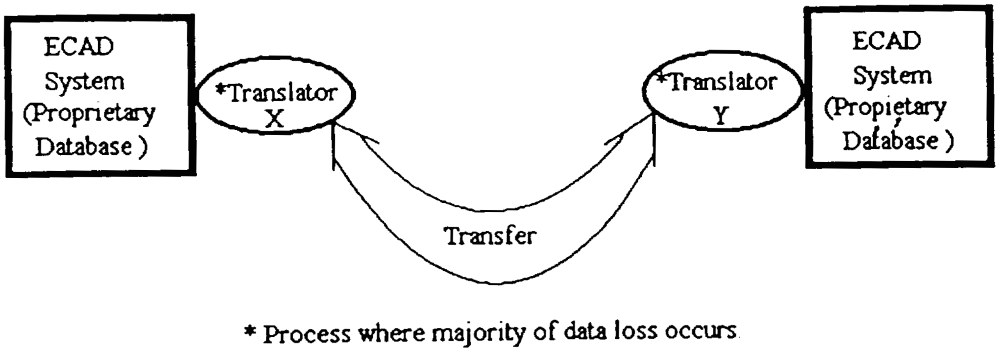


Figure 1: Electronic CAD file transfer process

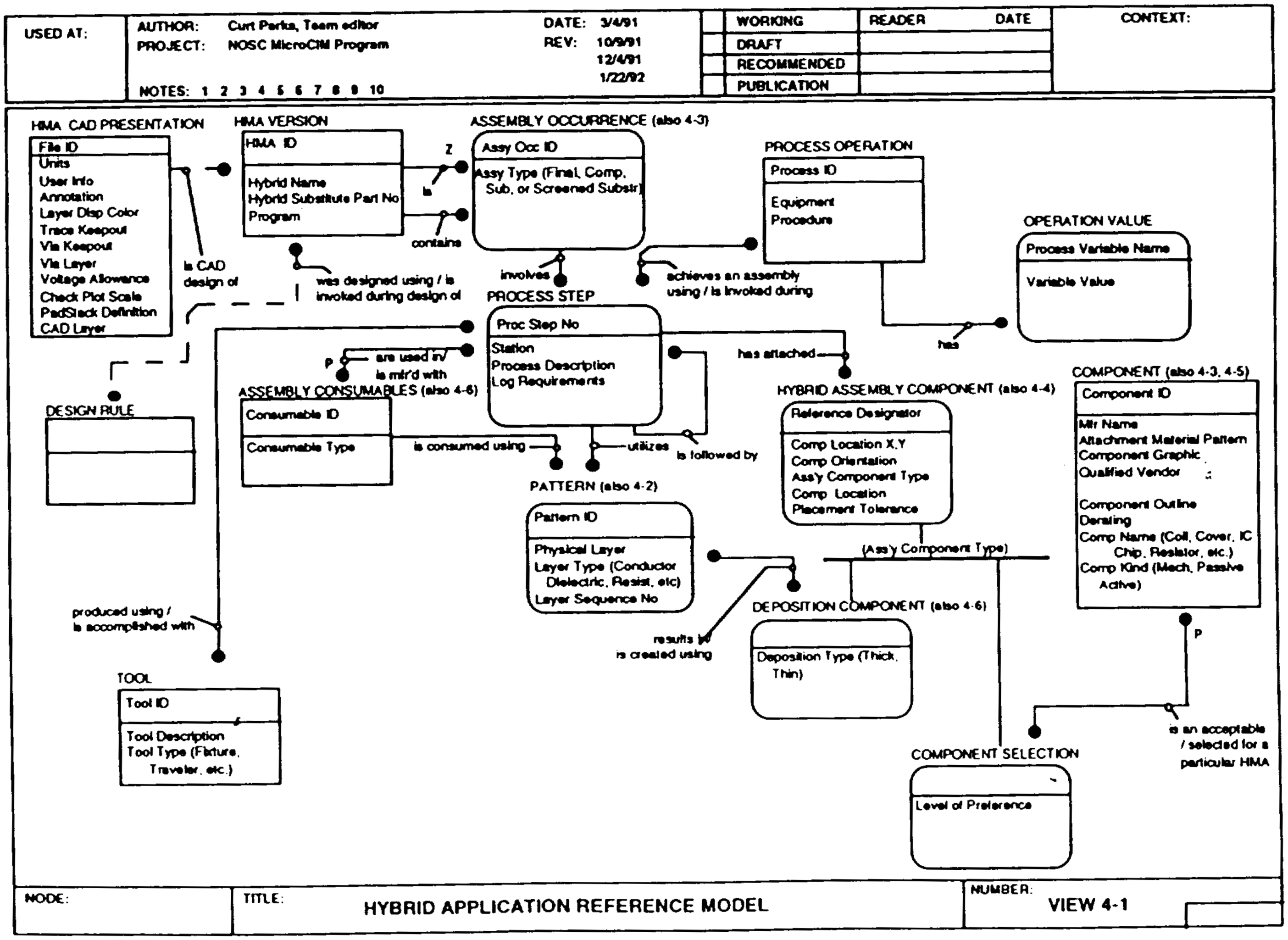


Figure 4: ARM diagram, top of model

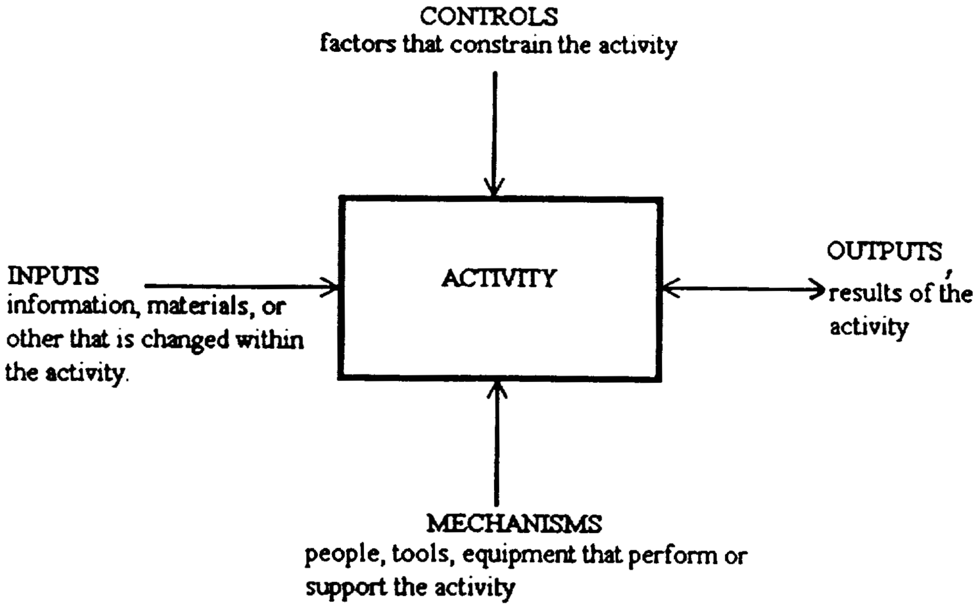

CONTROLS
factors that constrain the activity
INPUTS
information, materials, or other that is changed within the activity.
ACTIVITY
OUTPUTS
results of the activity
MECHANISMS
people, tools, equipment that perform or support the activity

Figure 2a: IDEF0 Methodology Diagram

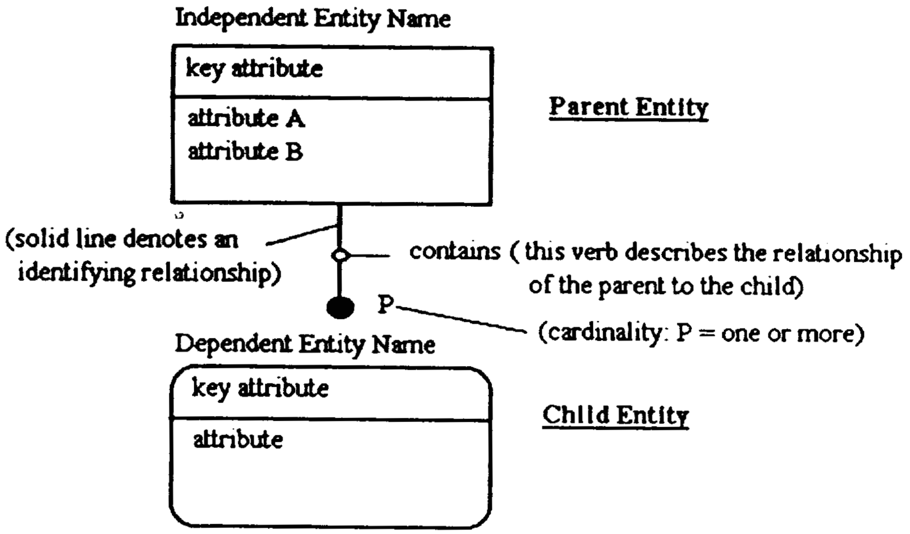

Independent Entity Name
key attribute
attribute A
attribute B
Parent Entity
(solid line denotes an identifying relationship)
contains (this verb describes the relationship of the parent to the child)
P
(cardinality: P = one or more)
Dependent Entity Name
key attribute
attribute
Child Entity

Figure 2b: IDEF1X Methodology Diagram

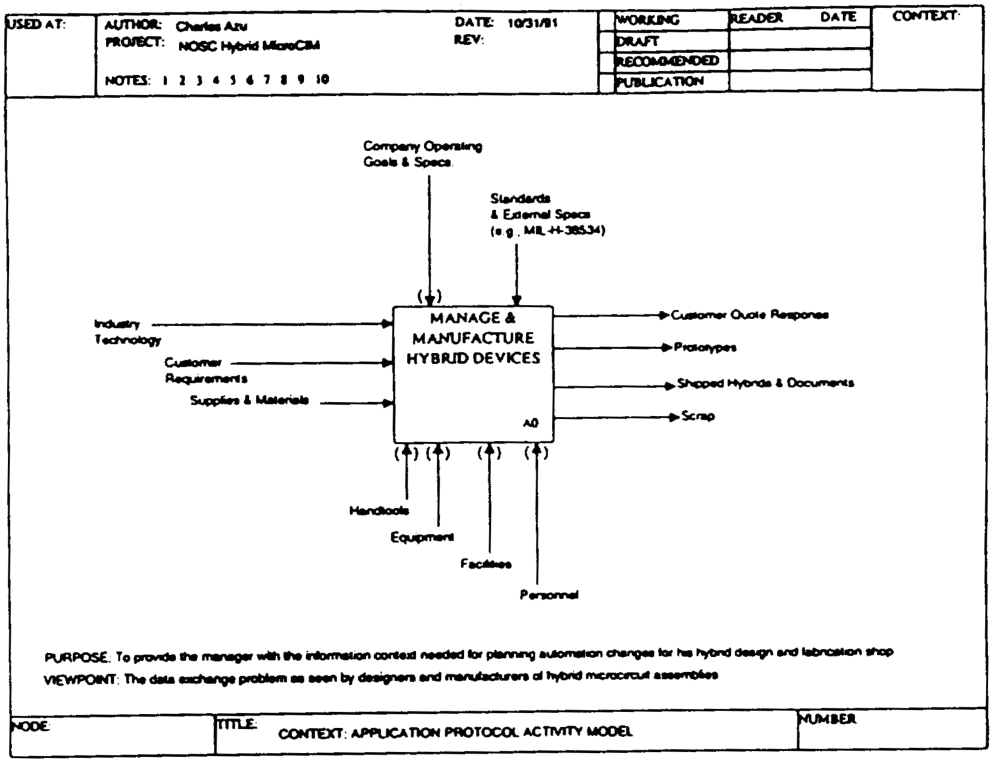

USED AT:
AUTHOR: Charles Azu
PROJECT: NOSC Hybrid MicroCIM
DATE: 10/31/91
REV:
NOTES: 1 2 3 4 5 6 7 8 9 10
WORKING
DRAFT
RECOMMENDED
PUBLICATION
READER
DATE
CONTEXT:
Company Operating Goals & Specs
Standards & External Specs (e.g., MIL-H-38534)
Industry Technology
Customer Requirements
Supplies & Materials
MANAGE & MANUFACTURE HYBRID DEVICES
A0
Customer Quote Response
Prototypes
Shipped Hybrids & Documents
Scrap
Handtools
Equipment
Facilities
Personnel
PURPOSE: To provide the manager with the information context needed for planning automation changes for his hybrid design and fabrication shop
VIEWPOINT: The data exchange problem as seen by designers and manufacturers of hybrid microcircuit assemblies
NODE:
TITLE: CONTEXT: APPLICATION PROTOCOL ACTIVITY MODEL
NUMBER

Figure 3a: AAM diagram A-0, top of model

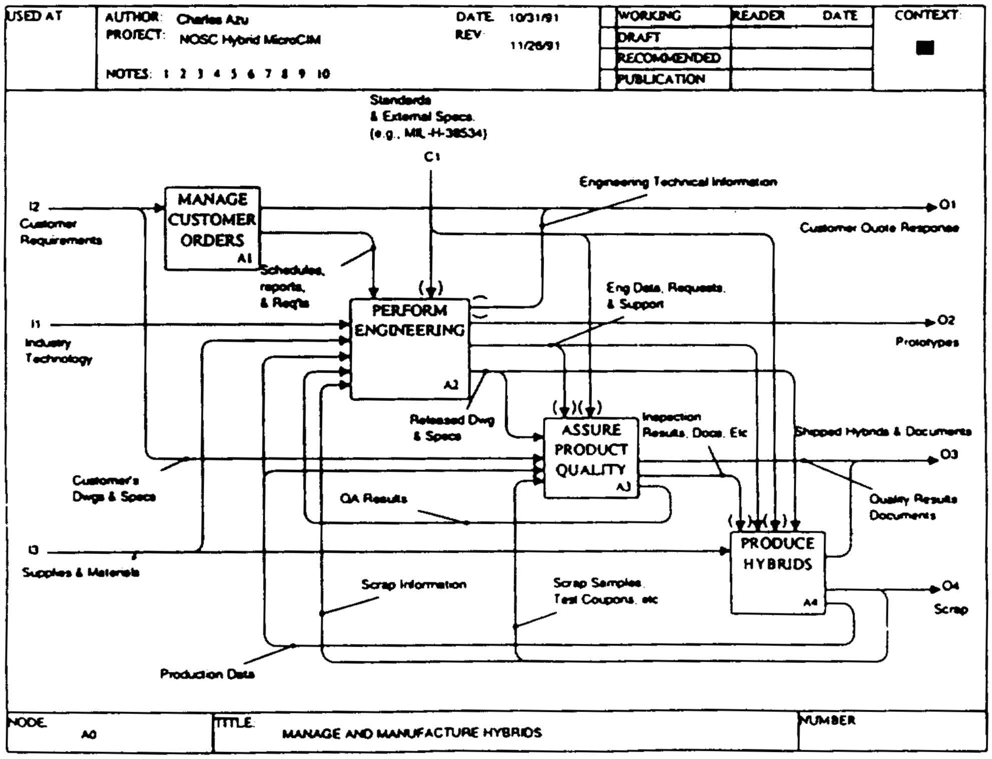
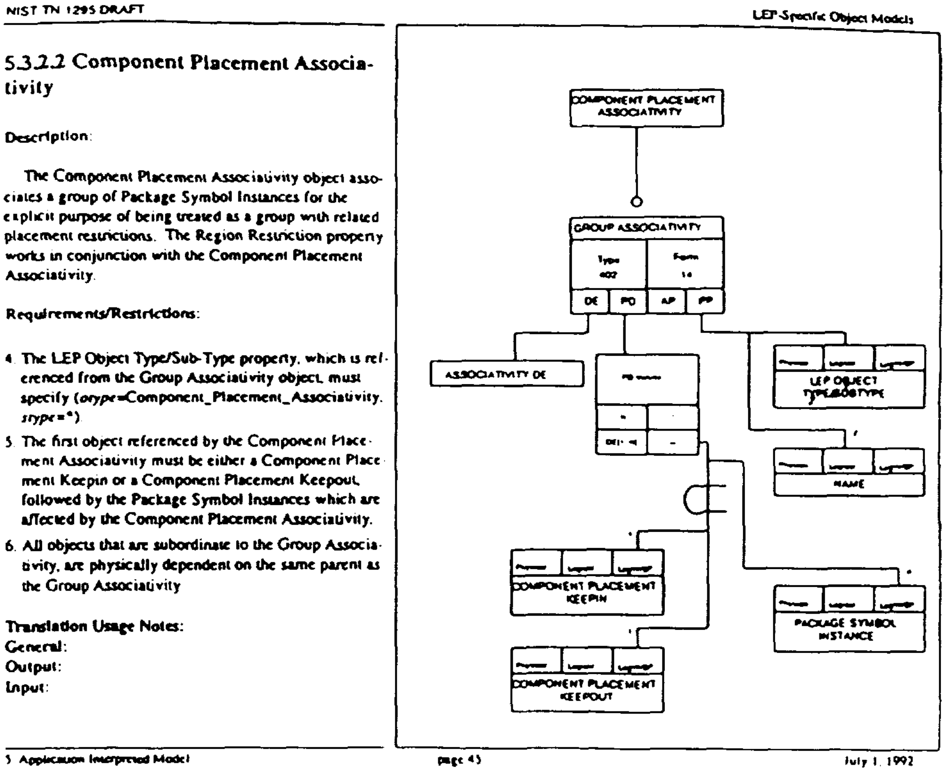


Figure 3b: AAM diagram A0



NIST TN 1295 DRAFT

LEP-Specific Object Models

5.3.2.2 Component Placement Associativity

Description:

The Component Placement Associativity object associates a group of Package Symbol Instances for the explicit purpose of being treated as a group with related placement restrictions. The Region Restriction property works in conjunction with the Component Placement Associativity.

Requirements/Restrictions:

4. The LEP Object Type/Sub-Type property, which is referenced from the Group Associativity object, must specify (*otype*=Component_Placement_Associativity, *stype*=*).
5. The first object referenced by the Component Placement Associativity must be either a Component Placement Keepin or a Component Placement Keepout, followed by the Package Symbol Instances which are affected by the Component Placement Associativity.
6. All objects that are subordinate to the Group Associativity, are physically dependent on the same parent as the Group Associativity

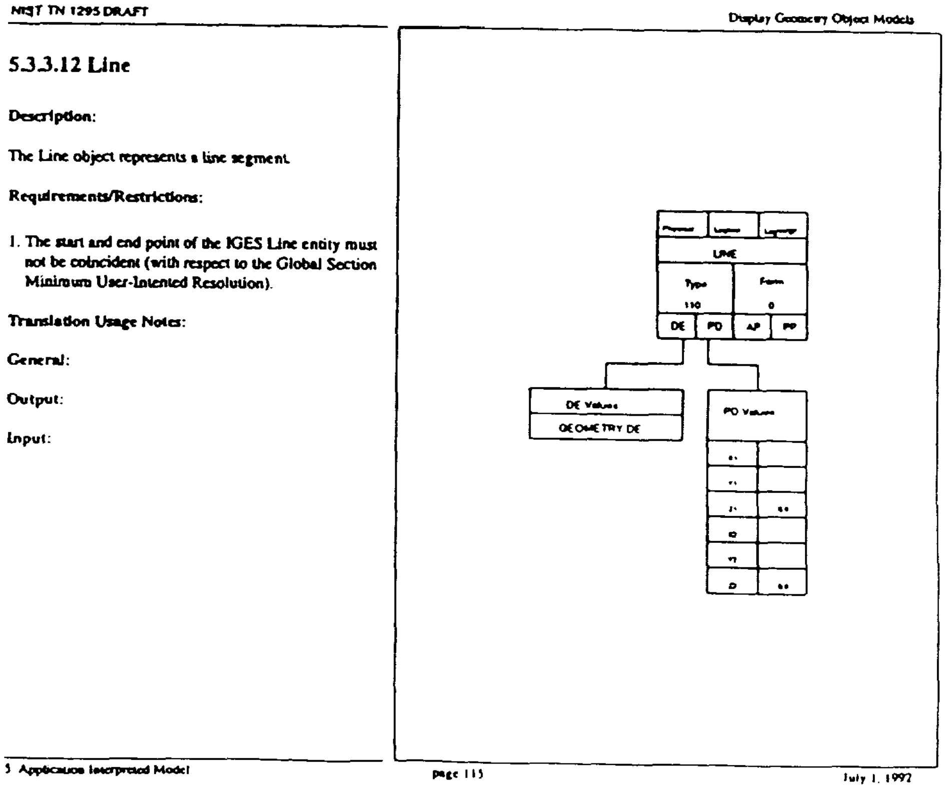
Translation Usage Notes:
General:
Output:
Input:

5 Application Interpreted Model

page 45

July 1, 1992

Figure 5a: Example object model specific to Layered Electrical Products (LEP)



NIST TN 1295 DRAFT

Display Geometry Object Models

5.3.3.12 Line

Description:

The Line object represents a line segment.

Requirements/Restrictions:

1. The start and end point of the IGES Line entity must not be coincident (with respect to the Global Section Minimum User-Intented Resolution).

Translation Usage Notes:

General:

Output:

Input:

5 Application Interpreted Model

page 115

July 1, 1992

Figure 5b: Example object model for Display Geometry

METHODS AND MEANS USED IN PROGRAMMING INTELLIGENT SEARCHES OF TECHNICAL DOCUMENTS

N93-22155

p- 8

David L. Gross
Computer Engineer
Analex Space Systems, Inc.
NASA Kennedy Space Center
Post Office Box 21206
Kennedy Space Center, FL 32815-0206
Phone: 407/861-5716
Fax: 407/861-5774

ABSTRACT

In order to meet the data research requirements of the Safety, Reliability & Quality Assurance activities at Kennedy Space Center (KSC), a new computer search method for technical data documents was developed. By their very nature, technical documents are partially encrypted because of the author's use of acronyms, abbreviations, and shortcut notations. This problem of computerized searching is compounded at KSC by the volume of documentation that is produced during normal Space Shuttle operations. The Centralized Document Database (CDD) is designed to solve this problem. It provides a common interface to an unlimited number of files of various sizes, with the capability to perform many diversified types and levels of data searches. The heart of the CDD is the nature and capability of its search algorithms. The most complex form of search that the program uses is with the use of a domain-specific database of acronyms, abbreviations, synonyms, and word frequency tables. This database, along with basic sentence parsing, is used to convert a request for information into a relational network. This network is used as a filter on the original document file to determine the most likely locations for the data requested. This type of search will locate information that traditional techniques, (i.e., Boolean structured key-word searching), would not find.

INTRODUCTION

The need to search technical documentation for desired information is a labor intensive activity. In the past, data searches have been restricted to human effort with limited computer searching, (generally Boolean key-word searching). This is primarily due to the type of information that is being searched and referenced. Technical documents are partially encrypted by the author's use of acronyms, abbreviations, and shortcut notations. At Kennedy Space Center (KSC), this problem is magnified further. A researcher who is searching for information based on an engineer's or a technician's notes is faced with notes that are usually more encrypted and/or abbreviated than those which are contained in the actual document. The problem is further compounded by the volume and dispersal of documentation that is produced during normal shuttle operations. The CDD addresses these problems. The commercial potential of this system is evident from the savings in man-hours alone. Any profession that devotes time to specific subject review and research would benefit greatly from this time-saving system, (e.g. legal, medical, information specialist, etc.).

BACKGROUND

In 1990 NASA funded a project to improve the data retrieval and dissimilation methods used by the Safety, Reliability & Quality Assurance (SR&QA) directorate. Systems and quality assurance reviews were identified as likely candidates for improvement. This procedure requires accessing a large number of technical documents and uses a large percentage of available man-hours. A project was initiated to develop a more time-efficient method of doing these searches. Several commercial packages were evaluated, but none met SR&QA's needs. Finally, a decision was made to develop custom software.

Software algorithms from the Artificial Intelligence (AI) field were used in an attempt to duplicate human search methods. The three methods that showed promise were:

1. Sentence parsing used in natural language processing
2. Confidence factors or weights from heuristic searching
3. Network connection and propagation from connectionism

Parsing analyzes the syntactic structure of sentences. To adapt this technique to technical data queries, parsing is used to identify word and phrase relationships such as subject-verb, verb-object, and noun-modifier (Figure 1). The parser uses knowledge of language syntax, morphology, and semantics. In technical document searches, sentence parsing is used to identify word types, (i.e. noun, verb, adjective) based on the context in which the abbreviation or acronym is used.

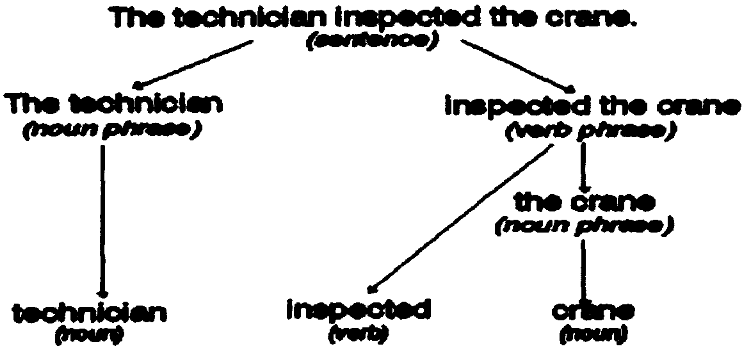


Figure 1. Sentence Parsing

Confidence factors or weights are normally used to measure the confidence level in a rule-based system. These factors combine to usefully measure uncertainty. In the CDD they are used to measure the probability that a given search parameter is correct. For instance, if a search query uses an acronym that has two possible meanings, a search for both would be performed using a lower weighing value than if the acronym had only one possible meaning.

Network connection and propagation refers to the construction of multi-layer networks and how the weights of the nodes are patterned. In this search technique, the nodes of the network are words (or phrases) with the pattern of the weighing determined by a set of heuristic rules. This network then can be used to develop a set of conditions that evaluate each area of a document.

METHODOLOGY

In any highly developed field, especially a highly technical one, there are a number of words, phrases, and acronyms that have specific meanings. These can be considered a specialized knowledge base for that particular field. Developing an intelligent search system for a specialized field must utilize that knowledge base, along with more general information of the English language.

In developing this knowledge base for NASA operations, a general database of acronyms, abbreviations, and synonyms was used as a starting point. Specialized acronyms and abbreviations used in normal shuttle operations were added to this database. In addition, word frequency tables were developed to identify the most commonly used words.

The first step in processing a query is to break down the sentence structure. Initially, the sentence or sentences are separated into individual word objects. These prime words form the first level nodes of the filter, with the order of the words maintained through the links between nodes (see Figure 2). The node object includes weighing variables for the word and for the links between nodes. The weighing variables for the prime nodes are set to a benchmark reference value of 100.

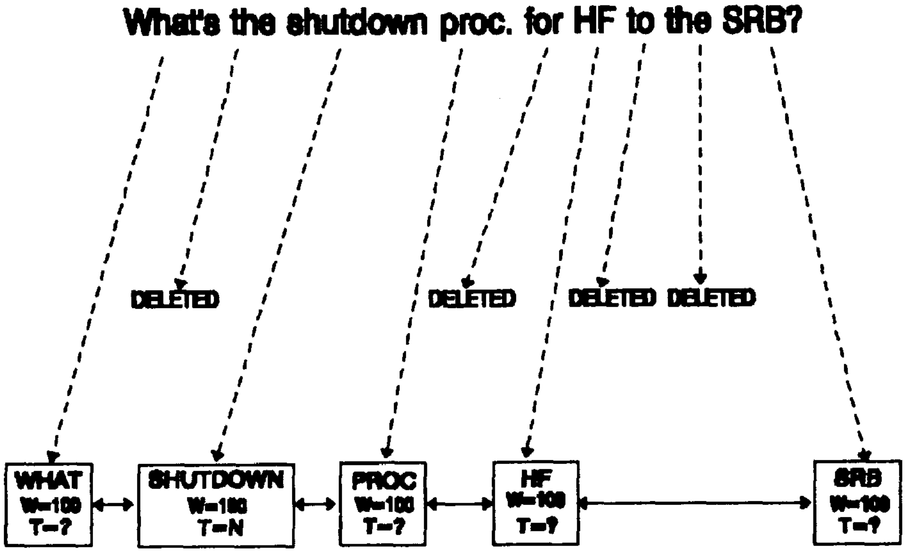


Figure 2. Generation of Prime Nodes

The knowledge base is used to expand the single level of prime nodes into a multilevel node network. Each word in the first level is referenced in the knowledge base. If a match is found, the reference values from the knowledge base become new nodes at a lower level. For example, in Figure 3, the node with the value "SRB" is matched in the acronym table with the value "Solid Rocket Booster". This value then becomes a new sub-node with links to the same nodes that "SRB" has. If more than one value is found, then more than one sub-node is created for each prime node.

In Figure 3, the parsing function identifies the prime node "PROC." as being used as a noun. A sub-node of this produced from abbreviation tables is "PROCEED," a verb. Since the word types do not match, the sub-node, "PROCEED" can be eliminated along with any synonyms produced from it. Eliminating the node this way would require assuming that the original query was structured syntactically correct. An alternate method is to reduce the weight of that node to indicate a much lower probability.

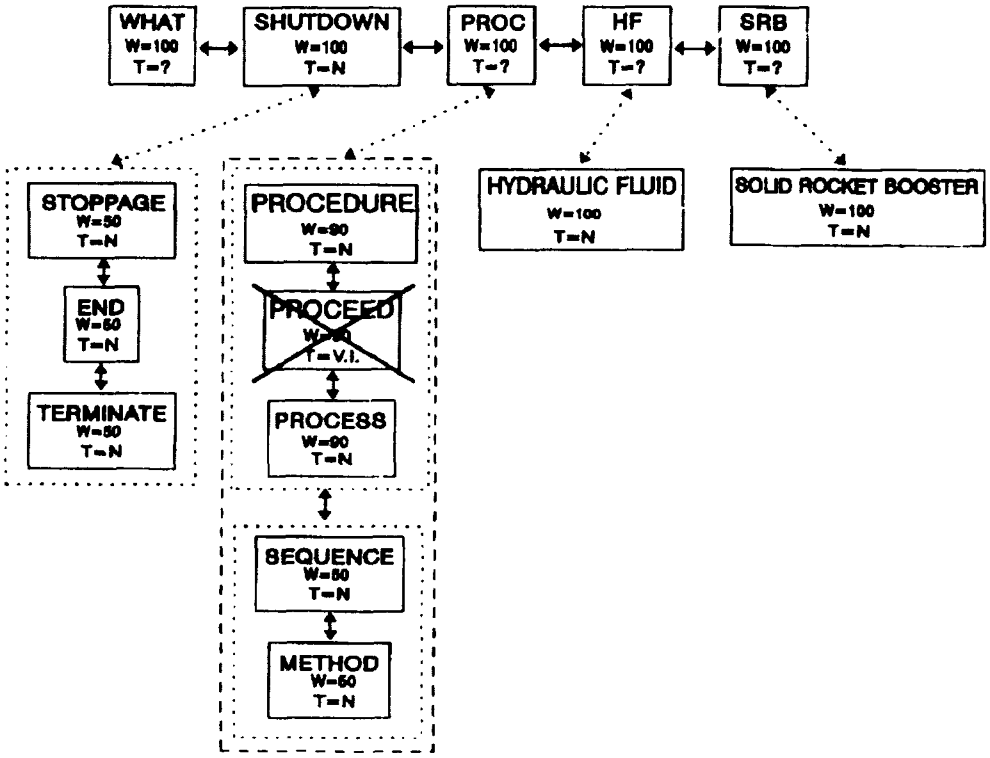


Figure 3. Generation of Sub-Nodes

After this multilevel network is created, a set of heuristic rules is used for setting the weights for each node. If the conditional part of the rule tests true, then the rule sets the weights of that sub-node, and can adjust the weight for the prime node(s) of the sub-node. Table 1 lists some of the heuristic rules used to set these values.

A Boolean search condition is then generated for every node in the net. This search condition is of the type:

if <node> exists then VALUE = VALUE + WEIGHT

The list of these conditions forms the filter function. The filter function generates a value for different areas of the document.

IMPLEMENTATION

The search technique described in the proceeding section was implemented, along with standard document handling techniques, into a system called the Centralized Document Database (CDD). This system has an extensive database of technical documents supported on a Local Area Network (LAN). The system provides a single point for accessing and searching technical documentation. The system is designed to access ASCII formatted files of various sizes and types and different physical storage locations.

#	Condition	Change
1	Word is a prime. Word is unique (not in any table of knowledge base).	Value = 400
2	Word is a prime. Word is not in frequency table.	Value = 200
3	Word frequency level > 10	Value = 0
4	Word frequency level > 5	Value = Value/2
5	Word (or phrase) is a sub-node. Phrase is only possible acronym meaning.	Value = Value of prime node
6	Word (or phrase) is a sub-node. Phrase is one of several possible acronym meanings.	Value = (Value of prime node) / (number of acronym meanings - 1)
7	Word is a sub-node. Word is a synonym.	Value = (value of prime node) / (number of synonyms)
8	Word (or phrase) is a sub-node. Phrase is only possible abbreviation meaning.	Value = Value of prime node
9	Word (or phrase) is a sub-node. Phrase is one of several possible abbreviation meanings.	Value = (Value of prime node) / (number of abbreviation meanings - 1)

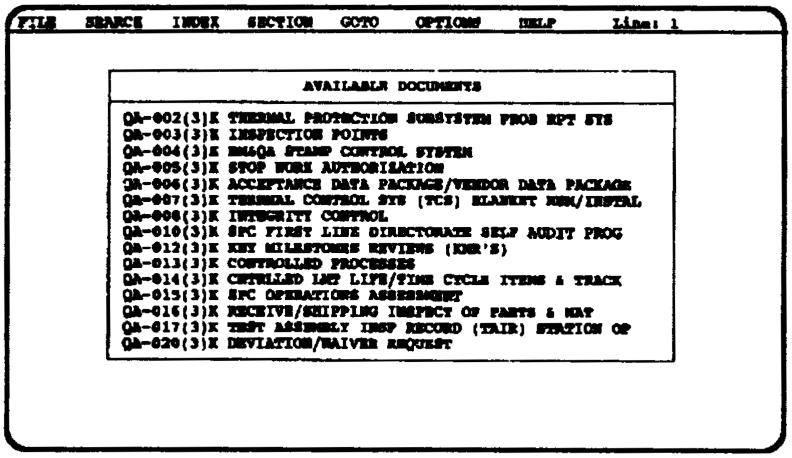
Table 1
Weighing Conditions

A basic menu system is used to call up and display all of the available document files (see Figure 4). It uses a number of filters for common word processors and mainframe printer formats. These are simple filters designed to mask the command codes used by the different application programs that produced the document. The end result is that a document in almost any format, (e.g. Word Perfect, Displaywrite, or mainframe redirected printer output), can be displayed (somewhat distorted) and used by the system.

The program provides immediate access to any part of a document through the use of special pointer files. The CDD program uses these pointer files to speed-up direct location access. These pointer files can be used for pages, sections, record numbers, or any string value. When a document is selected, the software will check for all available pointer files and add the options to the option menu. These pointer files are created outside of the CDD to fulfill specific needs within the SR&QA community. The CDD requires the pointer files to be in a particular format and location, but any programming language can be used to create them.

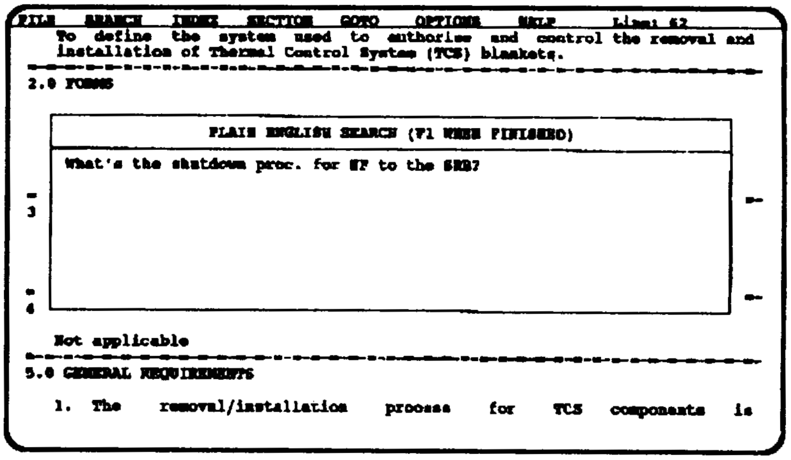
The CDD has the capability to perform several different types and levels of data searches. The simplest type is a basic Boolean key-word search. This type of search is a useful and fairly common type of search that can locate a specific string using standard AND/OR logic. The program provides an improvement to this type of search by expanding the Boolean logic to include any acronyms and abbreviations of the search strings from its built-in database.

The program has a fully operational version of the intelligence searching technique explained previously (Figure 5). The initial query is broken down into its related components (words and phrases). Network nodes are established and expanded through the methods described previously. A set of heuristic rules are used to assign weights for the nodes in the new levels.



```
AVAILABLE DOCUMENTS

QA-002(3)K THERMAL PROTECTION SUBSYSTEM PROB RPT SYS
QA-003(3)K INSPECTION POINTS
QA-004(3)K EM&QA STAMP CONTROL SYSTEM
QA-005(3)K STOP WORK AUTHORIZATION
QA-006(3)K ACCEPTANCE DATA PACKAGE/VENDOR DATA PACKAGE
QA-007(3)K THERMAL CONTROL SYS (TCS) BLANKET REM/INSTAL
QA-008(3)K INTEGRITY CONTROL
QA-010(3)K SPC FIRST LINE DIRECTORATE SELF AUDIT PROG
QA-012(3)K KEY MILESTONES REVIEWS (KMR'S)
QA-013(3)K CONTROLLED PROCESSES
QA-014(3)K CNTRLLED LMT LIFE/TIME CYCLE ITEMS & TRACK
QA-015(3)K SPC OPERATIONS ASSESSMENT
QA-016(3)K RECEIVE/SHIPPING INSPECT OF PARTS & MAT
QA-017(3)K TEST ASSEMBLY INSP RECORD (TAIR) STATION OP
QA-020(3)K DEVIATION/WAIVER REQUEST
```

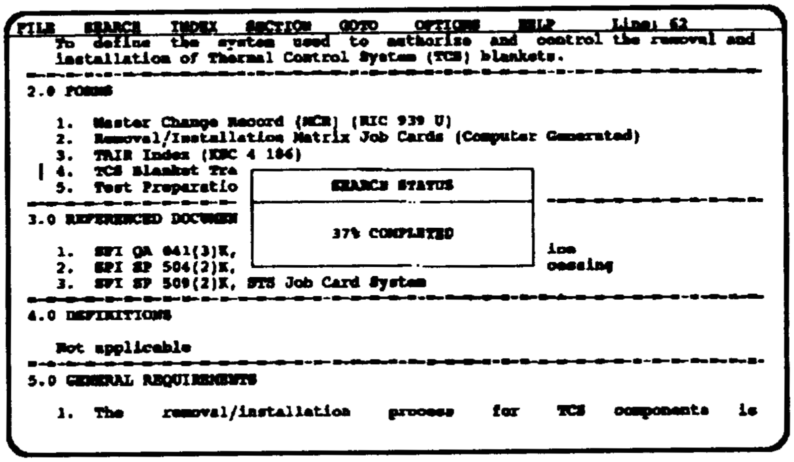
Figure 4. Document Selection in The CDD



```
  To define the system used to authorize and control the removal and
  installation of Thermal Control System (TCS) blankets.
------------------------------------------------------------------------
2.0 FORMS

      PLAIN ENGLISH SEARCH (F1 WHEN FINISHED)

   What's the shutdown proc. for HF to the SRB?
-
3                                                                      --

-
4                                                                      --

  Not applicable
------------------------------------------------------------------------
5.0 GENERAL REQUIREMENTS

  1.  The   removal/installation   process   for   TCS   components   is
```

Figure 5. Intelligent Search Query Screen

This network can identify and rank key areas of the document that are likely to contain the information requested. The program does this by filtering the document through the network. This relational network returns a weight value for the section of the document that is currently passing through the network. The user is appraised of the search status as the document is being processed (Figure 6). A list of pointers, to the sections of the document that had the highest values, is the final result of the filtering. The software will immediately display the area of the document that had the highest weighing (Figure 7). If the user does not find the needed information, the software will move to the next highest weighted area of the document.

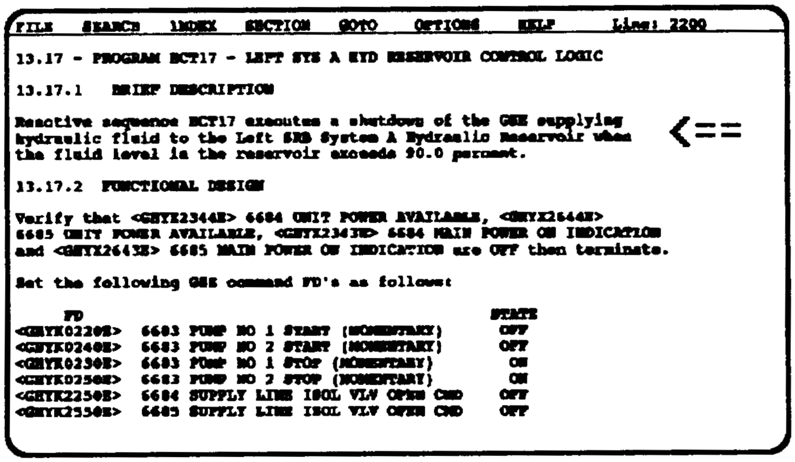


```
   To define the system used to authorize and control the removal and
   installation of Thermal Control System (TCS) blankets.
------------------------------------------------------------------------
2.0 FORMS

   1.  Master Change Record (MCR) (RIC 939 U)
   2.  Removal/Installation Matrix Job Cards (Computer Generated)
   3.  TAIR Index (KSC 4 186)
 | 4.  TCS Blanket Tra  ┌──────────────────────────┐
   5.  Test Preparatio  │      SEARCH STATUS       │
--------------------    ├──────────────────────────┤ -------------------
3.0 REFERENCED DOCUMEN  │                          │
                        │      37% COMPLETED       │
   1.  SPI QA 041(3)K,  │                          │ ion
   2.  SPI SP 504(2)K,  └──────────────────────────┘ cessing
   3.  SPI SP 509(2)K, STS Job Card System
------------------------------------------------------------------------
4.0 DEFINITIONS

   Not applicable
------------------------------------------------------------------------
5.0 GENERAL REQUIREMENTS

   1.  The    removal/installation    process    for    TCS    components    is
```

Figure 6. Search Status Screen

The results of any search are displayed on the terminal and can be exported to other applications. The program has the option to copy part or all of a document to a file or printer. The data is copied in standard ASCII format that can be imported into most word processor and database applications. The data also can be printed to any network printer.

Modifying the program to work in other fields (legal, medical, and business) would require the creation of a specialized database of words, acronyms, and abbreviations used in that field. In most cases this database already exists in the reference documentation used in that field.



```
13.17 - PROGRAM HCT17 - LEFT SYS A HYD RESERVOIR CONTROL LOGIC

13.17.1   BRIEF DESCRIPTION

Reactive sequence HCT17 executes a shutdown of the GSE supplying
hydraulic fluid to the Left SRB System A Hydraulic Reservoir when
the fluid level in the reservoir exceeds 90.0 percent.

13.17.2  FUNCTIONAL DESIGN

Verify that <GHYX2344E> 6684 UNIT POWER AVAILABLE, <GHYX2644E>
6685 UNIT POWER AVAILABLE, <GHYX2343E> 6684 MAIN POWER ON INDICATION
and <GHYX2643E> 6685 MAIN POWER ON INDICATION are OFF then terminate.

Set the following GSE command FD's as follows:

    FD                                             STATE
<GHYK0220E>  6683 PUMP NO 1 START (MOMENTARY)       OFF
<GHYK0240E>  6683 PUMP NO 2 START (MOMENTARY)       OFF
<GHYK0230E>  6683 PUMP NO 1 STOP (MOMENTARY)         ON
<GHYK0250E>  6683 PUMP NO 2 STOP (MOMENTARY)         ON
<GHYK2250E>  6684 SUPPLY LINE ISOL VLV OPEN CMD     OFF
<GHYK2550E>  6685 SUPPLY LINE ISOL VLV OPEN CMD     OFF
```

Figure 7. Results of Search

REFERENCES

[1] Luger, G. and Stubblefield, W. : Artificial Intelligence and the Design of Expert Systems, Benjamin/Cummings Publishing Company, Inc. 1989.

[2] Minsky, M. : Semantic Information Processing, The Massachusetts Institute of Technology Press, 1968.

[3] Sombe, L. : Reasoning Under Incomplete Information in Artificial Intelligence, John Wiley & Sons, Inc. 1990.

[4] Whittington, R.P. : Database Systems Engineering, Oxford University Press, 1988.

AN INTEGRATED INFORMATION RETRIEVAL AND DOCUMENT MANAGEMENT SYSTEM

L. Stephen Coles,
Group Chief Technologist,
J. Fernando Alvarez,
Technical Group Supervisor,
James Chen, William Chen,
Lai-Mei Cheung, Susan Clancy,
and Alexis Wong

Information Systems Integration Group
Institutional Data Systems
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109-8099

ABSTRACT

This paper describes the requirements and prototype development for an intelligent document management and information retrieval system that will be capable of handling millions of pages of text or other data. Technologies for scanning, Optical Character Recognition (OCR), magneto-optical storage, and multiplatform retrieval using a Standard Query Language (SQL) will be discussed. The semantic ambiguity inherent in the English language is somewhat compensated-for through the use of coefficients or weighting factors for partial synonyms. Such coefficients are used both for defining structured query trees for routine queries and for establishing long-term interest profiles that can be used on a regular basis to alert individual users to the presence of relevant documents that may have just arrived from an external source, such as a news wire service. Although this attempt at *evidential reasoning* is limited in comparison with the latest developments in AI Expert Systems technology, it has the advantage of being commercially available.

INTRODUCTION

Today, virtually all large organizations are inundated with data. In attempting to deal with the problem of storage and retrieval from this ever increasing volume of paper, all private and public institutions are exploring new methods for communicating information electronically beyond existing e-mail and Local Area Networks, especially now that the cost of optical storage and scanning technology is becoming more affordable. An additional problem, of course, that we will not consider here, is the graceful transition from the existing infrastructure using a paper data base and manual methods to an all-electronic form of doing business. Clearly, for at least some of the time, both types of systems will have to coexist side-by-side until the all-electronic form gains sufficient acceptance that most people essentially stop relying on paper, and we begin to treat our trees as endangered species rather than as mere commodities. As one example, the Defense Logistics Agency of the U. S. Defense Department, which accounts for about 70 percent of all U.S. Government acquisitions, is planning their CALS Program for Electronic Data Interchange (EDI) so that it will process the equivalent of 200 million sheets of paper per year in an all-electronic form by the end of 1995. The U.S. National Aeronautics and Space Administration (NASA) also retains millions of pages of documents accumulated over the last two decades that it expects to store and retrieve electronically.

This paper discusses the requirements for an intelligent document management and information retrieval system. JPL has now developed a prototype of such a system for NASA Headquarters in Washington, D.C. for the initial storage of 10,000 pages of documents that will be expanded to 1.6 million pages located in a large document archive. It is expected that new document pages will be accumulated initially at the rate of 300,000 pages per year. Access by key words from a full-text search index to both the ASCII form of these documents as well as their original raster-scanned images must take place in a matter of seconds on any of five standard PC, Macintosh, or UNIX Workstations operating concurrently over a coaxial ethernet LAN that will be upgraded in the next year to a fiber optic network (FDDI). The original scanned documents must always be available, since they may contain signatures or drawings that are essential to the documentation. To implement this system, we have chosen two open architecture 486 IBM-PC servers and a magneto-optical read/writable 88-cartridge jukebox operating over a Novell Netware and a LAN Manager Network. A high-performance scanner feeds pages from an automatic document reader at the rate of 34 pages per minute, compresses them according to a CCITT Group IV standard with a compression ratio of about 20:1, and uses a standard Optical Character Recognition (OCR) software package to convert them to ASCII text. Our experience shows that OCR error rates are quite variable depending on the quality of the source documents and are exquisitely sensitive to such idiosyncracies as the presence or absence of underlining. A high-resolution monochrome scanning station that can capture two full 8-1/2 x 11 pages side-by-side is being used for quality control during scanning and OCR. Color documents are scanned using a separate flat-bed scanner, since the time for scanning a single color document at high resolution typically exceeds five minutes per page. A commercially-available Relational Data Base Management System has been chosen for structured key-word retrieval and the maintenance of user-defined interest profiles.

Figure 1 shows an overview of the hardware configuration of the prototype system. Figure 2 shows the information flow of documents through the system from scanner to optical storage and subsequent retrieval. Figures 3-6 illustrate the additional processing that is needed for indexing files and the arrangement of Directory Structures on the magneto-optical jukebox.

HARDWARE SELECTIONS

The scanner selected was a Fujitsu 3093E (Calera CS100) with a speed of 34 pages per minute, although the Bell and Howell Copiscan II Model 3338 would normally be preferred for high volume work because of its greater capacity (42 pages per minute at 300 dots per inch). The color scanner chosen was the Advanced Vision Research Model AVR-8000/CLX, although any number of other models would have been acceptable including the HP ScanJet IIc or the Epson ES-300c. The compression board selected was from DISCORP, although a competing product from KOFAX would have been acceptable. The Calera WordScan Plus software was selected for Optical Character Recognition. The high-resolution monochrome display selected was the Sigma Design Multimode 120. The scanning workstation platform was a 486/50 MHz client PC with 16 MB of RAM. A Tricord Systems 486 Superserver was selected for the LAN Manager Server, while a 486/50 MHz DX server with 32 MB of RAM was chosen for the Novell image LAN. An HP LaserBank Library Jukebox was chosen as the 88-cartridge (55 GigaBytes) magneto-optical store. A stand-alone magneto-optical drive (680 MB) was attached to the scanning workstation for backup. The SQL full-text document search engine selected was *Topic* made by Verity, Inc. of Mt. View, California. *Topic* provides a comprehensive set of retrieval aids, such as concept-based queries, word proximity queries, synonyms, thesaurus, and so forth. A pair of 16.8 kbaud Courier HST modems from US Robotics were installed in Washington, D.C. and Pasadena running CoSession windows communication software to facilitate remote debugging.

REQUIREMENTS ON SOFTWARE DEVELOPMENT

An important requirement imposed by NASA on DRIMS (the JPL Document Retrieval Integrated Management System) was for hyperlinking raster-scanned images from the original documents to the ASCII text obtained by OCR from those documents after scanning. Thus, if the user identified a relevant document in connection with a structured query search of the full text of the files on the HP-88 jukebox and wished to examine, for example, an original signature specimen or an engineering diagram as it appeared in the original document, he or she should be able to do so quickly (with a single mouse click). Documents were separated at the request of NASA into various categories, including letters (correspondence), memos, proposals, reports, presentations, etc. In addition to scanning in color documents, it was also required that DRIMS have a procedure for handling double-sided documents or even bound documents like books or research papers (for which a wide flat-bed scanner surface is needed). The OCR process must have an accuracy of at least 96 percent (tolerating at most 4 illegible characters per 100), given an input of high-quality laser-printer text. In order to satisfy the requirement of a user-friendly human/machine interface, DRIMS was implemented in Microsoft Visual Basic under Windows 3.1. Because OCR is such a computer-intensive step in the processing of documents, it was a requirement that after scanning in up to 1000 pages per day of new material, OCR could be carried out during the evenings in a batch mode by passing only the name of a daily file directory with the day's scanned image files. Other requirements, such as user password security and new user set-up by the Data Base System Administrator (DBA) were also implemented in Visual Basic. Using LAN Manager, record-based document retrieval can also be performed efficiently from SQL applications. For example, documents can be rapidly searched by title, date, author, document owner, and so forth without having to search through the full-text *Topic* indices. Document retrieval should be possible from any of the following platforms: PC DOS character mode, PC Windows, Macintosh, and UNIX-based workstations (Sun, DEC, HP, etc.).

RESULTS

The initial users of DRIMS have been the staff of the Office of Space Science and Applications (OSSA) of NASA Headquarters in Washington, D.C. Two stages of data preparation were needed for DRIMS to accommodate the various types of documents. In the first stage, all documents were scanned as TIFF image files. Next, the second stage creates ASCII text files from these image files. Initially, 10,000 pages of documents were scanned and indexed. Documents are scanned-in as either one-page or multi-page documents (an automatic document feeder with a capacity of 50 pages is attached to one of the scanners) independently of their orientation (either so-called "portrait" or "landscape" orientation). A database for DRIMS was created through the combined entries for programs, organizations, events, persons, and other information in reference to the user-defined document types.

CONCLUSION

The transition to a paperless, all-electronic form of institutional work flow will be long and difficult, but the initial stages are now getting underway. Much more experience needs to be obtained with the problem of having humans and not machines correct errors in the OCR process, a labor-intensive, eye-straining, psychologically stressful activity, given that the quality of many of the source documents is quite poor. As one example, faded blue carbon copies of documents typed on a conventional IBM Selectric typewriter 15 years ago, where the registration of characters may not have been perfect, letters may touch one another, and subsequent ballpoint handwriting or official document stamps cut across the text on the document, cause considerable heartburn. Moreover, typing may have been done on a standard *form*, characterized by special boxes outlined by horizontal and vertical lines for input typing, and where the typist was not always careful to ensure that typing never spilled over the boundaries of a box, or distracting overtyping was done to correct typographical errors, and so on. There are also problems with newspaper text, multifonted text (boldface, italics, etc.), and multilanguage/multialphabetic texts. If the document contains tables of budget numbers, for example, the importance of 100 percent accuracy in OCR may be different than if the document contains only conventional text. This process of OCR correction has been known to "burn out" even the most energetic and determined of clerks who are unfamiliar with the domain of discourse. Increasing the resolution of scanning from 300 to 400 dots per inch has been shown to make an incremental improvement in reducing the OCR error rate, but ultimately this approach has diminishing returns. Only the most sophisticated computational linguistics techniques involving not just morphology, automatic spelling-correction, and grammatical correction, but semantic and pragmatic techniques will be needed to reduce the residual OCR error rate down to less than one percent.

Finally, another area for future growth will be provision to incorporate into such a system hooks for multimedia materials, such as voice annotation and full-motion video clips. For this purpose, it will be imperative to comply with the latest standards for color image compression that are currently evolving, such as JPEG, MPEG, DVI/RT (Digital Video Interaction/Real-Time from Intel and IBM), and SGML (Standard Generalized Markup Languages) now being defined by various industry, university, and professional society (ISO) groups.

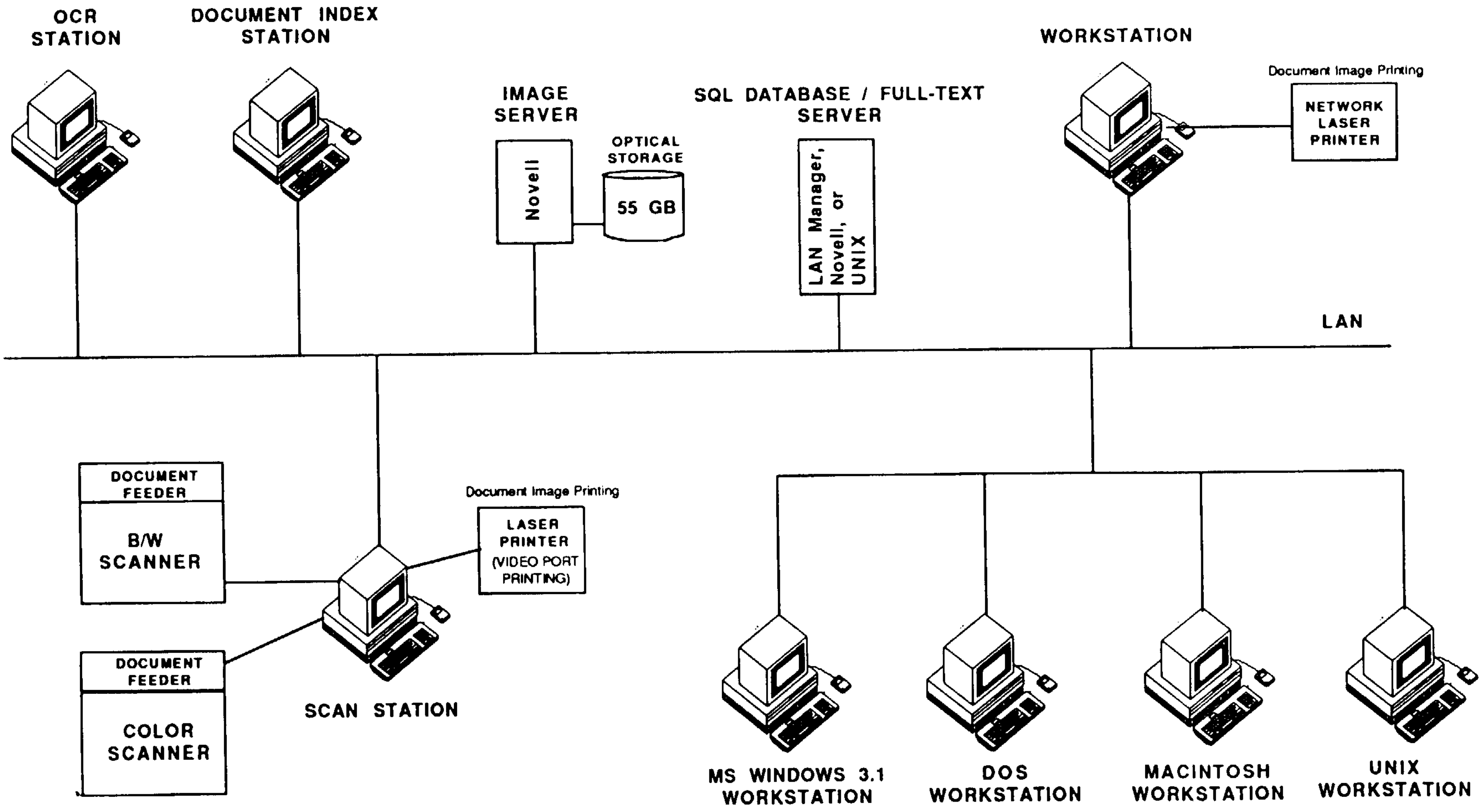
ACKNOWLEDGMENTS

The work 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.

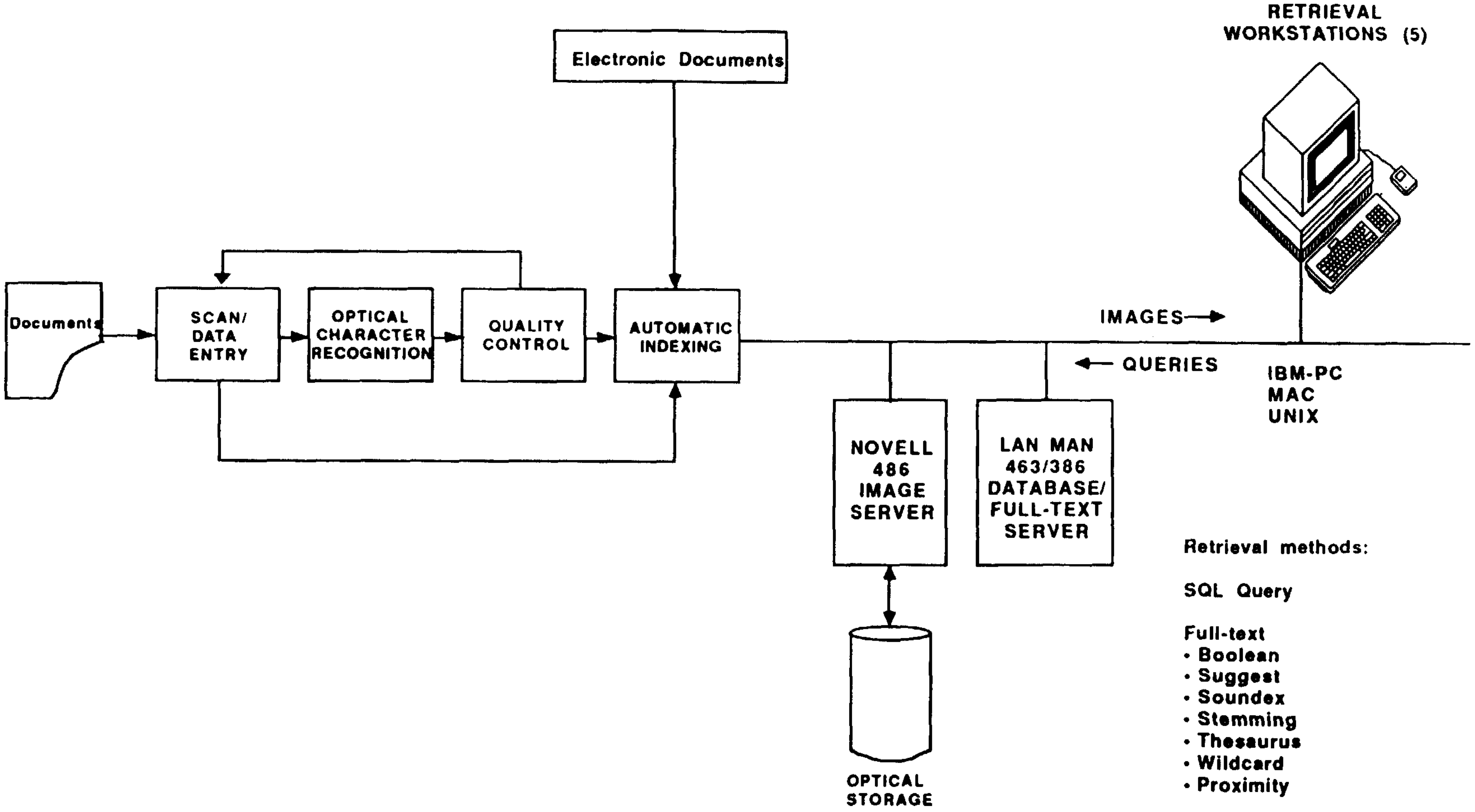
Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

JPL
Jet Propulsion Laboratory
California Institute of Technology

DRIMS HARDWARE LAYOUT



DRIMS DOCUMENT PROCESSING



Retrieval methods:

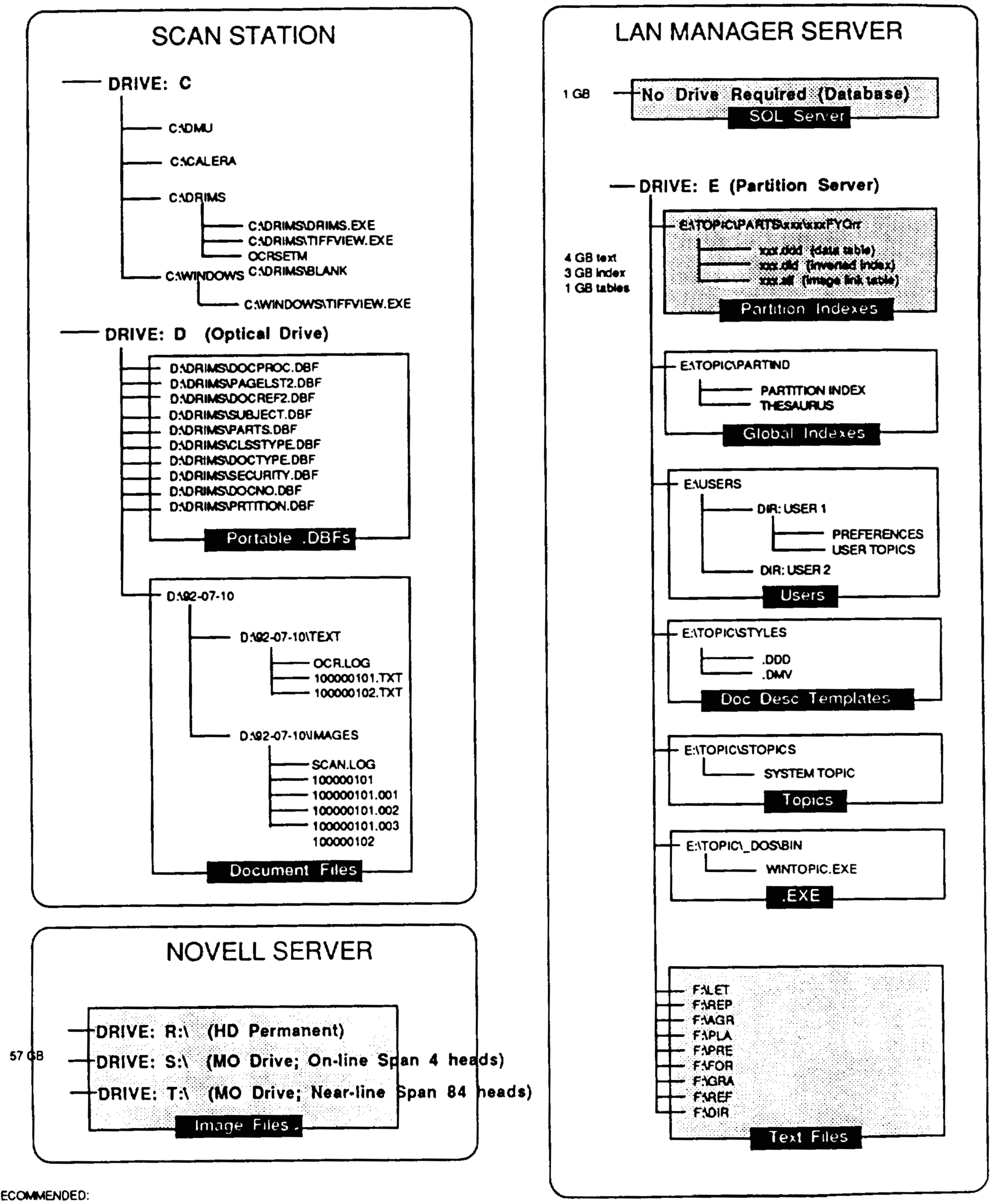
SQL Query

Full-text
- Boolean
- Suggest
- Soundex
- Stemming
- Thesaurus
- Wildcard
- Proximity

Object-oriented Concept Trees (Topics)

JPL
Jet Propulsion Laboratory
California Institute of Technology

DRIMS Directory Structure



SCAN STATION

DRIVE: C
- C:\DMU
- C:\CALERA
- C:\DRIMS
  - C:\DRIMS\DRIMS.EXE
  - C:\DRIMS\TIFFVIEW.EXE
  - OCRSETM
  - C:\DRIMS\BLANK
- C:\WINDOWS
  - C:\WINDOWS\TIFFVIEW.EXE

DRIVE: D (Optical Drive)
- D:\DRIMS\DOCPROC.DBF
- D:\DRIMS\PAGELST2.DBF
- D:\DRIMS\DOCREF2.DBF
- D:\DRIMS\SUBJECT.DBF
- D:\DRIMS\PARTS.DBF
- D:\DRIMS\CLSSTYPE.DBF
- D:\DRIMS\DOCTYPE.DBF
- D:\DRIMS\SECURITY.DBF
- D:\DRIMS\DOCNO.DBF
- D:\DRIMS\PRTITION.DBF

Portable .DBFs

- D:\92-07-10
  - D:\92-07-10\TEXT
    - OCR.LOG
    - 100000101.TXT
    - 100000102.TXT
  - D:\92-07-10\IMAGES
    - SCAN.LOG
    - 100000101
    - 100000101.001
    - 100000101.002
    - 100000101.003
    - 100000102

Document Files

NOVELL SERVER

57 GB
- DRIVE: R:\ (HD Permanent)
- DRIVE: S:\ (MO Drive; On-line Span 4 heads)
- DRIVE: T:\ (MO Drive; Near-line Span 84 heads)

Image Files

LAN MANAGER SERVER

1 GB — No Drive Required (Database)

SQL Server

DRIVE: E (Partition Server)

4 GB text
3 GB index
1 GB tables

- E:\TOPIC\PARTS\xxx\xxxFYQrr
  - xxx.ddd (data table)
  - xxx.did (inverted index)
  - xxx.ail (image link table)

Partition Indexes

- E:\TOPIC\PARTIND
  - PARTITION INDEX
  - THESAURUS

Global Indexes

- E:\USERS
  - DIR: USER 1
    - PREFERENCES
    - USER TOPICS
  - DIR: USER 2

Users

- E:\TOPIC\STYLES
  - .DDD
  - .DMV

Doc Desc Templates

- E:\TOPIC\STOPICS
  - SYSTEM TOPIC

Topics

- E:\TOPIC_DOS\BIN
  - WINTOPIC.EXE

.EXE

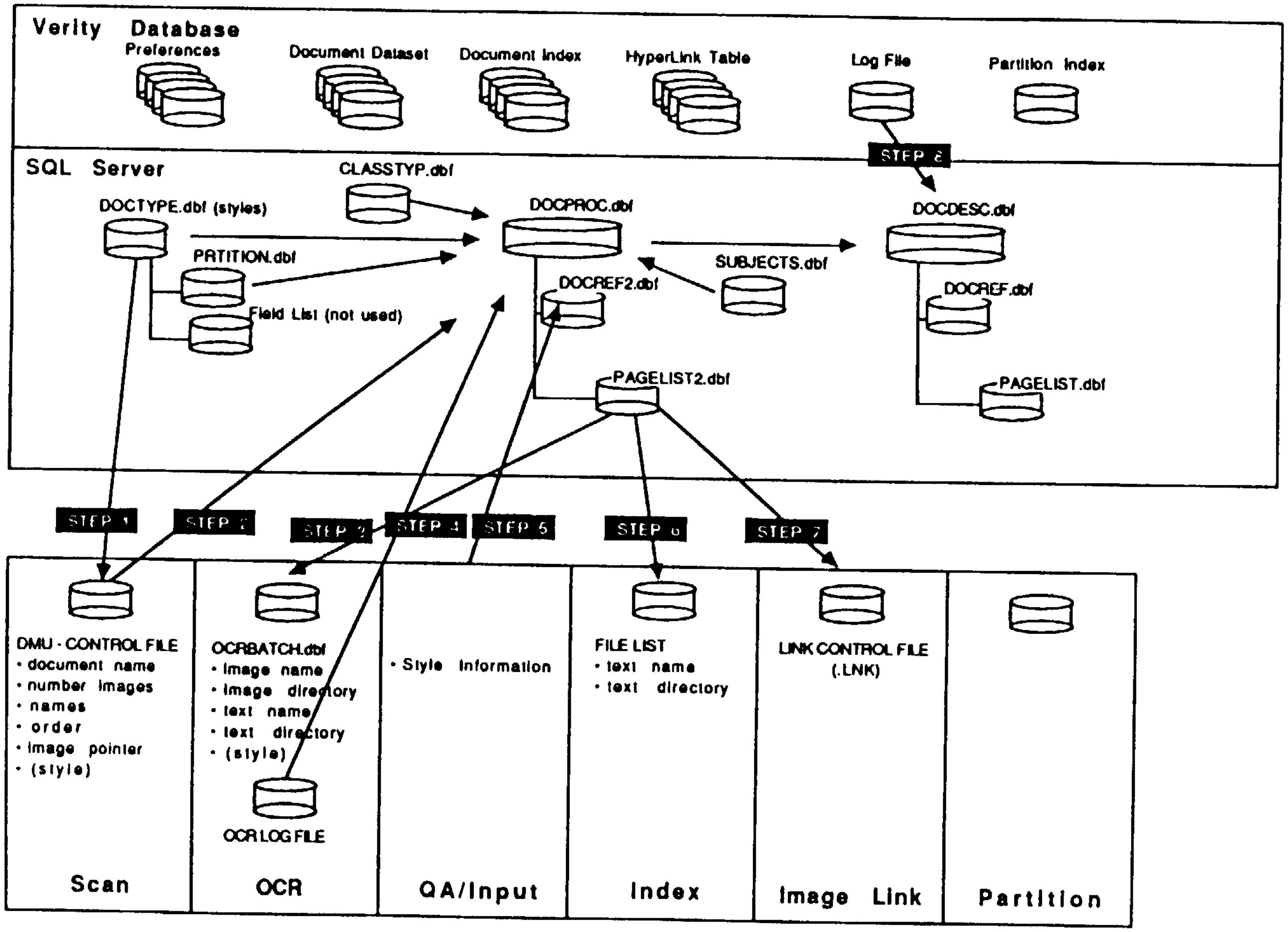
- F:\LET
- F:\REP
- F:\AGR
- F:\PLA
- F:\PRE
- F:\FOR
- F:\GRA
- F:\REF
- F:\DIR

Text Files

RECOMMENDED:
2-10 MB TEXT PER PARTITION
1-5K DOCUMENTS
2-10K PER DOCUMENT
UP TO 100 LINKS/DOCUMENT
10,000 LINKS/PARTITION

Last Update: 8-21-92
Previous Update 7/7/92
2/18/92

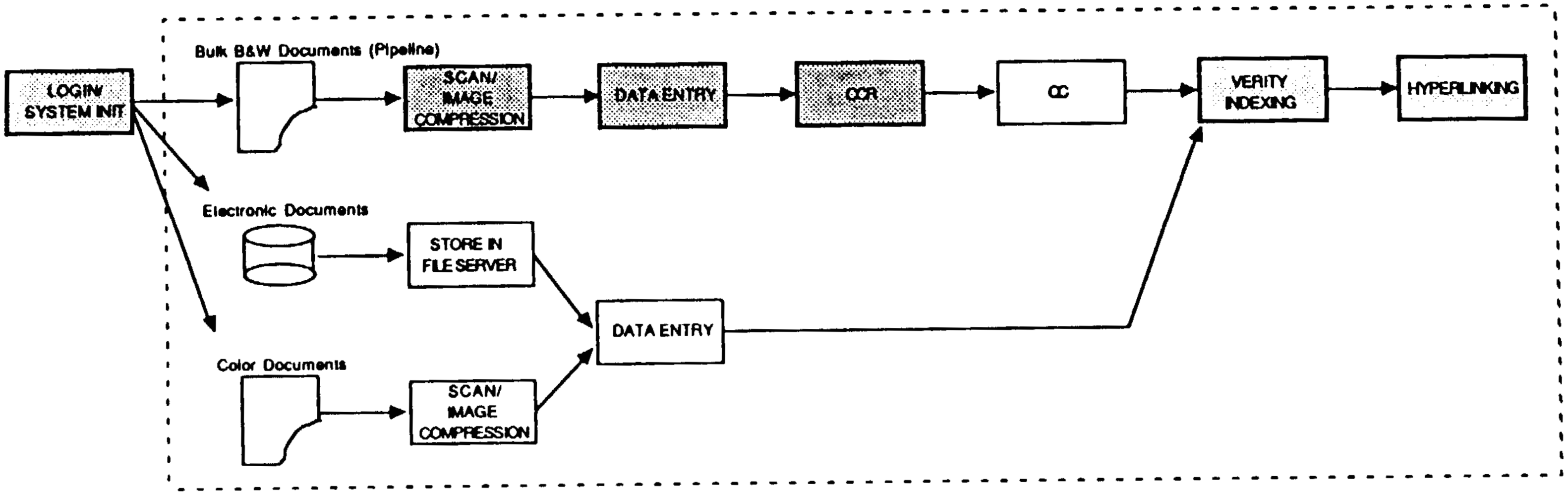
Database Entities



Last Update: 4/22/92
Previous Update: 4/17/92
Created: 2/13/92

DRIMS HIGH-LEVEL DATA PREP FLOW

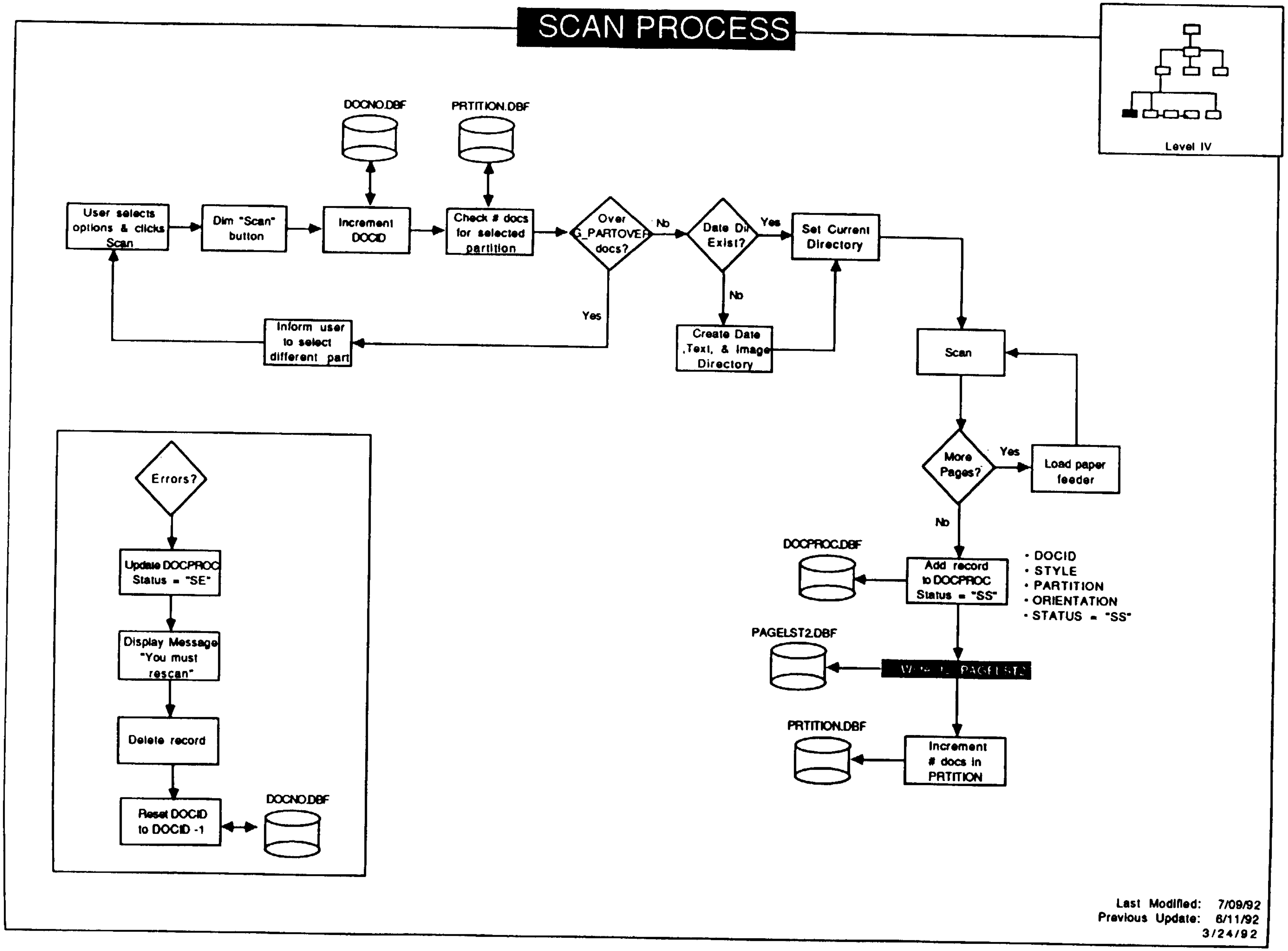
Level 1



PROGRESS

- ▨ UPDATES BEING IMPLEMENTED
- ▨ MODULE COMPLETED & DELIVERED
- □ PROGRAMMING IN PROGRESS
- □ DESIGN COMPLETED

Last Update: 6/8/92
Previous Update: 6/4/92
Created: 4/07/92


SCAN PROCESS
Level IV
User selects options & clicks Scan
Dim "Scan" button
Increment DOCID
DOCNO.DBF
Check # docs for selected partition
PRTITION.DBF
Over G_PARTOVER docs?
No
Yes
Inform user to select different part
Date Dir Exist?
Yes
No
Create Date ,Text, & Image Directory
Set Current Directory
Scan
More Pages?
Yes
Load paper feeder
No
Add record to DOCPROC Status = "SS"
DOCPROC.DBF
• DOCID
• STYLE
• PARTITION
• ORIENTATION
• STATUS = "SS"
Write to PAGELST2
PAGELST2.DBF
Increment # docs in PRTITION
PRTITION.DBF
Errors?
Update DOCPROC Status = "SE"
Display Message "You must rescan"
Delete record
Reset DOCID to DOCID -1
DOCNO.DBF
Last Modified: 7/09/92
Previous Update: 6/11/92
3/24/92

AN INVESTIGATIVE STUDY OF MULTISPECTRAL DATA COMPRESSION FOR REMOTELY-SENSED IMAGES USING VECTOR QUANTIZATION & DIFFERENCE-MAPPED SHIFT-CODING[1]

S. Jaggi, Senior Scientist
Advanced Sensor Development Laboratory
Lockheed Engineering & Sciences Company[2]
Stennis Space Center, MS 39529

ABSTRACT

A study is conducted to investigate the effects and advantages of data compression techniques on multispectral imagery data acquired by NASA's airborne scanners at the Stennis Space Center. The first technique used was vector quantization. The vector is defined in the multispectral imagery context as an array of pixels from the same location from each channel. The error obtained in substituting the reconstructed images for the original set is compared for different compression ratios. Also, the eigenvalues of the covariance matrix obtained from the reconstructed data set are compared with the eigenvalues of the original set. The effects of varying the size of the vector codebook on the quality of the compression and on subsequent classification are also presented. The output data from the Vector Quantization algorithm was further compressed by a lossless technique called Difference-mapped Shift-extended Huffman coding. The overall compression for 7 channels of data acquired by the Calibrated Airborne Multispectral Scanner (CAMS), with an RMS error of **15.8 pixels** was **195:1 (.041 bpp)** and with an RMS error of **3.6 pixels** was **18:1 (.447 bpp)** . The algorithms were implemented in software and interfaced with the help of dedicated image processing boards to an 80386 PC compatible computer. Modules were developed for the task of image compression and image analysis. Also, supporting software to perform image processing for visual display and interpretation of the compressed/classified images was developed.

INTRODUCTION

The exceedingly high data rates of remote sensing instruments have prompted needs for rapid data retrieval, transmission, storage and subsequent processing. These instruments acquire multispectral data for each ground scene element. Thus, several images are created from one spatial scene. These multispectral images stretch the demand on image processing techniques and equipment. Depending on the rate of data acquisition, the volume of data being generated can exceed the available capacities and technologies for data transmission. With continually increasing demands for improved spectral and spatial resolution, the requirements for data handling are likely to become more stringent. Two possible ways to solve this problem are on-board data compression for near-real-time processing and ground-based compression for data archiving. The near-real-time processing could be in the form of reducing the bandwidth of the image with a view towards performing some operation such as unsupervised clustering. This demands that the implementation of the compression scheme be capable of performing fast operations. The implementation of ground-based compression schemes is not as limited as speed is not a constraint.

[1] This work was performed under NASA contract NAS 13-315, Job Order No. 019-CAX01 under the direction of Mr. Bruce Spiering, NASA Project Manager, Advanced Sensor Development Laboratory, at the John C. Stennis Space Center.

[2] Former affiliation. Correspondence should be addressed to: Dr. S. Jaggi, Sverdrup Technology Inc., Engineering and Science Department, Applied Math and Physics Section, John C. Stennis Space Center, MS 39529.

Multispectral data has been tremendously useful in solving problems related to classification of objects through the use of remote sensing techniques. As the sensor systems provide more channels, it becomes increasingly critical to develop and implement methods that reduce the amount of processing without losing any advantages arising from increased amounts of information.

Data compression is a useful tool for the purpose of reducing the bandwidth requirements during data transmission or the memory requirements during data storage. Various techniques have been introduced for this purpose. These techniques can be classified into two broad categories - lossy and lossless. Lossy techniques exploit statistical and spatial correlations in the data to remove the redundancies, thus retaining most of the information. Lossless techniques attempt to perform the same task subject to the constraint of having to perfectly reconstruct the data. Lossy techniques allow for greater compression ratios by reconstructing the data within an error bound. Typical examples of lossy techniques include vector quantization, transform coding, and least means square (LMS) filter. Examples of lossless compression are run-length coding, contour coding and quadtree coding.

This paper describes a study to investigate the compression of multispectral visible and thermal imagery data. The aim is to implement algorithms that perform lossy and lossless data compression on multispectral remotely sensed images. In the case of lossy compression, the objective is to determine the trade-offs associated with the degradation in data which results. In the case of the lossless compression, the degree of compression that is possible is examined.

After the algorithms were implemented on a computer they were tested using a typical set of images obtained by airborne scanners operated and maintained by the Advanced Sensor Development Laboratory at NASA's Stennis Space Center. The scanneres are the Thermal Infrared Multispectral Scanner (TIMS) and the Calibrated Airborne Multispectral Scanner (CAMS). The original acquired images were subjected to the algorithms to investigate the amount of compression that is possible. This amount of compression depends on the nature of the data and the number of channels of the imagery. The objectives were different for lossless and lossy techniques.

In the case of lossless compression, only the compression ratio possible needed to be determined. In the case of lossy techniques, the ratio was programmable. The higher the compression ratio, the greater is the degradation in the reconstructed images. This study addresses the problem by studying the amount of degradation that results versus the compression ratio.

Various criteria were used for studying this degradation. One criterion is the difference between the original and the reconstructed sets of images. The closer the two sets of images are, the lesser is the degradation. The degradation is represented as the RMS error between the original and the reconstructed images. Another criterion is to compare the eigenvalues of their covariance matrices. This reveals the amount of correlation fidelity that is lost due to the compression.

Since lossless data compression can always be done after the lossy technique, it is available as an enhancement to the compression performance. However if the data needs to be interpreted in its compressed form, it may not be possible to do so after lossless compression. This is because lossless techniques usually store the data in a form that has no immediate perceptible relation to the original data. This issue is also addressed in this work, where the technique addressed will not only compress the data, but leave the result in a form that will allow the investigator to visually analyze the data. In short, it will not only compress but also give the added benefit of classifying the data in an unsupervised manner.

LOSSY DATA COMPRESSION

Background and Introduction

Several techniques have been introduced that reduce the spectral dimensionality of the data, thereby reducing the problem to a computationally manageable one. The spectral dimensionality of the data is defined as the number of channels of the imagery being used for analysis. Most of these techniques are modified forms of the Karhunen-Loeve Transform or the Principal Components Analysis, which by itself is computationally inefficient to implement. This transform exploits the band-to-band correlation of various regions of the imagery. Using these correlations, the redundancy in the multispectral set is eliminated. A classified image is the result of the process.

A singular value decomposition of the image yields the eigenvalues and eigenvectors of the image. The number of eigenvalues equals the number of channels in the sensor data. The total variance of the image set is the sum of all the eigenvalues of the covariance matrix. However, it is observed that for most images a significant portion of the total variance in the multispectral images is contained within a few dominant eigenvalues. Thus, the eigenvector images corresponding to these dominant eigenvalues are said to adequately represent the multispectral imagery with a minimum loss in total image variance.

The disadvantage with this method is that it is not yet possible to implement it real time. Also, the resultant eigenvector images are transformed versions of the earlier multispectral raw data. Thus, by themselves, they do not offer any clear interpretation of the imagery. Typical methods to perform multispectral classification are the maximum likelihood (ML) and the Principal Components (PC) Analysis. The ML method requires a priori knowledge of the behavior of the statistics of the data. Based upon those statistics, various regions in the images are classified. The ML method belongs in the category of supervised classification techniques.

The PC method decomposes the set of multispectral images into an equivalent set of orthogonal images, with each image being an eigenvector of the original set. Using the Karhunen-Loeve transform, the three eigenvector images corresponding to the three highest eigenvalues can be coded in a RGB format, to display a classification of the dominant correlations in the data. The PC method belongs in the category of unsupervised classification techniques.

In this section, we investigate a technique to be used for on-board data-compression of multispectral images obtained from airborne scanners. The airborne scanners of interest in this study are the TIMS and the CAMS which are NASA instruments operated by Sverdrup Technology at the Stennis Space Center. The TIMS is a thermal infrared sensor with six channels in the 8-12 micron region of the electromagnetic spectrum. The CAMS is a visible, near-IR and thermal sensor with eight channels in the visible and near-IR and one broad-band thermal channel. This paper only discusses the results of the CAMS imagery.

Here a method based on the vector quantization of the multispectral images is investigated for its use as a tool for unsupervised image clustering and subsequent data compression. The computational simplicity and the ease of implementation after the quantization parameters have been determined make vector quantization an attractive tool for data processing. The vector is defined in the multispectral imagery context as an array of pixels from the same location from each channel. The rate of compression is programmable depending on the size of the codebook. However, the higher the compression ratio, the greater is the degradation between the original and the reconstructed images. The compressed images are reconstructed and compared with the original set using the PC method. The eigenvalues of the covariance matrix obtained from the reconstructed data set are compared with the eigenvalues of the original set. This comparison is done for varying degrees of compression which are obtained by varying the size of the vector codebook. Also, the error obtained in substituting the reconstructed images for the original set is compared for different compression ratios. The effects of varying the size of the vector codebook on the quality of the compression and on subsequent classification are also presented. The eigenvalues of the covariance matrix of the original multispectral data-set were found to be highly correlated with those of the reconstructed data-set.

The section on lossy compression is organized into four parts. An introduction to vector quantization is followed by a description of the algorithm used to determine the optimal codebook. This is followed by a description of the imagery used for data analysis and the results. In conclusion a comparison of this technique is made with the results obtained from the PC technique.

Vector Quantization

Vector quantization is a scheme for mapping a large set of vectors into a much smaller set of vectors called codewords. In the case of multispectral data a vector is defined as an array of the pixel elements corresponding to a given location for all the channels of the imagery.

Goldberg et. al. , Gray, and Nasrabadi et. al. provide comprehensive details about the concepts. Hang et. al. and Ramamurthi et. al. discuss three variations on the algorithm.

Vector Quantization is defined as a mapping Q of k-dimensional Euclidean space $\mathbf{R}^k$ into $\mathbf{Y}$, a finite subset of $\mathbf{R}^k$.

$$\mathbf{R}^k \overset{Q}{\rightarrow} y \qquad y \varepsilon \mathbf{Y}$$

It can also be interpreted as an encoder-decoder operation. The encoder views an input vector $\mathbf{x}$ and generates the address or the code of the reproduction vector $Q(\mathbf{x})$.

$$Q(x) = y \qquad \{ x \text{ e } R^k : \; y \text{ e } Y \}$$

For multispectral data, k is the number of channels in the sensor. The decoder uses this address to reconstruct the reproduction vector y. The set of all y's, Y which is made up of all the constituent codes, is called the codebook and its elements i.e y are called the codewords or reproduction vectors. The criteria of choosing the codewords is to minimize the cost or penalty associated with reproducing the vectors y from x.

$$\text{Cost Function} = \| x - y_{cl} \|^2$$

where y_{cl} is a vector chosen from the codebook y such that it is closest to the vector x.

Thus for a 6 channel (k = 6) set of 512 X 512 pixel images, there exist 512*512 = 262,656 vectors, each of length 6 elements. This entire set of vectors is partitioned into a finite collection of subsets N. A codeword is chosen as a representative vector for each subset. A collection of these codewords is called a codebook. The codebook is assumed to be representative of the entire subset, such that any vector in the subset can be represented by one of the codewords in the codebook, with a minimum error between the original vector and its reproduction vector. Figure 1 shows the block diagram of a VQ implementation. An input vector is quantized by choosing the closest codeword from the codebook. Once the codebook has been determined, the encoder maps each input vector to one of the codewords in the codebook. For the purpose of storage or transmission, the only necessary information is the index of the codeword in the image and the codebook. For the purpose of regenerating the images, the decoder compares the index of each pixel location with the corresponding vector in the codebook. By choosing the codeword corresponding to the transmitted index, the output vector is created.

It has been shown by Shannon, that a vector quantizer will always give a better noise performance than a scalar quantizer. Among the advantages of VQ is its easy implementation. Once the codebook has been determined it can be easily developed into a hardware implementation that just performs a look-up operation. Also, if a large training set is used to determine the codebook, this universal codebook can be representative of a wide variety of features in the imagery. This could be especially useful for terrestrial data which may not exhibit many unique features. Figure 2 shows the basic concept involved in vector quantization for multispectral imagery. The vector, which is the array of the pixel values of all channels for a location, is quantized and the resulting codeword is substituted for the original vector. Thus the entire reconstructed set is made up of a finite set of vectors that make up the codebook.

The compression ratio is dependent on the number of codewords needed to represent the multispectral image. Since the vectors are being created in a multispectral manner, only one image needs to be stored. This image contains a spatial distribution of the codewords that represent the multispectral set. However, instead of storing the codewords, the index of the codewords is adequate as a symbol of the codeword. The number of bytes occupied by the original multispectral set (which contains 1 byte per pixel) is given by

Original set size = No. of channels * No. of rows * No. of columns.

After the image is compressed only the index representing the spatial distribution of the codewords needs to be stored. The index can be the binary address of the codewords. Thus if there are N codewords, the number of bits needed to address them is $\log_2(N)$. This leads to the expression of the compression ratio using vector quantization for multispectral imagery data

$$\text{Compression Ratio} = \frac{\left(\dfrac{\text{No. of Channels}}{\log_2(\text{ No. of Codewords })} \text{ x No. of bits per origl. pixel} \right)}{1}$$

The number of bits per original pixel is 8. The compression ratio is also expressed as a bit-rate. This is the average number of bits required to represent the compressed image. The expression for the bit-rate is given as

$$\text{bits per pixel (bpp)} = \left(\frac{\log_2(\text{Number of codewords})}{\text{Number of channels}} \right)$$

A discussion of the algorithm used to obtain these codewords follows.

Algorithm

The main problem associated with VQ is the design of an optimal codebook so as to minimize the cost of substituting the original vectors with the reproduction vectors **y**. This cost measure is denoted as d(**x**, **y**). A common measure used to estimate this distortion is the Euclidean distance between the original vector and its reproduction. Fig. 3 shows the steps involved in performing vector quantization. The codebook is initialized with a set of vectors that are selected from the original vector set. The entire vector set is then classified into these codewords. Using an algorithm, these codewords are then modified to minimize the distortion between the original set and the reconstructed set. This procedure is repeated until the codewords are optimized.

A set of training vectors can be used to perform the minimization. The codebook thus obtained is optimal in the mean square error sense. Several algorithms have been proposed to obtain this codebook. The most widely used algorithm is commonly known as the Linde-Buzo-Gray (LBG) algorithm.

In the LBG algorithm, an initial reproduction vector set, i.e the codebook C[0], is chosen. The criteria for choosing this initial set can vary as well. In this study, it was decided to space the initial vectors a minimum distance apart. Vectors were picked up from the image at random until the required number, N were selected.

A threshold is assigned as the minimum acceptable cost or penalty between the original vectors in the training set **x** and the reproduction vectors **y**. When the average error due to the quantization of the vectors is equal to or below this threshold, the codebook is said to be optimal.

Each vector in the training set is quantized using the initial codebook C[0]. For every vector in the training set **x**, a vector **y** from the codebook that is closest to **x**, is chosen. The average error due to the quantization of **x** to **y** is computed. The number of vectors **x** is determined by the size of the training set. The number of vectors in the codebook N, is fixed depending upon the level of quantization required. Each codeword in the existing codebook is replaced by the centroid of all the training vectors assigned to the codeword. This process is repeated until the error goes below a prespecified threshold.

Imagery Data Analysis

Before beginning the analysis of the images, a criterion needs to be established for comparing the original and the reconstructed images. Two criteria were determined to be of critical importance.

In remote sensing, it is important to be able to determine the correlations between the various multispectral images obtained. These correlations reveal various facts about the data being mapped. Thus it is important that if the original data is being transformed (i.e being subjected to lossy compression), its statistical correlations be maintained as closely as possible. This leads to an analysis of the principal components of the original and the reconstructed images. The eigenvalues of the images were obtained by singular value decomposition. The total variance of the multispectral image-set is the sum of the diagonals of the covariance matrix.

It is also important that the reconstructed images be as similar to the original images as possible. The measure chosen is the root mean square difference between the original and the reconstructed images.

$$\varepsilon_{RMS} = \sqrt{\frac{\sum_{\forall \text{pixels}} \| x - y_{cl} \|^2}{\text{No. of pixels X No. of channels}}}$$

The data used for the analysis was obtained by airborne scanners operated and maintained by the Advanced

Sensor Development Laboratory at the Stennis Space Center for NASA. Figure 4 details the characteristics of the scanners. The CAMS data were acquired over western Puerto Rico in January 1990 over land and water. The aim was to study impacts of man-induced changes on land that affect sedimentation into the near-shore environment.

The algorithm was coded into a PC-based system used for the purpose of image processing. This system uses two plug-in boards for performing basic image processing operations. The first board DT-2861 is a frame grabber that also has a 16 image buffer for performing simultaneous on-board imaging operations without having to access the hard disk for different sections of the image. The other board DT-2858 is used to accelerate image processing operations by having them done in hardware.

Results of Vector Quantization

A channel of the original image-set is shown in Fig. 5 for CAMS. The codebooks for different codewords were obtained after 15 iterations. The components of the two sets of the data were thencompared to check for the accuracy of the vector quantization process. Figure 6 shows the RMS error for different compression ratios for CAMS. Figure 7 shows the total variance of the reconstructed set. The total variance is the sum of all the eigenvalues of the covariance matrix. Figure 8 shows the relative variance of each eigenvalue. The numerical values of these relative eigenvalue distributions are shown in tabular form in Fig.9a. Figures 9b-d show the percentages for the first, second and the third most dominant eigenvalues of the covariance matrix of the reconstructed data-set. From the two sets of matrices it is obvious that most of the information is contained in the two dominant eigenvalues for which there is significant correlation between the original and the reconstructed data set as the number of codewords is increased. Figures 10-11 show the corresponding results for the CAMS data-set. The images show the indexes for the codewords and are not related to the original images. Each index corresponds to a vector which can be used to generate the reproduced data-set. The nature of complexity in Fig. 11 corresponding to 128 codewords can be contrasted with that of 4 codewords in Fig. 10.

Hardware Implementation

The implementation of the lossy data compression technique involves two stages. In the first stage the codebook is determined. This is done using the algorithm as described earlier. The algorithm requires as its input the compression ratio desired and the multispectral image data set. Its output is the optimized set of codewords that represent the imagery set. Each codeword is a vector with its number of elements equaling the number of channels in the multispectral set. In the second stage, each pixel location in the multispectral set is replaced with the codeword that is closest to it. This is the process of quantization. The original vector made up of the value of the pixel at a specific location in all channels is replaced by another vector which is picked from the codebook such that, of all the vectors in the codebook, this new vector is closest to the original vector. Thus the original vector is quantized to the new vector. Since the codebook has a fixed number of vectors or codewords, each codeword in the codebook is given an index number. For example a codebook with 16 vectors can have indexes from 0 through 15. This index of the new vector is sufficient to represent the new vector. Thus, each vector in the original image is replaced by the index of the codeword in the codebook that is closest to it. The information as to which index is what vector, i.e the codebook, can be tagged along at a later stage.

Figure 12 shows a possible configuration of the implementation. Once the codebook has been determined, the operation is reduced to almost a look-up-table approach. The incoming vector is compared with the codewords in the codebook and the index of the codeword closest to the vector is transmitted. If each index is denoted by a color, this information can also be displayed on a real-time monitor. There is hardware available to perform these operations. The only problem that still remains is that of finding the optimum codebook in a near-real-time manner. This can be achieved by a parallel scheme. While the existing codebook in the board is being used to compress the data, another codebook is in the making in the background. This can be done by an independent processing unit, which is dedicated to implementing the LBG algorithm. After it has optimized on a codebook it then updates the codebook currently in use, with the new codebook. However, every time a new codebook is used it needs to be transmitted. All subsequent indexing will be in reference to the new codebook. A criterion could be introduced for the update. The existing codebook is replaced only if the difference between the existing and the new codebook exceeds a certain threshold. Since terrestrial data is low frequency in nature and does not change very rapidly this threshold will obviate the need to constantly update the codebook and transmit a new set of codewords with the data. Also, if the ground scene does indeed change, this will change the codebook significantly and the threshold will allow for the replacement of the codebook.

Conclusion

The vector quantization technique is an effective tool for data compression and classification. Amongst the advantages of this technique is that it is easy to implement. It also effectively exploits the redundancy present in the channel-to-channel correlations of the data to reduce the memory required for the storage of the images.

LOSSLESS COMPRESSION

Introduction

In this section a technique for coding the same remotely sensed data with 100% restoration is investigated. This technique involves difference mapping and shift-extended coding of the original data. The first step is mapping of the data with the use of a difference transform. This mapped data is then used to generate a set of symbols. These symbols are then coded through a Huffman coder. Figure 13 shows the high-level description of these main functions.

Mapping Transform

The lossless coding technique discussed here can be divided into two parts - a mapping function and a symbol generating function. The mapping function used for this study is a difference mapping. Each pixel is mapped as the difference between the present and the previous pixel. Denoting the 'i'th pixel by x[i] and the 'i'th map by m[i],

$$m[i] = x[i] - x[i-1]$$

The advantage of this mapping is that the data, which is not changing rapidly on a pixel by pixel basis, can be condensed to a smaller dynamic range. For an 8 bit system, the pixel has a range of 0 - 255. The mapped function theoretically has a range of -255 -> +255. Assuming that the pixels are not changing rapidly, most of the values of the mapped function can lie within a much narrow range (say -4 -> +4). This number is fixed arbitrarily based on the statistics of the data. For reasons stated later in the section, a size that is a power of 2 is chosen.

$$N = 2^{**}b$$

where b is an integer. It is called the bit-size of the code.

The aim is to code the entire mapped set into a series of symbols that can be easily interpreted. The symbols available for the coding are 0, 1, 2, ... N-2, N-1. However the output of the difference transform could also be negative, e.g., -4 -> +4. Hence the problem still remains as to the representation of negative numbers. The symbols are thus made to represent numbers that fall within the range: -N/2, (-N/2 + 1), (-N/2 + 2), ... 0, 1, 2....(N/2 - 2), (N/2 - 1). Also, there are numbers that are not going to fall within this range. This problem is solved by using extender symbols. Thus the symbol 0 is not used to denote -N/2. Instead, it is used to denote that the mapped number m[i] is less than or equal to -N/2. Similarly, the symbol N-1 is not used to denote (N/2-1). Instead it is used to denote that m[i] is greater then or equal to (N/2 - 1). The other symbols retain their previous assignments. Thus the symbols 0 and N-1 are shift extender symbols. The entire mapped function can now be coded into these series of symbols. Each of these symbols now has a value between 0 and N-1. Fig. 14 describes the detail operation.

Huffman Coding

If the symbols themselves were used for the codes, the result would be a fixed-length compression, i.e., a word 'b' bits in length could be used to represent every symbol. However all the N symbols may not be uniformly distributed over the data. It is possible that some of the symbols have a much higher occurrence than the other symbols. This fact could be exploited by using a Huffman code on the data. This code assigns a bit-length to each symbol that is inversely related to the probability of occurrence of the symbol in the data. Thus symbols that occur more often are assigned a shorter bit-length than the symbols that occur less frequently. This results in further compression. The Huffman code requires that N symbols with a bit-length of b bits (such that N = 2**b) be fed into the algorithm. This is the reason for choosing the number N earlier, such that it is a power of

2. An optimum assignment is done only if there is a finite probability that can be associated with each of the N possible symbols.

Results of Lossless Compression
The image that resulted from the lossy compression was further compressed by using this lossless technique. The results of the analysis are shown in Figs. 15a-b. The greater the bit-size, the higher the compression possible. However, the complexity of generating the code increases, thereby increasing the computation time. Usually, the level at which the rate evens out is an acceptable bit-size. The compression ratios are higher for images that had fewer numbers of codewords. However, as was discussed in the section on lossy compression, there is a trade-off between information lost and compression attained.

OVERALL COMPRESSION RESULTS

The overall compression is achieved by multiplying the compression ratios achieved by the lossy and the lossless algorithms. Figure 16 shows the results of compression on the CAMS data-set. The overall compression possible for the 7 channel CAMS data with an RMS error of **15.8 pixels** was **195:1** and with an RMS error of **3.6 pixels** was **17.8:1**.

SUMMARY

A feasibility study was conducted to investigate the advantages of data compression techniques on multispectral imagery data. Two different techniques were implemented on remotely sensed data acquired from airborne scanners maintained and operated by NASA at the Stennis Space Center.

The first technique called **Vector Quantization** was used for lossy compression . The vector is defined in the multispectral imagery context as an array of pixels from the same location from each channel. The total number of original vectors is equal to the size of the image. Each of these vectors is quantized to a set of optimum vectors. This set is called the codebook. Since the size of the codebook is much smaller than the total number of original vectors, significant compression results. The rate of compression is programmable. However the higher the compression ratio, the greater is the degradation between the original and the reconstructed images. The analysis for 6 channels of data acquired by the Thermal Infrared Multispectral Scanner (**TIMS**) resulted in compression ratios varying from 24:1 (RMS error of 8.8 pixels) to 7 :1 (RMS error of 1.9 pixels). The analysis for 7 channels of data acquired by the Calibrated Airborne Multispectral Scanner (**CAMS**) resulted in compression ratios varying from 28:1 (RMS error of 15.2 pixels) to 8:1 (RMS error of 3.6 pixels). The technique of Vector Quantization can also be used to interpret the main features in the image, since those features are the ones that make up the codebook. Hence, Vector Quantization not only compresses the data, but also classifies it .

The compressed images are then reconstructed and compared with the original set using the Karhunen-Loeve Transform through the Principal Components Analysis. The eigenvalues of the covariance matrix obtained from the reconstructed data set are compared with the eigenvalues of the original set. This comparison is done for varying degrees of compression which are obtained by varying the size of the vector codebook. Also, the error obtained in substituting the reconstructed images for the original set is compared for different compression ratios. The effects of varying the size of the vector codebook on the quality of the compression and on subsequent classification are also presented. The eigenvalues of the covariance matrix of the original multispectral data-set were found to be highly correlated with those of the reconstructed data-set.

The second technique, called **Difference-mapped Shift-extended Huffman coding,** was 100% lossless i.e it resulted in images that were capable of complete restoration. Initially, the data was mapped to a difference transform. This transformed image was then converted into symbols using a specific bit-size. These symbols were then coded using Huffman coding. The output data from the Vector Quantization algorithm was further compressed without any increase in the RMS error by subjecting it to this technique. The TIMS data resulted in additional compression of 5.33 (for 24:1 compressed image) to 1.28 (for 7 : 1 compressed image). The CAMS data resulted in additional compression of 7 (for 28:1 compressed images) to 2.22 (for 8:1 compressed images).

Thus, the overall compression possible for the 6 channel TIMS data with an RMS error of **8.8 pixels** was **128:1** and with an RMS error of **1.98 pixels** was **8.8:1**. The overall compression possible for the 7 channel CAMS data with an RMS error of **15.8 pixels** was **195:1** and with an RMS error of **3.6 pixels** was **17.8:1.**

The algorithms were implemented in software and interfaced with the help of dedicated image processing boards to an 80386 PC compatible computer. Modules were developed for the task of image compression and image analysis. These modules are very general in nature and are thus capable of analyzing any sets or types of images or voluminous data sets. Also, supporting software to perform image processing for visual display and interpretation of the compressed/classified images was developed.

REFERENCES

Goldberg, M. and Sun, H.F., 1986. Image Sequence Coding Using Vector Quantization. *IEEE Transactions on Communication*, v. Com-34, iss. pp. 703-710.

Gonzalez, R. and Wintz, P., 1987. Digital Image Processing. Addison-Wesley Inc. Reading, MA, pp. 255-330.

Gray, R. M., 1984. Vector Quantization. *IEEE ASSP Mag*azine pp. 4-29.

Hang, H.M., and Haskell, B.G., 1988. Interpolative Vector Quantization Of Color Images. *IEEE Transactions on Communication*, v. Com-36, iss. 4, pp. 465-470.

Hang, H.M., and Woods J.W., 1985. Predictive Vector Quantization Of Images. *IEEE Transactions on Communication*, v. Com-33, iss. 11, pp. 1208-1219.

Huffman, D. A., 1952. A method for the construction of minimum redundancy codes. *Proceedings IRE* v. 40, pp. 1098-1101.

Leger, A., Omachi, T., Wallace, G. K., 1991. JPEG still picture compression algorithm. *Optical Engineering*, v. 30, no. 7, pp. 947-954.

Linde, Y., *et al*, Buzo, A., 1980. An Algorithm For Vector Quantizer Design. *IEEE Transactions on Communication*, v. Com-28, iss. 1, pp. 84-89.

Nasrabadi, N.M., and King, R.A., 1988. Image Coding Using Vector Quantization: A Review. *IEEE Transactions on Communication*, v. 36, iss. 8, pp. 957-971.

Panchanathan, S., Goldberg, M., 1991. A content-addressable memory architecture for Image Coding using Vector Quantization. *IEEE Signal Processing*, v. 39, no. 9, pp. 2066-2078.

Ramamurthi, B., and Gersho, A., 1986. Classified Vector Quantization of Images. *IEEE Transactions on Communication*, v. Com-34, pp. 1105-1115.

Shannon, C.E., 1948. A Mathematical Theory Of Communication. *Bell Systems Technical Journal* 27, pp. 379-423, 623-656.

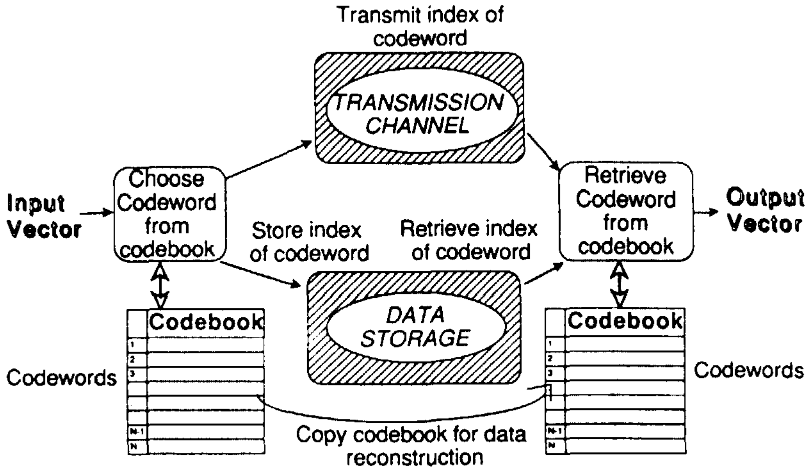

Transmit index of codeword
TRANSMISSION CHANNEL
Input Vector
Choose Codeword from codebook
Store index of codeword
Retrieve index of codeword
Retrieve Codeword from codebook
Output Vector
DATA STORAGE
Codebook
Codewords
1
2
3
N-1
N
Codebook
Codewords
Copy codebook for data reconstruction

Fig. 1 Vector Quantization - the concept.

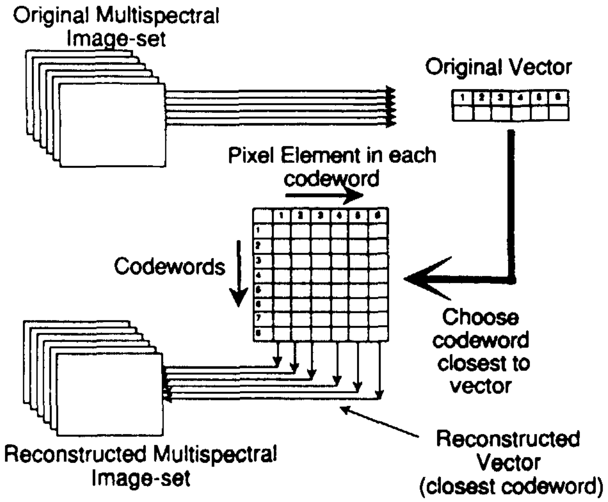

Original Multispectral Image-set
Original Vector
1 2 3 4 5 6
Pixel Element in each codeword
Codewords
Choose codeword closest to vector
Reconstructed Multispectral Image-set
Reconstructed Vector (closest codeword)

Fig. 2 Spectral Vector Quantization

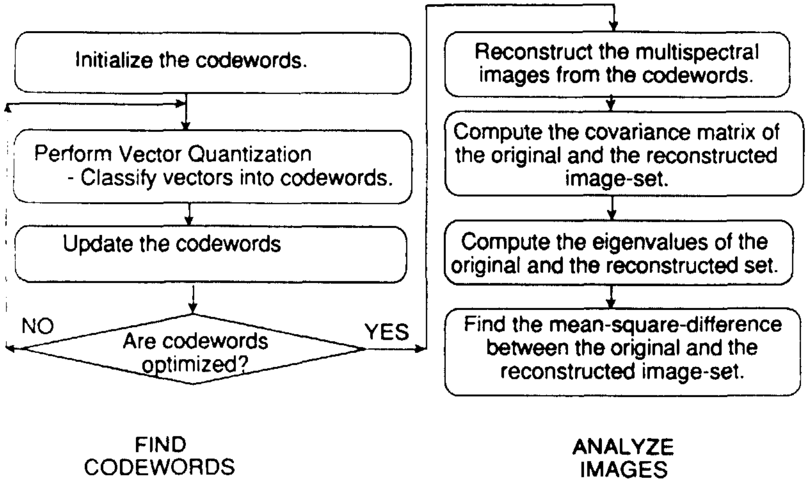

Initialize the codewords.
Perform Vector Quantization - Classify vectors into codewords.
Update the codewords
NO
Are codewords optimized?
YES
Reconstruct the multispectral images from the codewords.
Compute the covariance matrix of the original and the reconstructed image-set.
Compute the eigenvalues of the original and the reconstructed set.
Find the mean-square-difference between the original and the reconstructed image-set.
FIND CODEWORDS
ANALYZE IMAGES

Fig. 3 VECTOR QUANTIZATION - procedure & analysis.

ATTRIBUTES	TIMS	CAMS
Type of sensor	Airborne	Airborne
Type of radiation	Thermal infrared	Visible & infrared
Number of channels	6	9
Thermal bands	6 (8.2 - 12.2)	1 (8 - 12)
Visible bands	0	6 (.4 - .7)
Near Infrared bands	0	2 (.9 - 2.3)

Fig. 4 NASA Airborne Scanners at SSC.

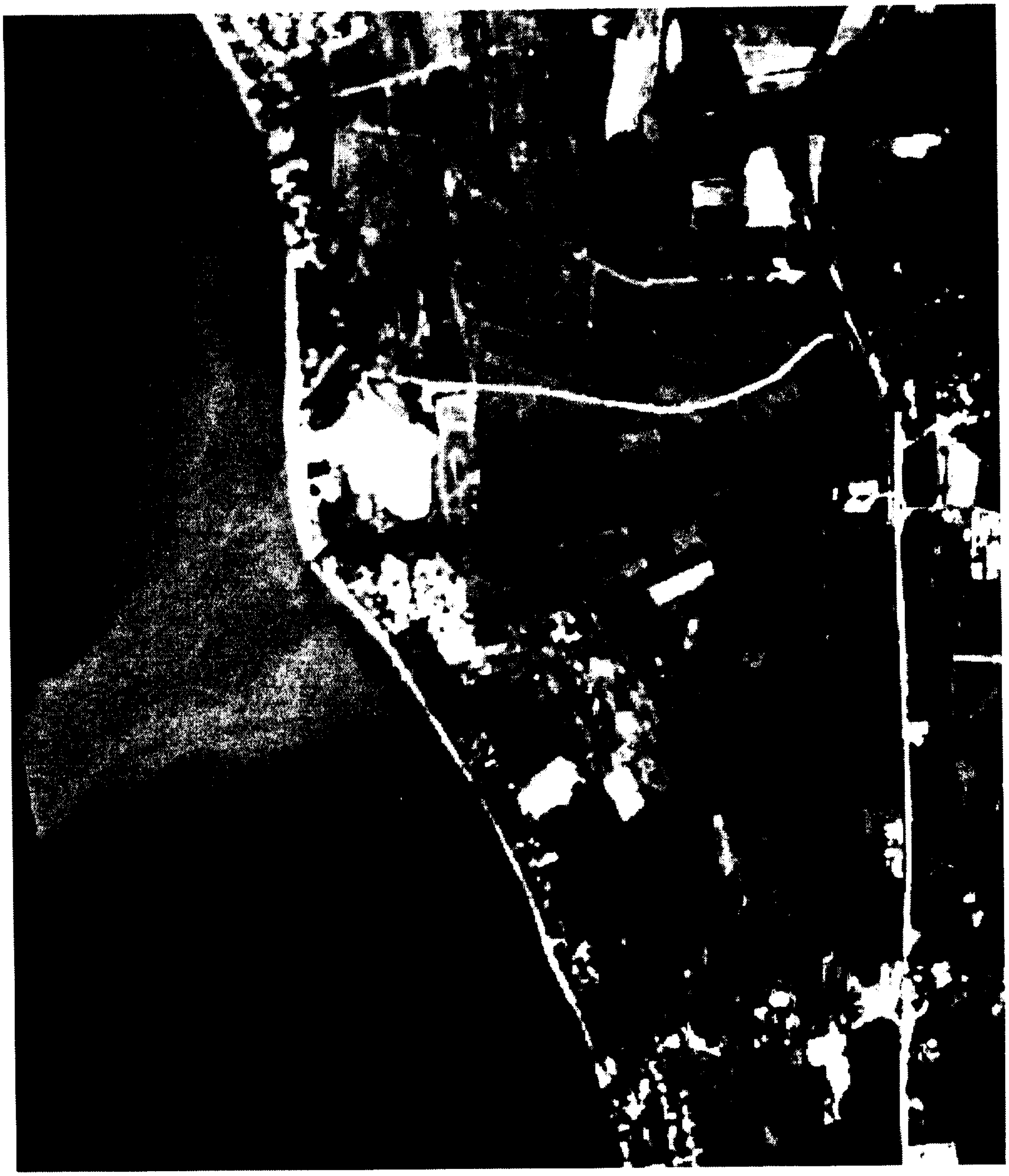


Fig. 5. CAMS acquired data - Channel 3.

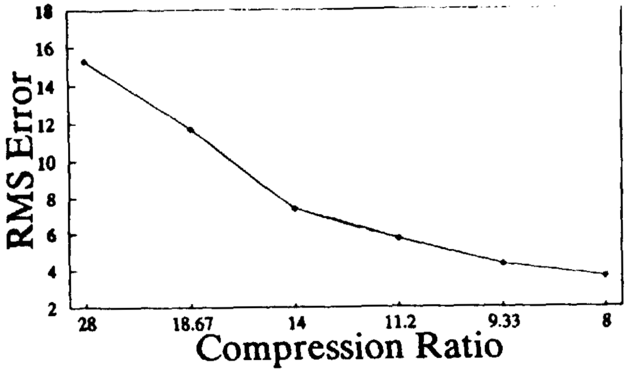


Fig. 6 RMS Error Between Original and Reconstructed Data

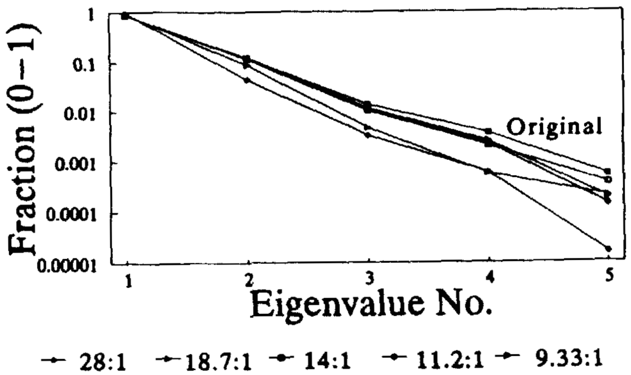


Fig. 8 Fraction of Total Variance in each Eigenvalue

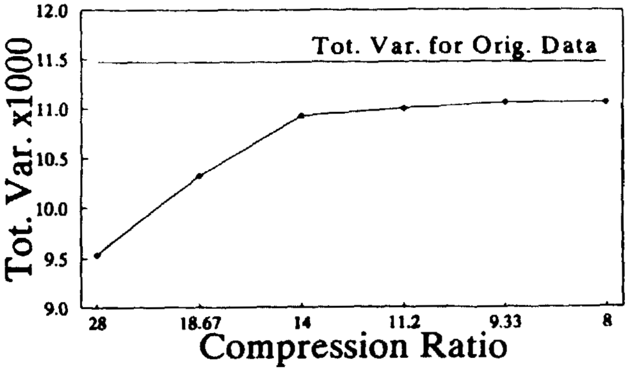


Fig. 7 Total Variance of Reconstructed Images

		Eigenvalues							
Cdwds	Ratio	1	2	3	4	5	6	7	SUM
128	8.00	.95134	.04472	.00321	.00059	.00001	.000057	.000056	100
64	9.33	.90651	.08800	.00463	.00057	.00020	.000027	.000027	100
32	11.20	.87112	.11606	.01030	.00202	.00036	.000065	.000065	100
16	14.00	.86408	.12271	.01070	.00221	.00013	.000079	.000079	100
8	18.67	.86178	.12382	.01140	.00245	.00018	.000188	.000188	100
4	28.00	.86033	.12429	.01242	.00258	.00026	.000051	.000051	100
ORIGINAL	1	.97754	.01617	.00481	.00099	.00028	.000190	.000170	100

Fig. 9a Relative distribution of variance in multispectral CAMS imagery data.

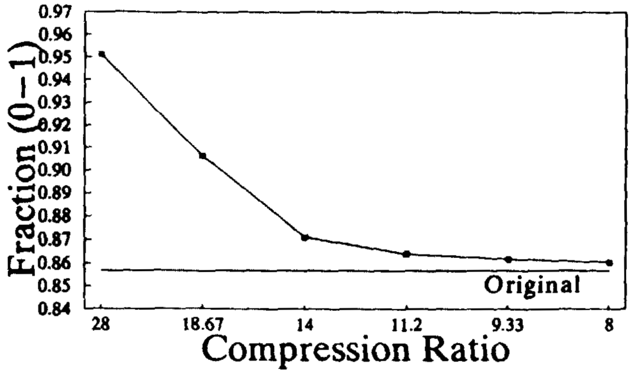


Fig. 9b Fraction of Total Variance in Eigenvalue No. 1

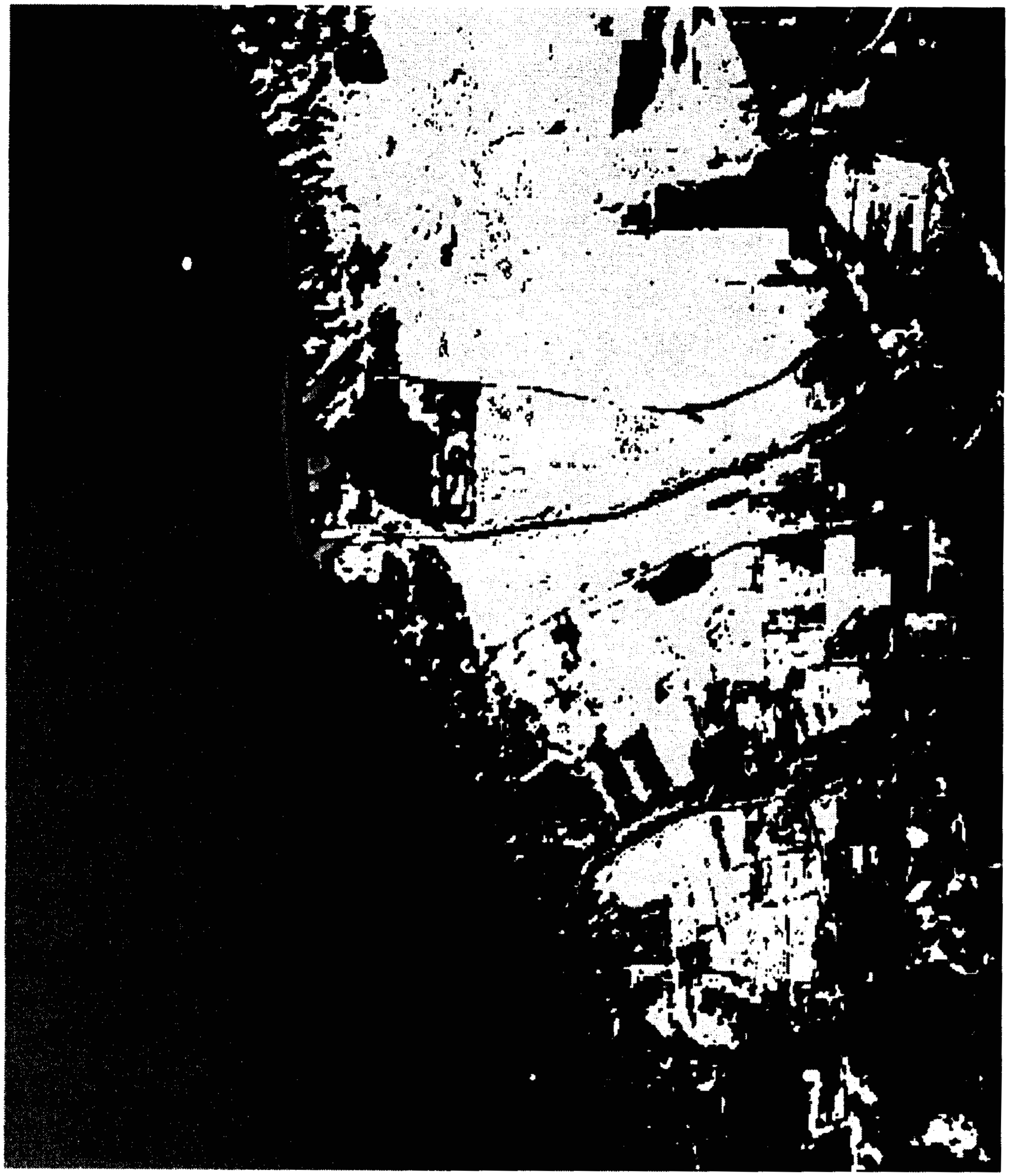


Fig. 10. Classified image - 4 codewords - compression ratio = 194.9.

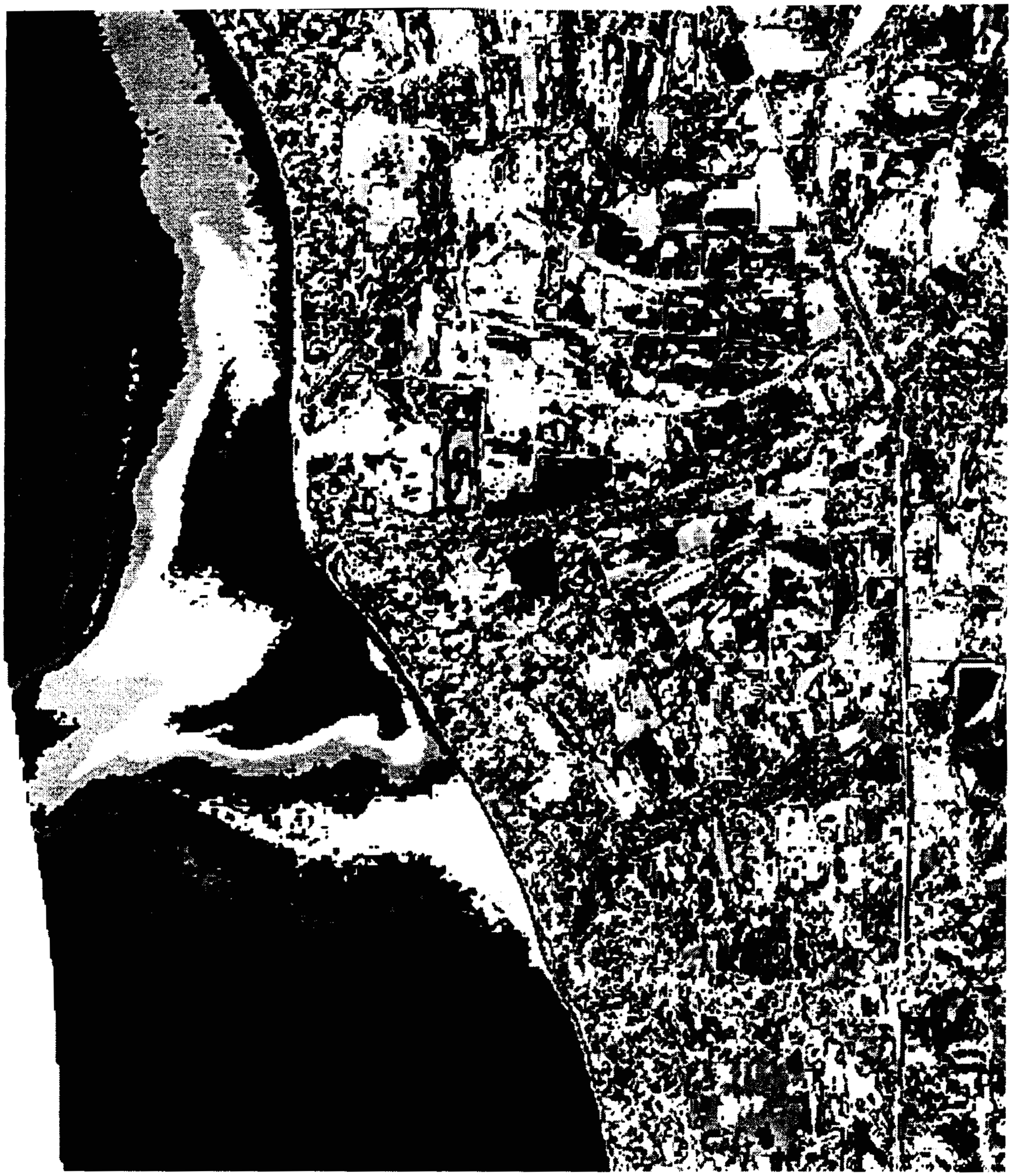


Fig. 11. Classified image - 64 codewords - compression ratio = 25.1.

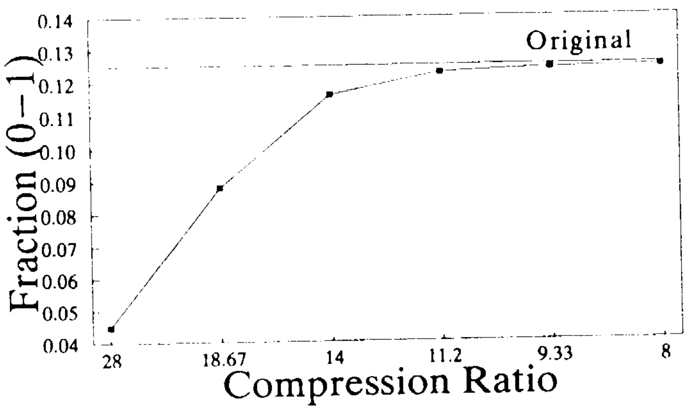


Fig. 9c Fraction of Total Variance in Eigenvalue No. 2

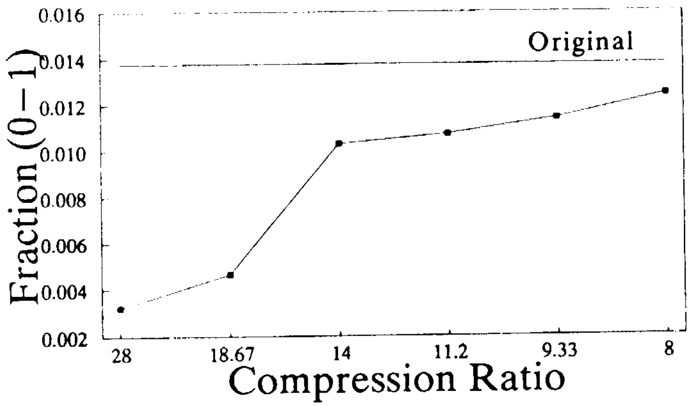


Fig. 9d Fraction of Total Variance in Eigenvalue No. 3

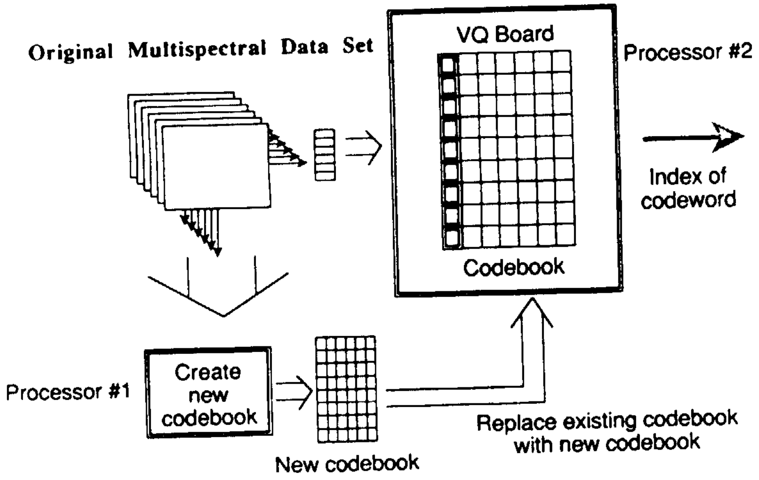


Fig.12 Implementation layout for vector quantization

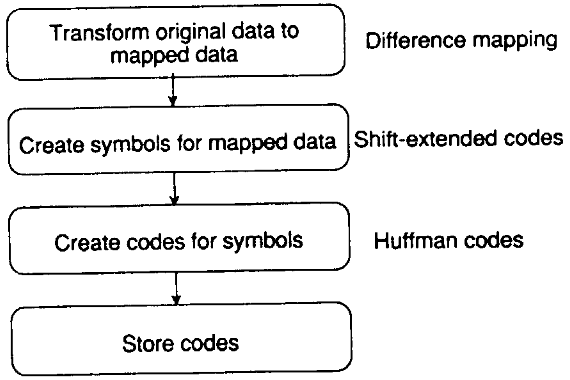


Fig. 13 Procedure for lossless coding.

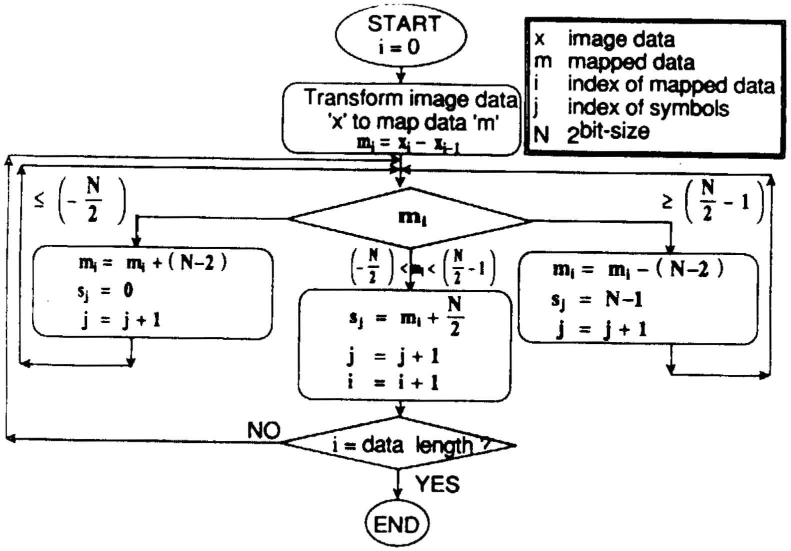


Fig. 14 Flow chart for difference mapping and shift extended coding

←CODEWORDS→

BITS	4	8	16	32	64	128
2	6.67	5.41	3.85	1.99	1.60	1.31
3	6.84	5.52	4.40	2.92	2.10	1.74
4	6.96	5.84	4.44	3.51	2.51	2.15
5	6.96	5.84	4.91	3.51	2.54	2.20
6	6.96	5.84	4.91	3.86	2.67	2.22
7	6.96	5.84	4.91	3.86	2.69	2.22
8	6.96	5.84	4.91	3.86	2.69	2.22

Fig. 15a Compression Ratios for lossless compression for CAMS images

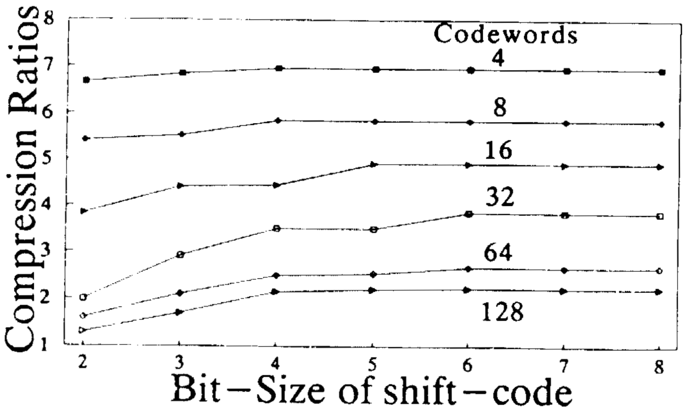


Fig. 15b Lossless coding on Vector Quantized images

<---VECTOR QUANTIZATION (LOSSY)--->LOSSLESS

CDWDS.	RATIO	RMS ERR	RATIO	RESULT.
4	28.00	15.28	6.96	194.88
8	18.67	11.69	5.84	109.03
16	14.00	7.39	4.91	68.74
32	11.20	5.67	3.86	43.24
64	9.33	4.29	2.69	25.10
128	8.00	3.63	2.22	17.76

Fig. 16 Resultant compression ratios for CAMS images

Smit

MANUFACTURING TECHNOLOGY PART 3: ROBOTICS

A Macro-Micro Robot for Precise Force Applications

Neville Marzwell
NASA- Jet Propulsion Laboratory
California Institute of Technology

Yulun Wang
Computer Motion, Inc.
Goleta, CA. 93117

Abstract

This paper describes an 8 degree-of-freedom macro-micro robot capable of performing tasks which require accurate force control. Applications such as polishing, finishing, grinding, deburring, and cleaning are a few examples of tasks which need this capability. Currently these tasks are either performed manually or with dedicated machinery because of the lack of a flexible and cost effective tool, such as a programmable force-controlled robot.

The basic design and control of the macro-micro robot is described in this paper. A modular high-performance multiprocessor control system was designed to provide sufficient compute power for executing advanced control methods. An 8 degree of freedom macro-micro mechanism was constructed to enable accurate tip forces. Control algorithms based on the impedance control method were derived, coded, and load balanced for maximum execution speed on the multiprocessor system.

Introduction

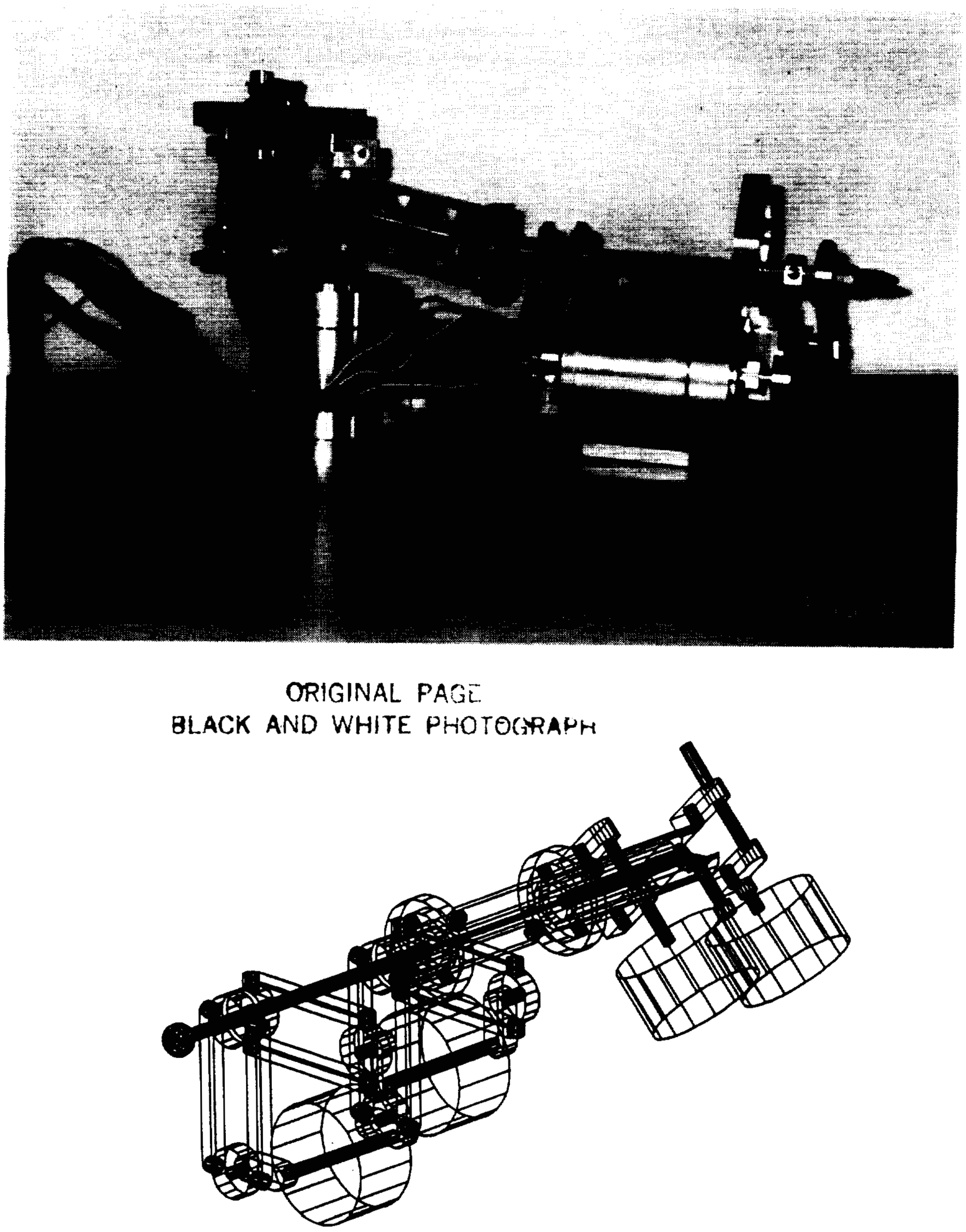
There are two main difficulties have made impeded the development of a high-precision force controlled robot. The execution of control strategies which enable precise force manipulations are difficult to implement in real time because these algorithms have been too computationally complex for available controllers. Also, a robot mechanism which can quickly and precisely execute a force command is difficult to design. Actuation joints must be sufficiently stiff, frictionless, and lightweight so that desired torques can be accurately applied.

The computational complexity problem has been addresses by building a high-performance real-time cost-effective multiprocessor system. This system is highly modular in structure, and was designed to support the needs of advanced robotic systems. The robot mechanism uses a macro-micro design, which allows the end-effector to have the properties of a small and light robot, yet preserves the workspace capability of a large robot. The following section discusses the mechanism design, and section 2.0 discusses the controller.

1.0 A Force Controllable Manipulator

A manipulator capable of delicate interactions with its environment must be designed differently from today's position controlled robots. It has been shown that a high-bandwidth, low effective end-effector inertia design is helpful for precise force control [1,2]. There are two design approaches to creating such a structure. One is to design the manipulator so that the entire structure is very light. This approach can be very costly since expensive materials and tight tolerances are required. The other approach is to attach a low-inertia small manipulator to the end of another larger and heavier manipulator. This macro-micro structure results in a combined structure with the low end-effector inertia of the micro robot and the large workspace of the macro robot.

The macro-micro design couples a 3 degree of freedom micro robot to the end of a 5 degree of freedom macro robot. A schematic and photograph of the micro design is shown in Figure 1, and a schematic of the macro design is shown in Figure 2. For the micro robot, the x and y directions are actuated with a parallel set of 5-bar-link mechanisms, one attached to each side end of the two motor shafts. The z motion is actuated by a fixed motor oriented perpendicular to the x and y motors. This motor is attached to the parallel link mechanism through a pair of universal joints. The range of motion is 2 centimeters along each axis. A fourth pneumatic motor, located furthest from the tip, rotates the tip through a series of transmissions at a constant speed for polishing, finishing, and grinding applications.



ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

Figure 1. The Micro Manipulator

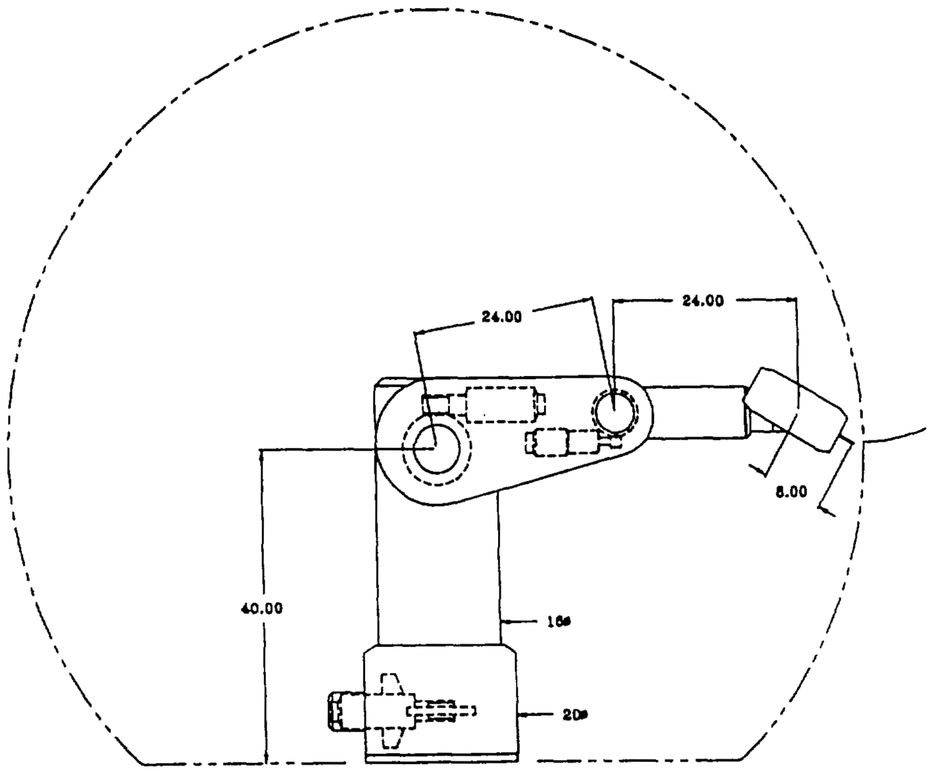


Figure 2. The Macro Manipulator

Since the macro and micro robots coordinate as a single system, many tradeoffs influence both designs. For example, the size of the micro robot's workspace influences the accuracy with which the macro robot must be able to position itself. The mass of the micro robot also influences the payload capability of the macro design. Our design strategy was to simplify the macro design by making the micro robot more capable. The main consequence of this decision is a large micro workspace, thereby allowing less accuracy and performance in the macro. However, the micro's workspace volume directly influences the overall mass and size of the design considerably. In our design, reducing travel along each dimension by a factor of two roughly reduces the size and mass of the micro robot by a factor of two.

The main objectives of the micro design were to minimize end-effector inertia, minimize joint friction, maintain tip orientation throughout the workspace, and support a maximum payload (i.e. force exertion) of 3 kilograms. The resulting tip inertia is roughly 250 gms. The joint friction was minimized by using direct-drive transmission and limited angle flex bearings at the joints. These limited-angle bearings offer virtually no friction. They do generate a spring force, however, which must be compensated for in the control law. Tip orientation is maintained by the parallel 5 bar link structure.

Secondary goals were to minimize the size and weight of the micro-manipulator. The final size is 35.5 by 19 by 17.8 centimeters, and the weight is 6.3 kilograms. Strain gages mounted on the links provide sensing for 5 degrees of freedom (as shown in Figure 1). Sensors for detecting a moment about the tip axis were not included.

The macro design is a 5 degree-of-freedom articulated manipulator, as shown in Figure 2. This manipulator supports the weight and continuous force exertion capability of the micro-manipulator throughout the workspace with 1g acceleration. A 1 meter reach was chosen as a reasonable workspace. The main features of this design are high mechanical rigidity, simple kinematics, large workspace volume, and cost effectiveness.

The 5 degree of freedom kinematic structure is very similar to the first five joints of a PUMA robot [3]. A 6th joint is unnecessary because the tip of the micro robot spins continuously. Link offsets, link lengths, and structural characteristics were designed to account for the size and mass constraints imposed by the micro-manipulator, however.

A variety of actuation methods have been considered. The options that were considered were direct-drive, harmonic drive, spur gear, worm gear, planetary gear, and different combinations of these. The goal was to maximize accuracy, resolution, and stiffness while staying cost effective. After various optimization procedures we decided on a harmonic drive - worm gear double reduction scheme for the first three joints. The last two joints, which carry a much smaller load, use harmonic drives.

The procedure for solving for the inverse kinematics equations of this robot is very similar to that of the PUMA robot and can be found in many of different robotics textbooks [4]. The kinematics and dynamic equations used for computed torque control can also be derived very easily using of the generalized formulations which have been developed [5]. However, because of the high reduction ratios of the transmissions, independent joint control is adequate.

2.0 A Modular Multi-processor Control System

A high performance multiprocessor system is used to satisfy the significant computational demands of controlling this robot. We designed this control system as a general purpose high performance controller with both hardware and software modularity as a key feature. The ability to easily rearrange and extend hardware and software modules to support different requirements for various tasks is particularly important in experimental projects such as this. Frequently designs are unable to accomodate even minor modifications without a major impact to the existing system configuration.

A schematic of the motion control system configuration is shown in Figure 5. The four basic units are the compute unit, the global memory unit, the position, velocity and digital I/O unit, and the A-to-D D-to-A unit.

The compute unit is based on Texas Instrument's TMS320C31 floating-point digital signal processor. In our earlier generation systems [6,7], we used a novel 3D computing processor which proved to offer much higher performance than DSPs or RISC processors on kinematic and dynamic calculations. However, due to the high cost of implementing this design using discrete datapath parts we opted to used an off-the-shelf processor. At a crystal speed of 33Mhz the TMS320C31 offers 33 MFLOPS of peak power. Each unit contains 2 Mbyte of program memory, 2 Mbyte of data memory, 2 programmable timers, interrupt capabilities for both the I/O Bus and the VME bus, and bus arbitration logic for accessing the I/O Bus. The memory is directly accessible by the host computer over the VME bus. Different levels of concurrency is provided to maximize execution speed. For example, the host may access data memory while the processor continues program execution. Programs are developed in either C or C++ on the host computer and downloaded to the appropriate unit before run time. Several libraries are provided to support program development. Remote procedure calls were provided so that UNIX services, such as printf(), scanf(), open(), and close(), are available for code development. Math functions, functions for accessing sensory data, and message passing functions for multi-processing are also provided.

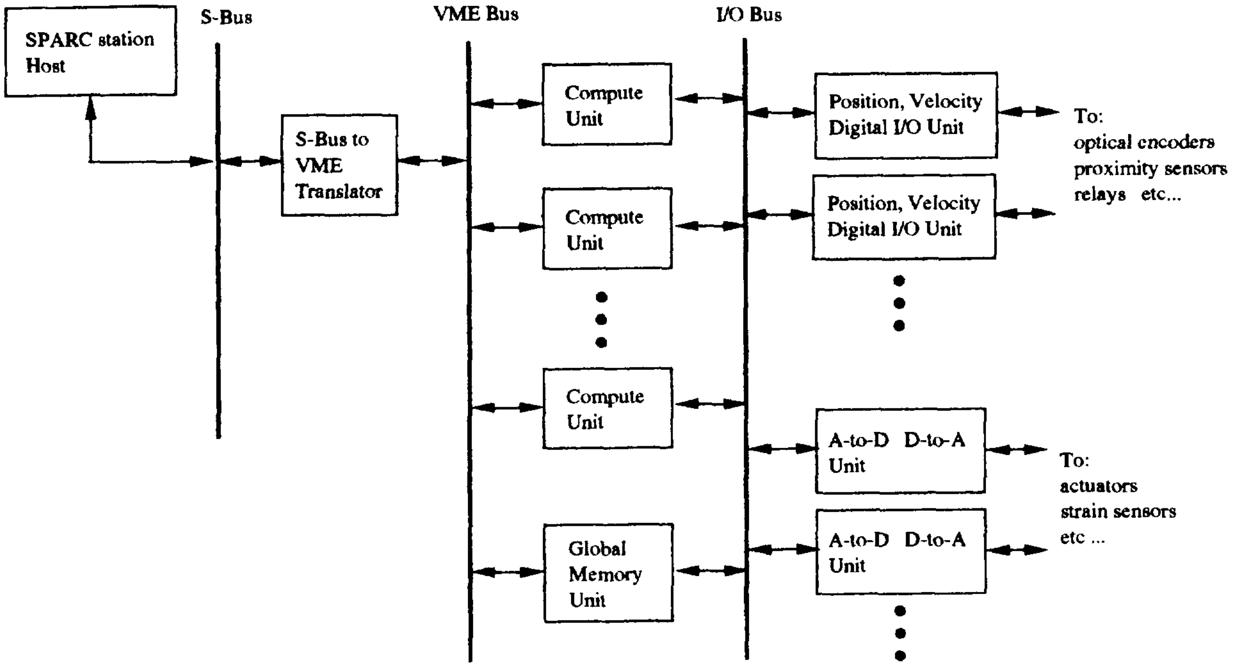

S-Bus
VME Bus
I/O Bus
SPARC station Host
S-Bus to VME Translator
Compute Unit
Compute Unit
Compute Unit
Global Memory Unit
Position, Velocity Digital I/O Unit
Position, Velocity Digital I/O Unit
A-to-D D-to-A Unit
A-to-D D-to-A Unit
To: optical encoders proximity sensors relays etc...
To: actuators strain sensors etc ...

Figure 5. The Motion Control System

The global memory unit contains 2 Mbytes of memory for passing messages between compute units, to and from the host, and to store global variables shared by multiple compute units. A mailbox message passing scheme is implemented to support multiprocessor communication. Information is passed from one compute unit to another compute by first acquiring the IO Bus, then writing the message into the target compute unit's mailbox, and then interrupting the target compute unit. The target compute unit reads its mailbox, and sends an acknowledgement to the sending compute unit. Hardware interlocking and interrupt mechanisms are included to achieve high bandwidth communication. Reading or writing a message requires ~3 ms overhead and another 180ns for each 32-bit word.

The position, velocity, and digital I/O unit accepts 6 channels of 2 channel quadrature encoder input and translates that into absolute position and velocity. Each channel also supports index pulse detection, which is generally used for position homing. Position is stored to 24-bit accuracy and velocity is stored to 10-bit accuracy. Thirty-two bits of digital input and 32 bits of digital output are included for instrumenting relays, proximity sensors, or other on-off type devices.

Velocity is generated by two different schemes, depending on the velocity range. At low speeds, velocity is generated in hardware by a free running counter which measures time between successive encoder counts. At high speeds, velocity is determined by calculating the number of encoder counts which have passed during the previous sample period. For each velocity read operation, the software automatically chooses between the two schemes by reading the velocity counter and comparing it with a threshold value. The result of this method is a more accurate velocity signal with minimized quantization effects.

Velocity is generated in hardware from the optical encoder signal by incorporating a free running counter chip which calculates the time between successive encoder pulses. Velocity is usually derived from a quadrature signal by subtracting the current position with the previous sample period's position. This subtraction may result in very quantized velocity signals especially at high sample rates, however. The hardware counter method produces a much more finely resolved velocity signal. There is still a problem, however, since at low speeds there may be significant time delay between new velocity acquisitions.

The A-to-D D-to-A unit provides 9 channels of 12-bit digital-to-analog output, and 8 channels of 12-bit analog-to-digital input. Separate digital to analog converters are provided for each output channel. A single analog-to-digital converter is multiplexed between the 8 input channels. Each channel requires 3 ms of conversion time. Software routines are provided to configure the card to only sample the channels which are in use. Conversion is performed continuously and asynchronously only on the channels being used. Therefore, the maximum delay from when the data was acquired to when it was read is 3 ms ¥ number of selected channels.

The software structure of the operating system level software is shown in Figure 6. Note that there is a clear separation between the real-time execution environment and the non-real-time UNIX environment. The UNIX environment is used for program development, user interface, and monitoring the real-time system. Because of the UNIX front-end, the robot interface must be carefully constructed such that the integrity of the real-time system is not lost. For example, UNIX service requests by the real-time system cannot be made while servoing since a real-time response from the UNIX process cannot be guaranteed.

Figure 7 shows the general hierarchy of the application software of the system. Macro calls provide fast access to the various hardware features of the system. C language routines provide the next layer, which support functions such as synchronizing multiple processes, remote procedure calls to the host, and algorithms for performing mathematic operations. At the highest level, object-oriented class libraries are supported in C++.

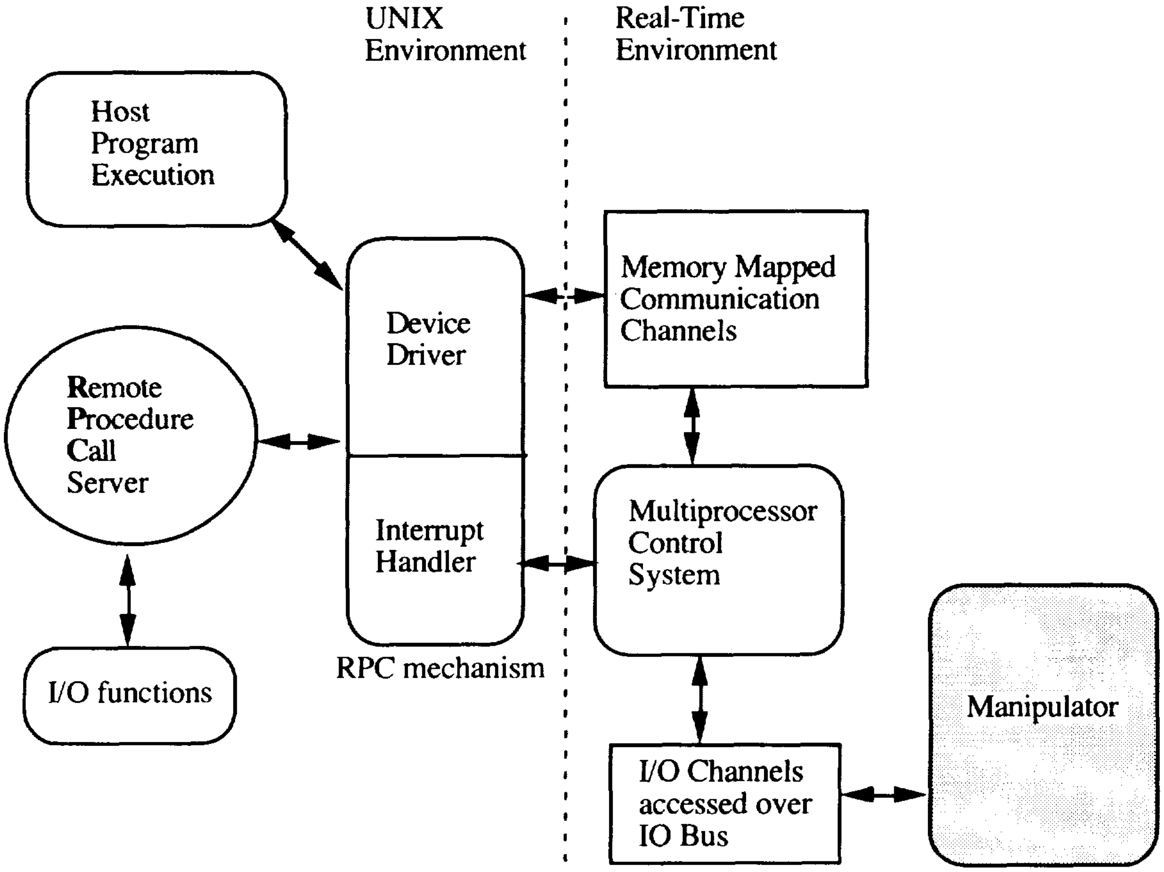

UNIX Environment
Real-Time Environment
Host Program Execution
Device Driver
Memory Mapped Communication Channels
Remote Procedure Call Server
Interrupt Handler
Multiprocessor Control System
I/O functions
RPC mechanism
Manipulator
I/O Channels accessed over IO Bus

Figure 6. System Software Structure

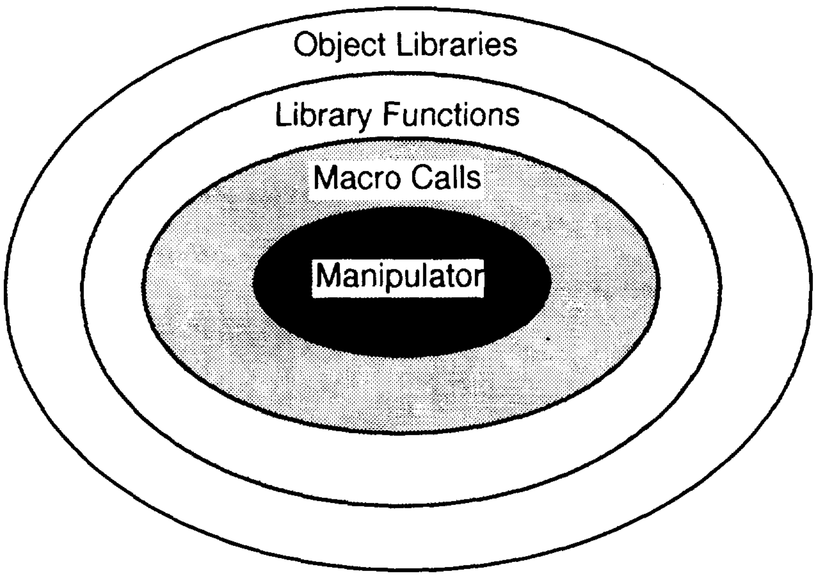


Figure 7. Software Support for Application Development

3.0 Impedance Control for a Macro-Micro Robot

The impedance control method enables a robot to interact with its environment in a well controlled and precise manner [8]. The manipulator's end-effector reacts to environmental disturbances in the same manner as a linear mass, spring, damper system. The mass, spring, and damper values are controlled electronically and can be different along different axes, and can continuously change during a trajectory.

This method is different from hybrid position/force control [9] since specific forces or positions are never specified. The control variable is the equilibrium point of the mass, spring, damper system without external forces. The advantage of this methodology is that a single control variable and control algorithm can be used to guide a robot through interactions with the environment. Hybrid position/force control, on the other hand, requires a switch in control methods and control variables whenever the robot changes the configuration in which it interacts with its environment.

Figure 8 gives an example of a trajectory specified by the equilibrium path where the manipulator comes into contact with a surface, slides across it, and then leaves the surface. Note that the nominal force exerted on the surface is proportional to the spring constant. By using the spring constant and surface location information, it is simple to calculate the equilibrium point's trajectory to produce a desired force across the surface. The force at the contact point will be influenced by contributions due to the mass and damper as well. Consequently, if precise force control is important, the smaller the mass and damper values are the better. The macro-micro design facilitates small mass values.

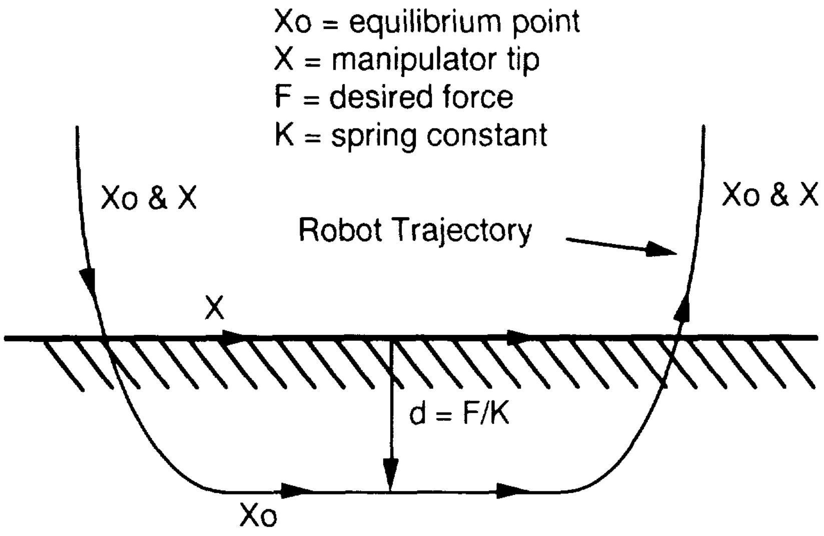


Figure 8. Manipulator trajectory specified by equilibrium point

The impedance equation can be written as follows:

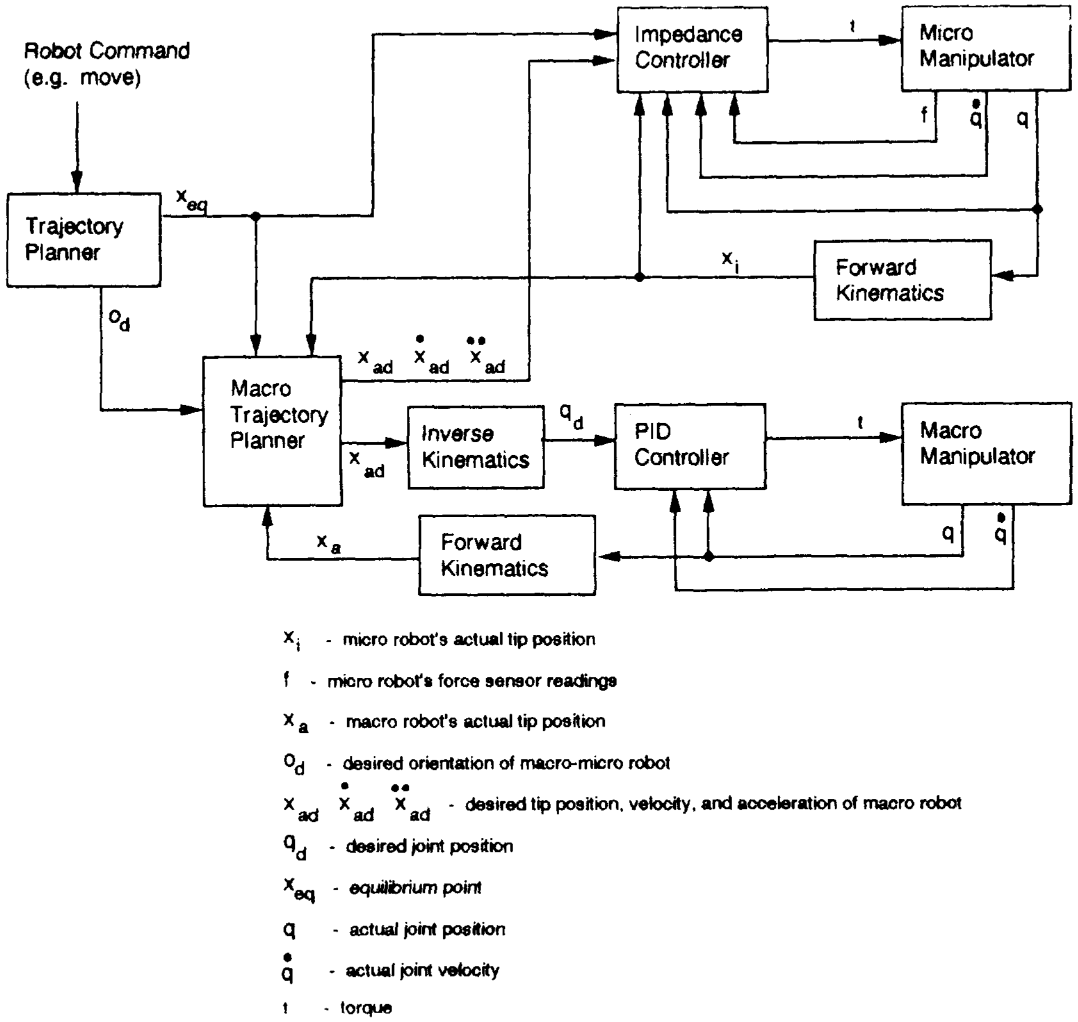
$$F_{ext} = M_s\left(\ddot{X}_R - \ddot{X}_o\right) + C_s\left(\dot{X}_R - \dot{X}_o\right) + K_s\left(X_R - X_o\right)$$

where

F_{ext}	external force applied to robot tip
X_R	tip position of macro-micro robot
X_o	desired equilibrium point of macro-micro robot
M_s	desired mass constant
C_s	desired damper constant
K_s	desired spring constant

Impedance control of a macro-micro design has the added complexity of managing the manipulator's redundancy to optimize force interactions by exploiting the micro robot's low tip inertia. In other words, the redundancy should be used to keep the micro robot from reaching its workspace limit, where one or more degrees of freedom would be lost. Our robot has 3 degrees of redundancy along the translational axes. Delicate interactions for translational motion is possible because of the micro robot. Orientation is left to the macro robot and is position controlled.

A block diagram of the control structure is shown in Figure 9. The impedance control law, which outputs torques to the micro robot, is derived by combining the desired impedance equation stated above with the equations of motion of the micro robot presented in section 1.2. Note that the servo control law for all 5 joints of the macro robot is a simple position controller without feedback from the micro robot. However, feedback from the micro robot is input into a real-time trajectory generator for the macro robot. This trajectory generator uses the robot's redundant degrees of freedom by constantly updating the macro robot's desired position such that the micro robot is centered in its workspace, and hence far from its workspace boundary. Consequently, entire manipulator can respond to external disturbances with the quick reaction of the micro robot over the entire workspace of the macro robot.



Robot Command (e.g. move)
Trajectory Planner
x_{eq}
o_d
Impedance Controller
t
Micro Manipulator
f
$\dot{q}$
q
Forward Kinematics
x_i
Macro Trajectory Planner
x_{ad} $\dot{x}_{ad}$ $\ddot{x}_{ad}$
Inverse Kinematics
x_{ad}
q_d
PID Controller
t
Macro Manipulator
q
$\dot{q}$
Forward Kinematics
x_a
x_i - micro robot's actual tip position
f - micro robot's force sensor readings
x_a - macro robot's actual tip position
o_d - desired orientation of macro-micro robot
x_{ad} $\dot{x}_{ad}$ $\ddot{x}_{ad}$ - desired tip position, velocity, and acceleration of macro robot
q_d - desired joint position
x_{eq} - equilibrium point
q - actual joint position
$\dot{q}$ - actual joint velocity
t - torque

Figure 9. Impedance control of macro-micro robot

The maximum distance which the micro will deviate from its center position is a relationship which includes the ratio of the maximum accelerations of the macro and micro, the magnitude and time of the maximum disturbance, and the reaction time of the servoing system. This information is important since it quantifies the critical tradeoffs between the micro's performance versus the macro's performance. We will obtain more insight into these relationships through experimentation of the robot.

With this control strategy, since the macro robot is purely position controlled it may be possible to apply this strategy to a micro connected onto the end of a commercial robot. However, the success of this approach is dependant upon the ability of the commercial robot to accept and quickly respond to new position commands. The requirements of a commercial robot used in this manner will become clearer with more experimentation on our robot.

5.0 Conclusion

An 8 degree of freedom macro-micro manipulator is controlled by an impedance-based controller, executed on a high performance multiprocessor control system. The manipulator's tip inertia is very low and can therefore react quickly to force disturbances. The control method compensates for manipulator dynamics, and can generate very precise torques. The multiprocessor offers sufficient compute power to meet the real-time demands of the control strategy.

Preliminary results show that this design will be capable of precise force control. More conclusive experimental results will be available at the end of the research effort in 1993.

Acknowledgements

The authors would like to thank Amante Mangaser, Keith Laby, Steve Jordan, and Jeff Wilson for their efforts in designing and implementing this robot. We would also like to thank Professor Yoshihiko Nakamura for his assistance in developing the control algorithm. 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. This work was funded from both the NASA/Jet Propulsion Laboratory under contract number NAS7-1149 and the National Science Foundation.

References

[1] Khatib, Oussama, "Augmented Object and Reduced Effective Inertia in Robot Systems," *Proc. of the American Control Conference*, Atlanta, Georgia, June 1988.

[2] Sharon, Andre, Neville Hogan, and David E. Hardt, "High Bandwidth Force Regulation and Inertia Reduction Using a Macro/Micro Manipulator System," *Proc. of the IEEE Conf. on Robotics and Automation*, Philadelphia, Penn., April 1988.

[3] Leahy, M.B. and et. al., "Efficient Dynamics for the PUMA-600," *Proc. of the IEEE Conf. on Robotics and Automation*, San Francisco, CA., 1986.

[4] Wolovich, William A., *Robotics: Basic Analysis and Design*, Holt, Rinehart and Winston, New York, 1987.

[5] Nakamura, Yoshihiko,"Unified Recursive Formulation of Kinematics and Dynamics of Robot Manipulators," *Proc. of the Japan - USA Symposium on Flexible Automation*, Osaka, Japan, July 14-18, 1986.

[6] Wang, Yulun, Amante Mangaser, Steven Butner, Partha Srinivasan, and Steve Jordan,"The 3DP: A Processor Architecture for 3-Dimensional Applications," *IEEE Computer*, January 1992.

[7] Wang, Yulun and Steven E. Butner, "RIPS: A Platform for Experimental Real-Time Sensory-based Robot Control", *IEEE Transactions on Systems, Man, and Cybernetics*, vol 19, no 4, July/August 1989, pp 853 - 860.

[8] Hogan, Neville, "Stable Execution of Contact Tasks Using Impedance Control," *Proc. of IEEE Int. Conf on Robotics and Automation*, Raleigh, North Carolina, 1987.

[9] Raibert, M. and J. Craig, "Hybrid Position/Force Control of Manipulators, *Journal of Dynamic Systems*, Measurement, and Control, vol. 102, pp. 126-133.

A FAULT-TOLERANT INTELLIGENT ROBOTIC CONTROL SYSTEM

Neville I. Marzwell
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109

Kam Sing Tso
SoHaR Incorporated
Beverly Hills, CA 90211

ABSTRACT

This paper describes the concept, design, and features of a fault-tolerant intelligent robotic control system being developed for space and commercial applications that require high dependability. The comprehensive strategy integrates system level hardware/software fault tolerance with task level handling of uncertainties and unexpected events for robotic control. The underlying architecture for system level fault tolerance is the distributed recovery block which protects against application software, system software, hardware, and network failures. Task level fault tolerance provisions are implemented in a knowledge-based system which utilizes advanced automation techniques such as rule-based and model-based reasoning to monitor, diagnose, and recover from unexpected events. The two level design provides tolerance of two or more faults occurring serially at any level of command, control, sensing, or actuation. The potential benefits of such a fault tolerant robotic control system include: 1) a minimized potential for damage to humans, the work site, and the robot itself, 2) continuous operation with a minimum of uncommanded motion in the presence of failures, 3) more reliable autonomous operation providing increased efficiency in the execution of robotic tasks and decreased demand on human operators for controlling and monitoring the robotic servicing routines.

INTRODUCTION

The reliability issue must be addressed before robotic systems can be dependably used in critical applications such as servicing the Space Station Freedom and patient monitoring and tending tasks in medical facilities. This paper describes a comprehensive approach which integrates the handling of hardware, software, communication, and operational errors in robotic systems. Although some work has been done on the handling of uncertainties and unexpected events during task execution [1, 2, 3, 4], there has been little research on the handling of system level hardware and software failures in robotics. The research addresses this void by means of a comprehensive strategy for integrating system level hardware/software fault tolerance with task level handling of uncertainties and unexpected events.

Table 1 shows our integrated approach to robotics fault tolerance. Errors are handled on two levels: the *system level* which includes the computers and other hardware, control software, and communications, and the *task level* which includes anomalies and uncertainties associated with the physical environment during task execution. Examples of faults, errors and recovery, together with the general fault tolerance strategy and our specific approach to handling them are shown for each of these levels.

In our terminology a *failure* is a difference between the actual behavior of a system and the expected behavior. An *error* is an undesired system state and a *fault* can be considered as a low-level failure of some subsystem. In other words, a fault causes the system to get into an error state and the failure behavior is a manifestation of the error state. For example, a faulty motor caused a servo gripper stuck at an erroneous open position which led to the failure to grasp an object.

There are four main classes of errors that can be identified in a robotic control system. These are hardware errors, software errors, communications errors, and operational errors [5].

- *Hardware errors* occur in all kinds of mechanical and electrical mechanisms, in control systems, in sensory devices, and in electronic and computer systems. They are caused either by component failure

Table 1: Integrated Approach to Robotics Fault Tolerance

Level	Class	General Strategy	Example Fault	Example Error	Example Recovery	Specific Approach
System	hardware	redundancy	processor stop	missed output	replaced by backup processor	extended distributed recovery block architecture
	software	design diversity	unanticipated singularity not handled	erroneous setpoint output	assertion check & alternate algorithm	
	communication	coding	line noise	data scrambled	data encoding check & retransmission	
Task	operation	intelligence	weak grip force	object slipped	replan & regrasp	knowledge-based system

or by design faults. A common technique for tolerating hardware failures and faults is the introduction of some form of redundancy. Key components of the system are replicated and work in parallel. If one of the replicated components fails, the remaining components continue to operate. The user does not notice the error, and the system continues to function correctly. The EDRB architecture described in this paper incorporates hardware fault tolerance in the form of a node pair and the associated fault detection and recovery software.

- *Software errors* occur through design faults in programs. With the increase in sophistication of robotic systems, software has become more significant and complex. The conventional technique for software reliability is extensive verification and validation. It is well known that software testing can only reveal the presence of faults, but not their absence. As a result, software fault tolerance techniques have been used to achieve high reliability for critical applications which may endanger human life or entail great financial loss. Tolerance to design faults relies on the application of design diversity, which creates diverse software components from a common requirement. Their diversity is introduced by the use of independent programmers, algorithms, programming languages, and tools. The goal is to increase the probability that software errors will be tolerated by diverse software components. The EDRB architecture incorporates software fault tolerance by using two diverse versions of the software coupled with an on-line acceptance test which can detect failures in either version prior to transmitting their output to the actuators.

- *Communication errors* occur in command and status information communicated between control computers, robotic controllers, and sensory devices. They are caused by transmission error due to noise, loss of synchronization due to timing errors, or loss of data due to hardware failures. The first two failures can be detected by the use of coding and recovered by retransmission. Hardware failures can only be tolerated by redundant communication links. The EDRB architecture addresses communication errors through encoding and redundant communication links.

- *Operational errors* are the physical errors that occur in the robot task environment. These are not software or hardware errors but refer to a range of faults due to uncertainties and unexpected events that happen during task execution. For example, an autonomous robot vehicle might unexpectedly find an obstacle in its path. Alternately, a robot might find that it has failed to grasp an object either because the object is not present or because the object slipped from its grip. Some failures due to defective components are also classified as operational errors because the conventional redundancy technique to tolerate their failures is not viable for them. For example, a "standby robot," even if economically possible, could not access all the operating space of a failed robot and therefore cannot be used to replace the failed robot. Operational errors are the types of error conditions that an intelligent robot must be designed to detect and recover from.

The next sections describe the system level fault tolerance provisions which tolerate hardware, software and communications faults; and the task level fault tolerance provisions which tolerate operational faults.

SYSTEM LEVEL FAULT TOLERANCE

A real-time fault-tolerant distributed architecture called the Extended Distributed Recovery Block (EDRB) [6] will be used to handle system level faults. The underlying fault tolerance algorithms and mechanisms are based on extensions to the distributed recovery block [7] which is in turn based on the classical recovery block [8] with real time extensions.

Figure 1 is a top level diagram of a robotic control system which incorporates the EDRB. This configuration is a typical teleoperated-autonomous dual-arm robotic system with supervised autonomy for space telerobotics [9]. Fault tolerance for hardware, software, and communications failures is provided for the Task Execution System because it is remotely located and must respond rapidly to these failures. Although other system elements are not shown as requiring fault tolerance in this example, nothing precludes the application of the EDRB for the Task Planning System, the interface, or other elements if such were required.

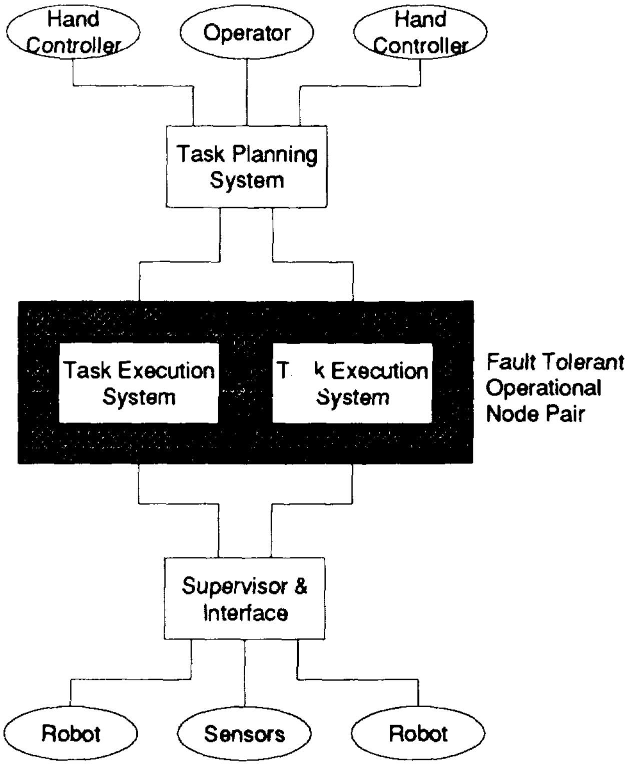


Figure 1: Fault-Tolerant Robotic Control System Based on the EDRB Architecture

In the general terminology of the EDRB, the replicated Task Execution System computers are collectively referred to as an operational node pair. One member of the node pair, called the *active* node, provides active 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 an indication of their states of health. 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***. 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 thȩ 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. For example, in a sensor processing application, the primary routine might use Kalman filtering whereas the alternate routine might use a moving average. After each processing iteration, an online acceptance test checks the validity of the output of both the primary and alternate routines. If the acceptance test shows that an error has occurred in the primary routine, the output will be taken from the alternate routine and control is passed to the shadow node.

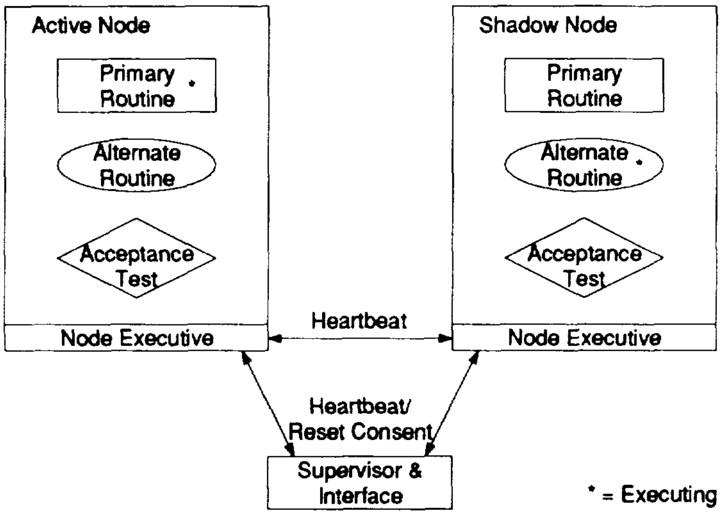


Figure 2: EDRB Software Structure

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 (by means of acceptance tests, timers, and alternate routines).
- Hardware or system software failures (by means of information encoding, timers, and redundancy)
- Communications link failures (by means of encoding, retransmission, and redundant communication links)
- Spurious recovery actions (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 times. 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.

The diversity to be achieved in the primary and alternate routines is highly dependent upon the application. The free motion of a 7-DOF redundant arm is used to illustrate how software diversity can be achieved. Two possible independent approaches to *configuration control* of redundant manipulators [10]: (1) the *Jacobian pseudoinverse* [11] which has good tracking but cannot handle singularity, (2) the *damped least square* [12] which is singularity robust but has bad tracking near singularity. The primary routine will use Jacobian pseudoinverse to ensure good tracking whereas the alternate routine will be based on the damped least square 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. On the other hand much of the "framework" coding (i.e., preparation of input, buffering of output, etc.) will be common to both modules because of the lower likelihood of failures. Should experience demonstrate that this is not the case, then these and other software components can also be made diverse.

The acceptance test is the single most critical element of the EDRB. If it fails to reject an incorrect result, or fails to accept a correct result, it comprises a single point of failure. As such, the acceptance test must be both simple and general. While this is a rigorous requirement, it is not impossible to meet in the context of 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 small, 3) the command joint angles are not close to joint limits, and 4) the observed force/torque values are close to the gravitational force of the grasped object.

TASK LEVEL FAULT TOLERANCE

The task level fault tolerance in the proposed design is a knowledge-based system which uses rule-based and model-based reasoning to monitor, diagnose, and recover from unexpected events that occur during the execution of robot tasks.

Most of the present robotic systems handle unexpected events by preprogramming error detection and recovery procedures for every probable error that can be perceived [13]. This approach is inefficient, and it is difficult to completely handle all failures. On the other hand, most of the artificial intelligence research efforts have focused on detection and recovery from failures in simulated robots. They made unrealistic assumptions about the real world and ignored performance and integration issues [3]. Other attempts at automatic error recovery without human intervention have not been used in real applications, because they could not handle the vast range of potential error conditions [2].

The approach outlined in this research addresses these problems as follows:

1. *Emphasis on the support of the robotic task execution system*:

   Most AI research on robotics has emphasized the task planning level. Experience from the NASA/JPL Space Telerobotics Program has shown that monitoring and recovery at the task execution level is both necessary and effective because of its quick response. Our design partitions the fault tolerance strategies into two levels: the *local level* which resides in the task execution system, and the *global level* which resides in the task planning system. The local level provides quick and simple monitoring and recovery actions, while the global level provides extensive and complete monitoring and recovery. The two levels complement each other in their efforts to monitor, diagnose, and recover from failures.

2. *Emphasis on the role of the operator in failure recovery*:

   It is doubtful that any strategy developed for automatic error recovery in a robotic control system can cover all potential failures. Even if such a strategy were developed, it would take some time before confidence in the automatic recovery capabilities would be gained. Therefore, it is necessary to develop a system in which the operator is integrated into the failure recovery process. The operator will always have the capability to approve, query, and intervene a recovery plan. Pertinent information is relayed to the operator with an emphasis on the human/computer interface design.

3. *Emphasis on the performance and integration of the knowledge-based system*:

   Most AI research has been done with familiar AI languages such as Lisp and Prolog in simulated application environments. Problems such as slow response time, communication difficulties, and interface incompatibilities were not addressed. In this research we use C and the CLIPS expert system shell (implemented in C) to enhance performance and minimize integration problems with the underlying UNIX operating system and the X Window based graphical user interface.

The following subsections describe the local and global task level fault tolerance strategies.

Local Fault Tolerance Strategies

Experience from the NASA/JPL Space Telerobotics Program has shown that local monitoring and recovery actions in the task execution system are both necessary and effective [9]. They are necessary because quick response time is always needed in emergency situations. At the task execution level, monitoring and recovery can be achieved in real time, e.g. at every sample at a rate of 200 Hz. Recovery actions can be initiated at the time of failure occurrence. This is especially crucial in ground/remote telerobot systems where the task planner is located at the ground station and the time delay is significant. Local monitoring and recovery are effective because most failures manifest themselves in excessive force, jerks, or undesired motion. Failures can be detected by monitoring the force/torque thresholds, joint velocities and limits without considering the robotic task context. The recovery action implemented in the JPL Telerobot System was to simply stop the arm, thereby preventing it from damaging the work site and itself. It was found to be an effective initial step for further recovery actions by the operators.

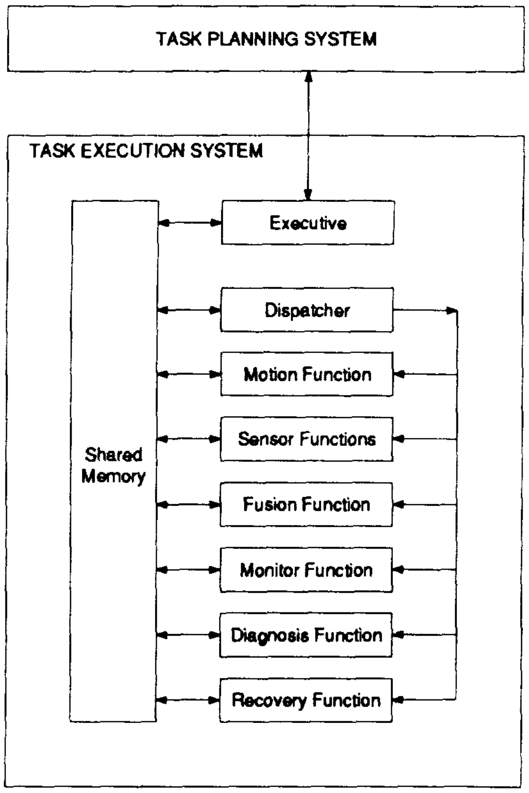


Figure 3: Task Execution System Architecture Supporting Local Fault Tolerance

Local monitoring and recovery are implemented in a task execution system such as the modular architecture [14] as shown in Figure 3. The Executive Process communicates with the Task Planning System to accept new commands and to return new statuses. The Execution Process consists of various modules that provide task execution, monitoring, and reflex capabilities. The *Dispatcher* starts and stops the execution of the various functions. The *Motion Function* sets up the kinematic relationships for interpolated motion. The *Fusion Function* combines the motion perturbations from each sensor with the nominal interpolated motion. These motion perturbations are calculated by the *Sensor Functions*.

This architecture is extended to support local task level fault tolerance as follows:

- *Monitoring* is done in the *Monitor Function*, which tests for various sensor values and conditions in every sampling period. Some examples are force/torque values, joint limits, joint singularities, and elapsed time. The Dispatcher is signaled when an anomaly is detected. The monitor rules implemented in the Monitor Function have the following general form:

  if <situation> **ensure** <condition>

  For example, for continuous collision testing, the monitoring rule is:

  if true **ensure** f/t < safety-threshold

  Another example in monitoring grasping is:

  if contact-sensor = CONTACT **ensure** finger-separation ≈ object-size

  The monitor rules not only allow us to test thresholds, they provide a means to monitor a sensor execution profile and test events that occur only in specific situations.

- *Recovery* is activated by the Dispatcher once an anomaly has been signaled by the Monitor Function. The objective of recovery at this level is to provide a fast reflex action to protect the arm and the work site. It is only the initial step of the whole recovery process. Although in most cases stopping the arm is appropriate to safeguard the hardware, there are situations where other recovery actions are needed. For example, if unstable conditions occurred during insertion, stopping the arm may still inflict damaging force to both the arm and the object. Other reflex actions such as relax and return to original position will be implemented.

- *Diagnosis* at this level is used to help the global recovery function to test, re-synchronize, and re-initialize sensors. Specific testing procedures will be devised for each sensor to help the Global Recoverer to determine if it has failed. For example, a force/torque sensor can be tested by comparing ten consecutive readings to ensure that the values are not fixed and that they are reasonable. Many sensor failures are due to communications being out of synchronization or in erroneous internal states. Functions that are able to re-synchronize and re-initialize the sensors will be implemented to assist the global error recovery strategy.

Global Fault Tolerance Strategies

Without the world model and required knowledge and the power to reason, local fault tolerance is limited to detecting errors and using simple reflex actions to protect itself and its work site. Global fault tolerance complements local fault tolerance provisions in that it makes extensive use of spatial reasoning, rule-based reasoning, and model-based reasoning to monitor, diagnose, and recover from failures. Figure 4 shows the architecture of the Task Planning System which supports global fault tolerance.

- *Monitor*: The global monitoring uses both rule-based and model-based reasoning to detect errors that cannot be detected at the local level. The rule-based reasoning is similar to that of the local level but is more sophisticated. For example, if the screwdriver does not seat correctly on the screw in bolt turning, it can be detected by comparing the execution force/torque profile with the force/torque signature stored in the knowledge base. The global monitoring also uses model-based reasoning. For example, a task which inspects the surface of a rectangular frame can detect arm motion errors based on the geometric model of the frame stored in the world model.

- *Diagnosis*: The global diagnosis decides what really occurred based on the raw data indicating an error. For example, a failed grasp may be caused by misorientation of a part/tool, a missing part, slippage of a part, a gripper that cannot close, incorrect compliance, collision, etc. Rules have been developed to perform the diagnosis. These rules use raw sensor data, the semantics and context of the failed task, and the physical behavior of the objects in the work site.

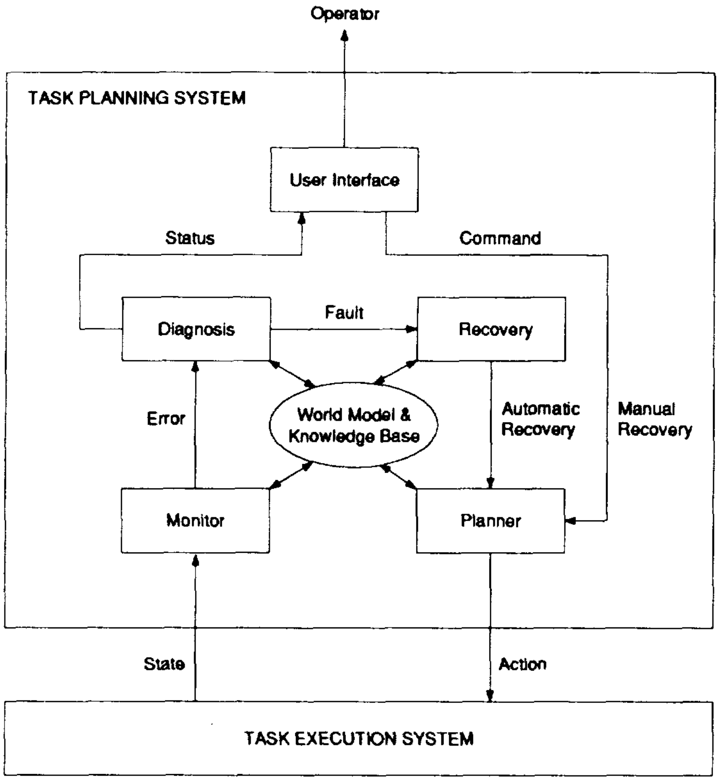


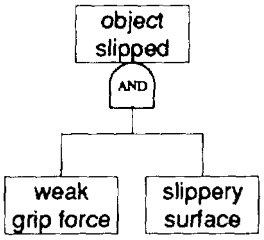
Figure 4: Task Planning System Architecture Supporting Global Fault Tolerance

- *Recovery*: After a fault has been identified, the global recoverer generates pertinent recovery actions according to the fault and the task context. Recovery actions can range from simple to complicated:
  - *Retry*: Faulty readings may be transient and a simple retry of the unfinished plan may be adequate.
  - *New Parameters*: Default parameter values in the task plan may not be appropriate for the situation; new parameter values are used to retry the failed action. An example is unstable compliant motion that becomes stable after the gain is reduced.
  - *Corrective Actions*: Extra actions are needed to correct the erroneous state. For example, a regrasp action is needed after an object slips during grasping.
  - *World Model Update*: An update to the world model is needed because it is found inconsistent with reality. For example, a missing object found during grasping should delete that object from the world model.
  - *Replan*: The original task plan has to be replanned. For example, a new path is used to avoid a collision.
  - *Reconfiguration*: Configuration of the available resources needs to be updated after the hardware has been found to have failed. For example, if a robot arm has failed, the planner should use the other arm to perform its tasks, if possible.
- *Planner*: The planner stores task plans for nominal tasks and contingency plans for failed actions. Although it is not the scope of this research, it would be useful if the planner could generate collision-free paths based on the world model.

  One important feature that has been added to the planner is the capability to check commands that are initiated by the operator. Routine and tedious tasks often cause human fatigue and boredom.

This can lead to human error and a resultant hazardous situation. The user interface ensures that no erroneous or unsafe commands are given by the operator. And the planner will check whether the preconditions of the actions are satisfied, and the postconditions are acceptable.

- *User Interface*: The user interface plays an important role in the handling of failures in the system. As the system becomes more intelligent, it is expected that the demand from the operator will be reduced. Nonetheless, the user interface will always allow the operator to approve, query, and intervene a recovery plan. However, it is not enough for the operator merely to take control. The operator needs information such as the robot's status, position, and previous activities. At the time an operator must take control, he may not know such information. The system condenses this information and relay the important data to the operator control station.
- *World Model*: Information representing the work environment is integrated and assimilated in the world model. It is used by all components in the Task Planning System. The model is updated every time new information about the environment is received as a result of robot actions, sensory data, physical processes, and fault diagnosis.
- *Knowledge Base*: The knowledge base is the repository of all the rules and actions for planning, monitoring, diagnosis and recovery. One of the critical problems in knowledge base construction is the acquisition of expert knowledge. A fault tree analysis of the target robotic system will be performed and the resulting fault trees will be used as the basis for creating IF-THEN rules. For example, the subtree in Figure 5a can be translated into the rule in Figure 5b. The fault tree is helpful because it provides a visual representation of the way in which failures are propagated in the system.



(a)

```
IF weak grip force AND
   slippery surface THEN
   object slipped
```

(b)

Figure 5: A Fault Tree and its Translated Rule

CONCLUSIONS

A comprehensive strategy is described which integrates system level hardware/software fault tolerance with task level handling of uncertainties and unexpected events in robotic control. A prototype of the EDRB has been implemented using PC/AT-386 computers, ARCnet, and the QNX real-time operating system. Extensive evaluation has concluded that the resulting system tolerates a broad range of hardware, system software, and application faults, with a 200 millisecond guaranteed response time. The system is currently being rehosted to a VME-based multiprocessor system using 68040 single board computers and the VxWorks real-time operating system. It is expected that the faster hardware and system software will achieve the 5 milliseconds response time requried by the Manipulator Control System of the JPL Remote Surface Inspection System [15] to control a 7-DOF redundant manipulator arm.

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.

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. This work was funded by NASA Small Business Innovation Research (SBIR) Contract NAS7-1172.

REFERENCES

[1] M. H. Lee, D. P. Barnes, and N. W. Hardy, "Knowledge based error recovery in industrial robots," in *Proceedings of the International Joint Conference on Artificial Intelligence*, (Philadelphia, PA), pp. 824–826, 1983.

[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. López-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. H. Lee, *Intelligent Robotics*. Open University Press, 1989.

[6] 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.

[7] 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.

[8] B. Randell, "System structure for software fault tolerance," *IEEE Trans. Software Engineering*, vol. SE-1, pp. 220–232, June 1975.

[9] S. Hayati, T. S. Lee, K. S. Tso, P. G. Backes, and J. Lloyd, "A unified teleoperated-autonomous dual-arm robotic system," *IEEE Control Systems*, vol. 11, pp. 3–8, Feb. 1991.

[10] H. Seraji, "Configuration control of redundant manipulators: theory and implementation," *IEEE Trans. on Robotics and Automation*, vol. 5, pp. 472–490, Aug. 1989.

[11] 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.

[12] H. Seraji and R. Colbaugh, "Improved configuration control for redundant robots," *Journal of Robotic Systems*, vol. 7, no. 6, pp. 897–928, 1990.

[13] I. J. Cox and N. H. Gehani, "Exception handing in robotics," *IEEE Computer*, pp. 43–49, Mar. 1989.

[14] P. G. Backes, K. S. Tso, S. Hayati, and T. S. Lee, "A modular telerobotic task execution system," in *Proceedings of 1990 IEEE International Conference on Systems Engineering*, (Pittsburgh, PA), pp. 511–514, Aug. 1990.

[15] S. Hayati, J. Balaram, H. Seraji, W. S. Kim, and K. Tso, "Remote surface inspection system," in *Proceedings of SOAR'92: The 6th Annual Space Operations, Applications, and Research Symposium*, (Houston, TX), Aug. 1992.

Multi-Beam Range Imager for Autonomous Operations

Neville I. Marzwell
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109

H. Sang Lee, R. Ramaswami
Science & Engineering Services, Inc.
4040 Blackburn Ln. Ste. 105
Burtonsville, MD 20866

Abstract

For space operations from the Space Station Freedom the real time range imager will be very valuable in terms of refuelling, docking as well as space exploration operations. For these applications as well as many other robotics and remote ranging applications, a small potable, power efficient, robust range imager capable of a few tens of km ranging with 10 cm accuracy.is needed The system developed is based on a well known pseudo-random modulation technique applied to a laser transmitter combined with a novel range resolution enhancement technique. In this technique, the transmitter is modulated by a relatively low frequency of an order of a few MHz to enhance the signal to noise ratio and to ease the stringent systems engineering requirements while accomplishing a very high resolution . The desired resolution cannot easily be attained by other conventional approaches The engineering model of the system is being designed to obtain better than 10 cm range accuracy simply by implementing a high precision clock circuit. In this paper we present the principle of the pseudo-random noise (PN) lidar system and the results of the proof of experiment.

1. Introduction

The pulsed laser ranging system has been limited in use because of its complexity, size, and cost. Generally the system consists of a high power pulsed laser(s), large optics, complex electronics and elaborate data systems. Consequently, the system is bound to be large in size and weight, consume high power, and prone to frequent breakdowns. The conventional cw FM laser radar technique is limited in terms of range and resolution due to limited frequency chirping (~100 GHz) and laser mode bandwidth (~200 MHz). The best resolution expected in this technique is only about a few thousandths of the range coverage. This technique relies on analog signal processing which requires a large return signal. The noise bandwidth is fully open in these techniques, unlike in the preferred pseudo-random noise (PN) modulation [Golomb, 1965] cw lidar system.

To overcome these difficulties and to develop a maintenance-free, operational system for mission oriented applications, a PN cw lidar system is being developed by Science & Engineering Services, Inc., MD using commercially available diode lasers. Unlike the pulsed laser radar, whereby the range is determined by the direct measurement of the transit time of a short (~ 10^{-9} sec) high power laser pulse, the disclosed PN cw laser radar measures the target distance by measuring the time shift of the return signal modulation sequence with respect to the reference modulation sequence which is the same as the transmitter modulation sequence. The time shift is measured by invoking the cross correlation of the return signal with the reference modulation sequence and determining the time shift needed to bring the return code train in phase with the reference modulation sequence. The cross correlation between the reference modulation sequence and returned signal train becomes maximum for the correct time shift, while is virtually zero for other time shifts. This correlation technique using PN code is a powerful method to measure the weak return signal buried in a random noise background because of the phase sensitive detection capability of the correlation process. In terms of the sensitivity of the measurement, this technique is the counterpart of the phase sensitive detection technique of differential signal measurement-the correlation technique is a most general case of the phase sensitive technique. The power of a 150 mW AlGaAs diode laser that is commercially available at present, is equivalent to the average power of 1.5 MW peak power pulsed laser operating at 10 Hz, commonly used in pulsed lidar systems. Thus, the maximum range measurable by the proposed PN cw lidar, in principle, is no less than the conventional high power pulsed lidar systems. A factor of 10 or more improvement in the diode laser power is anticipated in the near future. This offers a potential for a factor 10 improvement in range resolution for the same integration time, or reduction of the signal integration duration for the same resolution.

2. Description of the System

The sensor system block diagram shown in Fig. 1., consists of three sub-systems: a diode laser transmitter, scanner & receiver optics, and detector and signal processing electronics which are controlled by a system computer. In the baseline mode, the diode laser is digitally modulated by a pseudo random code of a given length (10 bit) and 1 µsec modulation time bin-width employing a modulator providing a 150 m quantization precision and 150 km unambiguous range coverage. After a fixed number of modulation periods(defined as a cycle in this disclosure), the modulation start time delay is readjusted by a prescribed manner using a delay generator to provide a phase shift between the cycles (typically an order of 10 periods) of the PN code. This phase information is then retrieved by the correlation calculation process and used to enhance the range resolution in the disclosed system in a unique way. The transmitter beam is deflected by the scanner to the target, and the return signal is deflected by the scanner to the receiver optics. This signal is measured by a detector/preamp unit to provide a low noise electronic signal. This signal is then further amplified and filtered by a postamplifier-filter unit. The gain of the post-amplifier is automatically adjusted to provide a proper signal level for the ADC by an AGC unit. The dynamic range of the ADC is not limited to be very low due to the specific features of the signal processing routine of the disclosed system whereby the ADC output is first summed at the accumulator. After accumulating the data for a preselected number of cycles, the time integrated digital signal from the accumulator is read into the correlator for the calculation of the correlation distribution and determination of the precise range.

The range is determined based on the correlation calculation as follows: The PRM cw Lidar relies on the delta function property of the autocorrelation function of a PN code. The cross correlations of the return signal with the reference modulation code are calculated for each of the n time bin shifts which cover the entire code length. Due to the delta function property of the PN code, the cross correlation is always zero except for the case in which a null phase difference between the transmitted code and received signal is realized. A null phase difference is realized for the time shift that corresponds to the round trip transit time between the transmitter and the target. Thus, the time shift that corresponds to the maximum cross correlation value gives the range of the target. Further analytical description of the technique used in this system is discussed below. Consider a cyclic, digital (1 or 0) nth order PN code represented by a_i. The cross correlation $_j$ of this code with another expression of PN code a_j' with elements 1 and -1, in place of 1 and 0, will then satisfy the following relationship:

$$\rho_j = \sum_{i=0}^{N-1} a_i\, a'_{j+i} = \begin{bmatrix} (N+1)/2 & for\ j = 0 \\ 0 & for\ j \neq 0 \end{bmatrix}$$

The return signal, s_i, acquired at the receiver is the convolution of the transmitter signal, $x_i = Pa_i$, and the target response R, thus

$$s(i) = \sum_{i=0}^{N-1} x_{i-j}\, R + b$$

where *b* represents the background and noise signal .Let P stand for the diode laser power. If we accumulate the data acquired for M periods of PR code cycle, the integrated signal, S, is

$$S(i) = \sum_{k=1}^{M} s_{i+(k+1)N}$$

The target signal and range is then derived by taking the cross correlation of the return signal with the reference modulation sequence which is shifted by *l* time steps relative to the return signal as

$$\rho_l = \sum_{i=0}^{N-1} S_i\, a'_{i-1} = M\left[P\,\frac{(N+1)}{2}\, R + b \right]$$

where $l = 2\ (d / c\ \Delta t)$, the number of modulation shifts during the round trip transit time to the target at distance *d*, and Δt is the modulation clock period which is referred to the time bin width in this disclosure. For any other shift the correlation becomes zero. Note since *b* is an uncorrelated noise signal, the correlation with "a' " becomes almost zero. Therefore, by finding the correct shift for the maximum correlation, the distance to the target can be derived.
However, the number of ones in a sequence of an M-code always exceeds the number of zeros by one as an intrinsic property. Therefore the background term *b* can not be neglected for excessively noisy data. A new modulation code, A-code [Nagasawa, 1990] which corrects this defect is implemented in this disclosure, by slight modifying the M-code. As far as the noise reduction is concerned, the correlation technique with this new code is identical to the case of the phase sensitive technique.

Although the diode output can be modulated much faster than a few GHz and signal processing system permits a speed as high as 100 MHz, the signal to noise ratio decreases as the modulation speed increases. Under this constraint, the baseline of the disclosed system is a 1 MHz modulation with a flexibility of successively switching it to 10 MHz and 100 MHz modulation for short range measurements. Based on 1 MHz modulation frequency, this technique will provide 150 m range quantization precision. Further improvement of range resolution beyond the quantization precision can be achieved without increasing the clock speed by employing a novel digital range resolution enhancement technique. This technique is implemented by introducing a known delay time at the start of each code sequence, then averaging the digitized signal to obtain a higher resolution than the quantization precision. The delay time can either be a systematic delay with higher timing precision (e.g. 100 nsec step for 1 MHz system) spanning 0 to a few time bins of the modulation, or a delay by a larger time step (e.g. 300 nsec) for spanning 0 to a few tens of modulation time bins. The correct range is calculated by multiplying the speed of light and the weighted average of the time shifts in the correlation calculation using the integrated signal. The correlation value is used as the weighting function in this averaging process. A small random time jitter which is characterized by a uniform distribution of delay can also be used for improving the range resolution. Due to the random distribution of the delay, the average range calculated from the L cycles of the PN code will then provide a higher range resolution, 15/L meter. With this technique, for a 10 msec averaging, the net resolution of the 10 MHz system, with 15 km unambiguous range, can be as high as 15 cm. For a smaller unambiguous range a shorter PN code may be used, thus facilitating a higher range resolution inversely proportional to the number of cycles L, achieving a 3 cm resolution for an 8 bit code.

A schematic of the timing diagram of transmitter modulation, analog to digital converter (ADC) synchronization as well as the accumulator and correlator operation is shown in Fig.2 The first trace 1 represents the uninterrupted master clock sequence of a given speed. The second trace 2, represents the sequence of the ADC start trigger pulse which repeats for every M cycles of the clock period. After K periods of the modulation (MxK clock pulse periods), the delay change is reset to a new value as shown in trace 4. This delay value will be remained fixed for the next K periods as shown in trace 3. For every ADC start pulse, the accumulator is re-phased as shown in trace 6 to integrate the return signal for the exact modulation period. The accumulator dump pulse is triggered after L cycles of delay change and synchronized with the immediately following ADC start pulse as shown on trace 7. Total L number of different delay values are reset for one correlation calculation and measurement of a high resolution range as shown in trace 8. The cross correlation can be calculated efficiently by a computer algorithm, if the maximum required range is less than a few tens of km's and the resolution requirement is relaxed. However, for a high resolution and large range coverage whereby the required information processing volume is prohibitive, a hardware correlator module such as TRW TMC2023 can be employed for high speed calculation. When the time delay is going through the change, the continuity of the transmitter modulation becomes interrupted for a few bin periods. However, the reference modulation which relies on the ADC start signal is still contiguous and the noise reduction property of the PN modulation technique remains valid with the disclosed technique. Only a small fraction of the return signal is lost in this process as indicated in the figure.

Based on the assumed configuration: transmitter power of 150 mW; 20 cm receiver aperture; and PN modulation of 1 μsec time bin, the signal from a target of 0.5 reflectivity at a distance of 15 km is estimated to be 5 photon per bin. This amounts to 4,400 photons for a 10 msec signal integration time. Any small detectable signal can be enhanced by the accumulation of data, while the random noise component is suppressed by the PN modulation correlation technique. Since the system is expected to be photon limited, the targets at a distance of 150 km is well within the ranging capability of the system simply by switching to a longer time-bin mode and integrating the return signal for a longer period. As the target approaches to a short distance (for example less than 15 km), the master clock is switched to 10 MHz to obtain 15 cm resolution for a 10 msec integration. The clock speed switches further to obtain better resolution as the target reaches a near field (<1.5 km). Since the signal amplitude varies as an inverse square function of the range, the integration time at near field can be shortened to less than a msec without increasing any other parameters of the sensor, thus, facilitating a faster scanning of the object at a close distance. The baseline receiver optics consist of a scanner, a 20 cm diameter telescope, a bandpass filter with less than 1.0 nm bandwidth for suppression of the background radiation, and an Si Avalanche Photodiode (SiAPD). The collimating lens is used as a narrow band filter and is followed by a focusing lens. The detector array which consists of a number of detectors arranged in a row vertical to the optical plane is located at the focal plane of the lens. In this way each detector collects the return signal from a predetermined field of view covered by one of the multi-beams of the transmitter. Thus scanning the multi transmitter beam and receiver FOV by the common scanner in a direction perpendicular to the detector array axis, this system scans a wide field of view (FOV) to provide range data within.

3. Proof of Principle Experiment

Proof of principle experiments have been carried out utilizing a target board at Goddard Space Flight Center and an existing lidar system at GSFC (P/T lidar). A PN modulated diode beam of 10 mW at 790 nm is sent out to the target and the return signal is received by a 10x20 cm^2 aluminum mirror. A multi-alkali PMT is used to detect the signal with substantially deteriorated sensitivity from the that of optimum wavelength. The signal is then digitized by the modified P/T lidar data acquisition system which was designed for the pulsed system. The data is then averaged for 10 periods of modulation with different delay values respectively. The correlation calculation is performed on this data for accurate range measurements. The reference point near the transmitter is independently measured by placing a corner cube reflector at 2 m distance from the transmitter, thus the range value is the differential range between the corner cube and the target. The experiment was carried out using various modulation frequencies between 1 MHz and 10 MHz with various delay times.

A typical return signal from the target at 850 m range.is shown in Fig.3a For a given modulation, a set of delay times is introduced at the beginning of the modulation train and the return is averaged before the correlation calculation to imitate the signal processing of the real time system to be developed. The typical correlation values of the return signal are plotted in Fig.3b. The accurate range obtained from the correlation values are summarized in Table 1 and also plotted in Fig.4 We note that thus measured range is self consistent throughout the various modulation frequencies. It is very important to note that the range measured with 5 MHz modulation and 10 nsec step delay agrees with that of 10 MHz and 10 nsec delay within 1.2 m which is well within the maximum error due to the time resolution. This result clearly demonstrate that the ultimate range resolution is determined by the time resolution of the delay regardless of the modulation speed Furthermore it validiates the range resolution enhancement technique implemented A commercial time delay generator chip with 50 psec precision is readily available and is planned to be implemented in our engineering model.

4. Recommended Commercialization Approach

There is a unique advantage to our PN lidar system over other systems including pulse system and frequency chirped lidar system. These are namely the long distance ranging capability and the high range resolution capability with Science & Engineering .proprietary range enhancement technique. With a moderate diode laser power and moderate size optics, the PN lidar system can achieve a few tens of km's distance ranging while the resolution is kept to better than 10 cm level. The proof of concept experiment shows a robust performance of the system concept without optimization of the system parameters, indicating the robustness of the technique used here. The approach to the commercialization will begin with a single beam portable system development following this program for ground based ranging and speed detection applications in terms of automobile speed enforcement, altimetry, civil engineering survey, as well as collision avoidance where the range resolution of 10 cm is satisfactory over a few km distance. The next step in the commercialization is the further enhancement of the resolution to a few mm at near distance of less than a meter for process control in manufacturing industries. The range imaging and robotics application will follow this phase naturally. One of the important steps in this development is micro-packaging the system. Since the system is based on the digital concept, the electronics can be packaged into one board size with a special DSP microprocessor chip. For a near distance (a few km's) operation, the optics can be substantially small (~50 mm) and readily packaged into a portable size. We plan to draw JPL's expertise in this area to speed up this process.

5. Applications

A wide range of application is covered by this system. Although some applications require a relatively complex multi-beam system, many commercial applications need only a single beam baseline system. We envision a commercial product that is cost effective, due to the continuous reduction in laser cost, micro-packaging and microelectronic fabrication, as well as the anticipated growth of the market for the proposed product. Some potential commercial applications of this sensor are as follows:

i) Robotics Application - As an active range imager for recognition, pose estimation and ranging. Combined with passive camera video imaging, the data can facilitate a high speed high resolution processing. At short distances, range resolution of less than 1 mm is possible with this system.
ii) Ranging Application - This includes applications in commercial aviation as well as navigation of surface vessels. Future environmental disasters such as the Alaskan oil spill can be prevented by employing this type of sensor. A cost effective version of this system has enormous commercial potential as an automobile collision prevention device for autonomous navigation of ground vehicles using the GPS system. A further development of the system as a civil engineering survey instrument used for accurate remote measurement of distance is also promising, because of its low cost and portability unlike the other systems currently available. Use of this system in police speed guns for vehicle speed violation detection is also very promising due to its accuracy and long range capability without it being detected by a radar detector.

iii) Aerosol and Wind Sensing - This type of laser radar can be used to measure the wind and aerosol field within the range of a few km's. With Doppler frequency shift measurements that can be easily accomplished utilizing a stable diode laser frequency, the baseline sensor can be directly applied to the development of a compact Doppler wind sensor.
iv) Environmental Application - This sensor can also be the baseline of an environmental sensor for remote monitoring of industrial smoke and urban smog. This sensor can be further modified and developed to be used as an active gas sensor for measuring many environmental gases.
v) Communications Application - Some of the parameters of this system are of interest for a highly directional communication system, including the frequency stabilization, receiver system design, and modulation technique. Short range exclusive communications may also be possible by tuning the laser output frequency to a specific atmospheric trace gas absorption line thus shielding the signal beyond a certain range.
vi) Airborne altimeter applications - This sensor, when used in the altimeter mode, can measure the forest timber volume by integrating the tree height information. The airborne altimeter application is also very useful for an areal surface topography, mapping, and land and soil management when integrated with a GPS system.
vii) Industrial manufacturing applications -for surface inspection to determine surface flaws and component defects, like solder joints, cracks, dents; in metrology to measure hole diameters, lengths, widths, thickness; in guidance and control for part sorting, palletizing, pick and place operations, insertion and removal; in integrity and placement verification to determine if a feature lies within specified bounds.
viii) IFF application - as a small, low power, robust system for a highly effective Identification Friend or Foe system in the battle field.
ix) Medical application - in measuring fluorescence decay and photon migration signature of tissue cells for cancer detection and other disease monitoring.
x) Emerging space applications - reconnaissance and docking operations, health status monitoring of space vehicles.

Acknowledgement

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.This work was funded by a NASA Small Business Innovative Research Contract, NAS7-1186 The development and commercialization is being carried out at the SESI, Burtonsville, MD.

REFERENCES

Golomb, Solomon W., "Digital Communications with Space Applications", Peninsular Publishing Company, Los Altos, California, 1964.

Nagasawa, C. M., Abo, H. Yamamoto, and O. Uchino, "Random modulation cw lidar using new random sequence," Applied Optics, Vol.29, No.10, 1990.

Table I. List of measured target ranges for various system parameters.

Measured Range (m) vs Number of delays

# DELAY	F_T = 1 MHz	F_T = 5 MHz	F_T = 10 MHz
0	803.4 ± 165 854.3 ± 30	831.0 ± 16.5 842.9 ± 30	834.2 ± 16.5 846.7 ± 30
10	876.7 ± 30	842.5 ± 3 853.5 ± 16.5	841.2 ± 3 848.1 ± 16.5
20	N/A	N/A	853.3 ± 15.7

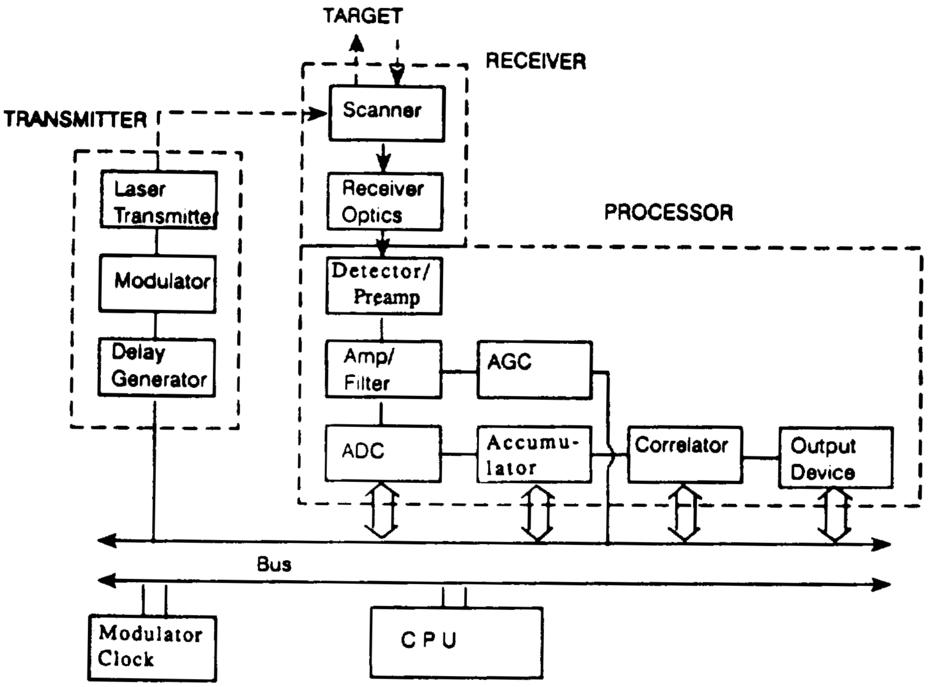


Fig. 1 Schematics of PN lidar system for accurate range imaging applications.

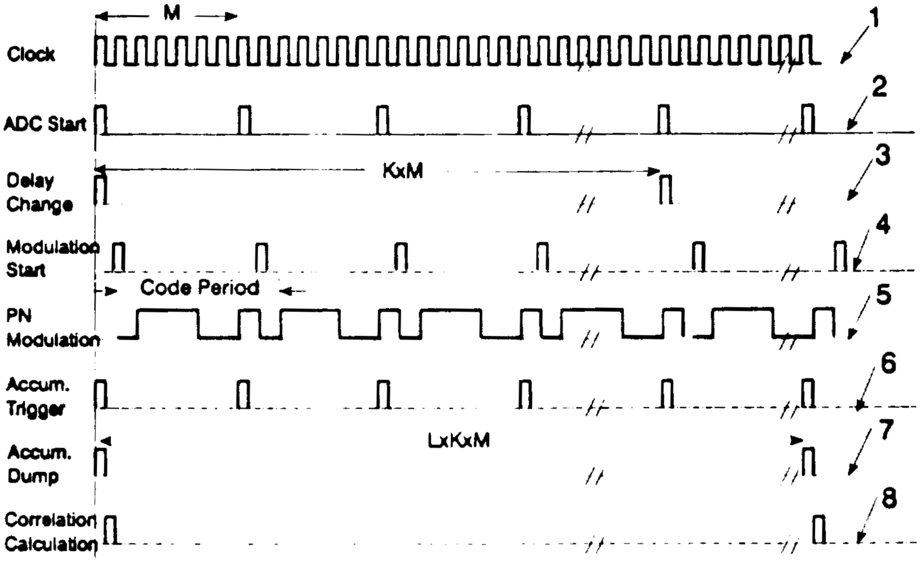


Fig. 2 Timing diagram for PN lidar system implementing the range-resolution-enhancement technique.

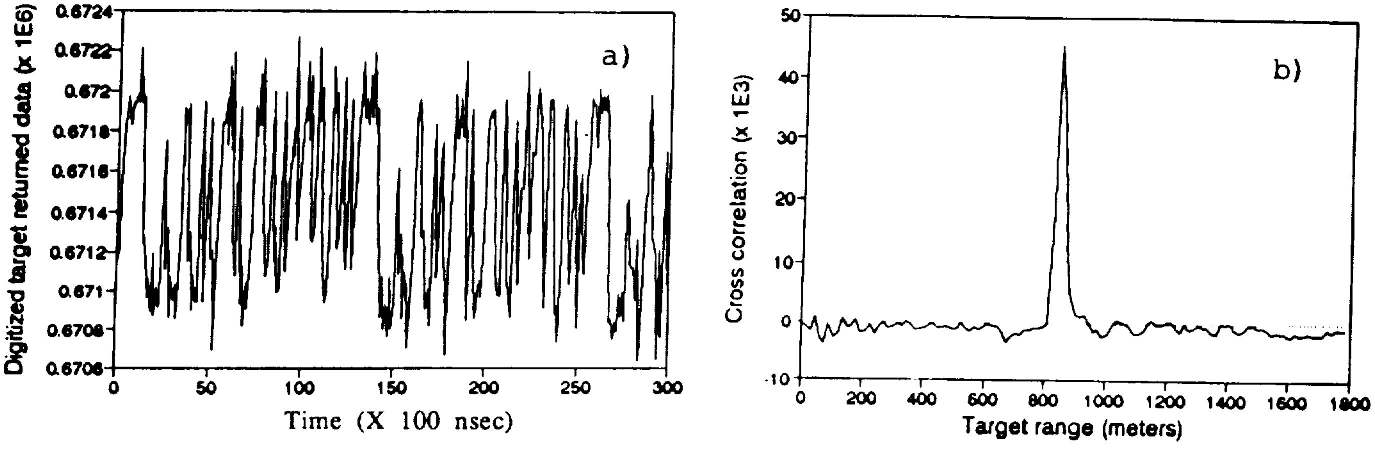


Fig. 3 a) Measured return signal from a target at a distance of 841.7 m, using a 10-mW diode laser 10 x 20 cm^2 optics.

b) Correlation calculated from the measured data shown above.

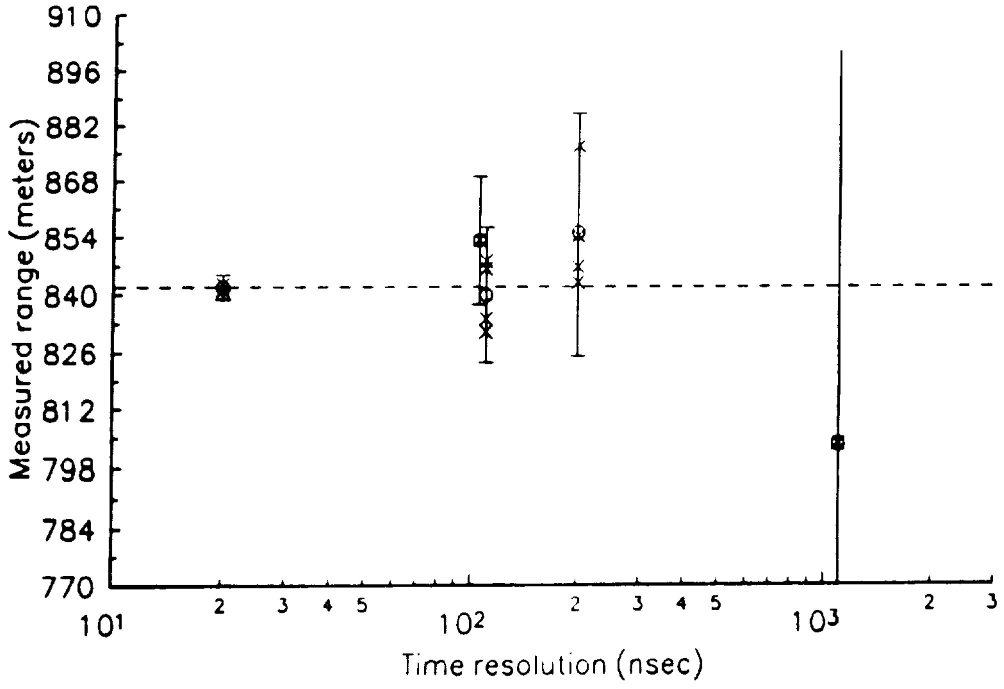


Fig. 4 A plot of measured target range for various system parameters. The "x" points correspond to each data set, and "o" symbols with the indicated error bar represent the average range value for a given time resolution. The time resolution is accomplished either by the system clock frequency or by the delay precision used in the range-enhancement technique. The two data points on the left that are in agreement within 1.3 m correspond to 10-MHz and 5-MHz clock frequency, respectively, and 10-nsec delay precision.

ROBOSIM: AN INTELLIGENT SIMULATOR FOR ROBOTIC SYSTEMS

Kenneth R. Fernandez, Ph.D.
NASA/MSFC, AT01
MSFC, AL 35812

George E. Cook, Ph.D.
Csaba Biegl, Ph.D.
James F. Springfield
Vanderbilt University
School of Engineering
Box 1826, Station B
Nashville, TN 27235

ABSTRACT

The purpose of this paper is to present an update of an intelligent robotics simulator package, ROBOSIM, first introduced at Technology 2000 in 1990. ROBOSIM is used for three-dimensional geometrical modeling of robot manipulators and various objects in their workspace, and for the simulation of action sequences performed by the manipulators. Geometric modeling of robot manipulators has an expanding area of interest because it can aid the design and usage of robots in a number of ways, including: design and testing of manipulators, robot action planning, on-line control of robot manipulators, telerobotic user interface, and training and education. NASA developed ROBOSIM between 1985-88 to facilitate the development of robotics, and used the package to develop robotics for welding, coating, and space operations. ROBOSIM has been further developed for academic use by its co-developer Vanderbilt University, and has been used in both classroom and laboratory environments for teaching complex robotic concepts. Plans are being formulated to make ROBOSIM available to all U.S. engineering/engineering technology schools (over three hundred total with an estimated 10,000+ users per year).

INTRODUCTION

In the development of advanced robotic systems concepts for use in space and in manufacturing processes, researchers at NASA and elsewhere have traditionally verified their designs by means of expensive engineering prototype systems. This method, though effective, frequently required numerous re-design/fabrication cycles and a resulting high development cost. To reduce costs ROBOSIM, a graphics-based simulator for robotic systems, was developed to allow rapid prototyping of robotic systems, including robot manipulators, multi-axes positioners, other motion-controlled mechanisms, and objects and structures in the operating environment. The reduced costs obtainable with the graphical simulator allows experimentation with many design alternatives prior to implementation in hardware resulting in improved end-item designs.

Researchers at Vanderbilt University (co-developers of ROBOSIM) have further refined the simulation package to fully exploit its various capabilities in the teaching of courses on robotics, mechanisms, industrial control, and advanced automation at both the undergraduate and graduate levels. ROBOSIM provides the students with a graphical means of visualizing objects and their relationships in 3-dimensional space. The relationships between coordinate frames that have been translated and rotated relative to one another are immediately provided in the form of transformation matrices, and they may be displayed and viewed from different viewing positions. With the simulation system, students may actually graphically construct every manipulator they study in the classroom or in class assignments, and then they may operate the manipulator and use it to grasp, move, and release objects, and develop and test action plans for using the manipulators in practical settings. The costs of doing this with physical hardware would of course be quite prohibitive in terms of both dollars and time. ROBOSIM allows the students to be exposed to numerous robotic systems without any of the constraints placed on use of a specific physical laboratory robot. Furthermore, without any fear of injury the students may operate their manipulators at any hour from their personal computer performing tasks ranging from arc welding to robotic care of hospitalized patients, all within the confines of the interactive simulation environment provided by ROBOSIM!

With ROBOSIM, students construct a robot manipulator link by link and assign coordinate frames to the joints of each link. ROBOSIM automatically checks and validates the frame assignments and provides error messages stating the type of error and where it occurred if mistakes have been made. Once the links of the manipulator have been constructed, a single command automatically assembles the links into the complete manipulator. Any manipulator so constructed, regardless of its complexity and structure, may be immediately moved and controlled on a joint-basis with a single *drive* command specifying the manipulator by name and the joint values. Straight-line cartesian motion requires the inverse kinematics, which must be provided by the student, if not contained in the family of manipulators provided in the basic package. Once the inverse kinematic solution is linked into the program, the students may verify their solution by actually driving the manipulator based on their inverse kinematic control. With inverse kinematic control, higher-level commands may be issued to the robot, such as *move-to-grasp*, which commands the robot to move to grasp an object at a predefined position and orientation relative to the object. If obstacles are in the way, the command *find-path* may be issued instructing the robot to employ built-in artificial intelligence-based heuristics to search for a collision-free path.

In addition to the advanced features of collision detection and artificial intelligence-based collision avoidance routines, ROBOSIM includes other advanced features such as configuration management and composite objects. The latter, for example, permits the student to assemble parts into a composite assembly which can be operated on as a unit while preserving the separate accessibility of its components. This permits parts to be joined by welding or bolting, for example, and then transferred to other processing operations in the manufacturing process.

All of these capabilities of the graphical simulation approach to the study of robotics and automation provide the student with an almost limitless range of possibilities without the prohibitive cost of physical hardware, without the prohibited time required to use such hardware, and without the potential danger associated with the use of physical hardware.

There are over 300 engineering and engineering technology schools in the U.S. which currently offer courses in robotics, mechanisms, industrial control, and advanced automation. It is estimated that the number of students taking such courses exceeds 10,000 per year. Each of these students is being targeted as a potential user of ROBOSIM in a distribution plan that is being jointly formulated by NASA and Vanderbilt University. Additionally, while emphasis here has been placed on the use of ROBOSIM as a teaching tool in the academic community, virtually all of the features that make it useful there are equally applicable and useful to the industrial user. Plans are being formulated for distribution of ROBOSIM to the industrial community as well.

TECHNICAL OVERVIEW OF ROBOSIM

The purpose of ROBOSIM, as it was originally conceived, was to provide a means of constructing and viewing three dimensional models of robot manipulators and various objects in their workspace, and simulating action sequences performed by the manipulators [1-4]. Geometrical modeling of robot manipulators is an expanding area of research because it can aid the design and usage of robots in a number of ways [5]:

- *Design and testing of manipulators:* The purpose of the modeling is to study different approaches to satisfy the design specifications of the manipulator.

- *Robot action planning:* The modeling environment is used to build a representation of the robot(s) and the objects in the workspace for creating and validating action plans by simulating the effect of these actions in the model space.

- *On-line control of robot manipulators:* The simulated action plans generated in the model space are transmitted (after validation) to the attached robot manipulators for execution.

- *Telerobotic user interface:* In applications where the operator of the robot has to be at a large distance from the workcell (nuclear facility, radioactive waste treatment, space, etc.) realistic graphical simulations can be used for better interaction with the manipulator.

- *Training and education:* Robotic simulation packages provide an inexpensive and safe way to teach the theory and operation of robot manipulators.

Design of Robot Manipulators: The ROBOSIM Modeling Environment

Similar to other solid modeling software tools, ROBOSIM models the three dimensional geometrical objects using lists of their bounding polygons. The ROBOSIM modeling language is used to specify complex geometric shapes which are used as manipulator links or as passive objects in models of robotic systems. All shapes are built from elementary geometric types like boxes, cylinders, cones, extruded polygons, etc. Translational, rotational and scaling geometrical transformations are used to combine these objects to form the desired shape. Link coordinate frames can also be added thus making it possible to specify the geometric transformations associated with a manipulator arm.

Once the ROBOSIM language interpreter finishes the processing of the code describing a solid shape, the resulting polygon list is converted into a named object (or robot link) in the robot simulator package's workspace. Thus the modeling of robot arms and objects and geometric scenarios is a two step process as follows:

- Modeling the geometric shapes, robot links, etc., which are used to build the scenario.
- Creating one or more named, distinguishable object instances of these shapes in the simulator's workspace.

These object instances can be individually manipulated upon, moved around, etc., from the interactive environment of the simulator, as described next.

Operating the Models: The ROBOSIM Simulation Environment

The ROBOSIM package provides an interactive simulation environment where every command entered by the user is immediately executed and the results are displayed on a graphics screen. From this interactive environment users can change the simulation scenario and operate the robot manipulator models in the system. The commands available can be grouped as follows:

- *Environment configuration:* Besides the modeling services discussed above additional commands are available for the setting of global parameters like camera position, display mode, light sources, etc. The graphics display module of the simulator supports different display options like wire-frame, hidden line, solid filled and shaded graphics depending on the capabilities of the hardware platform.

- *Manipulator control:* There are commands available for moving the models of manipulator arms in various modes: joint interpolated, straight line, rotation about an arbitrary axis, etc. Manipulator coordinates can be specified both in joint and world coordinates. The simulator has a built-in iterative inverse kinematics algorithm, but the user can also specify an explicit inverse kinematics method for his or her manipulator if such a method is available. Additionally, the objects in the workspace can be grasped, moved, and released by the robots. If the scenario contains several manipulators these can be operated in parallel.

- *Status reporting:* Reports about different aspects of the simulator's operation (arm positions, collision situations, etc.) can be obtained by using one of the appropriate commands from this group.

The command language of the ROBOSIM simulation environment has been designed with two goals in mind: (1) to provide an interactive user interface, and (2) to be usable as the interface to a higher-level task planner program. In the second application the task planner and the robot simulator are typically interfaced using some kind of pipe mechanism and the task planner outputs similar command sequences as entered by users in interactive applications. For this reason the command language has intentionally been kept simple.

Advanced Features

Although the basic modeling and simulation environment described above perform quite satisfactorily as a robot modeling tool, ROBOSIM's capabilities were greatly enhanced with the completion of various extensions to the basic package. Some of these extensions use heuristic, rule-based programming techniques. The extensions include:

- *Composite objects:* The ROBOSIM simulation environment also supports the linking of separate objects into a so-called composite object which can be operated on as a unit while preserving the separate accessibility of its components. A good example of this is a drawer with various objects in the drawer. During the course of the simulation a manipulator may have to pull out the drawer (the drawer and its contents have to be treated as a unit) and then pick up a single object from the drawer (now a part of the composite object has to be accessed individually).

- *Configuration management:* When a manipulator arm is programmed using world coordinates it is typical to have several valid solutions (sets of joint coordinates) by which the manipulator can reach the desired position. For example, in the case of manipulators similar to the Unimation PUMA 560 robot this manifests in left or right handed and elbow up or elbow down configurations for the arm. ROBOSIM permits the user to choose any configuration and stay with it for the duration of the simulation. However, this approach is not optimal when the manipulator has to move distances comparable to the limits of its envelope. For such cases the simulator provides an automatic configuration selection mechanism which is based on a set of rule-based heuristics.

- *Collision detection:* ROBOSIM also provides a way to check for collisions during a simulation run. The collision detection is based on the detection of intersections of solid object (passive objects or manipulator links) bounding polygons. For efficiency reasons the simulator maintains a rectangular bounding box for every object and invokes the more complex polygon-based collision detection method only if the bounding boxes intersect.

- *Collision avoidance:* The simulator also provides a heuristic path planning algorithm which is capable of recovering from collision situations. The collision avoidance is based on heuristic rules describing actions to try in various collision situations. Some of these rules are generic, others are manipulator-specific. An example generic rule is the "minimal volume rule" which is usable for large displacements of the end effector. It specifies a path to reach the desired target which includes a midpoint where the arm is folded in a way which minimizes its reach. Users can attach other heuristic rules specific to their manipulator models.

- *Interface to control real robots:* The simulation environment is also capable of generating command sequences for real robot controllers. In this mode only those motion commands are output which have been verified with the built-in collision checking to be safe. Currently only the PUMA 560 robot with the Unimation controller running the VAL II robot control system [6] is supported, but additional output modules can be added.

PLATFORMS

Implementations of ROBOSIM are currently available for the Hewlett-Packard HP 9000/300 and 9000/800 graphics workstation families, Silicon Graphics workstations, Intergraph workstations, and 386 or 486-based PC compatibles with EGA or VGA displays. The display mode may be wire-frame or shaded solid modeling for the workstation implementations. Currently, the PC versions display in wire-frame mode only. While currently limited to wire-frame displays, the PC versions are quite fast, offering displayed manipulator motions at speeds well in excess of the hardware being simulated. It is the ability to run ROBOSIM at high speeds on a PC that makes the package an attractive tool for all engineering students studying robotics, kinematics, industrial automation, mechanisms, and advanced automation. Armed with their personal computer, students can design and operate quite complex robotic systems, gaining an excellent appreciation for the many factors that must be taken into consideration in industrial automation. The PC version will also open the door to simulation technology for many small businesses that can not afford expensive systems tailored to high-end workstation implementation.

PLANS FOR DISTRIBUTION

The American Society for Engineering Education's *1992 Directory of Engineering and Engineering Technology Undergraduate Programs* lists 261 engineering schools and 339 engineering technology schools in the U.S. [8]. The total undergraduate enrollment in the engineering programs was 358,095 with 61,318 B.S. degrees awarded in 1992. The total graduate enrollment in the engineering programs was 70,980 M.S. students with 26,006 degrees awarded in 1992, and 34,647 Ph.D. students with 5,582 Ph.D. degrees awarded in 1992. The total undergraduate enrollment in the engineering technology programs was 33,797 with 7,458 B.S. degrees awarded in 1992.

Based on an analysis of students who would be taking courses in robotics, mechanisms, kinematics, industrial automation, advanced automation, other courses involving design, control, and operation of mechanisms and structures in three-dimensional space, the number of potential users of ROBOSIM in the academic community is estimated to be in excess of 10,000 per year. NASA and Vanderbilt University are currently formulating a distribution plan to service this need. Plans are likewise being formulated to distribute an industrial version, which is anticipated to have strong appeal to small businesses engaged in special equipment design and other related areas where graphical simulation with fast motion display on a PC would be a valuable tool.

FUTURE ENHANCEMENTS

The available experience with the ROBOSIM package suggests that it is a flexible and powerful tool for modeling and simulating robotic systems [9]. It has proven easy enough to understand and use in introductory courses on robotics, kinematics, industrial automation, and mechanisms, but its advanced features also make it possible to use in complex large-scale systems [10-13].

Some areas where future improvements are planned include:

- Programming language: A future version of ROBOSIM will be embedded into a general purpose interactive programming language interpreter. This new version will also support the old command syntax for backward compatibility, but the use of a general purpose language instead of the current command interpreter will offer several advantages:
    - Program flow control statements in simulations.
    - An easier way for users to specify inverse kinematic routines for their manipulator models.
    - An easier way for users to specify arm configuration selection and collision avoidance heuristics for their models.
- Development of a graphical user interface (GUI) to control simulation options: There are some system parameters in the simulation environment which are especially suited for control by GUI methods (while keeping the current interactive commands to control them as well). These include the camera setup, lighting model, and other similar options.
- Interface to common CAD packages for importing shape designs to be used in simulations.

- Continued improvements on the path planning and arm configuration management heuristics.

CONCLUSIONS

The ability of U.S. industry to compete globally will depend on an adequate supply of engineers and technologists trained in the application of robotics and automation to the problems of industry. The current slump in the U.S. robotics industry is due in part to this lack of a trained work force. Opportunities for students to take courses in advanced automation are currently centered primarily in major school having the resources to equip an expensive robotics laboratory. Additionally, safety issues and the need to continually upgrade and maintain these facilities has further limited the number of schools with world-class facilities. The wide use of ROBOSIM will allow all schools to provide students with the ability to study and develop advanced robotic systems in safety and at greatly reduced expense. Furthermore, as stated before, virtually all of the features that make ROBOSIM appealing to the academic community are equally useful to the industrial user, particularly small businesses. NASA and Vanderbilt University are jointly formulating plans to distribute ROBOSIM to both the academic and industrial communities.

REFERENCES

[1] Fernandez, K.R., *Robotic Simulation and a Method for Jacobian Control of a Redundant Mechanism with Imbedded Constraints*, Ph.D. Dissertation, Vanderbilt University, 1988.

[2] Springfield, J.F., Cook, G.E., Andersen, K., and Fernandez, K.R., "ROBOSIM: A Simulation Package for Robots", *University Programs in Computer-Aided Engineering, Design, and Manufacturing*, Eds: K.P. Chong, B.R. Dewey, and K.M. Pell, American Society of Civil Engineers, 1989, pp 239-246.

[3] Springfield, J.F., *ROBOSIM Workstation Extensions*, Master's Thesis, Vanderbilt University, Spring 1989.

[4] Wilson, S.L., *Interfacing of a Robot Simulation Program with Graphic Utilities of an Intergraph Interpro 360 System*, Master's Thesis, Vanderbilt University, August 1990.

[5] Biegl, C., Cook, G.E., Fernandez, K.R., and Smith, M.K., "ROBOSIM: An Intelligent Robotics Simulator", *University Programs in Computer-Aided Engineering, Design, and Manufacturing*, Ed: D. Stone, Tennessee Technological University, 1992, pp 209-216.

[6] Mirolo, C. and Pagello, E., "A Solid Modeling System for Robot Action Planning", *IEEE Computer Graphics and Applications*, January 1989, pp 55-69.

[7] Unimation, Inc., *User's Guide to VAL II*, Danbury, CT, 1986.

[8] *1992 Directory of Engineering and Engineering Technology Undergraduate Programs*, American Society for Engineering Education, Washington, DC, 1992.

[9] Fernandez, K.R., "The Use of Computer Graphic Simulation in the Development of Robotic Systems", *Acta Astronautica*, Vol. 17, No. 1, Pergamon Press, January 1988, pp 115-122.

[10] Fernandez, K.R. and Cook, G.E., *A Generalized Method for Automatic Downhand and Wirefeed Control of a Welding Robot and Positioner*, NASA Technical Paper 2807, 1988, 54 pgs.

[11] Fernandez, K.R., Cook, G.E., Andersen, K., Barnett, R.J., and Zein-Sabattou, S., *A Generalized Method for Multiple Robotic Manipulator Programming Applied to Vertical-up Welding*, NASA Technical Paper 3163, October 1991, 22 pgs.

[12] Fernandez, K.R. and Cook, G.E., "Use of Computer Graphic Simulation Techniques for Robot Control System Development", *Proceedings, The Eighteenth Southeastern Symposium on System Theory*, IEEE

Computer Society No. 710, New York, April 1986, pp 433-438.

[13] Cook, G.E., Fernandez, K.R., and Levick, P.C., "Robot Simulation", *Exploiting Robots in Arc Welded Fabrication*, Ed: J. Weston, The Welding Institute, 1989, pp 132-135.

BIOTECHNOLOGY AND LIFE SCIENCES
PART 3

A THREE CHANNEL TELEMETRY SYSTEM

Jeffery C. Lesho and Harry A. C. Eaton
The Johns Hopkins University
Applied Physics Laboratory
Laurel, Maryland 20723

ABSTRACT

A three channel telemetry system intended for biomedical applications is described. The transmitter is implemented in a single chip using a 2 micron BiCMOS processes. The operation of the system and the test results from the latest chip are discussed. One channel is always dedicated to temperature measurement while the other two channels are generic. The generic channels carry information from transducers that are interfaced to the system through on-chip general purpose operational amplifiers. The generic channels have different bandwidths: one from dc to 250 Hz and the other from dc to 1300 Hz. Each generic channel modulates a current controlled oscillator to produce a frequency modulated signal. The two frequency modulated signals are summed and used to amplitude modulate the temperature signal which acts as a carrier. A near-field inductive link telemeters the combined signals over a short distance. The chip operates on a supply voltage anywhere from 2.5 to 3.6 Volts and draws less than 1 mA when transmitting a signal. The chip can be incorporated into ingestible, implantable and other configurations. The device can free the patient from tethered data collection systems and reduces the possibility of infection from subcutaneous leads. Data telemetry can increase patient comfort leading to a greater acceptance of monitoring.

INTRODUCTION

The three channel telemetry system was designed to fulfill a need in diagnostic medicine to measure physiological variables at the point of origin without prolonged invasive procedures. A generic system was developed to be interfaced with many types of sensors. The system was designed onto a 2.3 mm by 2.3 mm custom BiCMOS chip fabricated through the MOSIS service. The small size of the device permits both implantable and ingestible applications.

Circuit Operation

Figure 1 shows a block diagram of the three channel telemetry transmitter. The device transmits temperature and two user defined signals to the external receiver. The low frequency and high frequency channels are band-limited from 0 to 250 Hz and 0 to 1300 Hz respectively. The chip employs both frequency modulation and amplitude modulation to transmit the signals. The operation of a single channel will be detailed as the two channels differ only in the center frequency of the current controlled oscillator.

The descriptions that follow can be traced in the block diagram. The sensor output is amplified and conditioned by an operational amplifier. The on chip op-amps have low gain by traditional standards, but they are adequate for closed loop gains less than 50 provided allowance is made for the finite open-loop gain. The inputs transistors are MOSFET and so have inherently high input impedance and low bias current. The amplified sensor signal is converted to a current to drive the current controlled oscillator (ICO). The voltage to current converter has a differential input with the positive lead internally connected to the op-amp output, and the negative lead available for user connection to a reference point. The reference input is also a high impedance MOSFET input. The differential input voltage is driven across an on-chip 100 kΩ resistor to produce an output current. A fixed offset current (nominally 10 μA), derived from a band-gap generated voltage across an on-chip resistor is also added to the current output in order to set the ICO center frequency and allow both positive and negative deviations. This type of input circuitry provides significant flexibility; however, some types of sensors may require more sophisticated front end circuitry.

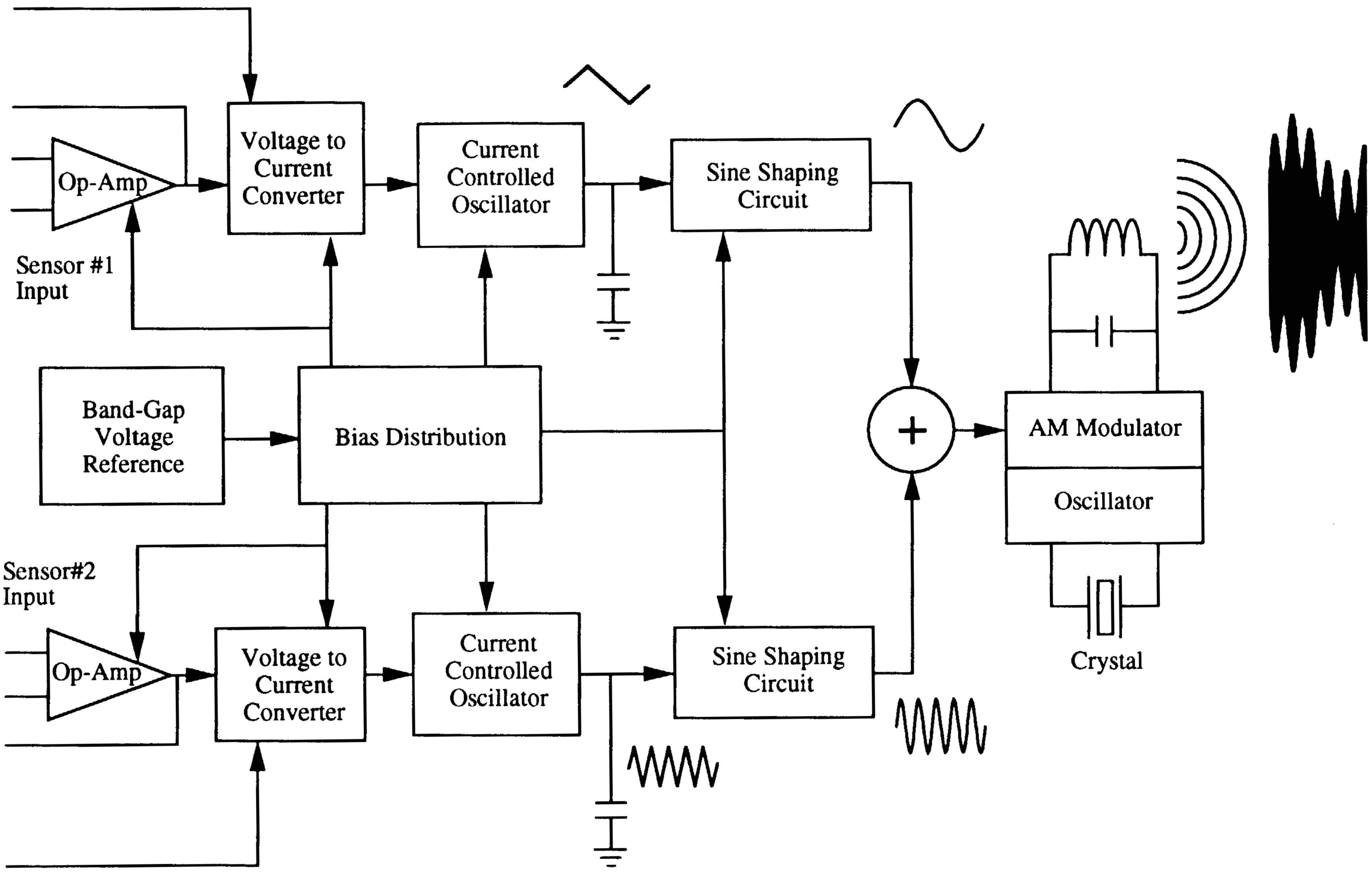

Voltage to Current Converter
Current Controlled Oscillator
Sine Shaping Circuit
Op-Amp
Sensor #1 Input
Band-Gap Voltage Reference
Bias Distribution
+
AM Modulator
Oscillator
Sensor#2 Input
Op-Amp
Voltage to Current Converter
Current Controlled Oscillator
Sine Shaping Circuit
Crystal

Figure 1. Telemetry chip block diagram

The current into the ICO programs the magnitude of the current through an external capacitor. This external capacitor is charged and discharged between two fixed voltages. When the capacitor voltage reaches one of the voltage trip points, the direction of the capacitor current is reversed. The resulting output of the ICO is a triangle wave between two voltages (nominally 400 mV and 800 mV) set by the band gap reference. Using voltages referenced to the band-gap reduces sensitivity to both supply voltage variations and temperature variations. The instantaneous frequency of the ICO output is described by equation 1:

$$f = (1.0 + Vd)/(0.8\ RC) \qquad (1)$$

where Vd is the differential voltage at the voltage to current converter input, R is 100 kΩ and C is the value of the external capacitor. Equation 1 shows that the signal of interest frequency modulates the ICO waveform. Frequency modulation allows constant voltages to be measured. The ICO's are designed for wide-band FM modulation in order to maintain dc accuracy. Constant voltages or currents are produced by a variety of sensors including strain gauges, chemical sensors and others.

The triangle wave contains odd harmonics of the ICO's fundamental frequency. The harmonics from the low frequency channel will fall in the band allocated to the high frequency channel and cause interference. Because of this, both of the channels are put through a sine shaping circuit. The sine shaping circuit substantially attenuates the odd harmonics present in the triangle wave. A shape circuit is used because it uses much less chip area than filter circuits, requires no external parts, and provides maximum flexibility for configuring the sub-carrier channels. The two shaped signals are summed to produce a combined signal that contains all of the information from the two sensors. The signals can be summed because the center frequencies and frequency deviations are chosen to limit the amount of cross-talk between channels.

The third channel is the temperature channel. The transducer is a temperature sensitive crystal which sets the frequency of the carrier oscillator. The carrier oscillator output is then amplitude modulated by the summed ICO signals. This final signal, which contains the two FM signals AM modulated onto the temperature signal drives a small coil which acts as an antenna. The coil generates a magnetic field that can be picked up by an external receiver. The temperature sensing crystal used to generate the AM carrier resonates at approximately 262 kHz at 25 °C and nominally varies 9 Hz/°C. The temperature measurement can be calibrated to better than 0.1 °C.

External Receiver

The received signal is picked up by a small ferrite core antenna coil and amplified by an automatic level controlled amplifier. Level control prevents the amplifier from saturating thus preserving the AM modulation when the pickup coil is close to the transmitter. When the transmitter is far away the automatic level control increases the gain so that the AM signal can be discriminated from the noise. The amplified antenna signal is split into two paths, one for temperature and one for the other sensors. The temperature signal is sent to a zero-crossing detector which feeds a computer controlled frequency counter. The other two channels are extracted by AM demodulating the signal received on the non limited antenna signal. This signal is equivalent to the summed signal in the internal system. Bandpass filters are used to extract the individual FM channels. The separated signals are then individually FM demodulated by phase locked loop circuits. A final filter to remove the FM carrier frequencies provides the measured physiological variables.

SYSTEM CONSIDERATIONS

Design Philosophy

Many of the design decisions were driven by the requirement that the final product be easy to manufacture. The custom IC was designed to use the minimum number of external parts in order to lower costs. The only external parts that must be separately assembled are the battery, sensors, ICO tuning capacitors, coil tuning capacitor, and the transmitter coil. Depending on the intended application there may also be gain setting resistors, signal conditioning capacitors, and other circuitry. This allows the user to customize inexpensive chips for different uses or for doing prototype work. If the volume of a particular application is high enough, some of the application specific circuitry could be integrated onto the chip to reduce costs. Another design consideration was to provide as many features as practical, while maintaining minimum chip area.

Practical Issues

The ICO center frequencies are determined by the nominal bias point of the input amplifier, the voltage to current converter reference voltage, the offset current, and the external capacitor. The offset current is determined by the band-gap voltage and an on-chip resistor. Because of the wide variation (± 30%) in the absolute value of the on-chip resistors, the ICO capacitors will generally have to be selected or trimmed in order to accurately set the center frequencies. Trimming the ICO center frequency also trims the frequency sensitivity of the ICO because the voltage to current converter uses an on-chip resistor that is matched to the resistor that sets the offset current. Applications that require low dynamic range and limited signal bandwidth may be able to use fixed value ICO capacitors.

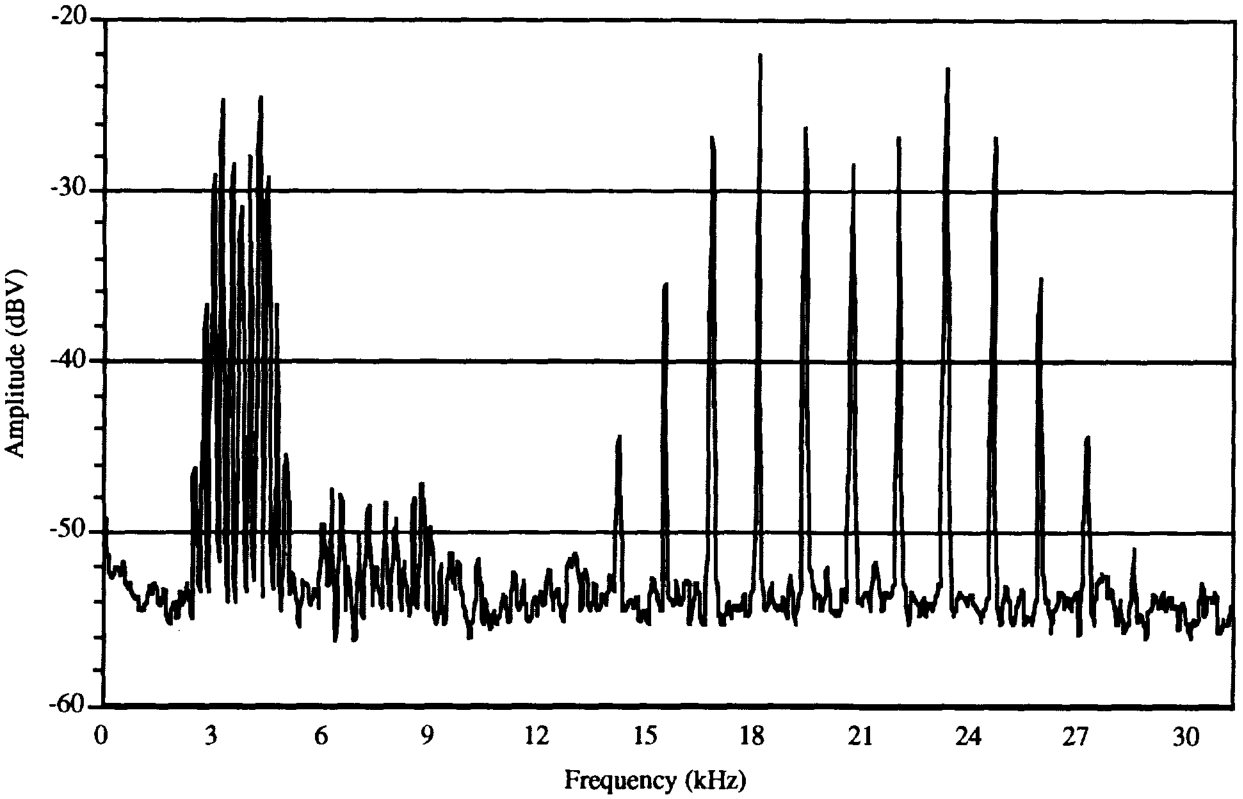


Figure 2. Frequency spectrum for summed FM channels under maximum bandwidth conditions.

Another possible trade-off is between signal bandwidth and dc accuracy. For wide bandwidth applications, the maximum frequency deviation of the ICO's should be around ± 20 % of the center frequency. This range allows the second harmonic caused by distortion in the sine-shaper and receiver circuits to fall outside of the passband of the FM signal, while still allowing for reasonable dc accuracy. Using ±20 % frequency deviation allows a total frequency error of 0.4 % not related to the input signal while maintaining a 1 % of full scale dc accuracy. Many different channel assignments are possible for this device. Our test system uses center frequencies at 3.8 kHz and 20 kHz. When modulated with maximum amplitude (± 20 % frequency deviation), maximum frequency (250 Hz and 1.3 kHz) inputs (corresponding to maximum channel bandwidth conditions) the measured frequency spectrum for the summed signal shown in Figure 2 results. Second harmonic distortion from the sine shaping circuit appears as a weak signal that falls between the passbands assigned to the channels. The sine shaping circuit minimizes the third and fifth harmonics in order to keep cross-talk to a minimum. Allowing larger than 20 % frequency deviation can improve the dc accuracy, because this tends to reduce the relative importance of battery voltage, and temperature on the center frequency. If the maximum frequency deviation is raised much beyond the 20% level there will still be little cross-talk, but the second harmonic distortion will fall into the signal's own passband resulting in distortion. This is not a problem for dc signals, and is unimportant if the amplitude of high frequency inputs is small.

The operational amplifiers have a variety of limitations. The input voltage range is from 800 mV below the battery voltage down to ground. The output voltage can cover the same range, but the gain drops and non-linearity increases when the output falls below 200 mV. The reference input range is the same as the op-amps, but behaves slightly non-linearly for voltages below 200 mV. A Schottky diode connected to ground and biased to a forward voltage drop of about 500 mV is available on the chip and can serve as a "virtual ground" for sensor connections. This "virtual ground" allows many types of sensors to be easily interfaced while maintaining the input and output signals above 200 mV. If the voltage to current converter is normally operated with zero differential input, then a change of 200 mV will cause a 20% frequency deviation. Some applications may benefit from having the voltage to current converter operate with a nominal differential voltage different from zero. This can either increase or decrease the change necessary for a 20% frequency deviation depending on the sign of the offset.

MEASURED CHARACTERISTICS

The measurements below represent data taken from 4 functional telemetry chips and generally reflect worst-case measurements.

Power Supply

Maximum supply currents is 690 µA at 2.5 V, 770 µA at 3.0 V, and 870 µA and 3.6 V. The transmitting coil current increases approximately linearly with increasing supply voltage, so that maximum transmission range is a function of supply voltage.

Input

Operational amplifier offset voltages from 8 amplifiers were 0.78, 0.36, 2.1, 6.1, -0.16, 0.87, -0.21, and 2.6 mV. This is too small a sample to realistically determine the bounds of the offset voltage, but it suggests that selecting devices with less than 1 mV offset voltages should be possible. Measured open loop gain was 850 minimum. The high frequency roll-off and slew rate limiting are well beyond the limits imposed on the signal bandwidth from considerations of the FM channel allocation. The offset voltage for the voltage to current converter circuits cannot be distinguished from the variation in offset currents added to the converter's output; however, based on the IC layout, it should be similar in magnitude to the op-amp offset voltage. The complete input circuit has a worst-case supply voltage to output current coefficient of 4.3 % /V, which limits either the range of useful battery voltage, or the dc accuracy of the signal transmission. For example, if a dc accuracy of 1 % is necessary , then the battery voltage should probably not be allowed to vary more than 100 mV over the course of measurement in order to leave room for other dc errors. This poor characteristic is the result of insufficient output impedance of the voltage to current converter. The circuit which performs the voltage to current conversion has been modified to improve this characteristic on the next generation chip.

Modulator

Maximum non-linearity for ICO output frequency verses differential input voltage for unity gain connected op-amp is less than 0.5 % of full scale for reference voltages from 800 mV below supply voltage down to 200 mV. The worst case non-linearity rises to 1% when the reference input is grounded. Severe non-linearity occurs for reference voltages above the specified input range. The maximum amplitudes of the harmonics at the output of the sine shaping circuit over the entire supply voltage range and ICO operating frequency range from 2 kHz to 25 kHz relative to the fundamental amplitude were: -20 dB for the second harmonic, - 29 dB for the third harmonic, -36 dB for the fourth harmonic, -38 dB for the fifth harmonic, all remaining harmonics were below -40 dB. The supply voltage coefficient of output frequency is dominated by the voltage to current converter coefficient for a total worst case of 4.3 %/V. The AM modulator circuit maintains the percent modulation between 50 % and 75 % over the power supply range.

Crystal Oscillator

The supply voltage coefficient of oscillator frequency was immeasurable. Variations of center frequency from unit to unit was no more than 2 Hz.

Transmitted Signal

Figure 3 shows the peak field strength of the transmitted signal from a coil having 300 turns in four layers 8.6 mm in diameter by 8.6 mm tall. The transmitter was operating with a nominal 3 V power supply. Range in a typical commercial environment is greater than 60 cm. Strong interfering signals can reduce the usable range. The transmitting coil current and field strength are proportional to supply voltage. This feature allows the power drain on the battery to diminish as the battery voltage falls resulting in slightly longer battery life. The test coil used in this measurement is applicable to ingestible applications where the telemetry system is packaged in a capsule form. Many other coil arrangements for different applications are possible.

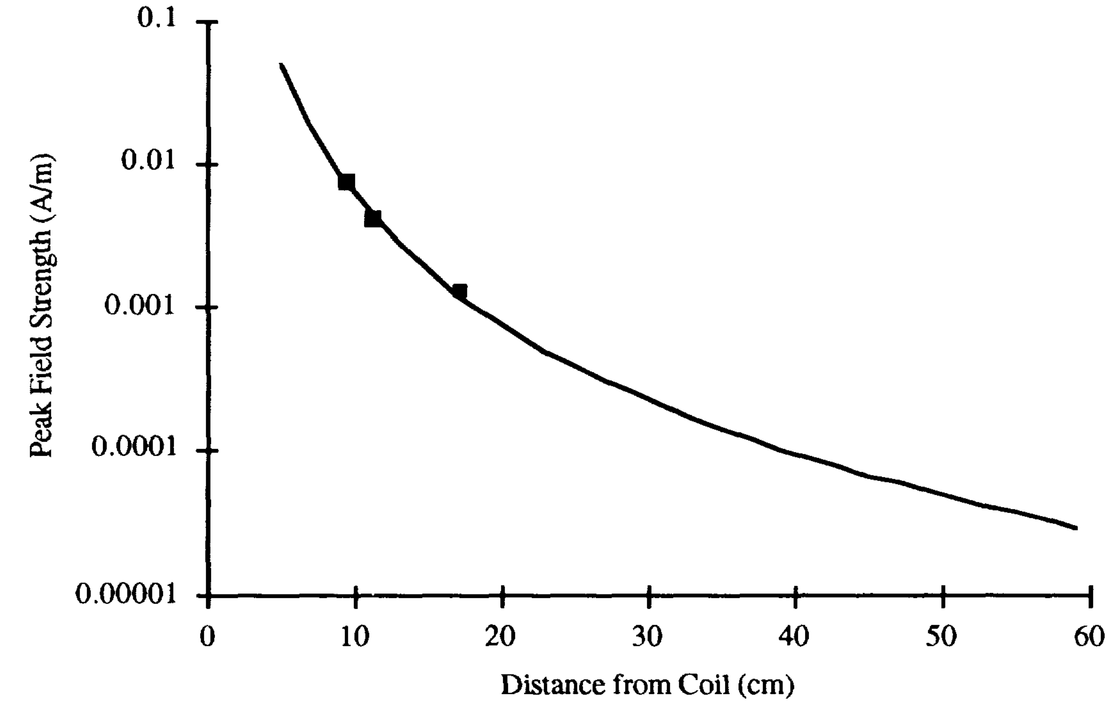


Figure 3. Axial field strength vs distance from coil. Symbols are measured values and curve is theoretical field strength.

Temperature Characteristics

The temperature coefficient of the FM channel center frequencies was + 0.48 %/°C, and was essentially equal to the temperature coefficient of the frequency sensitivity. This poor characteristic is basically the difference between the nominally 0.63 %/°C temperature coefficient of the on-chip 100 kΩ on-chip N-well resistor , and the nominally - 0.15 %/°C temperature coefficient of the external chip capacitors used in the test. This figure could be substantially improved by the use of N5700 dielectric capacitors, which will better compensate for the temperature variation of the on-chip resistors. Another possibility is to use a temperature correction algorithm on the received signal since the temperature of the telemetry system is always available. Future generation chips will probably utilize polysilicon resistors which use more chip area, but have only a + 0.09 %/°C nominal temperature coefficient in order to improve this characteristic.

EXAMPLE APPLICATIONS

Temperature is one variable that is always measured. The temperature transducer is a crystal that was designed to have a large linear frequency verses temperature characteristic. Because this transducer is required to generate the AM carrier for the other two channels, temperature is always available from the system.

One version of the telemetry system was customized to measure tendon force and the electromyogram of the muscle attached to the tendon. This implanted configuration used a strain-gage based force buckle to measure the tendon force. The experimenter could measure the amount of force, and muscle EMG produced by a given nerve stimulus. Such a device could be used to monitor the performance of a nerve stimulator.

Another example would be the measurement of pressure in the gastrointestinal tract. A capsule the size of a large vitamin pill could contain the telemetry system, pressure transducer, and a small battery. A bridge type pressure sensor could be used on one channel, while the other channel telemeters battery voltage in order to compensate for changes in excitation voltage. An ingestible sensor of this type could measure peristalsis of the gut, and also can pick up the respiratory rate.[1]

Another application could measure the partial gas pressures of oxygen and carbon-dioxide in the gastrointestinal tract, also in the form of a swallowable capsule. CO_2 and O_2 electrodes inside the pill could measure gas concentrations that diffuse through the silicone rubber outer coating of the pill. This measurement could be useful in detecting ischemia of the gastrointestinal tract. When blood flow to the gut is reduced, carbon dioxide builds up and oxygen is depleted from the tissues. These gasses diffuse through the tissues, so the ambient concentration of these gases in the lumen of the gut change with ischemia.[2]

The system is not limited to biological measurements. It could be modified to transmit data from any container whose integrity must be maintained. An example is a pressure gauge for vacuum systems that does not require a feedthru in the chamber. The pressure transducer would be interfaced to the telemetry chip and then the signal transmitted through one of the chamber viewing ports. The other channel could be used to transmit an analog or digital signal from the experiment in the chamber.

CONCLUSIONS AND FUTURE PLANS

A very small telemetry system has been designed and tested. The system is ready to be integrated with existing sensors to produce a device customized for a specific need. A hybrid substrate is being designed to make the final outline of the ingestible device as small as possible. The projected dimensions of the ingestible device are 9 mm diameter by 15 mm long.

A version of the chip which is currently under production contains new digital functions to send a calibration/identification signal to the external receiver. With the calibration embedded in the data stream the receiver will automatically read it and apply it to the data. The calibration factors are added at the factory and are stored in the telemetry chip until it is activated. Other changes include a band-gap regulated output voltage for sensor excitation, various circuit improvements to reduce power consumption and increase power supply rejection ratios, and on-chip power on/off circuitry. Many other future improvements are planned for this device including the ability to operate from a single 1.5 V battery.

Acknowledgement

This work was supported by the Office of Technology Transfer at the Goddard Space Flight Center under contract NDPR92635.

REFERENCES

1. Mackay R S, *Telemetering from within the Body of Animals and Man: Endoradiosondes*, Caceres C A, ed., Biomedical Telemetry, Academic Press, 1965; Chapter 9.

2. Hartmann M, Montgomery A, Jönsson K, et al: Tissue oxygenation in hemorrhagic shock measured as transcutaneous oxygen tension, subcutaneous oxygen tension, and gastrointestinal intramucosal pH in pigs. Crit Care Med 1991; 19:205

IMPLANTABLE STIMULATOR SYSTEM RESTORES MOTOR FUNCTION

This paper was withdrawn from presentation

AUTOMATED SYSTEM FOR ANALYZING THE ACTIVITY OF INDIVIDUAL NEURONS

514-52
150658
p-8

Isaac N. Bankman*, Kenneth O. Johnson#, Alex M. Menkes*,
Steve D. Diamond*, and David M. O'Shaughnessy#
The Johns Hopkins University
*Applied Physics Laboratory
#Neuroscience Department

N93-22163

ABSTRACT

This paper presents a signal processing system that i) provides an efficient and reliable instrument for investigating the activity of neuronal assemblies in the brain and ii) demonstrates the feasibility of generating the command signals of prostheses using the activity of relevant neurons in disabled subjects. The system operates on-line, in a fully automated manner and can recognize the transient waveforms of several neurons in extracellular neurophysiological recordings. Optimal algorithms for detection, classification, and resolution of overlapping waveforms are developed and evaluated. Full automation is made possible by an algorithm that can set appropriate decision thresholds and an algorithm that can generate templates on-line. The system is implemented with a fast IBM PC compatible processor board that allows on-line operation.

INTRODUCTION

Two equally important reasons have recently increased considerably the significance of processing the signals generated by neurons.

1) The expanding applicability of neural networks to diverse engineering problems has raised the interest in neurophysiological investigations that aim to study the collective behavior of neuronal assemblies. It seems clear that advances in areas such as pattern recognition, memory storage, speech processing, computer vision, and control will be possible by a better understanding of biological neural systems, especially the human brain. This promise has been recognized in the Congressional resolution designating the 1990's the "Decade of the Brain". Furthermore, the National Academy of Sciences indicated that neuroscience stands at the threshold of a significant expansion. However, as neuroscientists in The Johns Hopkins Medical School and other leading universities agree, an important prerequisite is instrumentation for observing the electrical activity of individual neurons and their assemblies.

2) In aiding the disabled, various kinds of prostheses have been effective and have led to increased attention to the field. In current prostheses, the motion command for a joint assisted or replaced by a prosthesis is generated by another joint: for example, the command for the legs of a paraplegic is initiated with switches operated manually by the subject. A recent direction of research, named *neural prostheses*, aims to generate the commands in a more convenient, natural and potentially more effective manner. In a neural prosthesis, the goal is to obtain the command directly from neurons or muscle fibres that are relevant to the target joint. If this can be achieved by obtaining the commands from the very neurons that once controlled the disabled joint, the most direct and natural link between the motion intent and the target joint can be developed. Here again, the prerequisite is instrumentation for obtaining the activity of neurons.

Sensory or motor information is processed by biological systems in the form of a distributed representation supported by a large number of neurons interconnected with excitatory and inhibitory synapses. The functional activity of an individual neuron depends on the strength of the synapses that provide excitatory or inhibitory input from other neurons, and on the activity of these neurons. When the net input is excitatory, above a threshold level, the neuron fires and generates an action potential across its membrane, that lasts about 1 ms. Ongoing activity in a neuron is manifested by a sequence of action potentials that can be recorded with an extracellular electrode. The main advantage over an intracellular electrode is the ability to record from more than one neuron at the same time, but the extracellular electrode also allows recording without damaging the neurons. The cost of these benefits is the requirement for sorting the interleaved neural spike trains (Fig. 1) to determine the firing instants of individual neurons.

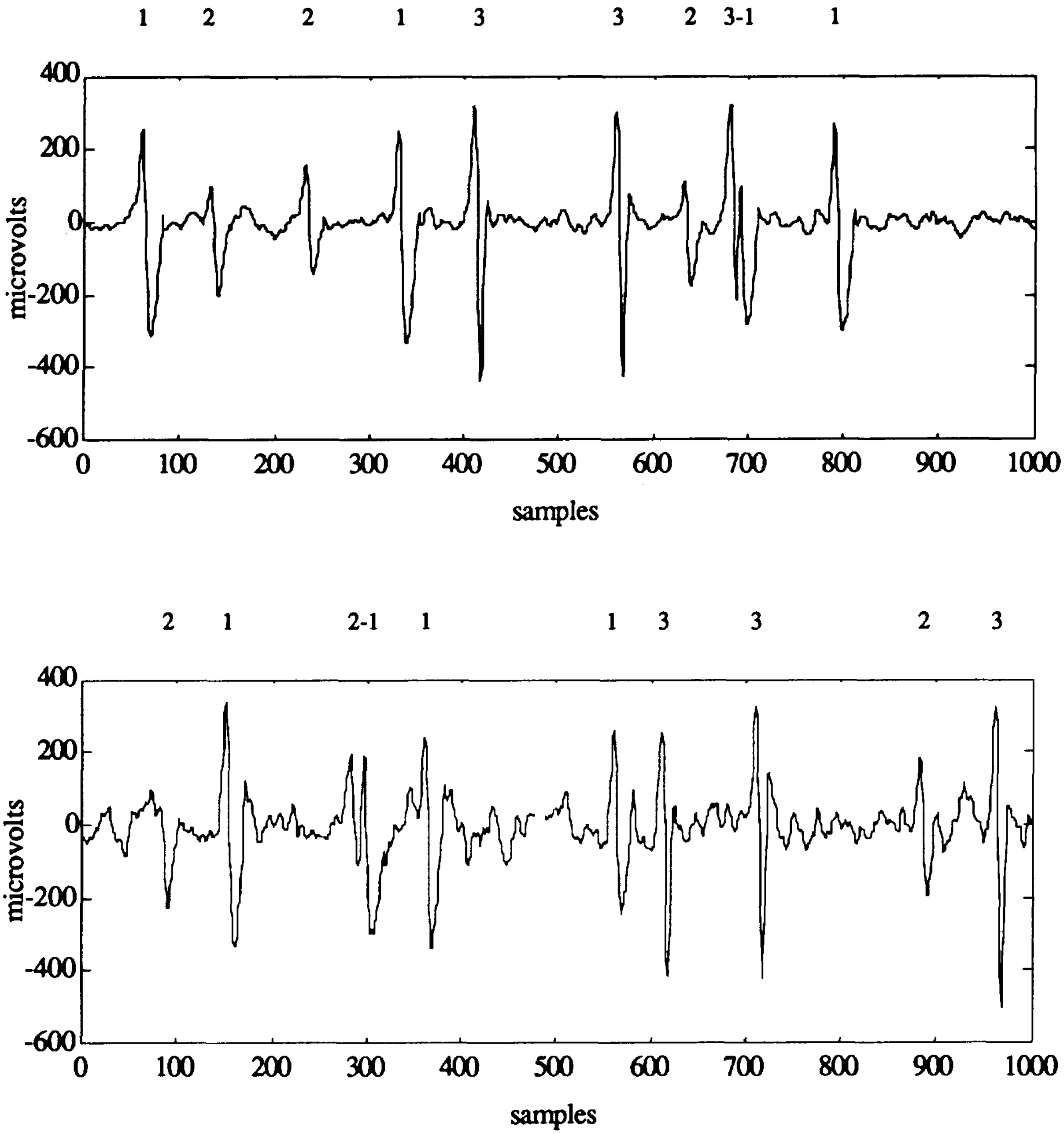


Fig. 1. Recordings at two different noise levels. The number on each spike indicates the corresponding neuron. The SNR of a spike in this study is defined as the rms value of the spike in units of standard deviation of noise. The SNR of the smallest spike (2) is about 4 in the top trace and about 2 in the bottom trace. Superpositions of pairs of spikes are labeled as 3-1 on the top trace and 2-1 on the bottom trace. The sampling rate was 32 KHz.

Different neurons generate distinct spike waveforms in the recording, due to differences in their dendritic geometry and the impedances of the medium connecting them to the electrode. The activity of individual neurons can be determined by sorting the different types of neural waveforms. An essential challenge in extracellular recordings is the relatively low signal-to-noise ratios (SNR) that can occur in many cases. The background noise is mainly due to the activity of a large number of distant neurons resulting in a considerable overlap between the spectra of waveforms of interest and noise. Furthermore, since the noise process is primarily made of the accumulation of low amplitude spikes, the autocorrelation function of noise has significantly high coefficients at lags as large as the average duration of a neural spike (about 1 ms). Moreover, two neurons often fire concurrently, their spikes overlap in the recording (Fig. 1) , causing additional difficulty in recognizing the activity of individual neurons. The neural waveform recognition problem is a typical example of detection and classification of transient patterns embedded in colored noise.

In recordings for brain research, it is preferable to achieve neural spike sorting on-line because immediate feedback on the recording allows better control of experimental conditions and recording quality, reducing the time requirement for the neuroscientist and the animal subject. In neural prosthesis applications on-line operation is essential. This requires a signal processing system that can operate in a fully automated manner, without human supervision. We developed signal processing and pattern recognition algorithms as well as a hardware implementation that operate with theoretically optimal recognition performance, on-line, in a fully automated manner [1-5].

The data used in the development of this analyzer were recorded from the primate cortex using a filter pass-band of 10 Hz to 10 KHz, a 12 bit A/D converter and a sampling rate of 32 KHz.

ALGORITHMS

In order to determine the activity of individual units, an automated system should perform three main tasks: i) discrimination of waveforms from noise (detection), ii) discrimination between waveforms of different units (classification), and iii) separation of overlapping waveforms (resolution of superpositions).

Detection

The analyzer performs detection by computing the power *p(k)* of the signal *N(k)* within a running window of length *n* :

$$p(k) = \sum_{i=0}^{n-1} N(k-i)^2 . \qquad (1)$$

A waveform is detected when the power exceeds an appropriate acceptance threshold which is set with an automated algorithm. Power detection yielded considerably better results than the widely used amplitude detection in tests comparing the detection performance as a function of SNR. The SNR was defined as the ratio of the signal rms value (computed with n samples) to the standard deviation of noise in the record. With the detection thresholds set for a false positive rate of less than 0.1%, power detection was 100% and 94% correct at SNRs of 3 and 2, while amplitude detection provided 95% and 71% correct detection respectively. The system generates, on-line, a template for each unit to be used in classification. When the templates become available, the detection of the corresponding waveform types can be improved with matched filtering. The improvement obtained with matched filtering can be as high as 40% more correct detection than the power detection technique, at low SNR levels.

Classification

Several methods for neural spike classification, ranging from amplitude discrimination to principal components and minimum mean-square-error, have been suggested [6-13]. In previous comparison studies [7] principal components and template matching using Euclidean distance were found to be the best for spike sorting in noisy data. Because the statistically optimal Bayesian classification can be achieved by template matching and because current technology allows its on-line implementation, we focused on neural spike sorting by template matching. In view of the diversity of applicable classification techniques, the emphasis of this project was placed on developing the theoretically optimal approach that could provide minimal noise sensitivity. Therefore, the optimal template matching approach was investigated and evaluated.

In the template matching method, each waveform is represented by *n* consecutive samples digitized on the waveform and it can be viewed as an *n* -dimensional vector. The waveform vector of a given neuron will vary somewhat each time that this neuron fires, due to additive random noise. The waveforms of the same neuron will form a cluster in sample space, around the centroid that would be observed in the absence of noise. When the distribution of noise amplitude is Gaussian, as in neural recordings, the optimal Bayesian classification can be achieved by computing the mean of the cluster of each neuron and by setting a decision boundary around each mean with a distance metric that depends on the covariance matrix of noise.

The probability density p(x) of a multivariate Gaussian cluster distribution is:

$$p(x) = \frac{1}{(2\pi)^{n/2}|C|^{1/2}} \exp[-(x-m)^T C^{-1}(x-m)/2], \qquad (2)$$

where, x is the n-dimensional random waveform vector, C is the covariance matrix of x vectors in that cluster, $|C|$ is the determinant of C, and m is the mean vector in the cluster. In the multiclass problem, each class has its own cluster and probability density.

Let the data have K different classes represented by w_i with the class index i = 1,...,K. Multiclass Bayesian classification is performed with discriminant functions that are based on the class densities and *a priori* probabilities of the classes. A convenient choice of discriminant function is:

$$g_i(x) = \log(p(x|w_i) + \log P(w_i), \qquad (3)$$

where, $g_i(x)$ is the discriminant function, $p(x|w_i)$ is the probability density of class i, and $P(w_i)$ is the *a priori* probability of class i. The class to which a candidate pattern belongs is determined by computing the values of the discriminant function using the pattern's vector x for each class. The pattern is assigned to the class with the highest discriminant function value.

When each class has a multivariate Gaussian distribution, the expression for $g_i(x)$ becomes:

$$g_i(x) = -(x-m_i)^T C^{-1}{}_i(x-m_i)/2 \; - (n\log 2\pi)/2 - \; \log|C_i|/2 + \log P(w_i). \qquad (4)$$

This expression can be further simplified because the $(n\log 2\pi)/2$ term is the same for each class, the covariance matrix C_i is the same for all classes and the *a priori* probability of each class is assumed to be the same. Therefore, classification with this discriminant function becomes equivalent to classification according to the minimal value of the quantity:

$$d_i^2 = (x-m_i)'C^{-1}(x-m_i) \qquad (5)$$

which is known as the squared Mahalanobis distance between x and m_i. Constant Mahalanobis distance contours are ellipses centered around the mean of each class; these ellipses coincide to equal density contours on the multivariate distributions. In most applications, it is desirable to have the option of leaving some patterns unclassified or rejecting them. To do so, only patterns that have a distance below an acceptance threshold are classified. This is equivalent to setting an elliptical decision boundary around the mean of each class.

None of the reported neural spike sorting applications have used the Mahalanobis distance in classification because of its computational burden. The Euclidean distance which has been commonly used, is equivalent to setting a circular decision boundary and provides suboptimal results. The extent of performance loss due to the use of Euclidean distance depends on the covariance matrix of noise, the noise level of the data and the similarity between different classes.The Euclidean distance can provide satisfactory results regardless of the covariance matrix of noise if the noise level is relatively low and the clusters of different classes are sufficiently apart. Our recent studies, using 32-sample templates and 5 different neural spike classes embedded in typical neural recording noise, showed that if the Euclidean distance between the means of the two closest clusters is more than 14 standard deviations of noise, perfect classification can be obtained. But as the clusters get closer and/or the noise level increases, the performance drops and the loss, referred to the optimal case, can reach up to 25%.

The extent to which Euclidean distance deteriorates the classification performance depends on the covariance matrix. The higher the autocorrelation in noise, the higher the eccentricity of the elliptical distributions and the lower the performance will be with Euclidean distance. However, if the noise had no autocorrelation, the covariance matrix would be diagonal and the Euclidean distance would provide optimal classification. Therefore, the system that we developed whitens the data before classifying the waveforms. Whitening is achieved with a digital FIR or IIR filter whose coefficients are determined by modeling the noise in the recording. The model is an autoregressive moving average (ARMA) model [13] and its inverse provides the whitening filter.

The contribution of whitening to template matching performance was evaluated using a test data set with 5 different spike types and prior knowledge of exact templates. With the acceptance threshold set to provide fewer than 0.1% false positives, the correct classification performance on raw data was 96% and 74% correct, at SNRs of 2 and 1 respectively, while on whitened data the performance increased to 100% and 91% correct respectively.

Resolution of superpositions

When two neurons fire simultaneously their waveforms overlap and generate a complex waveform that does not match any of the templates of individual neuron waveforms. Such superpositions occur with considerable frequency depending on the number of neurons in the recording, their firing rates, the duration of the action potentials, and the timing relations between the action potentials of individual neurons. Failure to recognize the spikes that overlap can cause underestimation of the firing rates, and can affect severely the analytic measures of interevent timing, such as cross-correlation between neurons or inter-spike interval histograms. In a typical recording containing waveforms from 5 neurons firing at moderately high rates, about one third of the waveforms may overlap.

The neural waveform analyzer that we developed includes a superposition resolution algorithm that essentially subtracts each template from the complex waveform and attempts to classify the remainder. The resolution is performed on whitened data and provides an optimal template matching approach for this task. The performance of this algorithm was evaluated in tests where all templates were available. With the acceptance threshold set to provide fewer than 0.1% false positives, the correct resolution performance was 100% and 95% correct at SNRs of 3 and 1.5 respectively, on whitened data.

Noise segmentation

Besides the three main functions of detection, classification and superposition resolution, two other functions are needed for a fully automated on-line system. The first one is automated threshold setting that requires a segment of only noise in the record. The decision thresholds for the power value in detection and the Euclidean distance in classification have to be set in accordance to the level of noise in the recording. Therefore, in order to set the thresholds automatically, a section that contains only noise has to be segmented from the recording. Furthermore, this has to be done at the beginning of the process, without using any of the detection techniques mentioned above because thresholds are not yet available. We developed an iterative algorithm that can separate the noise from all other transient waveforms, without using templates or decision thresholds. This algorithm is based on the fact that the amplitude of the noise process is normally distributed. The noise segmentation algorithm provided adequate results in test data sets that had SNR levels ranging from 1 to 10 and total spike rates of 5 to 160 spikes per second.

Template generation

The second function required for full automation is template generation. The system should be able to observe the incoming data and determine a template for the waveform of each neuron. In the system that we developed, this is achieved with unsupervised clustering techniques that provide two different approaches: sequential or simultaneous. In the sequential clustering approach, the first spike detected becomes the first template, representing the waveforms of the first neuron (type 1). The second spike is compared to this template; if the distance is lower than the acceptance threshold, the second spike is classified as type 1 and the template is updated by averaging. If the distance is greater than the acceptance threshold, the second spike waveform is used as a different template, representing a second type of spike waveform. Each subsequent spike is classified similarly and either the corresponding template is updated or a new type is initiated. The sequential clustering algorithm provided appropriate templates in the data sets that we used for evaluating the system. We are currently investigating simultaneous clustering algorithms that have potential for increasing the clustering performance at very low SNR levels. Simultaneous clustering algorithms use clustering criteria applied simultaneously to a large number of waveforms obtained at the start of the recording. The templates generated by such algorithms do not depend on the detection sequence of waveforms and the effects of noise at low SNR levels are reduced.

Complete system

The block diagram of the complete system is shown in Fig. 2. At the beginning of the recording, an initial section of the digitized data is processed by the noise segmentation algorithm that provides segments of only noise to the system. The thresholds for detection, classification and superposition resolution are set automatically using the extracted noise segments. The noise segments are also used by the ARMA modeling algorithm which in turn provides the coefficients of the whitening filter. After these preprocessing steps that last several seconds, the remaining recognition functions are performed on-line using a double buffering input arrangement.

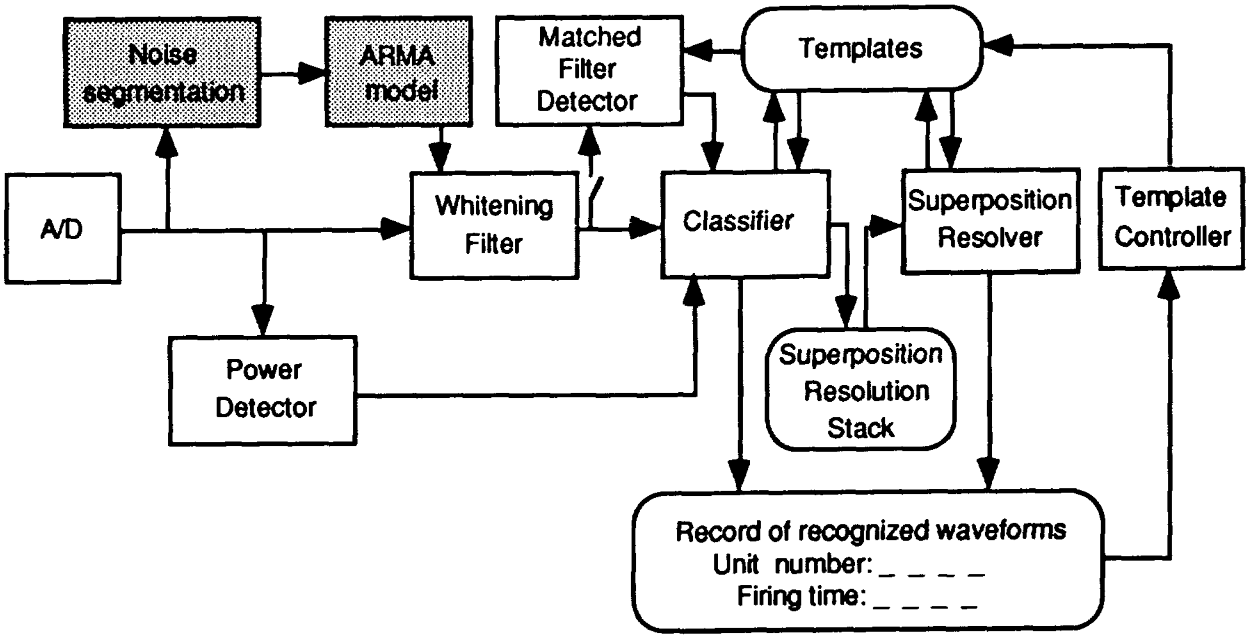


Fig. 2. Block diagram of complete system. Shaded components represent preprocessing algorithms.

At the start of the recognition process, since templates are not available, waveforms are detected with the power detector that passes indicates the occurrence times of the detected events to the classifier. The classifier applies sequential clustering to the whitened data to generate and update templates for each neuron in the recording and uses the templates for classifying the waveforms on the whitened data. Each time that a waveform matches a template, the corresponding type and occurrence time is recorded and the matching template is updated. If the waveform does not match any template, it is first assumed to be a superposition and placed in the corresponding stack.

The superposition resolution algorithm attempts to resolve each waveform in this stack by using the templates. If the unknown waveform can be resolved in terms of two templates, these two templates are updated and two spikes are recorded. If the waveform cannot be resolved then it is considered to represent a new neuron that started to fire and this waveform becomes a new template.

In some cases, waveforms that cannot be classified or resolved can be artifacts or spikes of very inactive neurons that are not worth pursuing. The template controller monitors the activity of each type by computing the number of times each template has matched a waveform in the recent past. If the activity of a given template is lower than a preset level, that template is eliminated.

When the template for a neuron is available, the detection of that type is performed by matched filtering, while power detection is kept active in order to detect new types as well as superpositions.

HARDWARE

The system is implemented on an IBM PC compatible, floating-point processor board developed in The Johns Hopkins Applied Physics Laboratory. This implementation allows about 40 MFLOPS operation for most of the functions of the system such power detection, whitening, classification, superposition resolution, matched filtering and template update, using assembly language. Each of the two 4 MBytes memory banks of the processor is connected to a parallel I/O port that can transfer data at a rate of 80 MBytes per second. The recorded signal is digitized with a commercial A/D board that stores the data directly on one memory bank of the processor using DMA through one of the parallel I/O ports, without passing from the IBM PC and without requiring time from the CPU of the processor. The on-line recognition results obtained by the processor are passed to the IBM PC through the second parallel I/O port for display and archiving. The human interface provides displays of the raw data, whitened data, running power values, matched filter outputs, templates, clusters projected on two dimensions, as well as measures of the data quality and the recognition difficulty. Further display functions such as correlation histograms, interval histograms, and raster plots of spikes can also be implemented.

CONCLUSION

We developed a fully automated system that can recognize the transient waveforms generated by several neurons in an extracellular recording. The most significant contributions of this system are i) theoretically optimal operation that provides minimal noise sensitivity, ii) the ability to generate all operational parameters (e.g. templates and thresholds) automatically and on-line, allowing its use with minimal human supervision, and iii) resolution of superpositions, providing an indispensable tool for more complete data acquisition.

Furthermore, since the processor is a general purpose computation tool, its use is not limited to only sorting the waveforms. After the recording, the same hardware programmed with a high level language such as C, can be used for investigating the collective behavior of many neurons. By allowing both waveform recognition and further neurophysiological investigation, this system provides a cost effective instrument for neuroscience research.

The fully automated and on-line operation is a unique property of this system that shows the feasibility of reliable on-line recognition of neural activity for neural prosthesis applications. In neuroscience research applications, fully automated on-line operation enables more efficient recording and processing. This is the result of the reduced human supervision and the immediate feedback that the system can provide to the user. By eliminating the need for the constant supervision that available systems require, the system that we developed allows more time and focus to the neuroscientist for the proper management of the experiment. Moreover, by recognizing and reporting immediately the activity of neurons, the system provides valuable feedback and guidance for the recording. Research decisions that are made several days after the recording, can be made during the recording owing to the immediate, automated results. This more efficient operation can lead to a reduction in the use of laboratory animals.

This system is an example of the transfer of military technology to civilian and commercial applications.

ACKNOWLEDGEMENT

This work was supported by the NIH Grant NS 07226, and partially by Department of the Navy, Contract No. N00039-91-C-0001.

REFERENCES

1. Bankman, I. N., Johnson, K. O., and Schneider, W., "Optimal recognition of neural waveforms," *Proc. of the 13th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 13, pp. 409-410, 1991.

2. Bankman, I. N., "Detection and Classification of Transient Signals: Sorting Neural Waveforms," *The Johns Hopkins APL Technical Digest*, vol. 12, pp. 144-152, 1991.

3. Bankman, I. N. and Menkes, A., "Automated Segmentation of Neural Recordings for Optimal Recognition of Neural Waveforms," *Proc. of the 14th Annual Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 14, pp. 2560-2561, 1992.

4. Menkes, A., Bankman, I. N., and Johnson, K. O., "Simulation of a Fully Automated System for Optimal On-line Recognition of Neural Waveforms in Extracellular Recordings," *Proc. of the 14th Annual Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 14, pp. 2562-2563, 1992.

5. Bankman, I. N., Johnson, K. O., Menkes, A, Diamond, S. D., and O'Shaughnessy, D. M., "Automated Analyzer for On-line Recognition of Neural Waveforms in Extracellular Recordings of Multiple Neurons," *Proc. of the 14th Annual Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 14, pp. 2852-2853, 1992.

6. M. Abeles and M. H. Goldstein, "Multispike train analysis," Proc. IEEE, vol. 65, pp. 762-773, 1977.

7. B. C. Wheeler and W. J. Heetderks, "A comparison o. techniques for classification of multiple neural signals," IEEE Trans. Biomed. Eng., vol. 29, pp. 752-759, 1982.

8. R. S. LeFever and C. J. DeLuca, "A procedure for decomposing the myoelectric signal into its constituent action potentials - Part I: technique, theory and implementation," IEEE Trans. Biomed. Eng., vol. 29, pp. 149-157, 1982.

9. E. M. Schmidt, "Computer separation of multi-unit neuroelectric data: a review," J. Neurosci. Meth., vol. 12, pp. 95-111, 1984.

10. R. M. Studer, R. J. P. DeFigueiredo, and G. S. Moschytz, "An algorithm for sequential signal estimation and system identification for EMG signals," IEEE Trans. Biomed. Eng., vol. 31, pp. 285-295, 1984.

11. M. Salganicoff, M. Sarna, L. Sax and G. L. Gerstein, "Unsupervised waveform classification for multi-neuron recordings: a real-time software-based system. I. Algorithms and implementation," J. Neurosci. Meth., vol. 25, pp. 181-187, 1988.

12. S. R. Smith and B. C. Wheeler, "A real-time multiprocessor system for acquisition of multichannel neural data," IEEE Trans. Biomed. Eng., vol. 35, pp. 875-877, 1988.

13. G. E. P. Box and G. M. Jenkins, "Identification of transfer function models," Time series analysis: forecasting and control, Holden Day, 1976.

IMPROVED INHALATION TECHNOLOGY FOR SETTING SAFE EXPOSURE LEVELS FOR WORKPLACE CHEMICALS

Bruce O. Stuart, Ph.D., D.ATS
Medical Department
Brookhaven National Laboratory
Upton, New York 11973

ABSTRACT

Threshold Limit Values recommended as allowable air concentrations of a chemical in the workplace are often based upon a no-observable-effect-level (NOEL) determined by experimental inhalation studies using rodents. A "safe level" for human exposure must then be estimated by the use of generalized safety factors in attempts to extrapolate from experimental rodents to man. The recent development of chemical-specific physiologically-based toxicokinetics makes use of measured physiological, biochemical, and metabolic parameters to construct a validated model that is able to "scale-up" rodent response data to predict the behavior of the chemical in man. This procedure is made possible by recent advances in personal computer software and the emergence of appropriate biological data, and provides an analytical tool for much more reliable risk evaluation and airborne chemical exposure level setting for humans.

INTRODUCTION

The Brookhaven National Laboratory (BNL), located in Upton, NY, has reactivated its Inhalation Toxicology Facility (ITF), under the auspices of a five-year collaborative research agreement with ManTech Environmental Technology, Inc. (ManTech Environmental), Research Triangle Park, NC. Coordinated under legislation set forth in the Technology Transfer Act of 1988, ManTech Environmental with Brookhaven National Laboratory will utilize these facilities and its resources for the conduct of toxicological research and testing to address a national need, i.e., providing inhalation biology studies to set safe levels of response to workplace chemicals. The ITF and its adjoining Veterinary Services Complex-Laboratory Animal Facility (VSC) are integral components of the BNL Medical Department.

The ITF, located in the eastern corner of the Medical Department complex (Figure 1), in its renovated configuration, encompasses a dual corridor system, to expose, monitor, and test virtually any gas, vapor phase, or aerosolized chemical of concern from a human health standpoint. BNL has conducted a variety of studies at the ITF on behalf of several governmental sponsors in the past, including the Department of Energy (DOE), the Department of Defense (DOD), and the National Institutes of Health/National Institute of Environmental Health Sciences (NIEHS) and the National Toxicology Program (NTP). The ITF/VSC performs toxicological research and testing of chemicals of interest to a variety of sponsors, including both regulatory and non-regulatory governmental agencies, civilian agencies, DOD, as well as sponsors from the private sector.

FACILITY DESCRIPTION

A complete description of existing and newly renovated areas of the ITF and supporting VSC are provided in this section. Supporting floor diagrams, flow charts, and equipment utilization plans also are included in order to better present this detailed information for review.

Floor Plans

A detailed floor plan of the ITF and VSC are shown in Figure 2. The 6,500-ft^2 ITF building is located at the eastern corner of the Medical Department and has a 1,800-ft^2 basement. The basement is devoted entirely to mechanical equipment to support the facility. New heating, ventilation, and air handling/air conditioning (HVAC) systems ensure that clean air flows out of the clean corridor at all access points (Figure 3). The ITF is divided into three areas: (1) access corridors (clean and dirty), (2) regulated inhalation exposure areas, and (3) the actual chambers and other equipment designed for containment. Two barriers are created between the access corridors and potential airborne hazardous chemicals in the chambers by maintaining the air pressure in the chambers at 5 mm water less than the pressure in the inhalation exposure area, which is in turn maintained about 5 mm less than the pressure in the clean access corridor.

The general access (non-regulated) area contains two offices, two laboratories, an electron microscopy suite, a glassware washing room, and several closets all opening from the "dirty" corridor. Two windows allow observation of the regulated exposure laboratories from the non-regulated access corridor.

The regulated inhalation exposure area consists of the clean corridor; two large inhalation exposure laboratories (A and B); generation, preparation, and analytical laboratories; cage and rack washing areas (air lock "A" and "B"); and two general purpose exposure rooms for acute and subchronic inhalation studies. The regulated area is separated easily into two distinct areas by locking a normally closed door in the hallway connecting the two chamber rooms. A hazardous material preparation laboratory, cage washing facilities, rack wash-down facilities, and the necropsy room open into Chamber Room A. An analytical laboratory, the necropsy room, and the general purpose exposure rooms open into Chamber Room B. The corridor outside these exposure rooms connects directly to the VSC clean corridors of the Medical Department. The installation of two air locks in the rack wash-down areas serving Chamber Rooms A and B ensures directional air flow and equipment flow toward the dirty corridor, but also prevents any airborne test material from escaping into the dirty corridor. As a standard operating procedure, all personnel must shower prior to entering the clean corridor and must shower out before exiting the laboratory into the dirty corridor using the two separate locker facilities connected to the air locks adjacent to Chamber Rooms A and B.

Heating, Ventilation and Air Conditioning

The non-regulated area, including offices, has a single pass air-handling system (Figure 4) with a common supply and local exhaust ventilation. The 5000 CFM air supply system is located on the roof with outside air being filtered, chilled (for dehumidification), heated, humidified, and HEPA-filtered before entering the ductwork. It is reheated in the ducts before entering each room, as required. The non-regulated area is maintained at a slight pressure positive to ambient pressure.

Four separate HVAC systems supply air to the building (Figure 4A, AC-1A, AC-1B, AC-2A, AC-2B). The main exhaust is through three laboratory hoods, each with its own separate HEPA filter and exhaust blower on the roof (Figures 4A and 4B). In addition, there is local exhaust ventilation provided in the laboratories, the showers, and in the chemical storage cabinets by individual blowers on the roof.

The air supply systems for the chambers (AC1, Figure 4A) and for the inhalation exposure rooms (AC2, Figure 4A) are located in the basement. Supply for the inhalation exposure room air is filtered through prefilters, HEPA filters, and charcoal filters. The 10,000 CFM supply is completely redundant up to the point where air enters the common air supply duct that goes up to the inhalation exposure rooms. The exhaust air from these rooms is removed through hoods and exhaust louvers in rooms (Figure 3). The air then enters a common exhaust system (E2) and is passed through HEPA filters and charcoal filters located on the roof prior to entering the environment.

The HVAC system for the inhalation chambers and glove boxes (AC1-E1, Figure 4A) is also redundant with supply and exhaust fans interlocked. Chamber exhaust air is cleaned immediately after exiting the chambers before entering the exhaust duct. This system also has backup filtration with both HEPA and charcoal filters on the roof in the event of an accident or a malfunctioning chamber filter. The supply air system (AC4, Figure 4B) for the clean corridor and the ITF is located on the roof. This system also provides air supply for the animal quarantine area and bedding storage. The supply systems and exhaust systems are completely redundant.

Electrical

Emergency power is available to all the HVAC systems associated with the inhalation exposure area and the chambers and glove boxes. Each room in the building has a single light fixture and an outlet (marked by a red cover) on emergency power.

Fire Protection

The building has sprinklers located throughout, and heat sensors are located in the supply air ducts. The building is constructed of foam sandwich panels on exposed steel (noncombustible universal building code type II-N). Ceilings in the inhalation exposure area are made of sealed plasterboard supported by a steel grid. Ceilings in the non-regulated area are made of fiber tile. Interior partitions are constructed of gypsum board on metal studs in the non-regulated area and of concrete block painted with epoxy coating in the inhalation exposure area.

Physical Design Features

Containment

The primary containment system in this facility consists of Chamber Rooms A and B, the analytical laboratory, the hazardous material preparation laboratory, the necropsy room, inhalation chambers, and glove boxes serviced by the AC1-E1 HVAC systems.

The HVAC system for the chambers and glove boxes is redundant with supply and exhaust fans interlocked. Filtered, conditioned air supplied at +8 cm water pressure (wg) is available in two ducts along the wall in each chamber room. Chamber air is exhausted through one of four high pressure (-30 cm wg) continuous welded stainless steel ducts, also located on the walls of the chamber rooms. Chamber air is cleaned immediately after exiting the chambers before entering the exhaust ducts. The air-cleaning devices are selected for the specific chemical under study. Each filter holder has provision for a HEPA filter and an appropriate vapor absorber. Vapors or gases are absorbed on charcoal, which is changed when necessary. Aerosols can be removed on HEPA filters. If other prefiltering devices are required, they will be installed in the chamber and the common exhaust system. This system also has backup filtration with both HEPA and charcoal filters on the roof in the event of an accident or a malfunctioning chamber filter. In the event of a failure of one system, the backup system will come on automatically within 5 sec, and an alarm will be sounded. The start-up also is accomplished with time delay relays to establish exhaust negative pressure first, in order to prevent the inhalation chambers from going positive with respect to the chamber room. All generation equipment is configured with "normally closed" solenoid valves that immediately terminate the flow of test chemical into the chambers in the event of power failure.

Under normal operations, no toxic or potentially carcinogenic chemicals will be airborne outside the primary containment system. All chemical standardization, generator loading, sample preparation, and chemical storage will take place inside the primary containment system. Whenever a chemical is handled in the secondary containment space or in the non-regulated area, it will be doubly contained in nonbreakable containers.

The secondary containment is the regulated area which is serviced by air supply and exhaust systems AC2-E2. Air for the regulated area is filtered through prefilters, HEPA filters, and charcoal filters. The 10,000 CFM supply is completely redundant up to the point where air enters a common air supply duct that goes up to the regulated area. The exhaust air from the regulated area is removed through hoods located in the Test Material Preparation Room (Figure 3) and the two air-lock exhaust systems serving chamber rooms A and B (Figure 3). The air then enters a common exhaust system (E2) and is passed through HEPA filters and charcoal filters located on the roof prior to entering the environment. The exhaust air-cleaning system consists of six sets of 2000 CFM HEPA-plus charcoal filters. Each set of filters can be isolated from its exhaust fan with manually controlled dampers for filter change-out. The exhaust fans, also in duplicate, are interlocked to the supply fans to form two independent systems. If any part of one system fails, the entire system shuts down and the backup system starts up independently of the other system. Start-up is arranged with time delay relays such that the exhaust system comes on before the supply system to prevent the regulated space from becoming positive. When a failure occurs, alarms are triggered to indicate that the primary system failed and that the backup system is now in operation An analog pressure sensor detects the differential air pressure and controls the supply air to maintain the required pressure gradient.

If the gradient falls below a set point, alarms (visual and audible) are activated. Alarms also are sounded when the differential pressure across the exhaust filters indicates that they should be changed, when the supply or exhaust damper is at a minimum, and when the supply duct air pressure falls below a set point. All alarms indicating a malfunction are sent to the local police headquarters which relays the alarm to the HVAC watch unit staffed 24 h/day.

Inhalation Chamber Operation

Conventional inhalation exposure chambers are usually square in cross section and have pyramidal tops and bottoms (Figure 4A). Air is introduced at the top and removed at the bottom. Windows are provided and one side usually consists of a door that allows animals to be introduced into or removed from the chamber. Such systems can be safely used to expose animals to chemical vapors (and to some aerosol particles) when certain precautions are taken. At 15 air changes/hour (the flow rate normally used), the chamber concentration will reach 1% of its equilibrium concentration less than 20 minutes after generators of test chemical vapors or gases are shut down. A possible problem is the potential for the animals to exhale some of the vapors or gases of test chemical that had previously been inhaled. This outgassing is investigated on a chemical by chemical basis.

These conventional systems are being used to expose animals to test chemical vapors such DMES (dimethylethoxysilane), styrene oxide, etc. Specific operational protocols are developed for each test chemical.

Exposure of rodents to potentially toxic test chemicals in the form of particulate aerosols is a more complex problem. Even after the concentration is no longer detectable by air sampling, the internal chamber surfaces including the cages and the coats of the animals themselves may be contaminated by the test chemical. If not properly controlled, handling the animals and cages could present an opportunity for skin contamination or the inhalation by technicians of any particles reentrained into the air. For this reason, inhalation exposures to hazardous particulate test chemicals must be done using nose-only exposure chambers. Such systems enable investigators to expose animals to potentially hazardous test chemical aerosols while constantly maintaining a physical barrier between the exposed animal and the personnel operating the system, and virtually eliminates test chemical contamination of the coats of the animal. This system is available in the ITF, as is shown schematically in Figure 4B. Exposures are carried out in a chamber that can expose up to 50 rodents via the nose only. Aerosol generators are located above the chamber, and the atmospheric clean-up takes place in filters or activated charcoal purifiers located in the exhaust stream. All exhaust air is filtered once more through absolute filters for secondary air purification. Chamber exhaust air cleaning may include electrostatic precipitation or air scrubbing as the primary clean-up procedure backed up by HEPA filtration.

Monitoring for test chemicals will be specific for the chemical in question. Many experiments will be conducted using four chambers; that is, three-dose level study plus a control. In the case of a five-dose level, four-dose level, or three-dose level study, one air analyzer, specific for the chemical in question, is used to sample the chambers consecutively. In addition, the system samples the chamber room and the common exhaust duct to verify that the room air is clean and that the air cleaning devices are functioning properly. Detectors will be chosen for their applicability to the chemical in question. Highly sensitive, dedicated instruments are available for many agents and will be used where applicable. Both gas chromatography and infrared spectroscopy are available with multiple valve switching for detecting vapors, and both techniques are capable of detecting more than one compound at a time. These instruments are calibrated with calibration gas mixtures. In addition, wet chemical methods may be used to verify the calibration.

Real-time particle analyzers incorporating optical scattering sensors are used where applicable to measure the concentration and particle size of aerosols in the chambers. Light-scattering devices can be used where deposition on internal surfaces does not interfere with calibration set points. Cascade impactor samples are taken to assess particle size at multiple points within the chamber.

By the use of such rigidly controlled test rodent exposure systems to provide much more reliable data, and computer-assisted techniques for physiologically-based toxicokinetic analyses of the data, valid determinations of allowable workplace air concentrations for specific chemicals can be obtained. As an example, the question was addressed of whether the change in workshift schedule from a standard 5-day, 8-hour/day work week to a different schedule involving longer shifts would affect the cumulative body burden of carbon tetrachloride in exposed workers. A physiological model for inhalation uptake of carbon tetrachloride in the rat was developed and then used to predict the kinetics of uptake in humans. It was found that the accumulation and removal of carbon tetrachloride in humans, based upon scaling up by the model using known biochemical and physiological parameters, followed a much slower time constant to cause greater retention of this airborne chemical at the end of the work week. This revealed that altered buildup patterns would occur with the longer work shift, requiring a longer between-shift recovery period. These techniques have also been expanded into more reliable risk estimate procedures for exposure of the general population to potentially carcinogenic volatile organics, such as methylene chloride.

ENERGY AND ENVIRONMENT PART 4: ENVIRONMENTAL TECHNOLOGIES

ACTIVE HYDRAZINE VAPOR SAMPLER (AHVS)

Rebecca C. Young
NASA DL-ESS-24
John F. Kennedy Space Center, Florida 32899

Charles F. McBrearty and Daniel J. Curran
I-NET, Inc.
John F. Kennedy Space Center, Florida 32899

ABSTRACT

The Active Hydrazine Vapor Sampler (AHVS) was developed to detect vapors of hydrazine (HZ) and monomethylhydrazine (MMH) in air at parts-per-billion (ppb) concentration levels. The sampler consists of a commercial personal pump that draws ambient air through paper tape treated with vanillin (4-hydroxy-3-methoxybenzaldehyde). The paper tape is sandwiched in a thin cardboard housing inserted in one of the two specially designed holders to facilitate sampling. Contaminated air reacts with vanillin to develop a yellow color. The density of the color is proportional to the concentration of HZ or MMH. The AHVS can detect 10 ppb in less than 5 minutes. The sampler is easy to use, low cost, and intrinsically safe and contains no toxic material. It is most beneficial for use in locations with no laboratory capabilities for instrumentation calibration. This paper reviews the development, laboratory test, and field test of the device.

INTRODUCTION

Hydrazine and monomethylhydrazine are widely used in space programs as rocket propellants. HZ is used in the Emergency Power Unit of the United States Air Force F-16 fighter planes. Commercially, HZ is used in applications such as a polymerization catalyst, boiler feedwater oxygen scavenger, blowing agent, and photographic developer. Hydrogen compounds are highly toxic and suspected carcinogens. In 1989, the American Conference of Governmental Industrial Hygienists (ACGIH) [1] proposed to reduce the HZ and MMH Threshold Limit Value (TLV) from 100 ppb and 200 ppb respectively to 10 ppb to protect personnel working with these substances. This reduction will significantly impact personnel safety monitoring because a near realtime, easy-to-use, commercial detector for measuring such a low level was not available. In response to the ACGIH proposal, the NASA Instrumentation Section at the John F. Kennedy Space Center (KSC) initiated an effort to develop the needed instrument. Contracts were awarded to three vendors for the development of electrochemical, ion-mobility, and paper tape technologies for a portable vapor detector. At the same time, the NASA Toxic Vapor Detection Laboratory (TVDL) initiated an in-house development for an AHVS for interim use. This paper reviews the development effort and provides the laboratory and field test results.

BACKGROUND

In 1991, techniques and prototype samplers capable of detection of 10-ppb HZ and MMH in air were developed [2]. The samplers were based on the use of a commercially available, intrinsically safe personal pump drawing ambient air through paper treated with a mixture of vanillin and phosphoric acid. Special holders were designed to facilitate sampling through the paper tape. The detection and quantification of this low-ppb concentration are based on the development of yellow color on the paper upon exposure to the HZ/MMH vapors. After laboratory development testing, two designs of the prototype were field tested and evaluated by KSC Environmental Health and Safety personnel. While the users found the detection capability of the prototype generally acceptable, they requested modifications that would minimize possible contamination and degradation of the chemically treated paper, allow easier use and documentation of results, and provide improved viewing of the color development during sampling. To achieve these requirements, new sample holders were designed and tested both in the laboratory and in the field at KSC, White Sands Test Facility, and Hill Air Force Base.

SAMPLER DESIGN

The AHVS shown in figure 1 consists of three parts:

(1) A commercial, intrinsically safe personal pump. Its flowrate is preset at 1 or 2 liters per minute.

(2) A card holder. The TVDL designed two card holders. The open-face design is for monitoring general areas, whereas the closed-face design with a viewing window is for sampling through a small opening for leak detection.

(3) An HZ/MMH card. The card is made of a strip of paper tape coated with vanillin that develops a yellow color upon contact with HZ or MMH. The chemistry is shown in figure 2. The intensity of the yellow color is proportional to the concentration of HZ or MMH. The paper tape is sandwiched in a thin cardboard housing with two 1.5-centimeter-diameter windows in the front and back. The windows are designed to align exactly with the air passage in the card holder. The card is enveloped in a special packet to ensure the integrity of vanillin chemistry until the card is ready for use. The NASA vanillin hydrazine card was obtained from GMD Systems Inc.

OPERATION

Three steps are followed to use the device: (1) attach the appropriate sample holder to the pump, (2) insert and clamp a card in the sample holder, and (3) turn on the pump to take the sample. For a 10-ppb concentration, the recommended sample time is 5 minutes at a 1-liter-per-minute sampling rate. Higher concentrations require less sample time. The color, as it is developing, shows on the front window of the card. After sampling, the card is removed from the holder, and the color is compared with a calibrated concentration estimator, which is a wheel consisting of five shades of yellow corresponding to five HZ/MMH concentrations. Using the color wheel, the approximate HZ/MMH concentration can be determined in the field. The color developed can also be measured in terms of a chroma reading using a Minolta Chroma Meter. Accurate concentration is determined by comparing the chroma reading to an HZ/MMH chroma calibration curve. The color wheel and chroma calibration chart are shown in figure 3. The color wheel was obtained from GMD Systems Inc.

LABORATORY TEST

Vapor Generation and Validation Equipment

The TVDL precision vapor generation system was designed to deliver known concentrations of HZ and MMH at controlled conditions of temperature (T) and relative humidity (RH) (figure 4). The system consists of four components: (1) a Kin-Tek Span Pac Model 361 precision vapor generator, (2) a Miller-Nelson Model HCS-301 flow/T/RH controller, (3) a sample vessel, and (4) a T/RH monitor.

The Kin-Tek vapor generator consists of three permeation devices housed in three temperature-controlled ovens. The permeation rate of the device is determined by the temperature of the oven, the length of the polymeric tube, and the polymeric material used. By first flowing small amounts of nitrogen through the permeation device and then diluting the hydrazine/nitrogen mixture with "conditioned" air from the Miller-Nelson unit, precise concentrations of HZ and MMH vapors are generated for use as standards. The TVDL uses a coulometric procedure for the validation of the standard vapor concentrations.

The coulometric procedure is simple and accurate. The hydrazine vapor is first collected in an impinger containing a 0.1-molar sulfuric acid absorbing solution. Following the vapor absorption, the amount of hydrazine in the solution is analyzed by constant-current coulometric titration. The procedure calls for dissolving a small amount of potassium bromide crystal in the absorbing solution. A direct electric current passing through a solution electrolyzes potassium bromide to form bromine, which rapidly reacts with hydrazine present in the solution. As long as hydrazine is present, the bromine concentration is undetected. At the moment all hydrazine has reacted, the

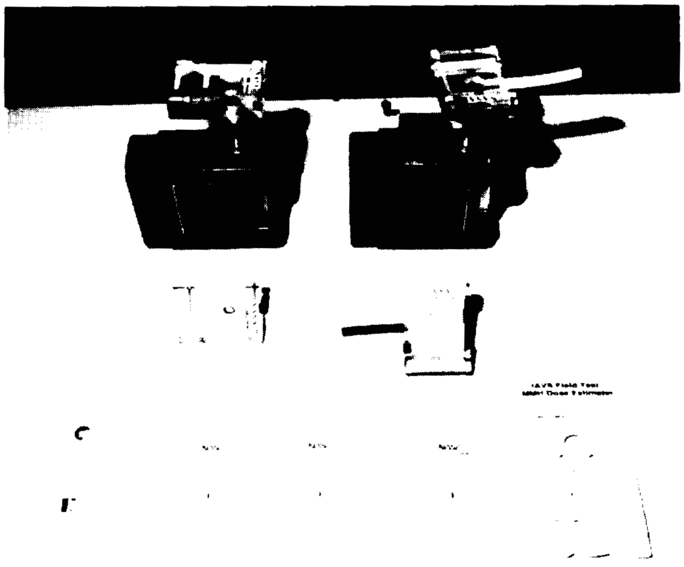


Figure 1. Active Hydrazine Vapor Sampler

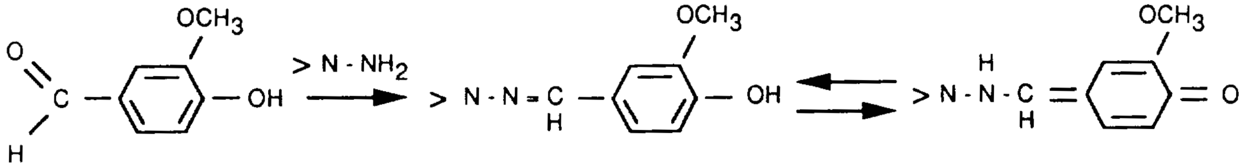


Figure 2. Reaction of Vanillin and Hydrazine

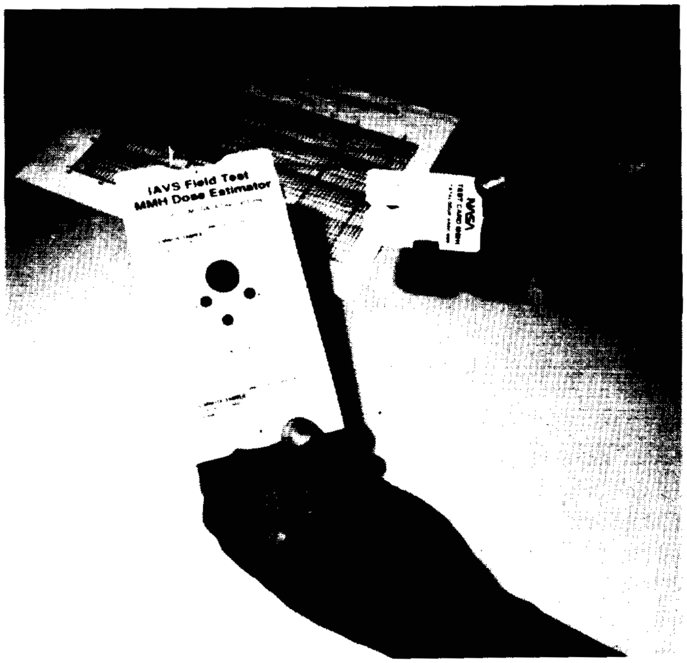


Figure 3. Color Wheel and Chroma Calibration Chart

bromine concentration increases to a detectable level, which signifies the end of titration. The length of titration time is determined by the amount of hydrazine. This titration result is used for subsequent calculation of hydrazine vapor concentration [3].

Color Measurement Equipment

For laboratory characterization of the hydrazine vapor sampler, a Minolta Chroma Meter Model CR-200 was used (figure 5). The Chroma Meter Luminosity Chroma Hue Angle (LCH°) color system uses cylindrical coordinates to measure color. For measuring the yellow color developed by the reaction of hydrazine and vanillin, the chroma variable was used because the luminosity and hue angle are fairly constant and the chroma readings are proportional to the vapor concentrations. In the lab, calibration curves are established using chroma readings and hydrazine vapor standards. Using the lab calibration curve and chroma readings corresponding to the five colors on the color wheel, concentration charts on the color wheel are established for field use.

Lab Test Parameters

The following test parameters and procedures were used:

(1) Precision and Linearity: The samplers were exposed to MMH concentrations of 9.2, 38.6, 296, and 950 ppb for four iterations at a standard laboratory vapor condition of 25 degrees Celsius and 45 percent relative humidity.

(2) Comparison of Open-Face and Closed-Face Samplers: Both sampler designs were subjected to the same precision and linearity tests and the results were compared.

(3) HZ Versus MMH Test: With all other test conditions held constant, the sampler was tested with an HZ or MMH vapor of comparable concentrations.

(4) Temperature/Relative Humidity Effects and Response Time: The sampler was exposed to a 10-ppb MMH vapor of a combination of temperature and RH conditions (0 to 84 percent RH and 5 to 40 degrees Celsius) for 3 minutes. Color measurements were taken every minute at 25 degrees Celsius and 45 percent RH until the readings were stable.

(5) Interference: Sunlight, ammonia, nitrogen dioxide, Freon, methyl ethyl ketone (MEK), and isopropyl alcohol (IPA) were tested for positive or negative color development.

(6) Shelf Life: Two batches of HZ/MMH cards were stored in the refrigerator and ambient laboratory storage area respectively for up to 42 days. During this period, cards were drawn from the batches and tested for integrity.

LABORATORY TEST RESULTS AND COMMENTS

The following laboratory test results were obtained:

(1) Precision and Linearity: Two data sets were obtained. For the lower concentrations, a 5-minute sample time was used; whereas for the higher concentrations, the sample time was decreased to 1 minute. For both sets, the sample rate was 2 liters per minute, and chroma readings were taken 1 minute after exposure. Tables 1 and 2 show the chroma readings of the respective sets of data. Figure 6 shows the linearity plot.

(2) Comparison of Open-Face and Closed-Face Samplers: Table 1 shows data for both sampler designs subjected to the same test conditions. The results showed no significant differences between the two designs.

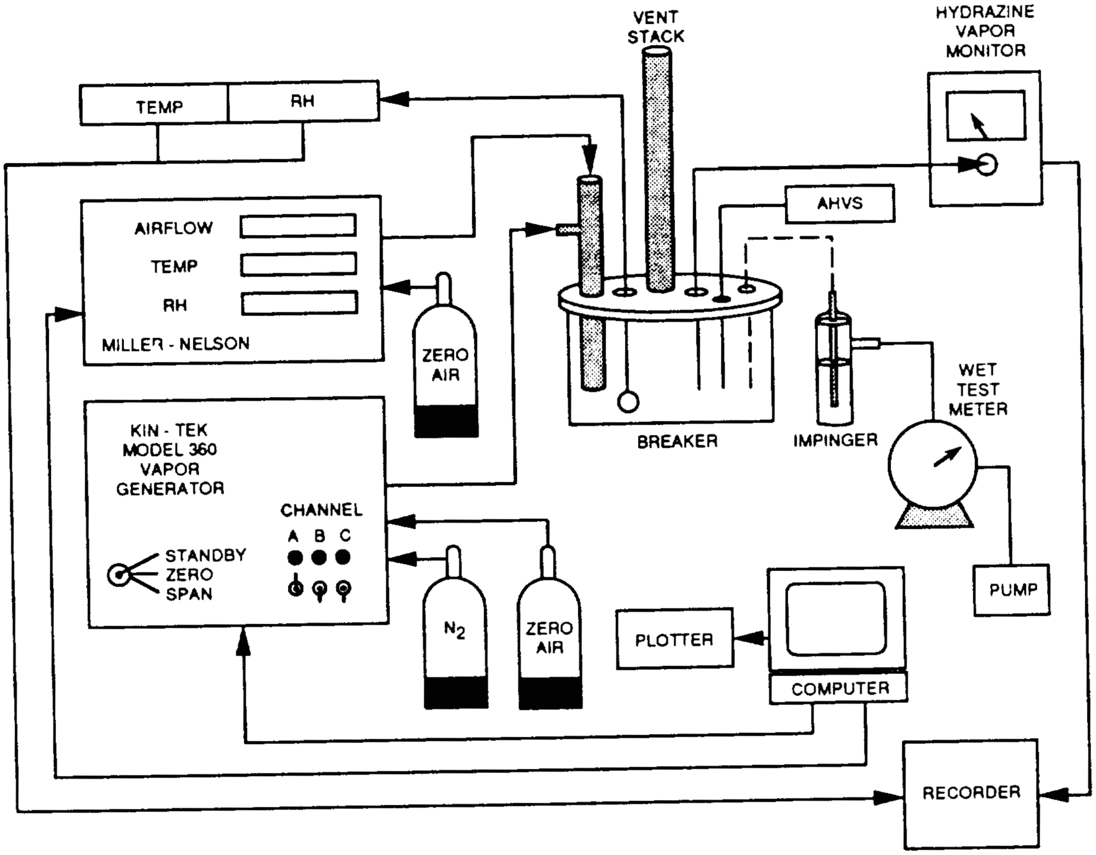


Figure 4. Hydrazine Vapor Generation System

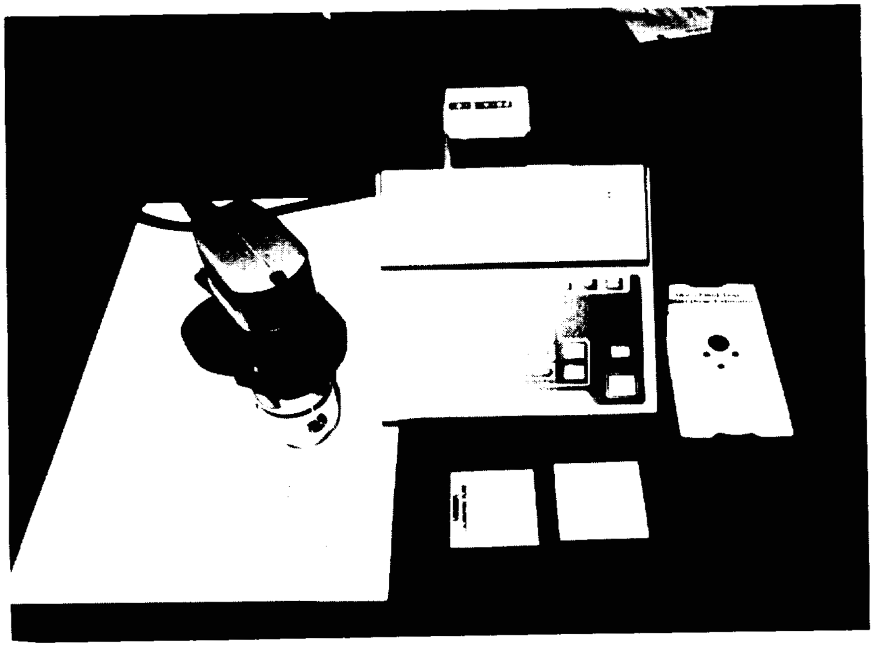


Figure 5. Minolta Chroma Meter

Table 1. Precision and Linearity Test (Five-Minute Sample)

Run	Chroma for 9.2 ppb	Chroma for 38.6 ppb	Chroma for 296 ppb
Open Face			
1	22.60	43.52	69.03
2	23.90	42.96	69.12
3	22.23	42.60	68.16
4	22.01	42.91	68.49
Mean	22.69	43.00	68.70
Standard deviation	0.85	0.38	0.45
Closed Face			
1	22.63	42.23	67.44
2	23.02	43.82	67.82
3	23.95	42.49	67.51
4	23.10	43.84	67.67
Mean	23.18	43.10	67.61
Standard deviation	0.56	0.86	0.17

Table 2. Precision and Linearity Test (One-Minute Sample)

Run	Chroma for 38.9 ppb	Chroma for 95 ppb	Chroma for 296 ppb	Chroma for 950 ppb
1	20.14	33.70	47.40	68.84
2	20.20	32.72	48.13	67.52
3	20.46	33.07	49.79	70.10
4	20.69	33.55	48.45	70.32
Mean	20.37	33.26	48.44	69.20
Standard deviation	0.25	0.45	1.00	1.29

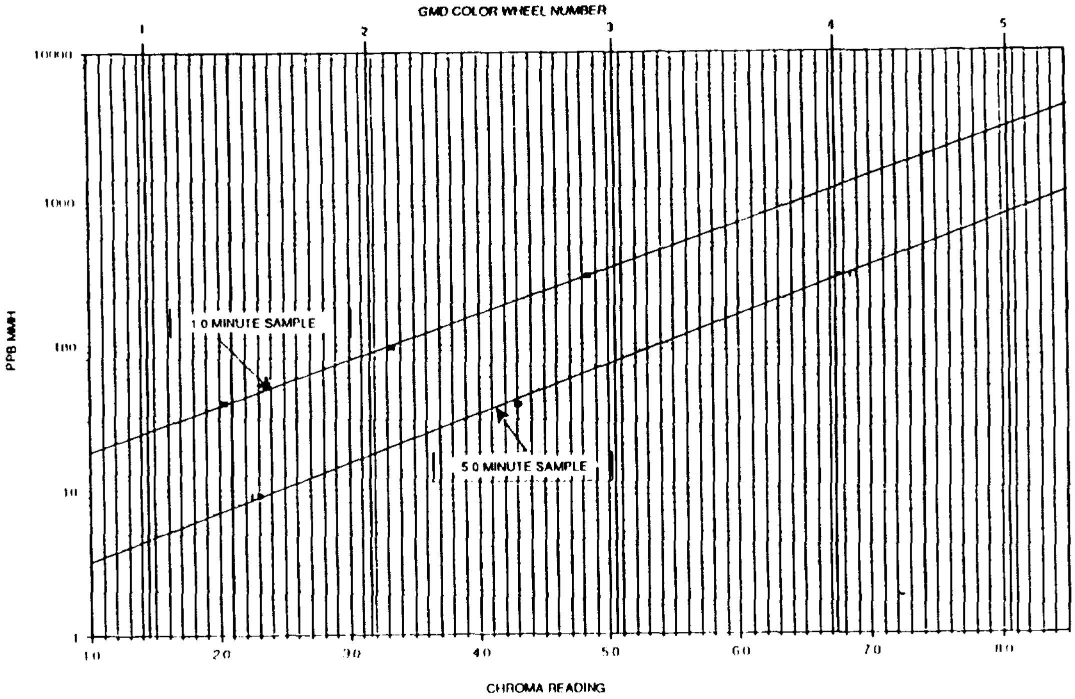


Figure 6. Linearity Plot

Table 3. HZ Versus MMH Test

MMH Run	Chroma for 9.2 ppb	Chroma for 38.6 ppb	Chroma for 296 ppb
1	15.55	33.93	66.82
2	16.09	34.65	65.36
3	15.67	33.63	68.23
4	16.06	33.78	64.91
5	14.81		66.52
Mean	15.64	34.00	66.37
Standard deviation	0.52	0.45	1.31
HZ Run	Chroma for 9.1 ppb	Chroma for 42.0 ppb	Chroma for 338 ppb
1	20.97	37.77	55.92
2	20.36	39.10	56.41
3	20.98	38.78	58.28
4	20.53	38.69	58.07
5	20.79	37.42	57.53
Mean	20.71	38.35	57.24
Standard deviation	0.31	0.72	1.04

(3) HZ Versus MMH Test: With all other test conditions held constant, the sampler tested with HZ or MMH vapor of comparable concentrations indicated a difference in results. Data are shown in table 3 and plotted in figure 7.

(4) Temperature/Relative Humidity Effects and Response Time: Chroma readings after exposure of each card to a T/RH condition were plotted (figure 8). The results indicated a minor temperature effect above freezing point and a minor RH effect. The color development was much slower at absolutely dry conditions. Color is developed as the vanillin-coated paper picks up moisture from the air [2]. At a 10-ppb level, a difference of two chroma units represents 1.5 ppb.

(5) Interference: No color is developed due to sunlight, ammonia, nitrogen dioxide, and Freon [4]. Table 4 shows MEK and IPA interference test results. The sample time was 5 minutes and the sample rate was 1 liter/minute. These tests indicated MEK and IPA do not interfere with MMH and HZ color development significantly.

(6) Shelf Life: Chroma readings obtained from the test are shown in table 5 and plotted in figure 9. There appears to be a small degradation of the chemical. Although slight degradation is detected by the Chroma Meter, the degradation probably will not be detected by the color wheel in the field. A study performed by the Naval Research Laboratory [5] indicated vanillin cards that passed the expiration date of November 1989 for approximately two years read an average of 20 percent lower compared with a fresh card.

FIELD TEST

The open-face and close-face AHVS was field tested by the industrial hygienists at KSC, White Sands Test Facility, and Hill Air Force Base. Although Hill Air Force Base has not completed the test, both KSC and White Sands users have reported the high sensitivity of the devices has greatly benefited them in the detection of extremely low levels of HZ/MMH vapors in near realtime. For example, the AHVS proved invaluable in the location and elimination of a serious, low-level source of contamination that had evaded detection using standard available monitors. The contamination source was identified as a pump used for the analysis of nonvolatile residue content in liquid hydrazine, which was moved from the fume hood to the lab's bench area in order to provide more space in the hood. As hydrazine vapor emitted from the pump due to hydrazine desorption, it contaminated the laboratory work area. KSC industrial hygienists used a conventional portable hydrazine instrument and a detection tube in an attempt to find the source. Both of these devices failed; however, AHVS quickly determined the pump as the culprit. Additionally, other devices failed to identify area contamination due to the cumulation of hydrazine vapor in a floor drain and a pin hold in a hydrazine storage drum. Both were identified by AHVS.

The field test has also verified that the AHVS is rather specific for HZ/MMH detection. Color development due to interferant was reported only in one instance that occurred when sampling a lunch room. In the lunch room, the AHVS developed a pink color. The pink color was later confirmed in the lab as interference due to cigarette smoke. White Sands has reported that chlorine gas does not interfere with the vanillin chemistry. This was found during a hyrazine spill cleanup where the decontamination process called for the use of Clorox.

Over all, the user's evaluations of AHVS are very positive. More units were requested by users from other NASA centers and Air Force organizations.

CONCLUSION

The Active Hydrazine Vapor Sampler is an important development because it is the only known device to date that can accurately measure 10-ppb HZ/MMH in less than 5 minutes. In addition, the device is lightweight, extremely easy to use, relatively inexpensive (less than $1,000 per unit), and contains no hazardous materials. Since the device requires no chemical calibration, it would be specially useful for organizations that have no chemical laboratory to support instrument calibration. The device is also versatile; in addition to using it as an area monitor or a leak detector, it can be easily adapted for use as a breathing zone monitor. Furthermore, the device is not

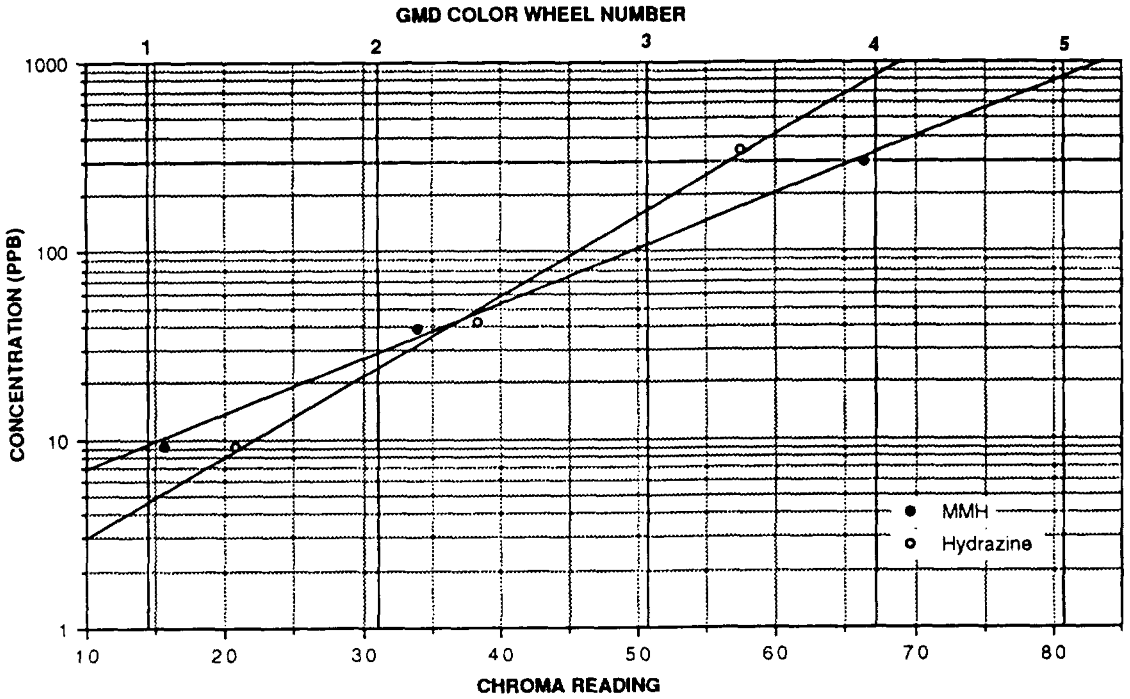


Figure 7. GMD Color Wheel Number

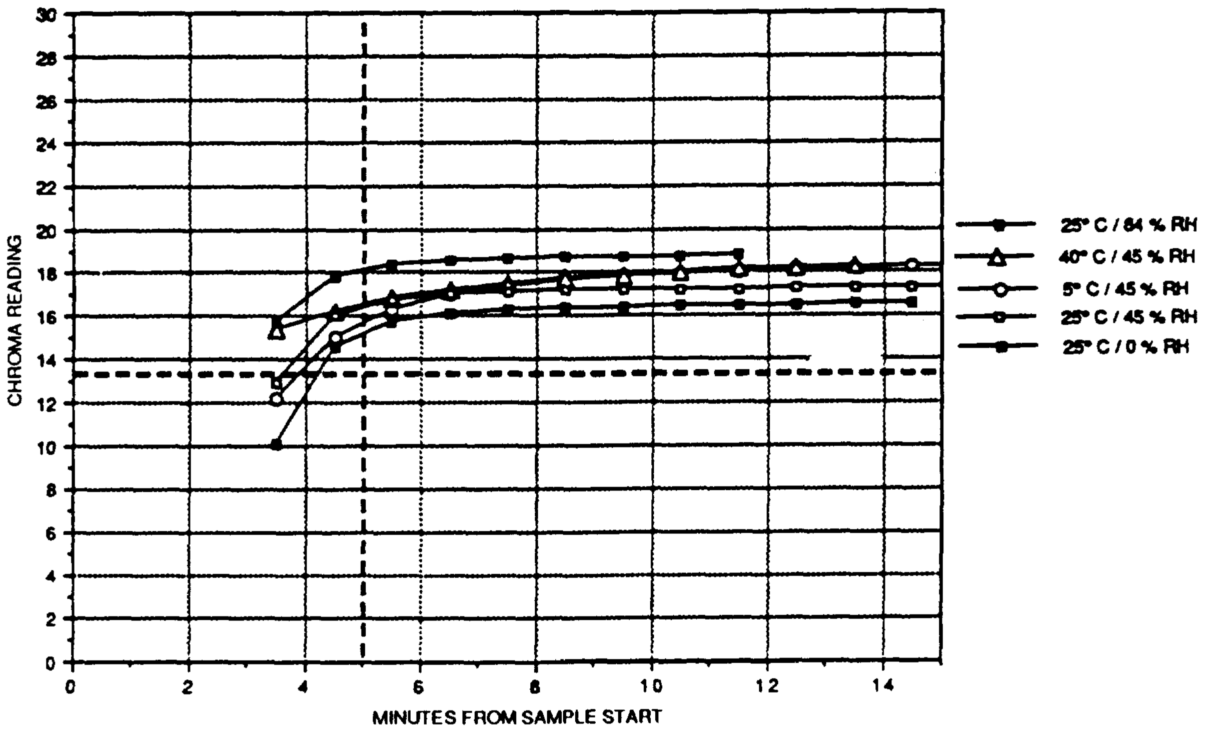


Figure 8. Response Time at Varying Temperature and Relative Humidity (10-ppb MMH, 3-Minute Sample Time, 2 Liters/Minute Sample Rate)

Table 4. Interference Test With Methyl Ethyl Ketone and Isopropyl Alcohol

Test Vapor	Concentration (ppb)	Interference Vapor	Concentration (ppm)	Chroma Reading
Hydrazine	24	-	-	16.49
Hydrazine	24	IPA	5,000	18.02
Hydrazine	15	MEK	4,000	13.35
Hydrazine	15	-	-	11.70
-	-	MEK	4,000	4.63
MMH	14	-	-	14.84
MMH	14	IPA	5,000	13.29
MMH	i4	MEK	4,000	19.55

Table 5. Shelf Life With Cards Stored in Refrigerator and Ambient

Day	Refrigerator Chroma Reading	Ambient Chroma Reading
8	50.39 50.64 52.01	50.34 51.19 51.49
Mean	51.01	51.01
15	51.26 52.04 50.97	53.50 51.89 52.58
Mean	51.42	52.66
29	49.98 49.73 50.91	51.48 49.98 51.24
Mean	50.21	50.90
42	48.06 48.36 49.48	48.43 49.00 48.72
Mean	48.63	48.72
Standard deviation	1.27	1.57

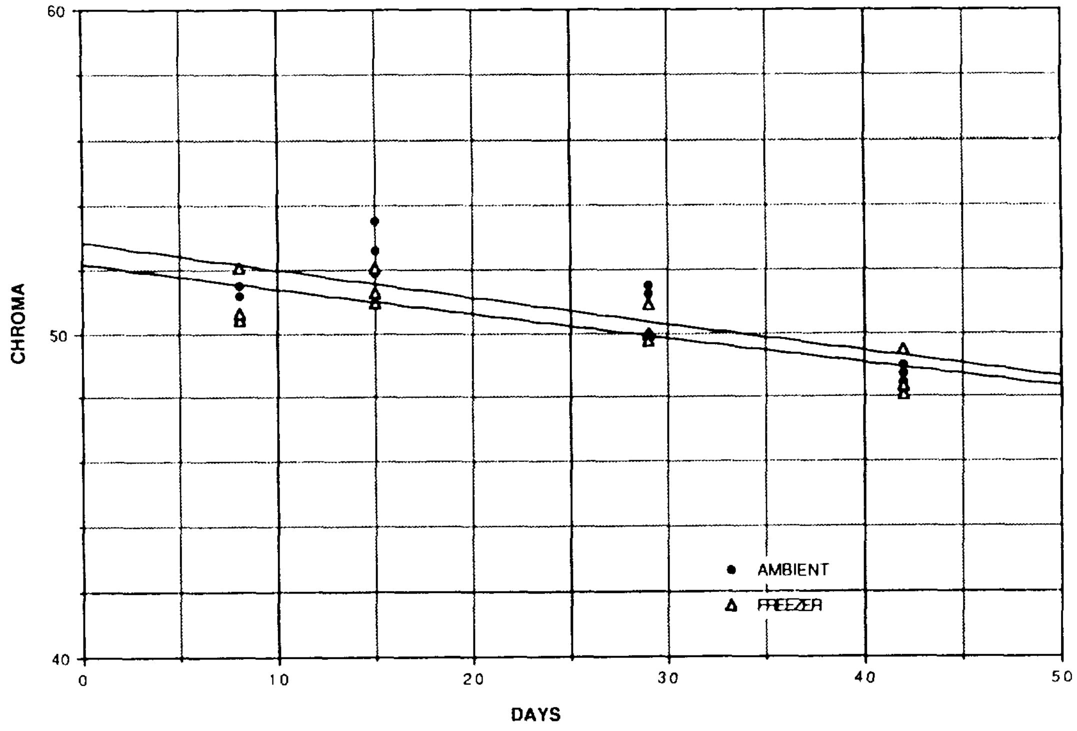


Figure 9. Shelf Life

limited to the measurement of HZ or MMH. Its application can be extended to measure numerous other toxic vapors by a simple change of the tape chemistry. The device has potential to significantly improve the response time for early detection of other toxic vapors and, thus, enhance protection of all personnel working with toxic substances.

REFERENCES

1. Threshold Limit Values and Biological Exposure Indices for 1989-1990, American Conference of Governmental Industrial Hygienists, page 44.

2. Blaies, D.M. et al., "Development of an Interim Active Vanillin Sampler Used To Detect 10 Parts Per Billion of Hydrazine and Monomethylhydrazine in Air," JANNAF Conference, Kennedy Space Center, Florida, August 1991.

3. Wyatt, Jeffrey, et al., "Coulometric Method for Quantification of Low Level Concentrations of Hydrazine and Monomethylhydrazine," American Industrial Hygiene Association Journal.

4. Crossman, Karen P., et al., "Laboratory Evaluation of a Colorimetric Hydrazine Dosimeter, "Naval Research Laboratory, Washington, D.C., Memorandum Report 6668.

5. Crossman, Karen P. and Rose-Pehrsson, Susan L., "Shelf Life Capacity of the Composite Colorimetric Hydrazine Dosimeter," Naval Research Laboratory, Washington, D.C. 20375-5000.

COP IMPROVEMENT OF REFRIGERATOR/FREEZERS, AIR-CONDITIONERS, AND HEAT PUMPS USING NONAZEOTROPIC REFRIGERANT MIXTURES

Douglas G. Westra
National Aeronautics and Space Administration
George C. Marshall Space Flight Center
Marshall Space Flight Center, AL 35812

ABSTRACT

With the February, 1992 announcement by President Bush to move the deadline for outlawing CFC (chloro-fluoro-carbon) refrigerants from the year 2000 to the year 1996, the refrigeration and air-conditioning industries have been accelerating their efforts to find alternative refrigerants. Many of the alternative refrigerants being evaluated require synthetic lubricants, are less efficient, and have toxicity problems. One option to developing new, alternative refrigerants is to combine existing non-CFC refrigerants to form a nonazeotropic mixture, with the concentration optimized for the given application so that system COP (Coefficient Of Performance) may be maintained or even improved. This paper will discuss the dilemma that industry is facing regarding CFC phase-out and the problems associated with CFC alternatives presently under development. A definition of nonazeotropic mixtures will be provided, and the characteristics and COP benefits of nonazeotropic refrigerant mixtures will be explained using thermodynamic principles. Limitations and disadvantages of nonazeotropic mixtures will be discussed, and example systems using such mixtures will be reviewed.

INTRODUCTION

Nearly 100 years have passed since the idea of using refrigerant mixtures was first proposed, however, the full potential of nonazeotropic mixtures in refrigeration systems is relatively unexplored [1]. Renewed interest in nonazeotropic refrigerant mixtures has developed in the last 15 to 20 years, but their use is limited mainly to the laboratory. With the sense of urgency being placed on the industry to come up with safe, efficient, and reliable alternatives to CFCs, an increased effort to develop their commercial potential is warranted.

The need to examine nonazeotropic refrigerant mixtures will be discussed as follows. First, the dilemma facing designers of refrigerator/freezers, air-conditioners, and heat pumps will be reviewed. Second, the uncertainties of the alternative refrigerants currently being developed to replace CFCs will be discussed. Third, the term "nonazeotropic" will be defined, and the characteristics and inherent benefits of nonazeotropic refrigerant mixtures will be described. The limitations and disadvantages of nonazeotropic refrigerant mixtures will also be identified. Next, research that has been conducted with nonazeotropic refrigerant mixtures, including a prototype heat pump developed for the Marshall Space Flight Center, will be explained. Finally, conclusions and recommendations will be discussed.

THE DILEMMA

Due to their ozone-depletion potential (ODP), CFC refrigerants R-11 and R-12 will be among the Group I refrigerants un-available beginning in 1996 [2,3]. Other CFC refrigerants in the accelerated ban include R-114, R-115, and R-500, but the remainder of this discussion will focus on the replacement of the two most common CFCs, R-11 and R-12.

The challenge of replacing R-11 and R-12 presents a great dilemma to manufacturers of home refrigerator/freezers since R-12 is a very popular working fluid for this application and R-11 is a common blowing agent used for refrigerator/freezer insulation [4]. In addition, home refrigerator/freezer manufacturers will be required to improve the efficiency of systems presently using these refrigerants by an average of 25% by 1993 (relative to 1990 standards) and possibly another 25% by 1998 [5]. The air-conditioning and commercial refrigeration industries also rely heavily on both R-11 and R-12 as working fluids [6]. The next section discusses the success of achieving these new goals with some existing alternative refrigerants.

ALTERNATIVE REFRIGERANTS

R-134a is becoming widely accepted as the replacement for R-12 in domestic refrigerator/freezer and automotive air-conditioning applications [7,8,9] and also as the replacement for medium pressure chillers which currently use either R-12 or R-500 [10]. R-123 is considered one of the leading candidates to replace R-11 in low pressure centrifugal chillers [10]. Unfortunately, these alternatives cannot simply be used as drop-ins for the CFCs they must replace; system enhancement, refurbishing, or component replacement is often necessary. Many challenges must be addressed before alternative refrigerants can be used on a full-scale basis.

First, R-134a is not compatible with the mineral oils commonly used for compressor lubrication [11]. There are a number of synthetic candidates being evaluated for use with R-134a, but none have been totally proven.

The refrigeration capacity and coefficient of performance (COP) of alternative refrigerants must also be established. Numerous investigations have been conducted to determine the capacity and performance of alternatives relative to their CFC counter-parts. A comparison of R-134a with R-12 in a residential heat pump [12] showed that approximately the same heating output was achieved with R-134a, but the COP of the system was approximately 15% less with R-134a than with R-12. Another series of tests [13] conducted at ARI (Air-Conditioning and Refrigeration Institute) Heat Pump Rating Conditions showed that R-134a exhibits a 6-11% increase in COP for moderate and warm rating conditions, while R-134a has a nearly identical COP to that of R-12 for a cold rating condition. In a test conducted for a household refrigerator/freezer [9], R-134a was shown to consume approximately 8% more power than R-12 and require more run-time, resulting in an energy consumption 45% greater than R-12. Tests conducted with R-123 indicate a 0 to 18% reduction in capacity compared to R-11 and a 0 to 15% reduction in COP [10]. These seemingly contradictory results indicate that more tests need to be conducted to establish the performance and capacity of alternative refrigerants. There is agreement, however, that systems need to be optimized for the replacement refrigerants to achieve adequate efficiency and capacity. This is due in part to different specific volumes at the compressor inlet, requiring compressor re-sizing.

The compatibility of alternative refrigerants with seals, bushings, gaskets, and other components is also still under investigation. R-123 may present some problems with elastomers made of Buta N or commonly used neoprenes, while R-134a has shown excellent material compatibility [10].

Toxicity is another major issue. The Southern Building Code Congress International (SBCCI) requires either ventilation, or a refrigerant detector and an alarm that sounds at a level of 10 parts per million (ppm) for R-123 [14,15]. There is less concern with the toxicity of R-134a, although it too requires either ventilation or detection because of its ability to displace oxygen [14,15]. Tests are presently being conducted to determine the carcinogenic effects of R-123, R-134a, and other refrigerants, with results expected by the end of 1992 [16].

The above discussion shows that much work has been done and is still being done to develop and test CFC pure-refrigerant alternatives. These alternatives have a lot of potential, but their total acceptance has not been established. Following is a discussion of another option, nonazeotropic refrigerant mixtures using existing non-CFC refrigerants.

NONAZEOTROPIC REFRIGERANT MIXTURES

Definition and Characteristics

The word "nonazeotropic" comes from the root word "zeotrope" with the double negative "non" and "a" as a prefix. Zeotrope is derived from the Greek words "zeo" (to boil) and "trope" (to change). Perhaps the easiest way to understand a nonazeotropic refrigerant mixture (NARM) is to examine an azeotropic mixture (normally labeled an "azeotrope"). An azeotrope, as illustrated in Figure 1, is made up of two or more refrigerants and occurs only at a particular composition. An azeotrope behaves as a pure refrigerant, undergoing no temperature change during condensation and evaporation (single negative "a" with "zeo", to boil, and "trope", to change; i.e., no change during boiling). The thermodynamic properties of an azeotrope are different than those of its two constituent fluids. Common azeotrope refrigerants are R-500 (73.8% R-12 and 26.2% R-152a) and R-502 (48.8% R-22 and 51.2% R-115).

A nonazeotropic refrigerant mixture is made up of two (or more) refrigerants of different volatility [17] and does not act as a pure fluid. When these two refrigerants are used in a vapor-compression cycle, the mixture changes composition as it boils or condenses. As a result of this change in composition, a temperature variation occurs during a constant pressure phase-change process. The magnitude of this temperature variation or "temperature glide" is a function of the properties and relative composition of the mixture constituents. The two-phase region with the bubble- and dew-point lines for a typical nonazeotropic refrigerant mixture is illustrated in Figure 2.

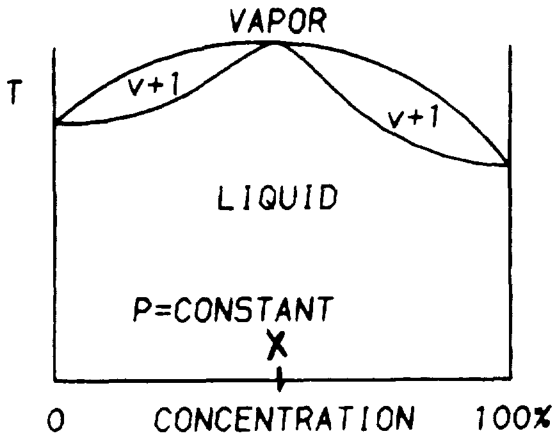


Figure 1. Two-phase Region for a Binary Mixture with an Azeotrope Occurring at Composition X.

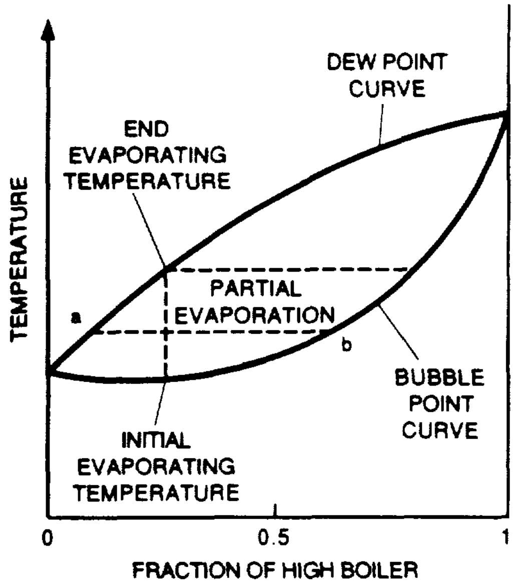


Figure 2. Two-phase Region for a Binary Mixture with no Azeotrope.

As liquid is vaporized during the evaporation process, the more volatile component (the refrigerant with the lower boiling point) will vaporize more rapidly than the less volatile component (high boiler). Therefore, the remaining liquid will be enriched with the less volatile component. From Figure 2, it is seen that as the liquid becomes more enriched with the less volatile component, the bubble point temperature increases. The bulk refrigerant temperature will rise as evaporation continues and will be at the temperature lying on the dew-point line at the completion of evaporation. Assuming that equilibrium exists at all points between the liquid and vapor phases, the concentration of the components in the two phases will be different. The points labeled (a) and (b) demonstrate this phenomena, marking the relative mass fractions in the two phases at some point along the length of the evaporator. The less volatile component is always more concentrated in the liquid phase than in the vapor phase.

COP Benefits

The potential benefit of a nonazeotropic refrigerant is illustrated in Figures 3 and 4. The arrows in Figure 4 indicate the flow direction in the heat exchanger. Also note that the same line nomenclature is used in Figure 4 as in Figure 3 (the solid lines represent pure refrigerant, etc.). It can be seen that using the nonazeotropic refrigerant mixture can reduce heat pump irreversibility and therefore, increase efficiency. The temperature of pure refrigerants is constant in the two-phase region. If the heat source or heat sink temperature varies throughout the heat exchanger, as is the case with a chilled water system or a hot-water heat pump, the irreversibilities associated with the heat transfer can be large. By utilizing the correct nonazeotropic mixture, the temperature difference between the refrigerant and the heat source/sink can be reduced, thereby minimizing the irreversibility produced during heat transfer.

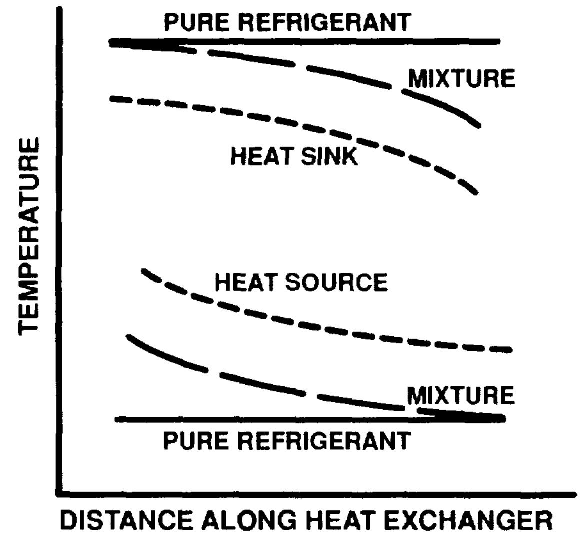


Figure 3. Temperature Variation Along Length of Condenser and Evaporator.

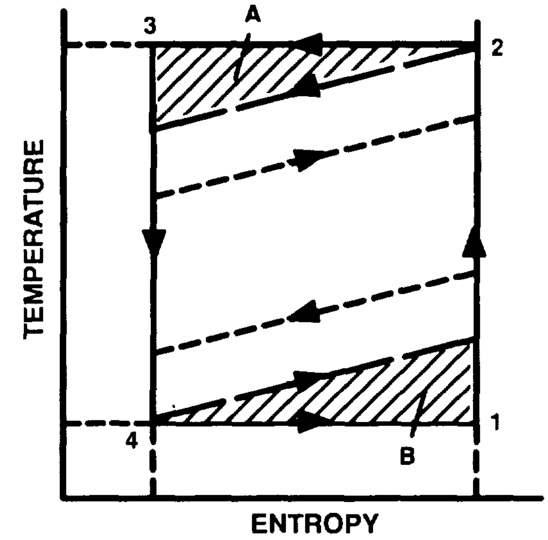


Figure 4. T-s Diagram of Counter-Flow Mixed Refrigerant Heat Transfer.

Assuming that the cycles for the pure refrigerant and the nonazeotropic mixture can perform the same duty, the ideal COP can be calculated for each cycle from the areas shown in Figure 4. The ideal (or carnot) COP of a heat pump using a pure refrigerant is simply the heat rejected by the condenser, Q_c, divided by the work of the compressor, W. For the pure refrigerant case, Q_c is the area below the condensation line, 2 to 3. The work W for the pure refrigerant is represented by the area enclosed by the points 1, 2, 3, and 4. Therefore, the ideal COP for a cycle with an isothermal phase change (pure refrigerant) is calculated as shown in equation 1. Areas A and B represent the difference in irreversibility between the two cycles. The ideal COP for a cycle with a non-isothermal phase change (nonazeotropic mixture) is calculated as shown in equation 2. It can be proven mathematically that the COP calculation in equation 2 is always greater than the COP calculation in equation 1.

$$COP = \frac{Q_c}{W} \tag{1}$$

$$COP = \frac{Q_c - A}{W - A - B} \tag{2}$$

Similar logic can be applied to a refrigeration or air-conditioning system. Equations 3 and 4 represent the ideal COP for each cycle as calculated from Figure 4. In these equations, Q_e is the heat absorbed by the evaporator for the pure refrigerant case, and is represented by the area below the evaporation line, 4 to 1. Inspection of equations 3 and 4 shows that the ideal COP of a cooling system operating with a nonazeotropic mixture is also always higher.

$$COP = \frac{Q_e}{W} \tag{3}$$

$$COP = \frac{Q_e + B}{W - A - B} \tag{4}$$

The potential increase in COP is the greatest in applications where the heat sink and heat source temperatures are approximately equal and of relatively large magnitude. The minimum requirements to achieve these performance improvements are: the selection of a mixture that yields the desired temperature change in both heat exchangers, a counter-flow heat exchanger that takes advantage of the temperature glide of the refrigerant, and minimized degradation of the heat transfer process. The magnitude of the phase change temperature glide is related to the differences in the normal boiling points of the mixture constituents.

Limitations and Disadvantages

Nonazeotropic mixtures should be selected for a specific operating condition, i.e., a heat source/sink temperature glide that remains constant. Any great variation from the design point could cause a decrease in COP, to the point where using the nonazeotropic mixture is less efficient than using a pure refrigerant.

Counter-flow or near counter-flow heat exchangers are necessary for both the condenser and evaporator to achieve benefits from nonazeotropic refrigerant mixtures. This is because a constant temperature difference between the refrigerant and the heat source/sink is necessary to minimize irreversibility in the condenser and evaporator. Ideally, the temperature glides of the condenser and evaporator must be similar in magnitude and slope.

Generally, nonazeotropic refrigerant mixtures exhibit lower heat transfer coefficients in the condenser and evaporator [18]. This is due partly to the concentration differences between the liquid and the liquid-vapor interface. Diffusion from areas of greater concentration to areas of lesser concentration tends to slow down the condensation (or evaporation) process.

When analyzing cycles for the pure refrigerant and the nonazeotropic mixture that perform the same duty, the heat exchange area of the condenser and evaporator must be greater for the nonazeotropic refrigerant mixture than for the pure refrigerant system. This is partly due to a reduction in log-mean-temperature-difference (LMTD) for the nonazeotropic system and also due to the reduced heat transfer coefficients.

Another disadvantage of nonazeotropic refrigerants is the added degree of complexity should a leak occur in the system. The more volatile constituent will leak first, leaving a higher concentration of the less volatile constituent. The entire system may need to be evacuated so that the correct mixture concentration may be re-established. This situation may become dangerous if a flammable refrigerant is mixed with a non-flammable refrigerant. When separation occurs, the flammable refrigerant may once again be a fire hazard.

TEST SYSTEMS USING NONAZEOTROPIC REFRIGERANTS

Automotive Air Conditioning

Tests on an automotive air-conditioning system were conducted to determine the feasibility of replacing ozone-depleting R-12 with a near-azeotrope blend of three refrigerants [8]. A near-azeotrope refrigerant is a mixture with a composition close to the azeotrope composition, acting nearly identical to a pure refrigerant (as does an azeotrope). The tests showed that the blend had a refrigeration capacity and COP slightly better than R-12. The blend also exhibited slightly better stability with compressor lubricants relative to R-12. The conclusions of the tests were that refrigerant mixtures show potential for use in automotive air-conditioning systems, although compressor replacement or modification is necessary.

Domestic Refrigerator/Freezer

A domestic refrigerator/freezer was tested using a nonazeotropic mixture of R-22 and R-142b [19]. Energy consumption was measured for a full day while the refrigerator maintained average temperatures of 5° F and 38° F (-15° C and 3.3° C) in the freezer and food compartments. To obtain the best performance with the nonazeotropic mixture, the condenser and evaporator were modified to achieve a cross-counter-current heat transfer arrangement. The local flow arrangement of an individual pass in this configuration remained cross-flow as in normal air-to-refrigerant heat exchangers. However, the overall flow arrangement became counter-current, therefore utilizing the advantages of the nonazeotropic mixture.

Results of the tests indicated a 3% increase in performance for a mixture containing 52% mass fraction of R-22 and 48% R-142b. The compressor was not modified or replaced for these tests, however, a synthetic lubricant was deemed necessary to achieve similar or better performance compared to R-12. Analysis indicated that an additional 5% increase in performance was possible by optimizing the motor and compressor design for the refrigerant mixture.

Air-to-Air Heat Pump

A mixture of R-22/R-152a showed a 5.5% COP increase relative to pure R-22 for a 3-1/2 ton high-efficiency (by 1985 standards) split heat pump [20]. This COP increase, verified using DOE test procedures, was obtained without changing the cross-flow heat exchanger configuration of the original heat pump design. There was a capacity decrease (7.9%) along with the COP increase, however, analysis predicted an 8% seasonal savings using the mixture. The performance increase was attributed to thermodynamic property variations. Further COP increases could be realized by employing a counter-flow heat exchanger or a cross-counter-current heat exchanger similar to the one referenced in the previous section. Even though the particular mixture tested may prove impractical due to the flammability of R-152a, the tests demonstrate the feasibility of nonazeotropic mixtures to improve the COP of air-to-air heat pump systems.

Water-to-Water Heat Pump

A prototype heat pump employing a nonazeotropic mixture of R-11[1] and R-22 was developed for the NASA Marshall Space Flight Center under a Phase II SBIR (Small Business Innovation Research) Contract. The prototype heat pump was developed to demonstrate the feasibility of using a nonazeotropic heat pump for crew hygiene[2] water heating on Space Station *Freedom* [21].

Figure 5 is a schematic diagram showing that the condenser side of the heat pump serves as the water heater (heat sink) while the evaporator side picks up waste heat from the thermal control system (TCS) of the space station (heat source). The hygiene water requires heating from 70° F to 145° F (21.1° C to 62.8° C) while the TCS water reaches 110° F (43.3° C) and is cooled to 50° F (10° C). The large temperature gradient required for the hygiene water system and the flexibility to adjust the TCS water temperature gradient make hygiene water heating on the space station an ideal application for a nonazeotropic heat pump.

To compare the performance of a nonazeotropic fluid to that of a pure fluid such as R-12, a theoretical analysis of the heat pump was performed from the thermodynamic properties of the two working fluids. The heat pump cycle for R-12 is shown in the T-s diagram of Figure 6 with a 45° F (7.2° C) saturated evaporator temperature

[1] During the SBIR contract, NASA did not require the contractor to consider CFC phase-out. Therefore, R-11 was an acceptable refrigerant candidate, and contract objectives were met.

[2] When the nonazeotropic heat pump contract was initiated, the hygiene water system was separate from the potable water system. Since then, the two systems have been combined.

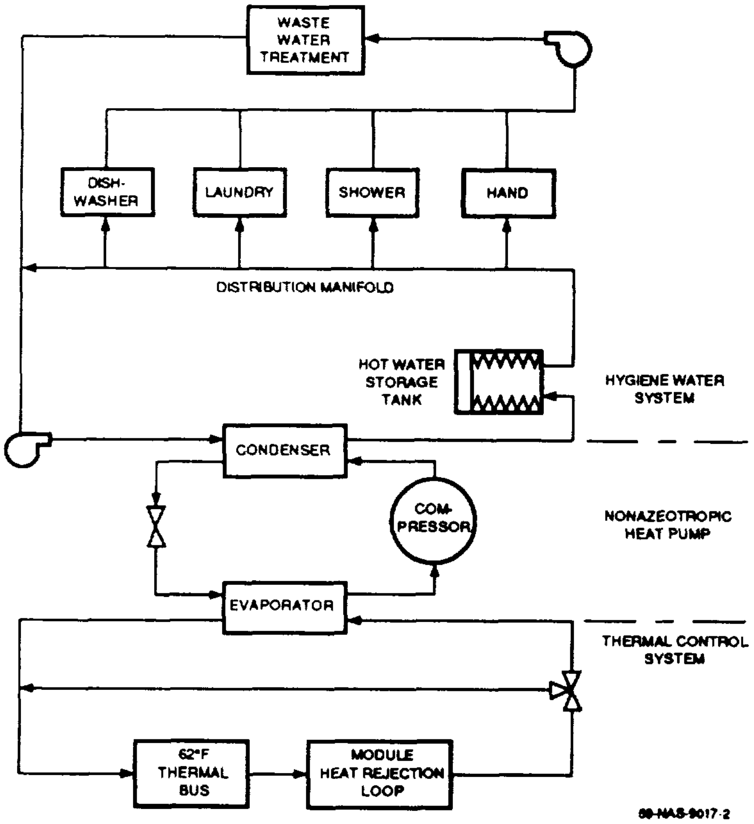
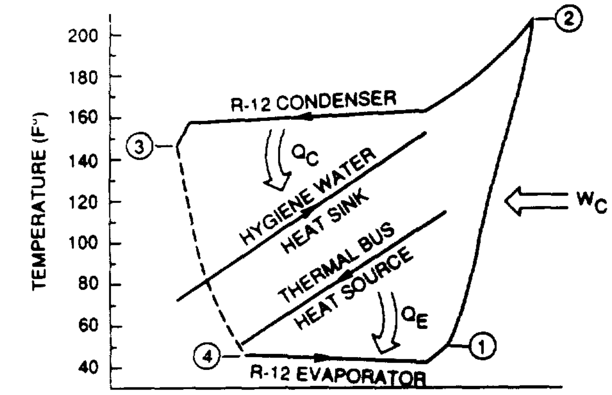


Figure 5. Water Heating Concept using a Nonazeotropic Heat Pump to Transfer Waste Heat from the Thermal Control System to the Hygiene Water System.

and a 150° F (65.6° F) saturated condenser temperature. The analysis assumed a 70% isentropic compressor efficiency, a 3 psi drop through each heat exchanger, 10° F (5.6° C) of condenser sub-cooling, and 5° F (2.8° C) of evaporator super-heat. In producing 3075 Btu/hr (900 W) of water heating, the R-12 heat pump consumes 263 W of power with a COP of 3.44.

A similar analysis was performed using the same heat sink and source for a mixture of R-22 and R-11. The T-s diagram of Figure 7 displays the results. The same compressor efficiency, heat exchanger pressure drop, condenser sub-cooling, and evaporator super-heat were assumed as in the R-12 case. The refrigerant mixture enters the evaporator at 45° F (7.2° C), similar to the R-12 cycle, however, the refrigerant mixture increases in temperature as it boils until it reaches 105° F (40.6° C) at the evaporator outlet. The condensing process begins at 150° F (65.6° C) as with the R-12 system, but the temperature glide of the refrigerant mixture results in 75° F (23.9° C) liquid at the condenser outlet. The nonazeotropic heat pump cycle produces the same amount of heating as the R-12 cycle, 3075 Btu/hr (900 W), while consuming only 102 W at a COP of 8.83. This dramatic theoretical COP increase is a direct result of the reduction in irreversibilities by using the nonazeotropic refrigerant mixture.

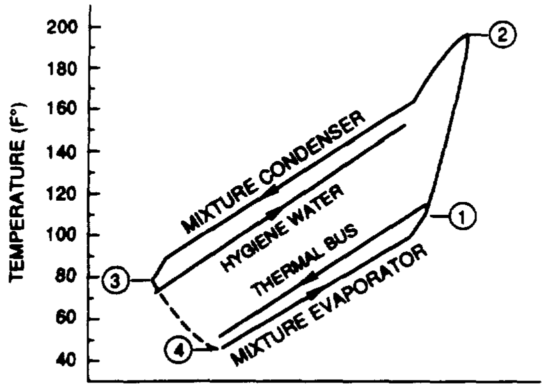
Some of the reduction in irreversibility can be attributed to a reduction in compression ratio. Note from Figures 6 and 7 the differences in system pressures between the two cycles, especially compressor discharge pressure (state point 2). The two theoretical compression ratios are 4.7 for the R-12 cycle and 2.2 for the mixture cycle. The reduced compression ratio not only allows significant power savings, but also increases compressor life.



	T	P	h
1	46.9	53.4	82.4
2	192.7	249.3	99.5
3	139.0	246.3	40.9
4	45.0	56.4	40.9

CONDENSER HEAT REJECTION	3075 BTU/HR
EVAPORATOR HEAT ABSORPTION	2178 BTU/HR
COMPRESSOR WORK	897 BTU/HR
COMPRESSOR POWER CONSUMPTION	263 WATTS
COEFFICIENT OF PERFORMANCE	3.44

Figure 6. R-12 Heat Pump Cycle.



	T	P	h
1	105.1	36.2	108.1
2	182.3	79.4	118.5
3	75.0	76.4	26.5
4	45.0	39.2	26.5

CONDENSER HEAT REJECTION	3075 BTU/HR
EVAPORATOR HEAT ABSORPTION	2727 BTU/HR
COMPRESSOR WORK	348 BTU/HR
COMPRESSOR POWER CONSUMPTION	102 WATTS
COEFFICIENT OF PERFORMANCE	8.83

Figure 7. Nonazeotropic Heat Pump Cycle.

Recall from the previous section that in order to accomplish the same amount of water heating with the nonazeotropic heat pump cycle described, a counter-flow heat exchanger with more area than that of the R-12 cycle is necessary. Prototype testing of the nonazeotropic heat pump resulted in measured COPs ranging from 70% to 90% of predicted values.

CONCLUSIONS

This paper presents reasons to continue investigation of nonazeotropic refrigerant mixtures. Due to CFC phase-out and higher efficiency standards, refrigeration and air-conditioning equipment designers face a dilemma. Uncertainties and problems associated with one possible solution, development of CFC alternatives, were reviewed. Another option, development of nonazeotropic mixtures using existing non-CFC refrigerants, was discussed, including definitions, characteristics, benefits, disadvantages, and limitations. Finally, tests of systems employing nonazeotropic mixtures were presented.

RECOMMENDATIONS

Because of the urgency to replace CFC refrigerants, it is recommended that the government and industrial sectors continue development of systems employing non-CFC nonazeotropic mixtures to the point of commercial and domestic application. It is suggested that this effort be conducted in parallel with the present effort to develop pure-refrigerant CFC alternatives. Possible non-CFC refrigerant combinations were not addressed in this paper. Reference 22, which contains a systematic approach for selecting ozone-safe refrigerants, is a good starting point for nonazeotropic mixture selection.

REFERENCES

1. Calm, J.M.; and Didion, D.A. "Research and Development of Heat Pumps Using Nonazeotropic Mixture Refrigerants." ASHRAE Annual Meeting, Honolulu, Hawaii. June, 1985. Pages 132ff.

2. Federal Register. "Protection of Stratospheric Ozone; Proposed Rule." Vol. 56, No. 189, Page 49553, Sept. 30, 1991.

3. The White House, Office of the Press Secretary. "Fact Sheet on Accelerated Phase-out of Ozone-Depleting Substances." Feb. 11, 1992.

4. Azer, Naim Z. "Recycling and Recovery of CFCs." ASHRAE Journal, February 1992, Vol. 34, No. 2, Page 47.

5. Vineyard, E.A. "The Alternative Refrigerant Dilemma for Refrigerator-Freezers: Truth or Consequences." ASHRAE Transactions, 1991, Vol. 97, Part 2, Pages 955-960.

6. Statt, Terry G. "Potential Ozone-Safe Refrigerants for Centrifugal Chillers." ASHRAE Journal, September 1990, Vol. 32, No. 9, Pages 46-51.

7. Bateman, David J. "Performance Comparison of HFC-134a and CFC-12 in an Automotive Air Conditioning System." SAE Technical Paper Series. No. 890305. February, 1989.

8. Bivens, D.B.; et. al. "Evaluation of Fluorocarbon Blends as Automotive Air Conditioning Refrigerants." SAE Technical Paper Series. No. 890306. February, 1989.

9. Vineyard, E.A.; Sand, J.R.; and Miller, W.A. "Refrigerator-Freezer Energy Testing with Alternative Refrigerants." ASHRAE Transactions, 1989, Vol. 95, Part 2, Pages 295-299.

10. Clark, Earl M.; et. al. "Retro-fitting Existing Chillers with Alternative Refrigerants." ASHRAE Journal, April 1991. Vol. 33, No.2, Pages 38-41.

11. Sanvordenker, K.S. "Durability of R-134a Compressors: The Role of the Lubricant." ASHRAE Journal, February 1991, Vol. 33, No. 2, Page 42.

12. Linton, J.W.; Snelson, W.K.; and Hearty, P.F. "Performance Comparison of Refrigerants R-134a and R-12 in a Residential Exhaust Air Heat Pump." ASHRAE Transactions, 1989, Vol. 95, Part 2, Pages 399-404.

13. Sand, J.R.; Vineyard, E.A; and Nowak, R.J. "Experimental Performance of Ozone-Safe Alternative Refrigerants." ASHRAE Transactions, 1990, Vol. 96, Part 2, Pages 173-182.

14. Standard Mechanical Code of the Southern Building Code Congress International, Inc. (SBCCI). 1991 Edition, Chap. 4.

15. Standard Mechanical Code of the Southern Building Code Congress International, Inc. (SBCCI). 1992 Revision, Chap. 4.

16. DuPont Company, Inc. "Program for Alternative Fluorocarbon Toxicity (PAFT) Update." June, 1990.

17. Merriam, R.L. "Design Concepts for Air-to-Air Heat Pumps using Nonazeotropic Refrigerant Mixtures." A.D. Little Company, Cambridge MA.

18. DeGrush, D.; and Stoecker, W.F. "Measurements of Heat-Transfer Coefficients of Nonazeotropic Mixtures Condensing Inside Horizontal Tubes." University of Illinois at Urbana-Champaign. ORNL/Sub/81-7762/6 &01. November, 1987.

19. He, X.; et. al. "Investigation of an R-22/R-142b Mixture as a Substitute for R-12 in Single-Evaporator Domestic Refrigerators." ASHRAE Transactions, 1992, Vol. 98, Part 1, Pages 150-159.

20. Galloway, J.E.; and Goldschmidt, V.W. "Air-to-Air Heat Pump Performance with Three Different Nonazeotropic Refrigerant Mixtures." ASHRAE Transactions, 1991, Vol. 97, Part 1, Pages 296-303.

21. Walker, David H.; and Deming, Glenn I. "Development of a Nonazeotropic Heat Pump for Crew Hygiene Water Heating." SAE Technical Paper Series. No. 911341. July, 1991.

22. Vineyard, E.A.; Sand, J.R.; and Statt, T.G. "Selection of Ozone-Safe Nonazeotropic Refrigerant Mixtures for Capacity Modulation in Residential Heat Pumps." ASHRAE Transactions, 1989, Vol. 95, Part 1, Pages 34-46.

NOVEL HOT WATER RECIRCULATING TECHNOLOGY CONSERVES ENERGY/WATER

This paper was withdrawn from presentation

VARIABLE-VOLUME FLUSHING (V-VF) DEVICE FOR WATER CONSERVATION IN TOILETS

Louis J. Jasper, Jr.
12389 Kondrup Dr.
Fulton, MD 20759

ABSTRACT

Thirty five percent of residential indoor water used is flushed down the toilet. Five out of six flushes are for liquid waste only, which requires only a fraction of the water needed for solid waste. Designers of current low-flush toilets (3.5-gal. flush) and ultra-low-flush toilets (1.5-gal. flush) did not consider the vastly reduced amount of water needed to flush liquid waste versus solid waste. Consequently, these toilets are less practical than desired and can be improved upon for water conservation. This paper describes a variable-volume flushing (V-VF) device that is more reliable than the currently used flushing devices (it will not leak), is simple, more economical, and more water conserving (allowing one to choose the amount of water to use for flushing solid and liquid waste).

INTRODUCTION

"Water is our most precious natural resource, but it is often taken for granted. We enjoy it, we waste it, we pollute it, assuming our need for water will be met in the future, as it has in the past, by merely turning on the faucet. Indeed it seems no thought is given to how water reaches the faucet and to whether or not the water will always be available when we want it. It's just there whenever and wherever we need it." The preceding quotation is from the *Water Conservation Book of the American Water Works Association* (AWWA) [1]. It sets the theme of this paper, which is water conservation by use of a more water-efficient toilet flushing device. Water shortages of varying degrees, due to droughts and/or inadequate development of water supplies, have occurred in almost all areas of the country. Moreover, it has been estimated that by the year 2000 more than 20 percent of the country wil' occasionally have serious water shortages [1]. It is obvious that increasing water-use efficiency is now mandatory in the United States because of population and industrial growth, along with increased demand for agricultural irrigation. The benefits of water-use efficiency are many, including energy savings, protection of the environment, wastewater flow reduction, and reduced cost.

This paper addresses residential indoor water use, and water conservation in the bathroom, specifically related to toilets. Decreasing toilet wastewater flow saves energy, protects the environment, and saves money. Inside the home, most water is used in bathrooms. Figure 1 shows the average inside water use for nonconserving homes [1]. Toilets use 28 percent of the water inside the home, or 22 gallons per capita per day. Toilets and toilet leakage account for over one-third of the inside water use for nonconserving homes. The V-VF device could significantly reduce the percentage of toilet water use.

The United States Environmental Protection Agency has a fact sheet entitled "21 Water Conservation Measures For Everybody." Three of the twenty-one conservation measures described in this fact sheet are related to the toilet. Conservation facts 10, 11, and 12 are as follows [2]:

(10) Repair leaky toilets to save more than 50 gallons of water per day. Add 12 drops of food coloring into the tank. If the color appears in the bowl one hour later, the unit is leaking.

(11) Install a toilet displacement device to save thousands of gallons of water per year, or 5 to 7 gallons per flush. Place 1 to 3 weighted plastic jugs into the tank, making sure the jugs don't interfere with the flushing mechanism or a suitable flow. Or instead of jugs, use toilet dams that hold back a reservoir of water during each flush, saving 1 to 2 gallons. Don't use bricks because they can chip and foul the flushing mechanism.

(12) When buying a new toilet, select a low-flush model that uses less than 1.5 gallons of water to flush, saving over 7,000 gallons per year per person.

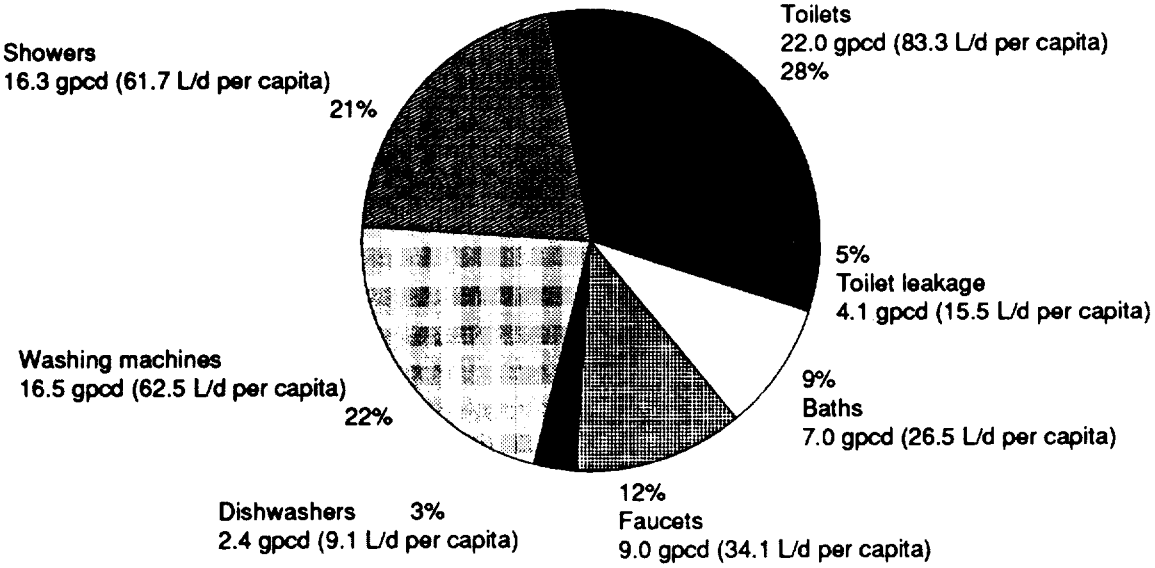


Figure 1. Average inside water use for nonconserving home.

I am in complete agreement with conservation fact 10. I am also in partial agreement with conservation facts 11 and 12. A toilet displacement device and a low-flush toilet can save thousands of gallons of water per year. However, this paper describes a V-VF device which I believe is a better approach to the toilet wastewater problem. It will conserve thousands of gallons more water per year per household than the low-flush toilets, ultra-low-flush toilets, and water displacement devices. The main reason for this belief is that the V-VF device allows the user to control the amount of water expelled from the toilet tank. Five out of six toilet flushes are for liquid waste, which needs only a fraction of the water volume per flush as compared to the amount of water required to flush solid waste. Table 1 shows 7-, 5-, 3.5-, and 1.5-gallon flush toilets compared with a 5-gallon flush toilet that has installed a V-VF device. The V-VF device is described in the next section of this paper. The comparison is made with six toilet flushes, five of which are for liquid waste and one which is for solid waste. It is assumed in this comparison that the user of the V-VF device uses 5 gallons of water to flush solid waste and 0.5 gallon to flush liquid waste. The above-mentioned numbers are based on the full tank capacity being used for flushing solid waste and a minimal amount of water used to replace the amount of water in the toilet bowl for liquid waste.

The data in table 1 show that the low-flush, ultra-low-flush, and V-VF toilets are far more water conservative than the conventional flush toilets, with the V-VF toilet the most water conservative. Several additional benefits can be realized by using the V-VF device in the toilet tank. Low-flush toilets are defined as toilets that use no more than 3.5 gallons per flush and meet performance standard A112.19.2M-1982 of the American National Standard Institute (ANSI). When low-flush toilets were first sold, reports indicated a need for double flushing; however, some claim this problem has now been solved. Experience indicates that frequent double flushing is still required for flushing solid waste with a low-flush toilet. This creates water waste. Ultra-low-flush toilets (1.5 gal. per flush) also have several less than desirable features. They use a supply of compressed air to assist the flushing action. Residential models cost about 50 dollars more per toilet than low-flush models. Concerns still exist for double flushing and

Table 1. Comparison of conventional, low-, and ultra-low-flush toilets with a V-VF device installed in a 5-gallon conventional toilet

Group toilets	Water used for 6 flushes (gal.)	% water savings normalized to 7-gal. tank
Conventional, 7-gal.	42	
Conventional, 5-gal.	30	29
Low-flush, 3.5-gal.	21	50
Ultra-low-flush, 1.5 gal.	9	79
Conventional 5-gal. with V-VF device	7.5	82

sewer-lateral clogging. The Housing and Urban Development (HUD) conducted a water conservation study [3] in which it surveyed leaking toilets. Of 188 toilets tested, 20% leaked. Low-flush toilets leaked considerably more than older conventional toilets. From an extremely small sample size, 60 percent of the low-flush toilets were found to leak. Although the small sample size reduces the accuracy of the statistics, it is my opinion that a significant number of low-flush toilets may still have this problem, especially as they age. Water savings that may be realized by repairing leaking toilets are significant and may average 24 gallons per day per toilet [1]. When the low-flush toilets are made more reliable (less leakage and no double flushing), then the V-VF device can be installed in these toilets to realize additional wastewater savings.

From the facts and statistics given above, one can understand that for water conservation it is important to have a toilet-flushing device that is reliable (will not leak), simple, efficient, and economical. The author offers the V-VF device as an option for reducing the water required to flush toilets.

DESCRIPTION OF THE V-VF DEVICE (PATENT PENDING)

There are three main objectives of the V-VF device:

- to allow the user of a toilet to control the amount of water expelled from the toilet flush tank according to the type of waste to be flushed.
- to achieve water conservation with a mechanism that is simpler than that used in conventional toilets or in other water conservative devices intended for use in conventional toilets.
- to provide a flushing mechanism that is more reliable (does not leak) and less expensive to manufacture than conventional flushing mechanisms.

These objectives are met by replacing the overflow tube and the flapper outlet valve assembly of the conventional toilet with a low-density polyethylene bellowed tube that is compressible to any chosen amount in order to allow water to flow through the tube into the bowl, thereby flushing the toilet. By the elastic resilience of low-density polyethylene, the compressed bellowed tube becomes restored to its original height after flushing and when pressure on the tube is discontinued.

Two versions of the V-VF device are described. One version (fig. 2) uses a handle protruding through the tank cover for compressing the bellowed tube in a plunger-like motion. Although this version of the device is simpler, a hole is required in the tank cover for installation. For this reason, it may be the least attractive of the two. The second version of the V-VF device (fig. 3), uses a handle on the front or side wall of the tank to cause rotation of a shaft which extends into the tank. By means of a cantilever beam, a vertical rod depresses the bellowed tube for flushing. This version may be sold as a replacement kit and, therefore, may be more readily introduced to the public in a cost-effective manner.

Figure 2 is a view of the V-VF device employing a flush handle that rests on top of the toilet tank cover. It shows a conventional design of a toilet flush tank. Note that the water inlet mechanisms and water level controls are not shown. The collapsible bellowed tube, which is made of a low-density polyethylene-type material in the form of a bellows, provides complete elastic recovery when an applied stress is released. One end of the tube opens into the water outlet. It is affixed to the bottom of the tank to form a water-tight seal. When the handle is pressed downward, the bellowed tube is compressed to a point below the water level in the tank. A controllable amount of water enters the bellowed tube, flows through the water outlet, and thereby flushes the toilet. The handle lies flat on the tank cover when not in use. The pivot joint on the handle allows the handle to be rotated into a vertical position for flushing. The handle is provided with markings to indicate the volume of water flushed. After flushing, the handle is released, the collapsible bellowed tube returns to its original uncompressed state, and water enters the tank up to the level controlled by the water inlet mechanisms (not shown in fig. 2) but no higher than the top of the bellowed tube. Therefore, the bellowed tube also is an overflow tube which prevents excess water from entering the tank in case the water inlet mechanisms fail. The cap on top of the bellowed tube has openings to allow water to flow into the tube. The sleeve, with slits surrounding the bellowed tube, forms a cavity which allows water to enter into the cavity through the slits so that water surrounds the bellowed tube for flushing. The sleeve also prevents transverse motion of the bellowed tube when it is compressed. This simple design minimizes friction between the bellowed tube and

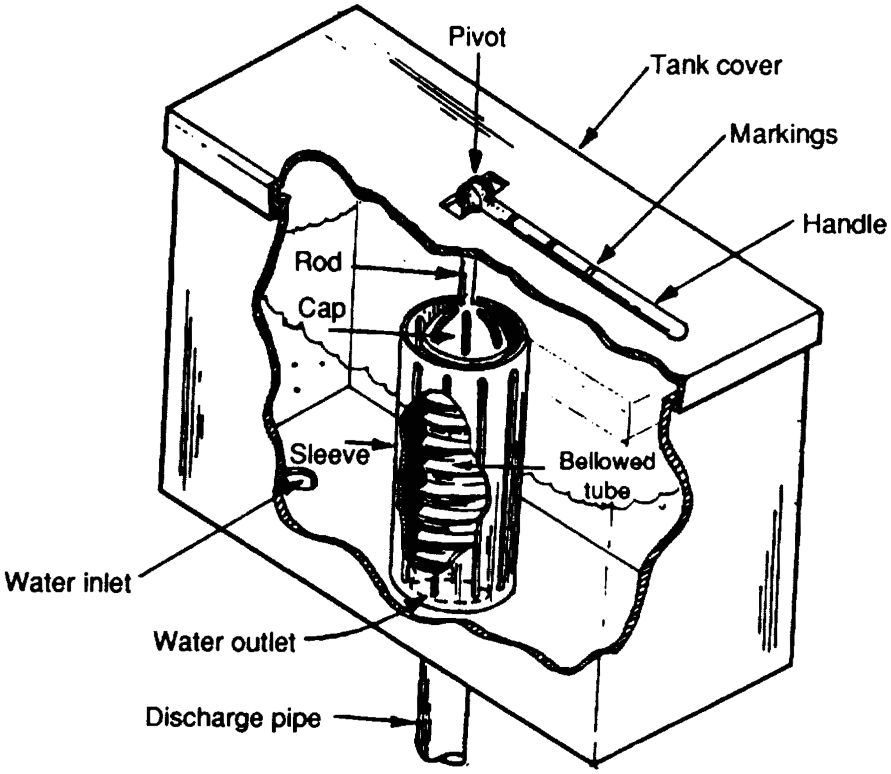


Figure 2. V-VF device inside tank having handle on top of tank cover.

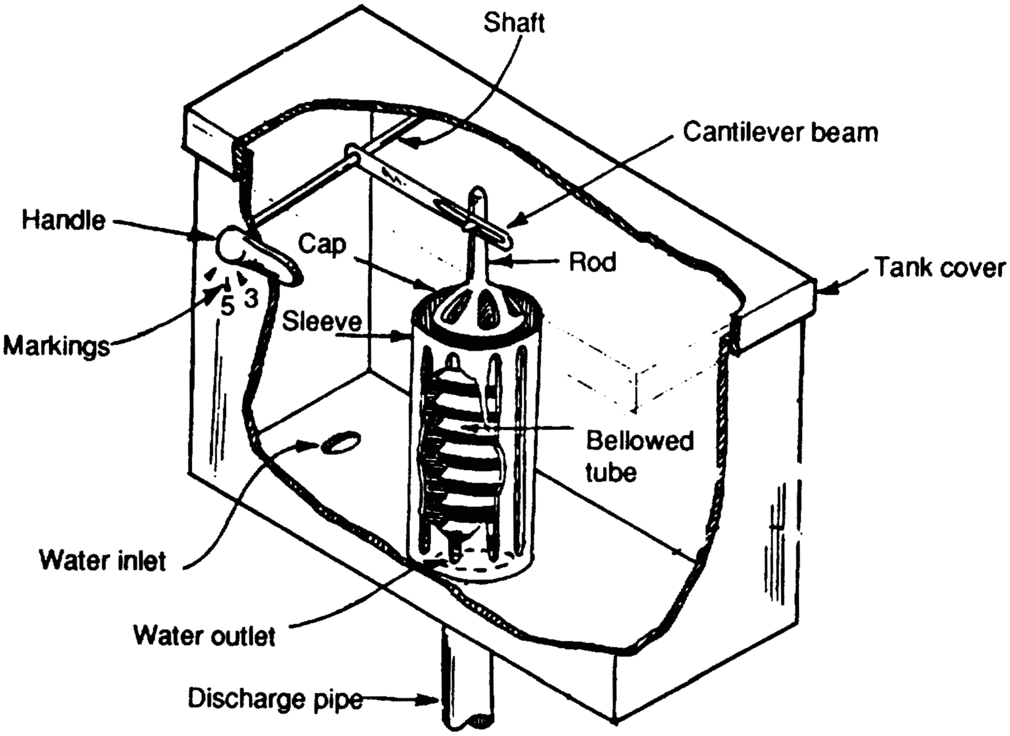


Figure 3. V-VF device inside tank having handle on front wall of tank.

sleeve so that the tube has complete elastic recovery when the applied stress is removed from the tube. Also, minimal components are used in the design to produce a highly reliable, low-cost flushing mechanism.

The second version of the V-VF device is shown in fig. 3. The flushing mechanism is actuated by the lever (on the front wall of the tank) and its attached shaft. The lever is rotated clockwise a chosen number of degrees to cause the cantilever beam to rotate in a similar fashion (move in an arc) and in turn causes the rod attached to the bellowed tube to move downward a given amount. Note that the arc-like motion created by rotation of the lever is converted into a vertical displacement of the bellowed tube. The user controls the volume of water flushed simply by rotating the handle a given amount. The rotation amount is shown by markings arranged in a circular pattern around the lever to indicate the volume of water flushed for a given rotation of the lever.

CONCLUSIONS

Two versions of the V-VF device were described: one version uses a handle that lies on the tank cover and another uses a handle on the front wall of the tank. Because it can be sold as a replacement kit for most conventional toilets, the latter version of the V-VF device is the most attractive for introduction to the public.

It was shown that the V-VF device has a minimal number of parts, serves both as overflow and flushing tube, and is designed and constructed to be low-cost, reliable, and readily adaptable to conventional and low-flush toilets. The V-VF device is more reliable than the conventional chain-flapper valve because it does not require a seal-tight contact between a movable flapper valve and the water outlet opening in the tank bottom. Also, most importantly, the V-VF device allows efficient control of toilet wastewater by enabling the user to choose the amounts of water for flushing both liquid and solid waste. Unlike the low- and ultra-low-flush toilets, double flushing is eliminated, and a more sanitary flush is accomplished with a reduced possibility of sewer-lateral clogging. Presently, it may be more advantageous to use the V-VF device in conventional 5-gallon-capacity toilets instead of low-flush models because of the double flushing, leakage, and sewer-lateral clogging problems that might exist with these models. As an example, if double flushing of the 3.5-gallon-capacity toilet is required 50 percent of the time for flushing solid waste, then in actuality the 3.5-gallon tank becomes a 5.25-gallon tank when flushing solid waste.

The variable water flushing feature, the elimination of double flushing, and reduced water leakage can save thousands of gallons of water per household per year. The average residence uses 107,000 gallons of water during a year, and an individual uses, on the average, about 123 gallons daily. One inch of rain on an acre of land generates 27,000 gallons of water. This means that 4 inches of rainfall is required per acre to satisfy the average yearly residential water requirement. The quotation from the AWWA book that was given in the introduction of this paper should be taken very seriously, with the aim of achieving efficient water use and wastewater flow reduction. The V-VF device accomplishes both aims.

REFERENCES

1. "Water Conservation," by William O. Maddaus, ISBN 0-89867-387-9, American Water Works Association.

2. "Fact Sheet": 21 Water Conservation Measures For Everybody," United States Environmental Protection Agency.

3. Residential Water Conservation Projects—Summary Report, Rept. No. HUD-PDR-903, Brown and Caldwell Consul. Engrs. Prepared for the Dept. of Housing and Urban Devel., Office of Policy Devel. and Res. (June 1984).

4. Water Trivial Facts For Use During National Drinking Water Week, prepared by the United States Environmental Protection Agency.

INFORMATION AND COMMUNICATIONS PART 4: COMPUTER SOFTWARE

AUTOMATED REAL-TIME SOFTWARE DEVELOPMENT

Denise R. Jones
NASA Langley Research Center
Hampton, VA 23681

Carrie K. Walker
NASA Langley Research Center
Hampton, VA 23681

John J. Turkovich
The Charles Stark Draper Laboratory, Inc.
Cambridge, MA 02139

ABSTRACT

A Computer-Aided Software Engineering (CASE) system has been developed at the Charles Stark Draper Laboratory (CSDL) under the direction of the NASA Langley Research Center. The CSDL CASE tool provides an automated method of generating source code and hard copy documentation from functional application engineering specifications. The goal is to significantly reduce the cost of developing and maintaining real-time scientific and engineering software while increasing system reliability. This paper describes CSDL CASE and discusses demonstrations that used the tool to automatically generate real-time application code.

INTRODUCTION

Advanced flight vehicles rely heavily on software to accomplish their missions. The cost of designing, developing, testing, and maintaining avionics software (vehicle guidance, navigation, and control) is becoming an increasingly larger part of total vehicle cost. Until recent years, the lack of appropriate tools to aid in the creation of software has led to nonuniform development techniques and software that is difficult to maintain, and is either unreliable or very expensive to verify. Computer-aided software engineering (CASE) is a technology aimed at automating the software development process in order to produce cost effective and reliable software systems.

One of the major problems associated with the development of real-time scientific and engineering software, aside from general software development issues, is communication among the developers. Typically guidance, navigation, and control (GN&C) algorithmic designs are created by engineers with formal educations in aerospace, mechanical, electrical, or other pure engineering disciplines, rarely from a software engineering discipline. These engineers use languages that are indigenous to their disciplines to specify and communicate their designs. These specifications are presented to software engineers who must interpret them, turn them into a software specification and, subsequently, flight code.

GN&C engineers are not programmers and vice-versa. Consequently, they speak different technical languages. Communication between these groups concerning avionics designs is difficult and prone to error. Usually, programmers must confer with engineers extensively before a software specification adequately represents algorithms as conceived by the engineers. One solution to this communication problem is to educate GN&C engineers to cope with a diverse set and growing number of programming issues. Most engineers these days can program to some extent, but to expect them to perfect the art to the degree that programmers have is unreasonable. Another solution is to educate programmers to be able to design GN&C algorithms or more readily assimilate engineers' GN&C algorithms, which are becoming more diverse and complex. This solution is also impractical. Even programmers who have been developing flight code for years, do not truly comprehend the many static and dynamic interactions that must occur for a flight vehicle to carry out its mission.

Problems are magnified when changes to the code are needed. Proper configuration management practices dictate that changes should be made first to the engineer's design, then to the software specification, and finally to the code. Unfortunately this is not always practiced, and documents become inconsistent and incorrect. A tool is needed that automatically transforms the design specifications developed by GN&C engineers into compilable flight code and the associated system documentation. A Computer Aided Software Engineering (CASE) tool being developed by the Charles Stark Draper Laboratory (CSDL) for the NASA Langley Research Center is such a system.

This paper discusses approaches to software development and defines CASE in general terms. The CSDL CASE software development tool is described and several demonstrations that used the tool are discussed.

APPROACHES TO SOFTWARE DEVELOPMENT

There are several methods of developing software systems, as shown in figure 1. Early software efforts (fig. 1a) were, for all practical purposes, manual. Objectives and algorithms were communicated via combinations of verbal instructions, hand written notes, and, at best, typed documents.

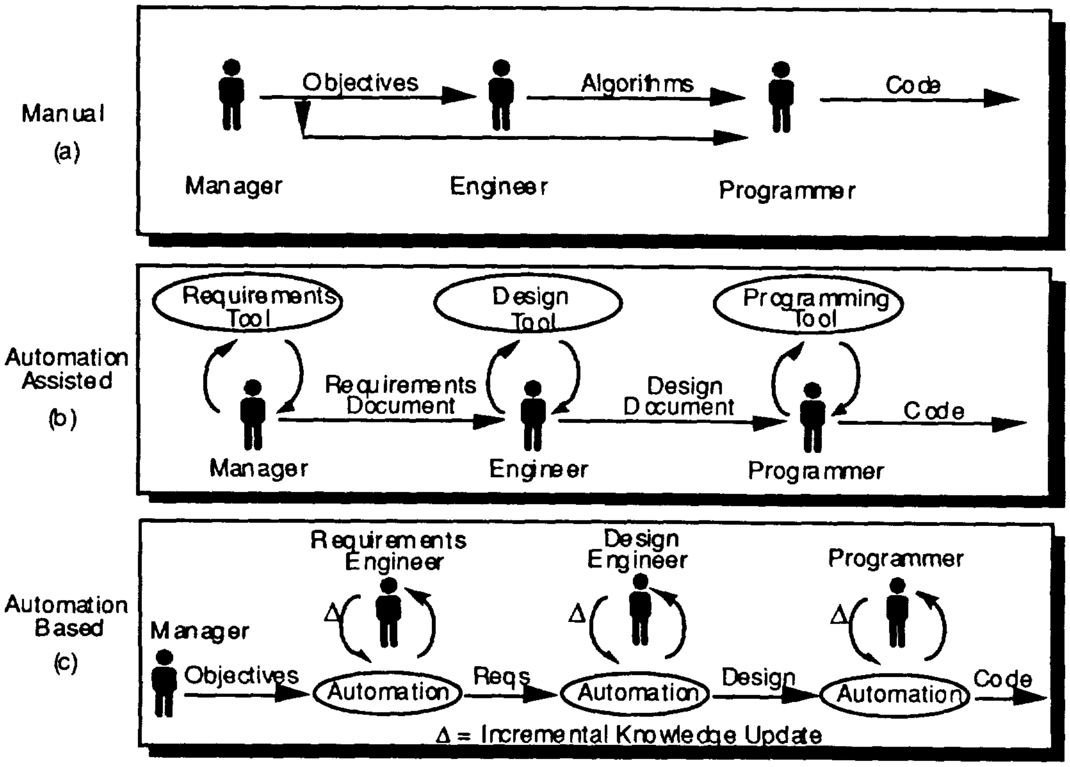


Figure 1. Software development methods.

As expectations, confidence, and perceived flexibility of digital computation grew, so did the notion of computer programs grow into the notion of software and the art of developing these programs into the science of software engineering. Today software development, because of growing size and complexity, is viewed largely as a managerial problem. NASA, DOD, and industry have established standard methodologies for designing, developing, and maintaining software. To support established methodologies, tools that assist the software development process through automation have been and are being developed (fig. 1b).

In order to achieve significant reductions in software costs it is necessary to treat the software problem not merely as a managerial problem but as a technological problem. By doing so, the software development process becomes automation based, as opposed to automation assisted. This automation based system is supported with knowledge acquired by experts as they interact with the system in operation. In this way, software development leverages accumulated knowledge that can be reused from application to application. Instead of people acting as bottlenecks in a flow to analyze functionality (fig. 1b), they serve to fine-tune accumulated knowledge as appropriate on new projects. Notice in figure 1c that people have been removed from the mainstream of software development. The primary function of the mainstream process is to automatically apply accumulated knowledge and secondarily to incrementally acquire knowledge. CSDL CASE takes this approach.

COMPUTER-AIDED SOFTWARE ENGINEERING

Computer-aided software engineering (CASE) is a software technology that began in the early 1980s. CASE is aimed at automating the software development process to improve software productivity and quality. The basic concept is to provide a set of well-integrated software tools and methodologies that automate the entire software development life cycle from analyzing application requirements to maintaining the resulting software system [1].

Hundreds of CASE tools have emerged over the last decade. Many of these tools are workstation-based using graphics and windowing capabilities to interact with the software developer. CASE tools automate a variety of software development and maintenance tasks such as creating structured diagrams and pictorial system specifications, generating executable code and system documentation, and error checking. Generally a CASE tool covers only a portion of the software life cycle. Most CASE tools also support one or more software development methodology, such as Yourden's structured design [2] or DeMarco's structured analysis [3]. These methodologies use notations that are composed of diagrams, graphs, charts, tables, and formal languages.

CASE offers many enhancements to the software development process. Most of these benefits are a result of the automated environment that CASE provides through an interactive graphic user interface. Some of these benefits are as follows [1]:

- enforces software/information engineering,
- makes prototyping practical,
- frees developer to focus on creative part of software development,
- enables reuse of software components (prototypes, data, code, functional designs),
- simplifies program maintenance and reduces maintenance costs,
- reduces software development time and cost,
- improves software reliability and quality, and
- increases productivity.

The introduction of CASE technology has changed the software development process. Traditional development emphasized the later phases of the software life cycle, with 65 percent of the effort placed on the coding and testing phases [1]. CASE automates much of the latter part of the software life cycle so more time can be spent on analysis and design. Table 1 shows the differences between traditional and automated software development [1].

Traditional Development	Automated Development
Emphasis on coding and testing	Emphasis on analysis and design
Paper-based specifications	Rapid iterative prototyping
Manual coding	Automatic code generation
Manual documenting	Automatic document generation
Software testing	Automated design checking
Maintain code	Maintain design specifications

Table 1. Differences between traditional and automated software development.

CHARLES STARK DRAPER LABORATORY CASE TOOL

Background

The foundation of the Charles Stark Draper Laboratory computer-aided software engineering (CSDL CASE) system was developed under the Draper Laboratory's internal research and development (IR&D) program [4]. It was supported by the Advanced Launch System (ALS) Advanced Development Program under the direction of NASA Langley Research Center from 1988 - 1989 and was known as ALS CASE [5]. The advancement of CSDL CASE continues through the sponsorship of the NASA Langley Research Center and the Draper Laboratory Corporate Sponsored Research (CSR) program.

System Description

Although there are many CASE tools available commercially, most support general purpose systems development. Those that do support real-time scientific and engineering software applications generally do so by providing "extensions" to the notations used to support the general purpose systems development. CSDL CASE supports real-time scientific and engineering software systems development from an engineering perspective.

CSDL CASE views software design, development, and maintenance from the viewpoint of the application engineer [6]. The engineering design or functional specification *is* the software specification from which the source code is automatically produced. The functional specification is input into the CSDL CASE system in the form of hierarchical engineering block diagrams and algebraic equations, notations familiar to avionics and control systems designers. The resulting source code is then maintained by modifying the specification through the block diagrams, therefore, eliminating the need to maintain functional specifications, source code, and documentation separately. This approach enforces consistency, eases maintenance, and increases productivity and reliability since modifications are made solely to the functional specification through an interactive graphical user interface with the software system generated automatically.

The architecture of the CSDL CASE tool is shown in figure 2 [7]. This architecture, which is implemented on an engineering workstation platform, consists of a user interface, an automatic software designer, automatic code generators, and an automatic document generator. The user interface facilitates specification of algorithms and systems of algorithms as engineering block diagrams. The automatic software designer converts the graphical specifications into a machine-dependent design. Code generators automatically apply the syntax of the target language to the software design in order to produce source code. The document generator automatically constructs and prints hard copy documentation of the specification according to a specified format. The following sections describe this architecture in more detail.

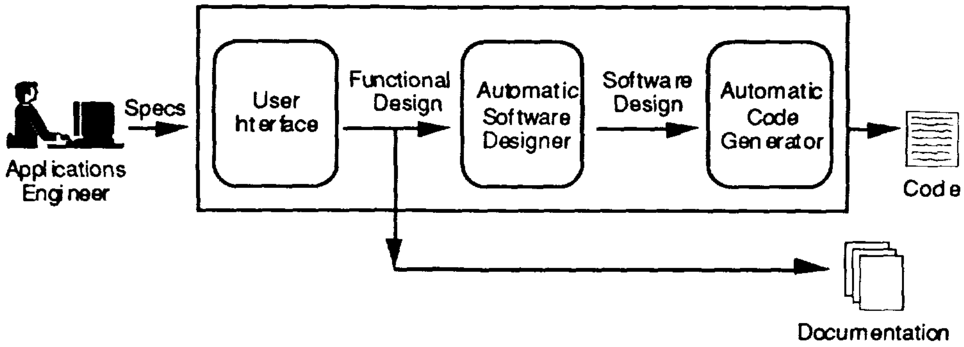


Figure 2. CSDL CASE system architecture.

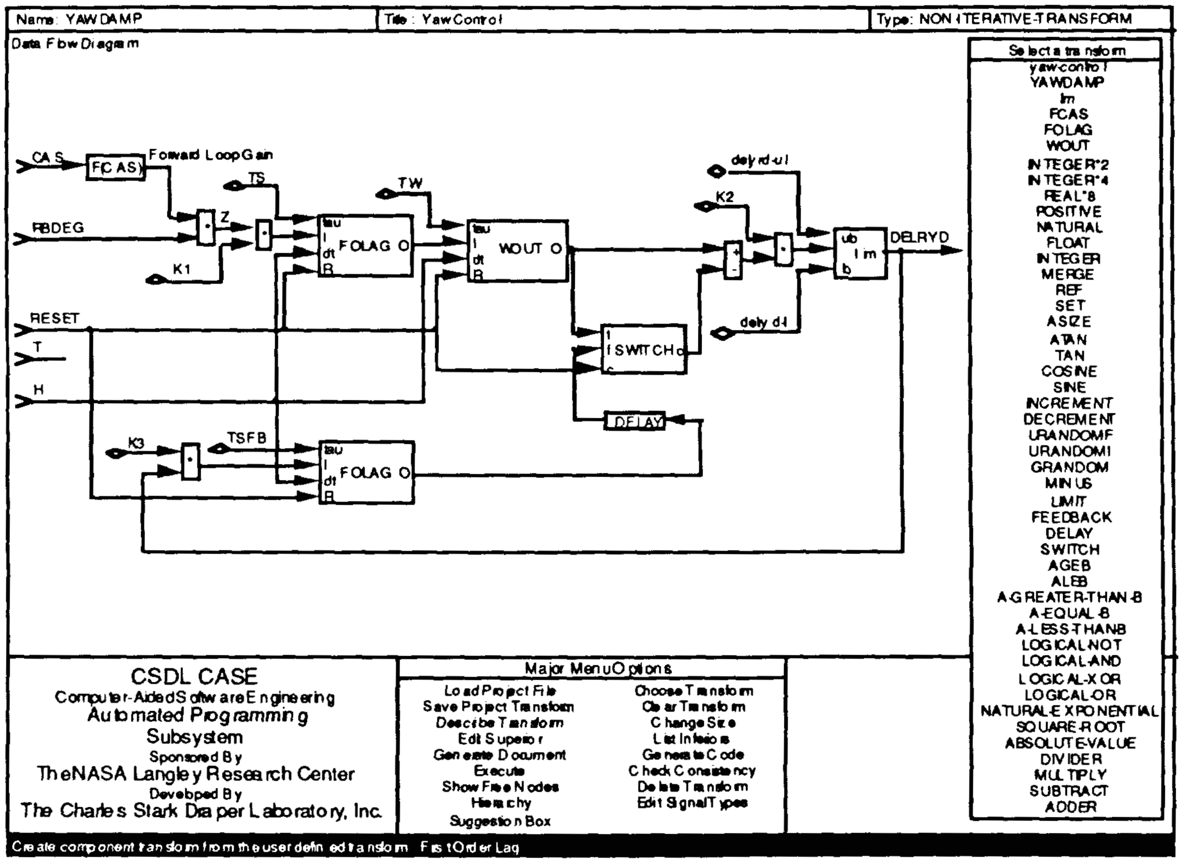
User Interface

The user interface was designed with the purpose of specifying systems of engineering algorithms as opposed to software designs. This specification emphasizes the functionality and interaction of algorithms in order to convey their meaning to other engineers. Software designs by contrast do not clearly convey the meaning of algorithms and systems of algorithms. Instead, software designs dominantly reflect constraining characteristics of the execution environment such as computational resources, memory resources, input/output resources, and resource connectivity.

Systems of algorithms are specified through extensive use of engineering block diagrams [8]. The computational aspects of a diagram are referred to as "transforms" since they explicitly transform inputs into outputs with no hidden side effects. The data aspects of a diagram are "signals" that carry information from one transform to other transforms. Hierarchies of both transforms and signal types can be built either bottom-up or top-down. For bottom-up design, predefined sets of building blocks for both transforms and signal types are supplied. In the case of transforms, these are called primitive transforms and are comprised of such things as add, subtract, multiply, divide, absolute value, switch, etc. For signal types, these are called predefined types and include integer, float, character, string and boolean. The user can also create his own signal types such as arrays and records. For top-down design, the engineer needs only to specify the input and output characteristics of a transform before using it in a block diagram. The details of the transform's data flow and processing can be deferred until a later time, or a body of existing code can be referenced rather than automatically generating code.

A sample CSDL CASE user interface window is shown in figure 3. The window consists of many subwindows, called panes. The center pane can display either a block diagram or a description of a signal type. The three panes across the top of the window display the name, title, and type of the block diagram that is being edited. The additional three panes located at the bottom of the window display, from left to right, update messages, a menu of major specification options, and detailed responses to the selection of a major menu option.

Figure 3 also gives an example of an engineering block diagram specification. The diagram is a graphical representation of functionality that is captured in a centralized, object-oriented knowledge base as a result of developing the diagram. The diagram is created by selecting and placing primitive transforms, user-defined transforms, input terminals, output terminals, and constant value terminals and then connecting them with signals. A transform is selected from a pop-up menu (see figure 3) and the resulting graphical representation is placed on the diagram. Input, output, and constant terminals are selected and placed by again calling up a pop-up menu, specifying detailed definitions for them, and then positioning them in the diagram.



Name: YAWDAMP

Title: YawControl

Type: NON ITERATIVE-TRANSFORM

CSDL CASE
Computer-Aided Software Engineering
Automated Programming
Subsystem
Sponsored By
The NASA Langley Research Center
Developed By
The Charles Stark Draper Laboratory, Inc.

Create component transform from the user defined transform First Order Lag

Figure 3. CSDL CASE user interface window.

Automatic Software Designer

The automatic software designer takes as input the object-oriented, functional form of the specification and determines a generic, procedural form which takes into consideration characteristics of the execution environment. During the automatic design process, engineering diagrams are converted into procedures, functions, or in-line code depending on their usage. Each transform in a diagram becomes either a statement or a block of statements. A statement consists of an assignment to one or more variables or a call to a procedure or function. Input, output, and constant terminals are converted into their respective classes of variables. The automatic designer also generates variables when it is necessary to implement state. State variables appear as a result of feedback loops in the graphical specification. In addition, local variables may be needed to reflect the connectivity in a diagram; these variables are also generated by the software designer. Connectivity is also used to determine the execution order of statements. Traversals of diagram connectivity from both outputs to inputs and inputs to outputs are performed to determine which statements must execute in sequence, which may execute in parallel, and which are conditionally executed.

Automatic Code Generator

Ada and C source code generators have been developed. Each automatic code generator takes as input the generic, procedural form of a software design produced by the automatic software designer and creates source code in the syntax of the corresponding target language, while exploiting the functionality of the chosen language.

The generated Ada code is hierarchically structured and contains one Ada function and procedure for each function and procedure in the software design format. Since C is not a hierarchical language and recognizes only function program blocks, the C generator flattens the definition hierarchy and creates functions for both procedures and functions in the software design.

Automatic Document Generator

The documentation process consists of three distinct steps: generation, formatting, and printing. The document generator takes as input the same object-oriented, functional representation of the specification used to generate a software design. The generation process creates both the text and graphics for the title page and body of the document which reflects the block diagram specification. Although the generation is performed in CSDL CASE, the formatting and printing is performed on an engineering workstation using a commercial publishing software package. The final product is a fully collated document which includes a title page, table of contents, list of figures, sections and subsections, appendices, and an index. The entire document is composed and assembled with no manual intervention.

APPLICATIONS OF CSDL CASE

CSDL CASE has been used on both small and moderate size applications [9]. The small applications are a Boeing 737 yaw damping system, a Martin Marietta Lateral Acceleration Sensing System for the Titan IV, and a General Dynamics electromechanical actuator model. The moderate-sized applications are an autoland control system for the Boeing 737 aircraft, an autonomous exploration vehicle simulation, and a guidance and control system for a planetary lander. Each of the small applications resulted in the automatic generation of hundreds of lines of code and tens of pages of documentation. Each of the moderate size applications generated thousands of lines of code and hundreds of pages of documentation. The 737 autoland system is described below as a representative CSDL CASE application.

Boeing 737 Autoland System

Ada code and documentation for a Boeing 737 autoland flight control system [10] were generated using CSDL CASE. The CSDL CASE specification for this control system was reverse engineered from a FORTRAN implementation and documentation for that implementation. This documentation contained both block diagrams and flowcharts. The autoland system controls pitch, roll, yaw, and throttle for the B737 aircraft from about 5000 feet altitude until touchdown. Ada code and documentation were automatically generated using CSDL CASE [11], resulting in approximately 3000 lines of Ada code and 200 pages of documentation.

Execution of the Ada code was tested against execution of the FORTRAN code. The objective of this experiment was to determine deviations in the outputs of the automatically generated Ada code and the manually generated FORTRAN code. The FORTRAN implementation had previously undergone extensive testing. Tests were structured so that all paths within the design would be exercised. A test was conducted by subjecting FORTRAN code to a time varying set of inputs as it executed. Both inputs and outputs were recorded. The Ada code was then executed, using the recorded FORTRAN inputs. Outputs were compared to the FORTRAN outputs.

Thirty-three tests were performed. Eleven discrepancies were detected. Nine discrepancies were traced to the CSDL CASE user; these were errors in the block diagrams that had been specified to the automatic programming system. Two discrepancies were traced to errors in the FORTRAN code. No errors were traced to either the automatic software designer or the automatic Ada code generator.

As a byproduct of this activity, a strategy for testing code that is based on engineering block diagrams was developed [11]. This strategy prescribes a method for designing tests that cover the functionality expressed by block diagrams while minimizing the number of tests. The test design is dependent on the types of transforms in a block diagram and their connectivity.

As an extension of this activity, the Ada code was regenerated and targeted for a flight critical computer system. The objective of this demonstration was to show that the code produced with CSDL CASE is not only suitable for execution on a general purpose computer for prototyping or simulation purposes, but that the code is also appropriate for a typical flight critical architecture. The architecture used for this demonstration was the Advanced Information Processing System (AIPS) [12] being developed at the Draper Laboratory for NASA Langley.

AIPS is a set of hardware and software building blocks that can be configured in various ways to form fault-tolerant distributed computer system architectures. An AIPS configuration can be used for a broad range of applications, including highly-reliable flight critical applications. An AIPS node can consist of a single Fault Tolerant Processor (FTP), duplex FTPs, or triplex FTPs, depending on the function criticality and needed reliability. A node can even contain one or more Fault Tolerant Parallel Processors (FTPPs), if large capacity throughput is needed for the application. Hardware fault detection provides low fault tolerance overhead. System services are provided by an operating system written in Ada. The system complexity is hidden from the applications; therefore, the applications developer does not have to consider redundancy during application development.

The objective of this demonstration was met. The feasibility of using unaltered CASE-generated code on a flight critical architecture was demonstrated. First, enhancements were made to CASE to allow the AIPS FTP to be one of the architectures available for code generation. This enhancement entailed altering the automatic code generator to reflect AIPS system dependencies, particularly to include the appropriate math library instantiations and communication routines.

The steps involved in the actual demonstration are illustrated in figure 4. Specifications for the Boeing 737 autoland system were entered into CSDL CASE using the interactive graphical interface (see figure 3). Ada source code targeted for the AIPS FTP and detailed documentation were generated. The Ada code was compiled on a Digital Equipment Corporation MicroVax using the XDAda compiler [13] and linked with the AIPS operating system. The automatically generated code was successfully executed on the AIPS FTP while interacting with a dynamic simulation of the Boeing 737 executing on the MicroVax. Proper performance was verified visually during execution by a Macintosh display of aircraft state. The displays graphically depicted the aircraft trajectory, attitude, and attitude rates. These displays are shown in figure 5.

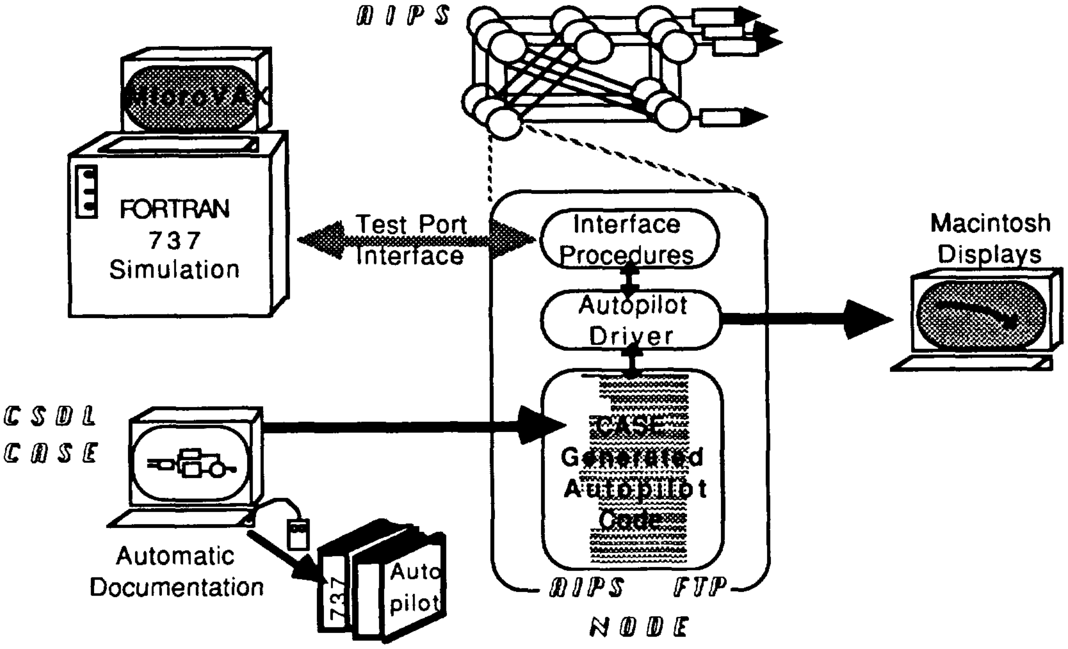
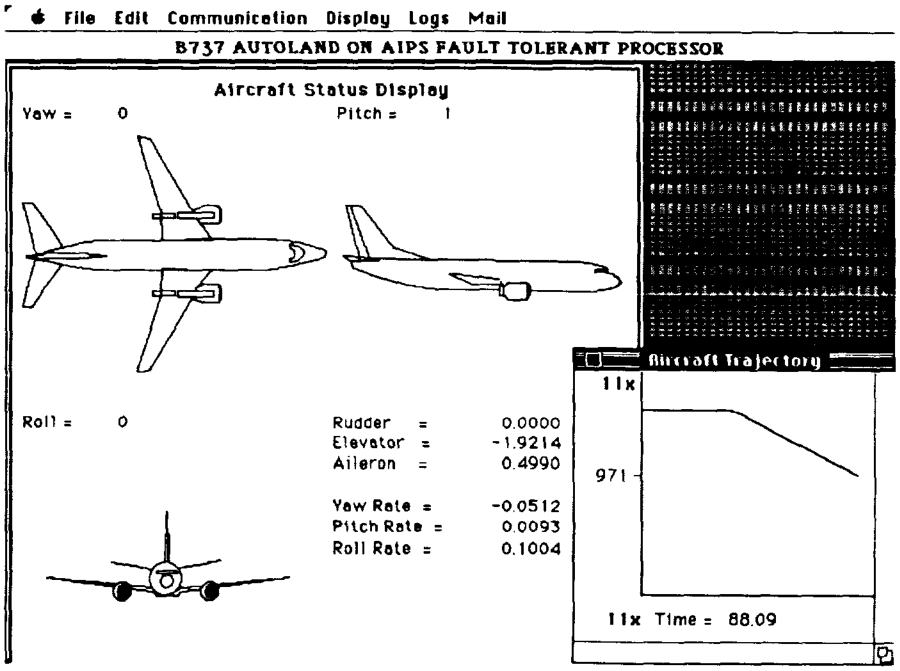


Figure 4. Boeing 737 autoland demonstration.

Table 2 contains data from one run of the demonstration. Initially, the aircraft was flying level at 1500 feet, approximately 118 knots, and was aligned with the runway. The glideslope was intercepted at 34.8 seconds into the simulation and took approximately seven seconds to capture. As is graphically depicted in figure 5, there was no overshoot during glideslope capture and no oscillation along the trajectory. During flare, the system removed all but an acceptable amount of vertical velocity (~3ft./sec.).


File Edit Communication Display Logs Mail
B737 AUTOLAND ON AIPS FAULT TOLERANT PROCESSOR
Aircraft Status Display
Aircraft Trajectory
11x
971
11x

Yaw = 0
Pitch = 1
Roll = 0

Rudder = 0.0000
Elevator = -1.9214
Aileron = 0.4990

Yaw Rate = -0.0512
Pitch Rate = 0.0093
Roll Rate = 0.1004

Time = 88.09

Figure 5. Aircraft status and trajectory displays.

Environmental Conditions:

Atmosphere = STD62	Heading = 67.33 deg.
Winds = None	Elevation = 9 ft.
Runway = Langley 07	Desired Glideslope = -3 deg.

Flight Conditions:

	Initial Condition	Glideslope Intercept	Glideslope Capture	Near Mid-descent	Begin Flair	Touchdown
T (sec)	0.00	34.8	42.4	109.6	175.31	181.41
Alt (ft)	1500.00	1499.75	1461.00	738.73	56.04	18.87
Gamma (deg)	0.00	0.004	-2.995	-2.99	-2.99	-0.87
CAS (kts)	117.85	117.73	119.75	118.01	117.84	108.54
Yaw (deg)	67.33	66.95	67.17	66.90	67.29	67.33
Pitch (deg)	3.97	3.52	-0.17	0.90	------	------
Roll (deg)	0.00	-0.07	0.16	0.03	-0.76	0.26
HDOT (ft/s)	0.00	0.01	-1.16	-10.42	-10.34	-2.76

Table 2. Boeing 737 autoland experiment data.

Advanced Guidance Demonstration

In a follow-on effort to the Boeing 737 autoland experiment, the process of demonstrating automatically-generated Ada code on a *distributed* fault-tolerant computer architecture is being pursued. This demonstration is a more extensive experiment to further test this technique for developing a flight critical computer system. The application for this demonstration, an advanced adaptive guidance algorithm for a launch vehicle, is much more ambitious.

During analysis, engineers converge on a single algorithm that they consider to be the baseline design. This baseline design is then transcribed into the form of a software specification, since the software specification is typically different from the design form used during analysis. Typically, errors are introduced during 1) the initial transcription from the analysis representation to the software representation and 2) during maintenance of both the analysis and software representations.

In order to avoid these errors and simultaneously reduce development time by eliminating duplicated effort, the algebraic specification capability of CSDL CASE is being modified to accept MATLAB™ scripts. MATLAB™ is an analysis and simulation tool used for the development and test of guidance and control algorithms [14]. This feature will allow both analysis and flight code to be developed from a single specification. When iteration on a design for analysis purposes is complete, the software specification is, therefore, also complete. The analysis and development of the guidance algorithm for the experiment is being performed with MATLAB™. This aspect of the demonstration is intended to show, with MATLAB™ as an example, that it is possible to generate Ada and C code from design specifications, and that design analysis and simulation tools will be usable within CSDL CASE.

Also being demonstrated in this effort is the ability to generate and execute code for a *distributed* architecture, as a network of AIPS FTPs and FTPPs (Fault Tolerant *Parallel* Processors) will be used for execution. As in the previous experiment, the code will be executed on the AIPS while interacting with a dynamic vehicle simulation executing on a MicroVax. The process for this demonstration is shown in figure 6.

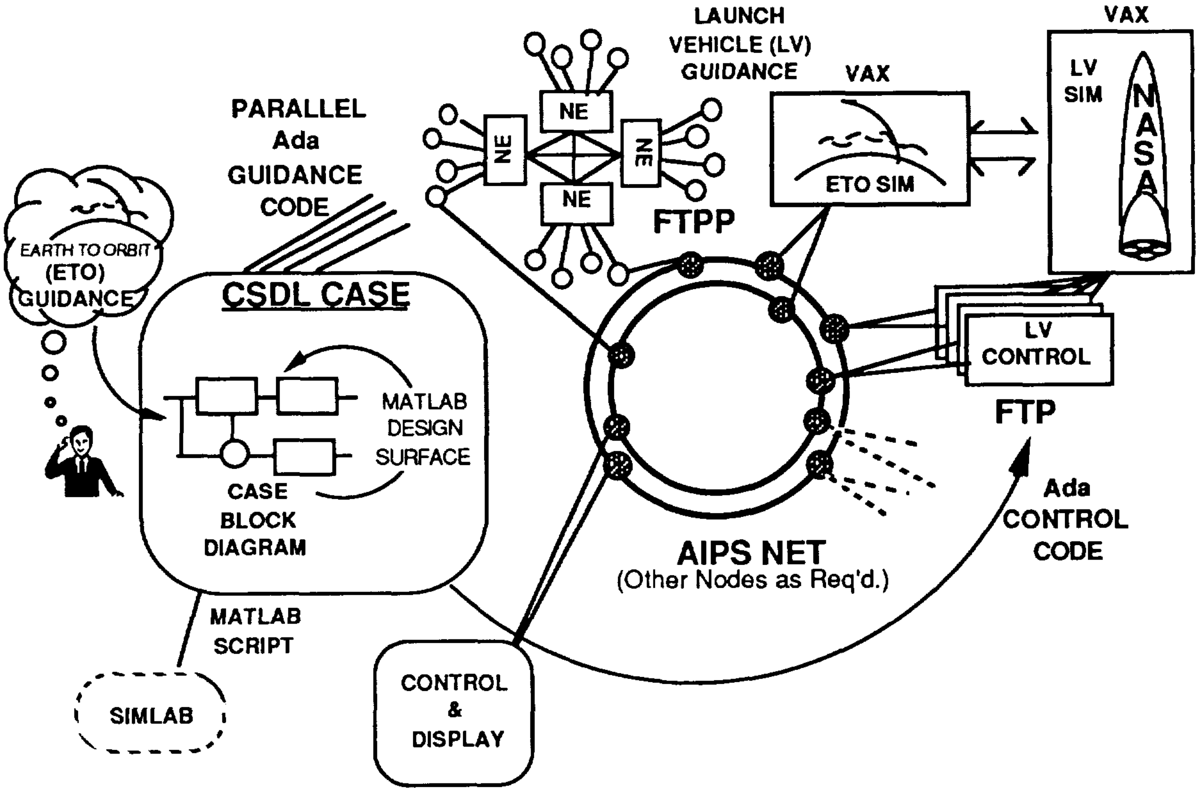


Figure 6. Advanced guidance demonstration on distributed AIPS.

FUTURE ACTIVITIES

Future activities include the development of an Automated Testing Subsystem, a Software Design Methodology User Interface, and the inclusion of formal semantic representations for software designs. Commercialization is also being pursued.

Automated Testing Subsystem

Automating testing would involve automating the five steps common to testing, namely: 1) design; 2) setup; 3) execution; 4) analysis; and 5) documentation. ***Design*** is the determination of how the code will be exercised and monitored in order to ensure that the code complies with its corresponding requirements. ***Setup*** is the process of collecting and assembling all of the components that are specified in a test design. These components can include source code, object code, load modules, simulators, input data, specifications on data to be collected during execution, and expected test results. ***Execution*** is the process of downloading a load module to a target environment, initiating execution, monitoring and collecting data, and terminating execution. ***Analysis*** is the process of determining if actual results agree with expected results. ***Documentation***, obviously, consolidates information in the previous four steps for dissemination and review. Clearly, automated design and thorough analysis are the most complex of the described steps. The remaining three steps and simple analysis are candidates for an initial prototype automated testing system.

Automated test setup will allow test data to be specified for each transform (or selected transforms) in an application design. Test driver code will automatically be generated for the supplied test data. For execution, a load module will be automatically developed for the target environment. A capability will be provided to automatically compile, link, and execute test code, and then monitor execution on the user's workstation while recording data. This process should be controlled from the user's workstation, even if the code is executed on another machine. Simple analysis will be accomplished by comparing achieved results with desired results. The results and discrepancies will be documented in a test report. The display screens and documentation could both vary in complexity from the output of numeric values to graphs, plots, or complex visualization techniques. Naturally, an initial prototype will employ the more simple means of data display.

Software Design Methodology User Interface

The objective of the Software Design Methodology User Interface (SDMUI) is to allow software designers to specify a software design methodology and automatically apply that specified methodology to a functional design. Software design methodologies that are specified through this SDMUI, will be able to be reused on different functional designs. It will be possible to reconfigure a functional specification from one software design into another. Reconfigurations retain the specified functionality of an application design but alter elements of a software design such as memory utilization, execution time, modularity, and compile time in order to optimize execution on a specific execution environment. Formal representations will be used to define software designs, system constraints, and software designing methodologies.

Formal Semantics

Formalisms for representing the semantics of software designs will be investigated. Formalisms will be used to develop a language for representing software designs and methodologies for manipulating software designs. All manipulated versions of a software design should contain the same functionality but would consume different computational resources (i.e., functionally equivalent designs would be represented in different, but equivalent, software designs dependent on the target machine). The design specified by a user of CSDL CASE is the engineer's view of a software application's functionality. As long as the top-level inputs and outputs are the same, it does not matter how the software implementation of this application is executed. It is the intent to be able to transform one method of executing a target software application into another and to be able to demonstrate with mathematical rigor that the two methods are equivalent.

Commercialization

It is the intent of NASA Langley and the Draper Laboratory to pursue commercialization of CSDL CASE in cooperation with a third party. Ideally, present capabilities would be available commercially, supported by a commercial software vendor. The tool would also remain a research vehicle for NASA Langley and the Draper Laboratory to pursue software automation issues like those described in this section.

SUMMARY

The CSDL CASE system views software design, development, and maintenance from the perspective of the application engineer. Unlike most code generators which address the generation of source code from software designs, this approach addresses automated code and documentation generation from specifications of functional application designs. The user interface provides a natural design technique for the specification of real-time software by allowing the functional specifications to be input as hierarchical engineering block diagrams, a notation familiar to application engineers. With this technique, rapid system development and prototyping are possible. Maintenance concerns are reduced since both the code and documentation are maintained by changing the graphical specification. Software reliability and productivity are increased because of the reduction in manual operations, the consistency and completeness checking provided by the system, the automatic translation of specification to code, and the support for reuse of specifications. The application of this CASE tool to many real avionics designs has driven the development of the tool and has proven the viability of the approach. It is felt that CSDL CASE has a place in the commercial arena, particularly in the control systems design environment.

REFERENCES

1. McClure, Carma: *CASE is Software Automation.* Englewood Cliffs, N.J., Prentice-Hall, 1989.

2. Yourdon, Edward; and Constatine, Larry: *Structured Design.* Englewood Cliffs, N.J., Prentice-Hall, 1985.

3. DeMarco, T.: *Structured Analysis and System Specification.* Yourdon Press, New York, 1978.

4. McDowell, M. E.: *Computer-Aided Software Engineering at The Charles Stark Draper Laboratory.* CSDL-P-2802, The Charles Stark Draper Laboratory, Inc., April 1988.

5. Turkovich, John J.; et al: *Advanced Launch System (ALS) Advanced Development Program 2501 Computer-Aided Software Engineering (CASE) Final Report.* NASA Contractor Report, to be published.

6. Turkovich, John J.: Automated Code Generation for Application Engineers. *AIAA/IEEE 9th Digital Avionics Systems Conference*, October, 1990.

7. Walker, Carrie K.; and Turkovich, John J.: Computer-Aided Software Engineering: An Approach to Real-Time Software Development. *AIAA Computers in Aerospace 7*, October, 1989.

8. *ALS CASE User's Guide.* The Charles Stark Draper Laboratory, Inc., Cambridge, MA, June, 1991.

9. Walker, Carrie K.; Turkovich, John J.; and Masato, T.: Applications of an Automated Programming System. *AIAA Computers in Aerospace 8*, October, 1991.

10. *Linear 737 Autoland Simulation for AIRLABS.* Sperry Corporation, SP710-032, June, 1985.

11. Lewin, A. W.; and Turkovich, J. J.: *Testing the ALS CASE Version of the Boeing 737 Autoland Flight Control System.* NASA Contractor Report, to be published.

12. Lala, J. H.; Harper, R. E.; and Alger, L. S.: A Design Approach for Ultra Reliable Real-Time Systems. *IEEE Computer*, May, 1991, pp. 12 - 22.

13. *XDAda Technical Summary.* Digital Equipment Corporation, June, 1989.

14. MATLAB™ High Performance Numeric Computation Software. *MATLAB™ User's Guide*, The MathWorks, Inc., January, 1991.

CONSTRAINT CHECKING DURING ERROR RECOVERY

53-61
5664
P-10

N93-22169

Robyn R. Lutz
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109

and

Johnny S. K. Wong
Department of Computer Science
Iowa State University
Ames, IA 50011

ABSTRACT

The system-level software onboard a spacecraft is responsible for recovery from communication, power, thermal, and computer-health anomalies that may occur. The recovery must occur without disrupting any critical scientific or engineering activity that is executing at the time of the error. Thus, the error-recovery software may have to execute concurrently with the ongoing acquisition of scientific data or with spacecraft maneuvers. This work provides a technique by which the rules that constrain the concurrent execution of these processes can be modeled in a graph. An algorithm is described that uses this model to validate that the constraints hold for all concurrent executions of the error-recovery software with the software that controls the science and engineering activities of the spacecraft. The results are applicable to a variety of control systems with critical constraints on the timing and ordering of the events they control.

INTRODUCTION

This paper presents a technique for checking constraints on the asynchronous software processes involved in error recovery aboard spacecraft. These autonomous processes are constrained at the event level by timing, precedence, and data-dependency rules. Violations of these constraints can jeopardize the spacecraft.

Analyzing all the potential process interactions during spacecraft error recovery is difficult and tedious. A single failure on the spacecraft may at times trigger several different processes whose actions must then be compatible. More than one failure may also occur at a time, causing several error-recovery processes to be invoked. In addition, there is at any time a unique sequence of uplinked commands (instructions to subsystems) executing on the spacecraft. These commands must also be compatible with the actions of the error-recovery software.

The spacecraft software is highly interactive in terms of the degree of message-passing among system components, the need to respond in real-time to monitoring of the hardware and environment, and the complex timing issues among parts of the system. Opportunities for unanticipated process interactions will grow in future missions with advances in hardware, distributed architectures, and "smart" science instruments.

As the opportunity for process concurrency increases, the ability to perform constraint checking on these concurrent processes also must increase. In order to detect hazardous error-recovery scenarios, an improved capability to model and analyze precedence, timing, and data-dependency constraints is needed.

Timing constraints usually have been defined in terms of periodic actions or deadline requirements. This definition is inadequate to model the timing constraints on spacecraft commands. The work described here extends the definition of timing constraints to represent allowable intervals between commands and the execution times of activities initiated by commands. These extensions of recent research results allow more accurate modeling of the required ordering and timing relationships among commands.

The model represents timing, precedence, and data-dependency constraints on spacecraft commands by means of a labeled graph in which the nodes represent commands and the edges represent constraints on those commands. This constraints graph, together with a number of potentially concurrent software processes (including the error-recovery processes and the current sequence of uplinked commands), are input to an algorithm called the Constraints Checker (see Fig. 1). The algorithm tests each edge (representing a constraint) in the constraints graph to determine whether any interleaving of the commands in those processes can fail to satisfy the constraint.

Any documented constraint which is not satisfied in every interleaving of the concurrent processes is recorded in a file for later analysis. The analysis may lead to a change in the documented constraint or to a change in the software. The Constraints Checker is then run again on the corrected input to verify that the constraint can no longer be violated during error recovery.

This work was performed in the context of the Galileo spacecraft, an interplanetary probe currently journeying to Jupiter. Preliminary results were reported in [8] and additional results in [9]. Ongoing research indicates that the results are applicable to a variety of asynchronous systems with precedence and critical timing constraints.

ERROR RECOVERY

System-level software onboard the spacecraft monitors and responds to failures. The standard definitions are used here of a *failure* as an event in which the behavior of the system deviates from its specification and of an *error* as an incorrect state of the system which must be remedied [12]. The error-recovery processes include those that respond to a loss of uplink or downlink communication, to thermal, power, or pressure anomalies, and to indicators regarding the health of the computers.

Instructions called *commands* instruct the different subsystems of the spacecraft to take specific actions at specific times. Error-recovery processes, composed of commands, are invoked in response to a detected failure. Individual commands are also assembled into groups of time-tagged commands, called *command sequences*, which are periodically sent to the spacecraft from the ground. Command sequences are stored temporarily in the spacecraft's memory until the time comes for each command to execute.

Some command sequences are so critical to the success or failure of the spacecraft's mission that they are labeled *critical sequences*. The command sequences used at launch or to direct Galileo's science and engineering activities at Jupiter are examples of critical sequences.

Should a failure occur during a non-critical sequence, a computer may cancel its activity. A critical sequence, however, must continue to execute even during error recovery. This requirement for concurrent execution of a critical sequence and error recovery drove much of the work presented here.

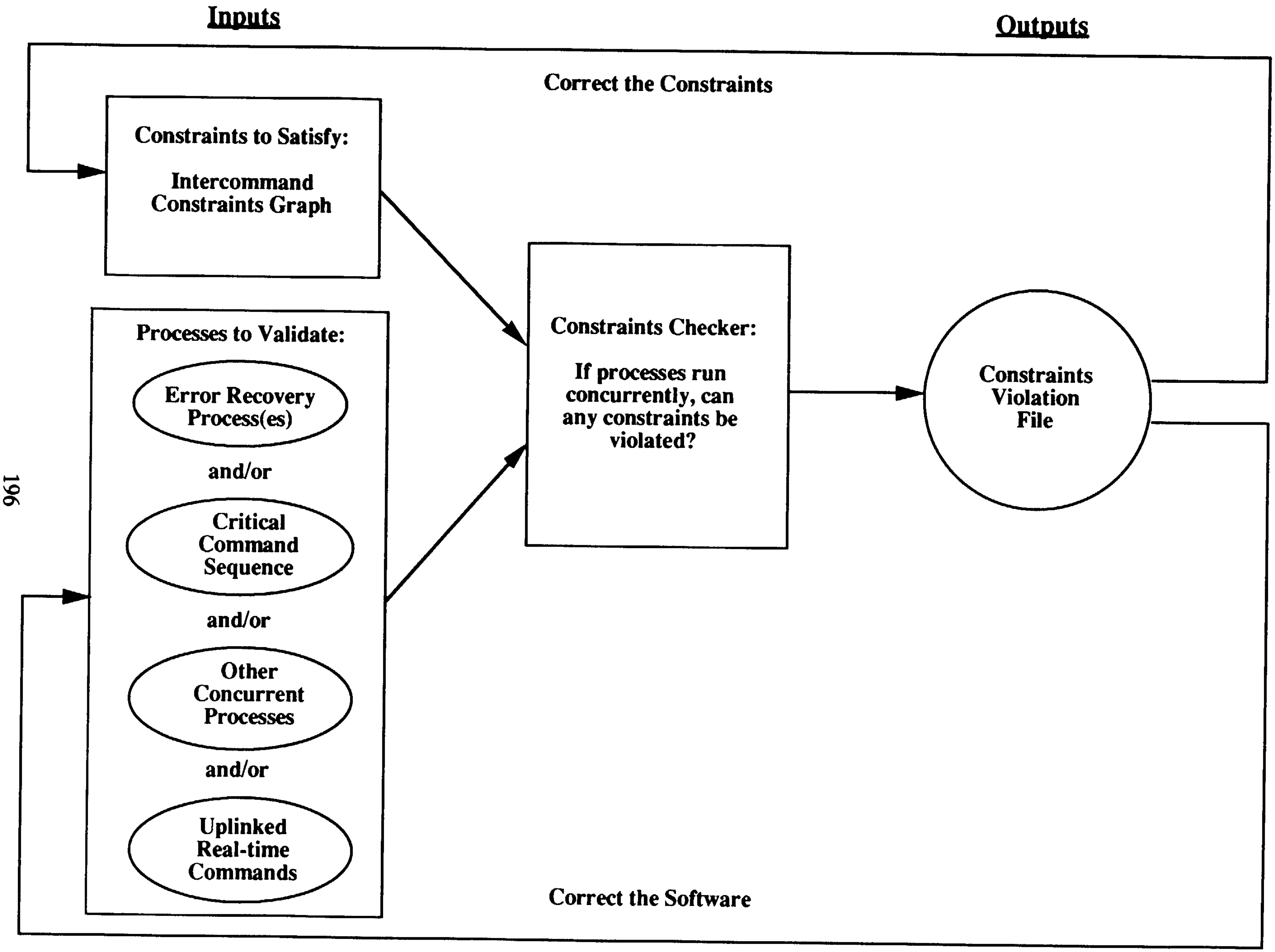

Inputs
Outputs
Correct the Constraints
Constraints to Satisfy:
Intercommand Constraints Graph
Processes to Validate:
Error Recovery Process(es)
and/or
Critical Command Sequence
and/or
Other Concurrent Processes
and/or
Uplinked Real-time Commands
Constraints Checker:
If processes run concurrently, can any constraints be violated?
Constraints Violation File
Correct the Software

Figure 1. Functional Block Diagram of Constraints Checker

Constraints are imposed on the commands in an effort to preclude conflicting interactions among the possibly concurrent processes. Some commands can interfere with the effect of other commands if they are executed too closely or too far apart in time. Certain commands must precede or follow other commands to accomplish the desired action. Some commands change the values of parameters used by other commands. Commands relating to power or propellant usage, to temperature or attitude control, to spacecraft or data modes, can endanger the collection of scientific data, a subsystem, or the spacecraft if intervening commands issued by another process leave the spacecraft in an unexpected state.

To restrict the interactions that are allowed, constraints are placed on the interleavings of commands. The flight rules as well as other documented system and operational specifications forbid certain interactions as unsafe. The constraints also describe ordering (precedence) and timing relationships between commands that must be maintained even when processes containing those commands execute concurrently.

Detecting undesirable interactions involving error-recovery processes is especially important since error-recovery software usually executes only when a failure has already been detected onboard the spacecraft. If a critical sequence is executing, the software must quickly (thus, autonomously) reconfigure the spacecraft to the state that is the precondition for the next activity in the sequence.

In addition, because error-recovery software is responsive, it is asynchronous [3]. It may begin execution at any time. In fact, because the spacecraft often is most taxed during the most critical science activity, error-recovery software is most likely to execute when the spacecraft is active. Error-recovery capabilities not only increase the number of concurrently executing processes, but also tend to be executed at the busiest (in terms of process interactions) times. It is also the case that hardware failures due to physical damage tend to be clustered in time [10].

APPROACH

The problem addressed here is how to check that the concurrent execution of the asynchronous processes that cooperate during error recovery satisfy the precedence constraints, maintain data-dependency (read/write) constraints, and satisfy the timing constraints.

The model which was developed allows the command's duration (the length of time required to execute the activity initiated by a command) to be attached to a command. Thus, a constraint of the form, "If command c_j occurs, then the activity initiated by the occurrence of commmand c_i must first have completed," can be represented. This type of constraint is common on both spacecraft and other real-time systems. If the duration of the commanded activity is variable, it is limited by a worst-case time which is represented in the model.

The model takes into account both precedence constraints and timing constraints. Precedence and timing are fundamentally different in that precedence does not require a notion of duration [11]. Most methods that currently exist to model precedence constraints do not incorporate timing requirements and so are inadequate for modeling the timing constraints on spacecraft. Similarly, many techniques that are currently available to model timing constraints tend to ignore precedence constraints.

A wide variety of powerful formalisms exists to model the specifications and behavior of real-time systems. See, e.g., [1, 4, 5, 6, 7, 13, 14, 15]. However, none of the available methods readily translates to the domain of validating error recovery on spacecraft. Some techniques consider both timing and precedence constraints but define timing constraints only in terms of periodic events (e.g., sampling rates), fixed execution times for events, and deadline or timeliness requirements. This provides too limited a model for the aperiodic and interval timing constraints on spacecraft commands.

Requirements are often modeled in terms of a lower time bound (after which an event may occur) and an upper time bound (by which the event must occur). In contrast, on the spacecraft there is often a need to model an operational constraint such as an interval within which an event is permitted to occur (but perhaps won't) and outside of which interval the event must never occur. This distinction between a "hard" time constraint (an interval within which the action should be taken) and a "soft" time constraint (an interval within which the action may occur) [2] is often absent in formal models.

The work described here brings together the study of real-time constraints with the study of precedence and data-dependency constraints.

MODELING THE CONSTRAINTS

The constraints that exist at the command level on the asynchronous execution of the spacecraft error recovery are modeled via a constraints graph. A constraints graph is a directed graph G=(V,E) in which each node c_i , c_j , ε V is labeled by a command and each edge e ε E is labeled by a constraint. The edge (c_i , c_j) is drawn as in Fig. 2.

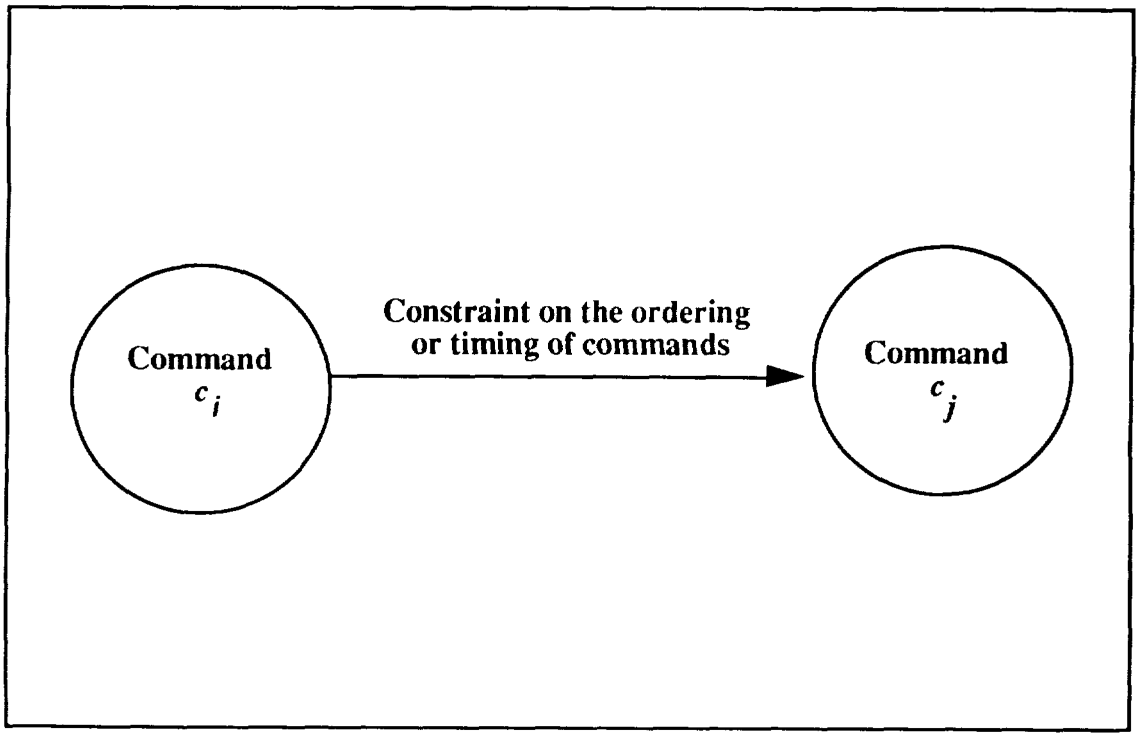


Figure 2. Modeling an Intercommand Constraint

Intercommand constraints are rules that govern the ordering or timing relationships between commands. There are three main types of intercommand constraints: timing constraints, precedence constraints, and data-dependency constraints. Fig. 3 presents an example of each of these three constraint types.

Intercommand timing constraints are safety properties. They impose a quantitative temporal relationship between the commands by asserting that "every c_i can only occur with timing relationship τ to c_j."

Precedence constraints enforce an ordering on the commands and so involve functional correctness, a concern of safety properties. Precedence constraints also involve liveness properties since they assert that if one command occurs, then another command must precede it: "If c_j occurs, then a c_i must precede it."

If a timing constraint exists between commands c_i and c_j, either command can legally occur alone. However, if a precedence constraint exists requiring command c_i to precede command c_j, then c_j cannot occur in isolation from c_i.

Data-dependency constraints involve restrictions placed on the order of commands when two or more processes access the same variable and at least one process changes the value of the variable. In such cases a concurrent execution of the processes can lead to a result different from the sequential execution of the processes. To forestall the data inconsistency that could result from this, a data-dependency constraint is used to specify the order in which the commands that read/write the variable must occur. Such a data-dependency constraint is represented as a precedence constraint.

CONSTRAINT CHECKING

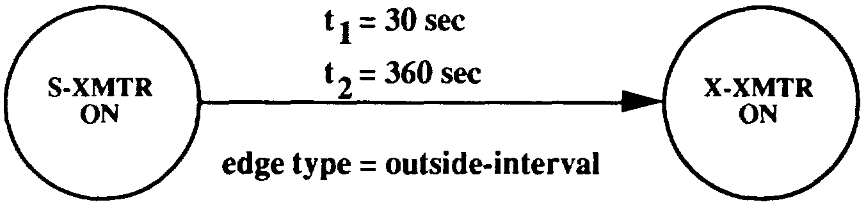
The constraints graph and a set of processes (time-tagged lists of commands) are input to the Constraints Checker. The Constraints Checker fixes one process' timeline and determines the range of start times that the fixed timeline and the constraint represented by each edge impose on the other processes. Each edge type is associated with an algebraic predicate which relates the time of issuance of the commands which are the edge's endpoints to the constraint represented by the edge. The Constraints Checker tests whether the appropriate predicate holds for each edge in the constraints graph. An edge which fails to satisfy the required predicate is flagged. In this case some scheduling of the processes can cause the constraint represented by that edge to be violated.

The Constraints Checker makes an assertion about the allowable start times of the process whose timeline is not fixed. It makes this assertion based on the edge type currently being surveyed, on the constraint represented by the edge, on the offset between the processes' start times and their issuances of c_i and c_j, and on the fixed start time of one process. If the Constraint Checker's assertion concerning when the other processes should start is inconsistent with the user-provided range of start times, then the edge is flagged.

If the edge is flagged due to erroneous information in one of the edge or node labels, the user can readily correct the input data and run the Constraints Checker again to verify the adequacy of the correction. If the edge is flagged due to a problem with the existing error-recovery response, the data output with the flagged edge helps the user identify the problem. The user responds by shifting the processes' timelines or by curtailing the concurrency that allowed the intercommand constraint to be violated. The goal is to adjust or limit the concurrent execution of the processes so that the edge will not be flagged in a subsequent run.

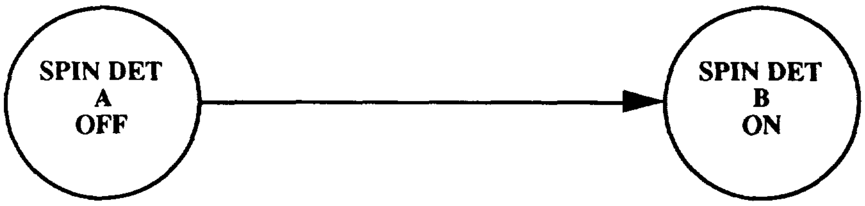
(a) Timing Constraint:

Documented constraint: "The time separation between powering on the S-Band Transmitter and powering on the X-Band Transmitter shall be either less than one-half minute or greater then 6 minutes."



(b) Precedence Constraint:

Documented constraint: "Spin Detector B can only be powered on after Spin Detector A is turned off."



(c) Data-Dependency Constraint:

Documented constraint: "The Commanded-Maneuver-Status Variable must be updated by the command sequence before the thruster burn. The error-recovery process reads the Commanded-Maneuver-Status variable in case of a thruster burn failure. The update of the variable by the command sequence must precede the use of the variable by the error-recovery process."

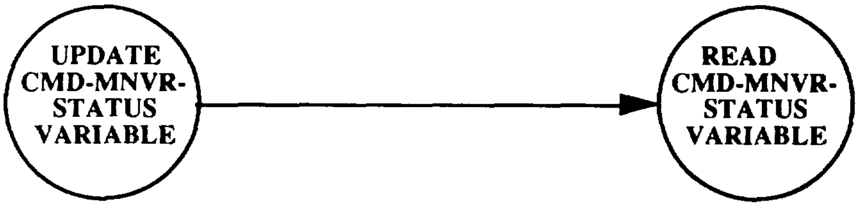

UPDATE CMD-MNVR-STATUS VARIABLE
READ CMD-MNVR-STATUS VARIABLE

Figure 3. Examples of Intercommand Constraints

After all the edges have been surveyed, the Constraints Checker computes for every distinct pair of processes the range of safe start times of one process relative to the other. Within this time interval the intercommand constraints are satisfied.

RESULTS

The error-recovery scenarios chosen to evaluate the Constraints Checker involved failures during the execution of the critical sequence of commands controlling the Galileo spacecraft's arrival at Jupiter. The activities of the processes that must cooperate during error recovery are highly constrained due to the complexity and time criticality of the engineering and science activity during the planetary encounter. There are thus many opportunities for constraint violations during error recovery.

Experience with this method indicates that the Constraint Checker's major benefit is early detection of significant inconsistencies between the design and the constraints. The error-recovery scenario that was analyzed addressed the complicated timing issues involved when a critical sequence of commands had to continue execution during recovery from a hardware anomaly. The constraints graph input to the Constraints Checker consisted of 40 nodes and 40 edges (20 precedence and data-dependency edges and 20 timing edges).

The results were as follows. The Constraints Checker flagged seven intercommand constraint violations (four precedence constraints and three timing constraints). One precedence edge was flagged because a global variable could be used before it was updated. Two other edges were flagged because a specific command that was required to precede another did not occur. (One of these violations was later traced to outdated documentation). Another flagged edge was, in fact, not violated but appeared to be, since the associated command appeared six hours earlier.

Two of the three timing edges that were flagged involved a timing discrepancy between the requirements and the code. The third timing edge that was flagged resulted from the unforeseen consequence of a data field taking on a value which was possible, but forbidden in operations.

Another eight errors, involving incorrect or inconsistent documentation, were identified during construction of the constraints graph. Seven of these eight errors were significant enough to have caused inaccuracies in the constraints graph and in the results of the Constraints Checker.

The intercommand constraint violations flagged by the Constraints Checker involved either discrepancies between the constraints and the code or unforeseen consequences of unlikely but possible error-recovery scenarios. The Constraints Checker appears to be useful in enhancing the developers' ability to visualize abnormal scenarios and in flagging constraint violations that occur only in some subset of the possible error-recovery responses.

The code analyzed here was a baseline version, rather than the most current flight version. This choice was made in order to provide code that had been well thought out but not tested. It is at this intermediate stage of the development process, when the intercommand constraints have been initially documented but the details of the design and the timing are still evolving, that the Constraints Checker may be most effective.

A version of the Constraints Checker underwent initial development for use on the later-canceled Comet Rendezvous/Asteroid Flyby (CRAF) spacecraft. The Constraints Checker's function was to serve as a software-development tool to analyze and validate scenarios involving the interactions among concurrent processes during error recovery.

POTENTIAL APPLICATIONS

The method outlined here is useful for addressing related problems in domains other than spacecraft. A primary concern in evaluating the safety of a complex, embedded system is whether error recovery can result in the violation of constraints on that system. More generally, the design of event-driven systems often involves analyzing whether the concurrent execution of processes with unpredictable start times can jeopardize the system. Such issues are readily investigated with this method. The Constraints Checker can also function as an extension to existing simulation or CASE tools to provide more accurate validation of complex timing constraints.

The Constraints Checker is also well suited to operational situations in which a portion of the control software is regularly replaced. On the spacecraft the sequences of commands are examples of this "temporary" software. On the space station, for example, procedures to sequence activities outside the astronauts' responsibilities will be regularly uplinked from the ground. The proposed method could ease the operational difficulties of quickly checking that new or temporary software is safe and will not conflict with the "permanent" software.

Commercial applications of this method could range from safety-critical process-control systems to event-driven flight control or command-and-control systems. Current tools often model and test only periodic or deadline timing constraints. The proposed method offers the capability to quickly and accurately model and check that even aperiodic and interval constraints among events will always hold in the system.

CONCLUSIONS

This paper has shown how to construct a constraints graph that models precedence, timing, and data-dependency constraints. The constraints graph provides a means of visualizing the intercommand constraints that must be satisfied by every concurrent execution of processes during error recovery. This paper has also described a constraint-checking algorithm that, given a constraints graph and a set of processes, detects violations of the modeled constraints.

The Constraints Checker is designed specifically to help answer the question of whether existing system-level error recovery is adequate. It offers a flexible, embeddable, and relatively simple alternative to simulation of error-recovery scenarios. In the context of the spacecraft, the algorithm identifies the unexpected effects resulting from the interleaving of error-recovery processes and mission-critical sequences of commands. In a broader context, the research presented here is part of an ongoing effort to investigate the behavior of concurrently executing processes subject to precedence and timing constraints.

ACKNOWLEDGEMENT

The first author thanks Chris P. Jones of the Jet Propulsion Laboratory for many helpful insights.

The work of the first author was started at Iowa State University, supported by grant NGT-50269 from the National Aeronautics and Space Administration, and was completed at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

REFERENCES

[1] R. Alur, C. Courcoubetis, and D. Dill, ``Model-Checking for Real-Time Systems," in *Proceedings of the 5th Annual IEEE Symposium on Logic in Computer Science*, Los Alamitos, CA: IEEE Computer Science Press, 1990, pp. 414-425.

[2] A. A. Bestavros, J. J. Clark, and N. J. Ferrier, ``Management of Sensori-Motor Activity in Mobile Robots," in *Proceedings of the 1990 IEEE International Conference on Robotics and Automation*. Los Alamitos, CA: IEEE Computer Society, 1990, pp. 592-597.

[3] R. H. Campbell and B. Randell, "Error Recovery in Asynchronous Systems," *IEEE Transactions on Software Engineering*, vol. SE-12, August 1986, pp. 811-826.

[4] S. Cheng, J. A. Stankovic, and K.Ramamritham, "Dynamic Scheduling of Groups of Tasks with Precedence Constraints in Distributed Hard Real-Time Systems," *IEEE Real-Time System Symposium*. New Orleans, LA: 1986, pp. 166-174.

[5] J. E. Coolahan, Jr., and N. Roussopoulos, "A Timed Petri Net Methodology for Specifying Real-Time System Timing Requirements," *International Workshop on Timed Petri Nets*. Silver Springs, MD: IEEE, 1985, pp. 24-31.

[6] T. A. Henziger, Z. Manna, and A. Pnueli, "Temporal Proof Methodologies for Real-Time Systems," in *Proceedings of the 18th ACM Symposium on Principles of Programming Languages*, 1991, pp. 353-366.

[7] F. Jahanian and A. K.-L. Mok, "Safety Analysis of Timing Properties in Real-Time Systems," *IEEE Transactions on Software Engineering* vol. SE-12, Sept. 1986, pp. 890-904.

[8] R. R. Lutz and J. S. K. Wong, "Validating System-Level Error Recovery for Spacecraft," *AIAA Computing in Aerospace* 8, vol. 1. Washington: AIAA, 1991, pp. 69-76.

[9] R. R. Lutz and J. S. K. Wong, "Detecting Unsafe Error Recovery Schedules," *IEEE Transactions on Software Engineering*, vol. 18, no. 8, August 1992, pp. 749-760.

[10] J. D. Northcutt, *Mechanisms for Reliable Distributed Real-Time Operating Systems, The Alpha Kernel.*. Boston: Academic Press, 1987.

[11] A. Pnueli and E. Harel, "Applications of Temporal Logic to the Specifications of Real Time Systems," in *Formal Techniques in Real-Time and Fault-Tolerant Systems.*, Ed. M. Joseph. Berlin: Springer-Verlag, 1988, pp. 84-98.

[12] B. Randell, P. A. Lee, and P. C. Treleaven, "Reliability Issues in Computing System Design," *ACM Computing Surveys*, vol. 10, June 1978, pp. 123-166.

[13] F. H. Vogt and S. Leue, "The Paradigm of Real-Time Specification Based on Interval Logic," in Proceedings of the Berkeley Workshop on Temporal and Real-Time Specification, Eds. P. B. Ladkin and F. H. Vogt. Berkeley, CA: International Computer Science Institute TR-90-060, pp. 153-178.

[14] J. M. Wing, "A Specifier's Introduction to Formal Methods," *Computer*, vol. 23, Sept. 1990, pp. 8-26.

[15] J. Xu and D. L. Parnas, "Scheduling Processes with Release Times, Deadlines, Precedence, and Exclusion Relations," IEEE Transactions on Software Engineering, vol. 16, March 1990, pp. 360-369.

SPINOFF TECHNOLOGY: ENGINEERING AND SCIENTIFIC COMPUTER CODES

This paper was withdrawn from presentation

FAILURE ENVIRONMENT ANALYSIS TOOL APPLICATIONS

Ginger L. Pack
NASA Johnson Space Center
Houston, Texas 77058

David B. Wadsworth
Lockheed Engineering and Sciences Company
Houston, Texas 77058

ABSTRACT

Understanding risks and avoiding failure are daily concerns for the women and men of NASA. Although NASA's mission propels us to push the limits of technology, and though the risks are considerable, the NASA community has instilled within it, the determination to preserve the integrity of the systems upon which our mission and, our employees lives and well-being depend. One of the ways this is being done is by expanding and improving the tools used to perform risk assessment. The Failure Environment Analysis Tool (FEAT) was developed to help engineers and analysts more thoroughly and reliably conduct risk assessment and failure analysis. FEAT accomplishes this by providing answers to questions regarding what might have caused a particular failure; or, conversely, what effect the occurrence of a failure might have on an entire system. Additionally, FEAT can determine what common causes could have resulted in other combinations of failures. FEAT will even help determine the vulnerability of a system to failures, in light of reduced capability. FEAT also is useful in training personnel who must develop an understanding of particular systems. FEAT facilitates training on system behavior, by providing an automated environment in which to conduct "what-if" evaluation. These types of analyses make FEAT a valuable tool for engineers and operations personnel in the design, analysis, and operation of NASA space systems.

INTRODUCTION

FEAT was developed as part of an effort to find ways to better identify and understand potential failures that threaten the integrity of NASA systems. Past and current methods of failure assessment consists of developing often enormous amounts of documentation in the form of Failure Mode Effect Analysis (FMEA) worksheets. Engineers create these worksheets by attempting to exhaustively enumerate potential system failures and consequences. Hazards analysis is performed in a similar manner; experts are gathered together and are asked to brainstorm about the hazardous manifestations of various failures. System knowledge and experience are necessary for ensuring the comprehensiveness of this approach. However there are troubling drawbacks to this technique. First, there exists the difficulty of anticipating every scenario. Analysis is also inherently constrained by the limits of actual experience. Further, such methods lack consistency and do not enforce a standard level of coverage. Although there is certainly much to be credited to knowledge acquired through experience, it is not sufficient to avoid unanticipated interactions which may lead unexpectedly to undesirable consequences. As many industries have learned, sometimes experience comes at too high a cost. Those at NASA have been looking for better ways to anticipate failure and for tools to assist in "designing out" potential problems. FEAT was developed to address this problem.

TECHNICAL APPROACH

FEAT is a software application that uses directed graphs or, digraphs, to analyze failure paths and failure event propagation. The behavior of the systems to be analyzed is represented as a digraph. Then, the digraph model of the system, is used by FEAT to answer questions concerning the cause and effects of events which are captured in the model. Therefore, the first step in using FEAT is to create the digraph model of the system in which one is interested. Once FEAT has analyzed the digraph, it has the information it needs to perform cause and effect analysis.

What are digraphs? Directed graphs are graphs that consists of a set of vertices and a set of edges, where there is an edge from one vertex *a* to another vertex *b* . The vertices are drawn as circles and the edges are drawn as arrows. The direction of the arrows indicates a causal relationship between the vertices (see figure 1). The vertex from which the edge begins, is called its source; and the vertex at which the edge terminates, is called its target. Direct graph theory is an accepted and established area of mathematical study. Therefore we will only introduce it in this paper, to the extent necessary for an understanding of how it is used in FEAT. The interested reader may find further information by consulting the literature.

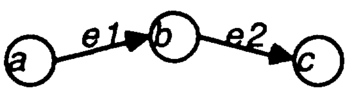


Figure 1

The structure of the digraph can be represented by a matrix, and consequently can be easily implemented in a computer. The conversion from digraph to matrix is straightforward and is illustrated below in figure 2. This matrix is called the *adjacency* matrix (reference 1), and is the basis from which other information about the graph can be derived. The matrix of the graph is obtained by entering either zero or one, depending on whether or not an edge connects two vertices. The presence of an edge from a to b in figure 1, indicates an entry of one (1) into the corresponding matrix entry. However, since there is no edge from a to c, a zero (0) would be entered in the corresponding matrix entry.

$$\begin{array}{cccc} & a & b & c \\ a & 0 & 1 & 0 \\ b & 0 & 0 & 1 \\ c & 0 & 0 & 0 \end{array}$$

Figure 2

Additional information can be added to the digraph, by applying logical operators to express conditional statements. FEAT uses AND and OR operators to accomplish its analysis. The AND operator is represented on the graph as a vertical bar with a horizontally placed arrow at its center. An OR operator is simply two or more edges whose target is the same vertex. Theses operators [figure 3], and their use in FEAT [figures 4 & 5], are described below.

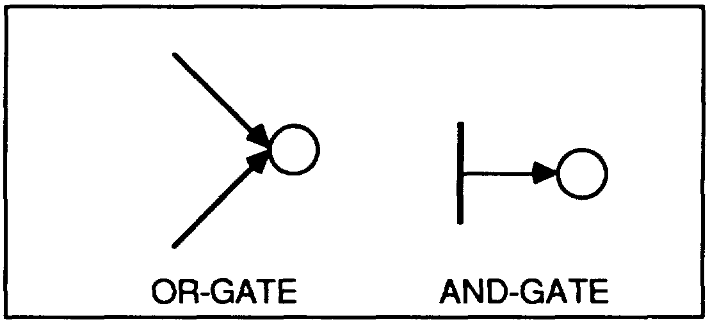


Figure 3

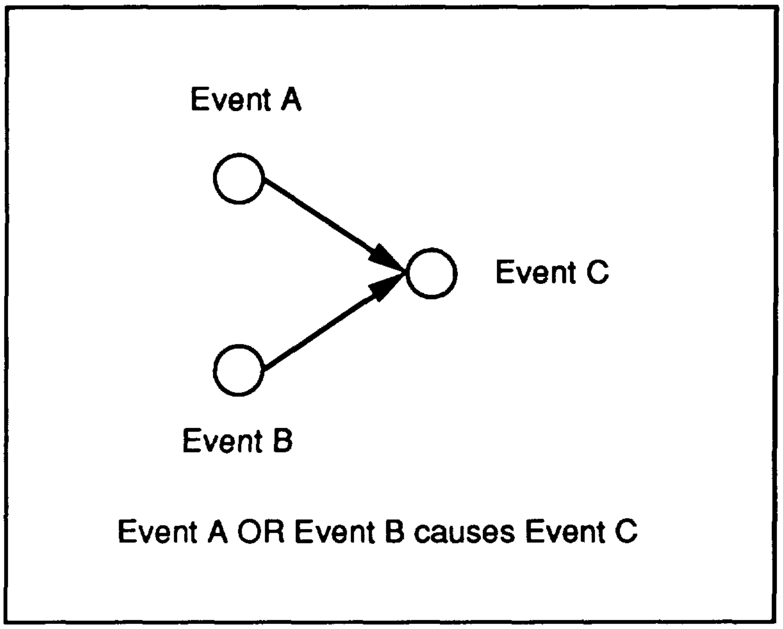


Figure 4

The "AND" gate is shown in Figure 5. The AND gate is used when both event A and Event B must occur in order for Event C to occur. Conversely, if only Event A occurs or, if only Event B occurs, then Event C does not occur.

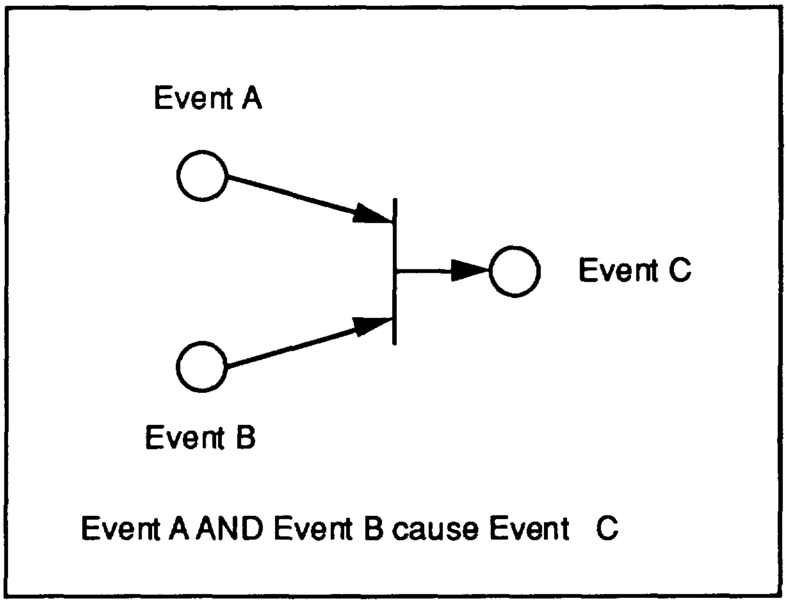


Figure 5

Analytical Capabilities The *reachability* of an event refers to whether there is a path by which other events in the digraph can be reached. A given event is said to reach another, if the first event can cause the second through some path of the graph. Using the adjacency information derived from the digraph, reachability can be computed for every event and pair of events in the digraph. Analysis can be conducted upstream or downstream from an event node. (References 2, 3 and 4 provide a much more detailed discussion of digraphs and reachability.)

Reachability information allows FEAT to answer the following questions about a modeled system:

A. What happens to the system if "Event A (and Event B and Event C and ...)" occurs?

B. What are the possible causes of "Event A"?

C. What common cause could account for the simultaneous indication of numerous events?

D. What is the susceptibility of the system to new events given that one or more events has already occurred, or the system has been reconfigured due, for example, to maintenance?

Digraph Example The following example demonstrates how a digraph might be implemented for a light and switch. The digraph provides a methodical way in which to express the topology and behavior of a system. It is worth noting that the digraph itself may have various constructions for the same information contained in it, depending on who created it. Different modelers may lay out the digraph differently. However, for a properly constructed digraph, the same information will be captured. In the following example [figures 6 & 7], power source A provides current to switch A which connects to the bulb. Similarly, power source B can energize the bulb.

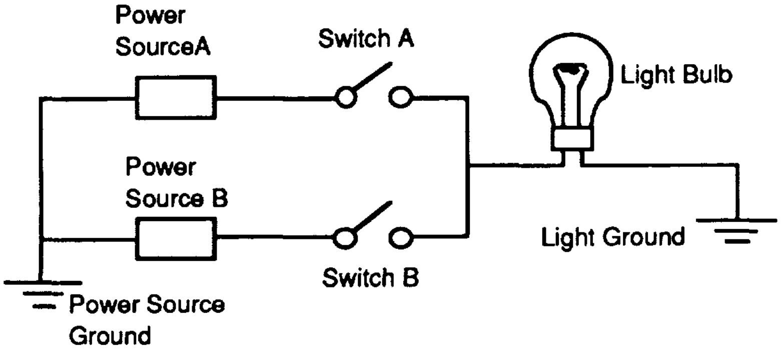


Figure 6
Light bulb and Power Source Schematic

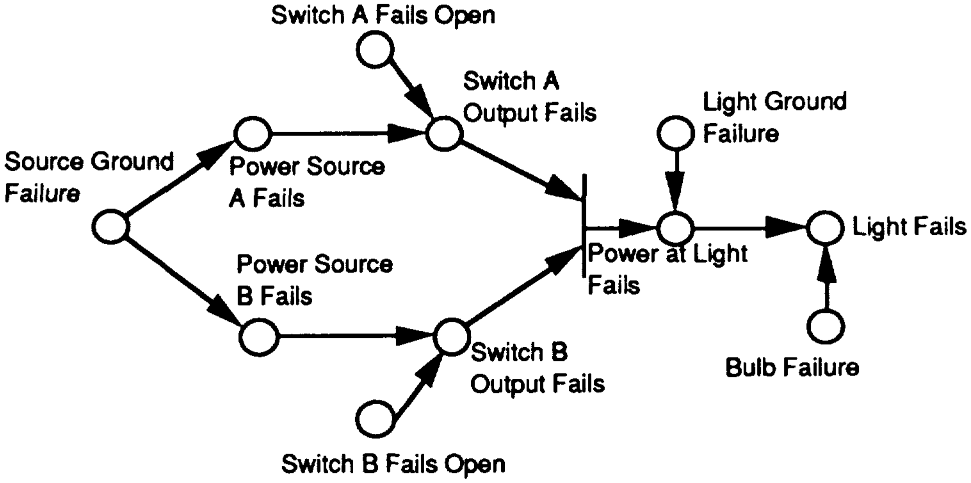


Figure 7
Digraph of Light bulb Schematic

- If "Power Source A Fails" or "Switch A Fails Open" then "Switch A Output Fails". This is an example of OR logic and is shown in the digraph by the arrows leading into "Switch A Output Fails".

- If output from both switches A and B fail, then they will cause the "Power at Light to Fail". This logic appears as an **AND** gate on the digraph (the vertical line). In this case, the **AND** gate reflects redundancy designed into a system.

Why digraphs?

Directed graphs are useful because they visually depict the logical topology and dependency relationships of physical and conceptual systems and processes. Because they capture causal effects between events, they can be used to describe system behavior. Directed graphs are also easily converted into a matrix and, because of this, can be readily analyzed in a computer. Creating and laying out the digraph of a system, also formalizes the method of evaluation during the analytical process, and provides a standard representation convention. Finally, digraph analysis is mathematically sound, since methods for determining connectivity paths of the digraph vertices can be mathematically proved.

DIRECTED GRAPHS AND FEAT

Digraph construction is facilitated by use of an editor specifically designed for the task. Such an editor is included in the FEAT package which consists of two programs: Digraph Editor and FEAT.

Digraph Editor

The Digraph Editor facilitates construction of the digraph model by allowing the user to create event nodes, edges, and the logic operators, and to connect and arrange them into a digraph. Event nodes and edges are laid out and connected using the logic operators. The pieces that make up a digraph are supplied in a digraph toolbox from which items may be selected. These items are placed on the screen and arranged to produce the system digraph.

Other information is needed to complete the digraph and to make it usable by FEAT. Event nodes have an associated text block, which includes information that will identify the event node to FEAT, describe the event for the user, and relate the event to a drawing which contains the component to which the event pertains. This information is extracted from tables that the user creates. Digraph Editor uses the tables to automatically generate a mnemonic reference that FEAT will use to identify the event.

Digraph Editor also provides a number of tools for validating and verifying the model as it is being developed. Digraph Editor will check tables for duplicate entries, check nodes for incorrect form, and determine whether a selected node has a duplicate in the digraph. Digraph Editor also contains an algorithm that allows the user to analyze small or incomplete digraphs, while still in the editor. Once the digraph is completed and the paths in it are analyzed, FEAT can return answers to questions regarding the behavior of the modeled system.

Currently, digraph models are created manually by selecting and arranging digraph components; the modeler must interpret drawings and other sources of information to generate the digraphs. This is a laborious task. Consequently, efforts are underway to develop methods to automatically translate schematics and drawings into corresponding digraph models.

Digraph Editor is currently only available for the Macintosh II class of computer.

FEAT

FEAT is the portion of the package that analyzes single or multiple digraphs, and graphically displays causes and effects of events. Propagation results are shown both on the digraphs and on an another associated graphical representation, such as a schematic or block diagram. FEAT uses a multi-step algorithm, described in Reference 2, to compute reachability for each event and pair of events in the digraphs. Events are identified to FEAT through the mnemonic that is generated by Digraph Editor. Queries about the behavior of the system are made by selecting events and telling FEAT to return all of the causes of that event (targeting), or by telling FEAT to return all of the effects of that event (sourcing). FEAT displays all of the single events, and all pairs of events that may cause a selected event. Multiple events may also be selected and analyzed. FEAT allows some events to be temporarily removed from the analysis so that answers can be obtained about a reconfigured system.

FEAT also contains a feature which allows users to attach to a schematic, formatted database information and graphics. In this way, component descriptions, parts lists, drawings, etc, may be displayed in conjunction with a schematic.

One of the major advantages of FEAT, as discussed in Reference 2, is that it allows the analysis of very large systems. Large systems can be digraphed by creating and connecting a series of smaller digraphs. FEAT understands when propagation occurs across the digraphs.

Planned enhancements to FEAT include the following: increasing the speed with which reachability is computed by improving FEAT's computational algorithm; provision of a method for computing and displaying probabilities of events occurring; and computation and display of the time it takes for an event to propagate through the graph.

FEAT is currently available for the Macintosh II class of computer and for UNIX/X-Windows/OSF-Motif systems. No programming skill is required to use FEAT. However, a course in digraph modeling is quite helpful in learning how to construct system models.

DIGRAPHS AT NASA

Why NASA chose digraphs

NASA's interest in digraphs began as part of the Shuttle Integrated Risk Assessment Project (SIRA). SIRA was initiated in the wake of the Challenger accident, in an effort to find better ways of assessing risk and preventing failure. Digraphs support such analysis by providing end to end cause and effect analysis of modeled systems. Digraphs also provide a standard and methodical approach for conducting safety analysis and risk assessment. Digraphs capture information in an easily retrievable format, and facilitate the transfer of design information. FEAT takes advantage of these characteristics in a way that aids engineers and analysts with design, assists safety engineers with risk assessment, and promotes understanding of system behavior, thereby making FEAT a good tool for training inexperienced persons.

What has been done at NASA?

The first system to which digraph analysis was applied was the Space Shuttle Main Engine System (SSME). Since then, acceptance of digraphs and the use of FEAT has extended in several directions. Most recently, FEAT has been formally released to the Space Station Freedom Program (SSFP) Technical Management Information System (TMIS), as Digraph Data System (DDS) Release 1.0. DDS will, through TMIS, be available to SSF Engineering and Integration, SSF Combined Control Center, and the various Work Packages and their contractors. A Macintosh Powerbook version of FEAT will be deployed as a Development Test Objective (DTO) on the STS-52 flight scheduled for October 1992. Reliability and Maintainability personnel at NASA-JSC, are using FEAT to construct a model of the Simplified Aid for Extra-Vehicular Activity (EVA) Rescue (SAFER). FEAT is also being used to model the redesigned Servo Power Amplifier (SPA) for the Remote Manipulator System (RMS).

Proponents have used FEAT for a variety of analytical tasks, such as Fault Tolerance Analysis and Redundancy Management (FT/RM), Fault Detection, Isolation, and Recovery (FDIR), and "What-If" analysis. Within the Space Station Freedom Program, FEAT is being used in the performance of Integrated Risk Assessment for the station, which includes Failure Mode and Effects Analysis (FMEA), Hazards Analysis (HA), and FT/RM. FEAT has also been established as a baselined tool in the Mission Operations Combined Control Center, where flight controllers will use FEAT models to assist with real-time monitoring tasks. FEAT's role is expanding in both Space Station and in Space Shuttle.

Space Station The Space Station Engineering Integration Contractor (SSEIC), is using FEAT to perform integrated risk assessment. This tasks consists of performing the analysis to assure the station design is safe, reliable, and has an acceptable level of risk (reference 5). The space station design consists of modules designed and built by the United States, and of modules which will be designed and built by NASA's international partners. The work to be performed by NASA is divided into four Work Packages distributed among different centers. Additionally, a variety of contractors are working in support of the Work Packages. Consequently, system integration is a paramount concern of the program. SSEIC is tasked with ensuring the integration of these various factions and is using digraph-based FEAT, to work the integration problem. Specifically, FEAT supports the following areas of the Integrated Risk Assessment process:

1. Reliability Analysis
2. Safety Analysis
3. Integrated Risk Analysis
4. Integrated Risk Assessment

The models being developed for the station Integrated Risk Assessment will eventually be provided to Mission Operations personnel for use in FDIR of the on-orbit station.

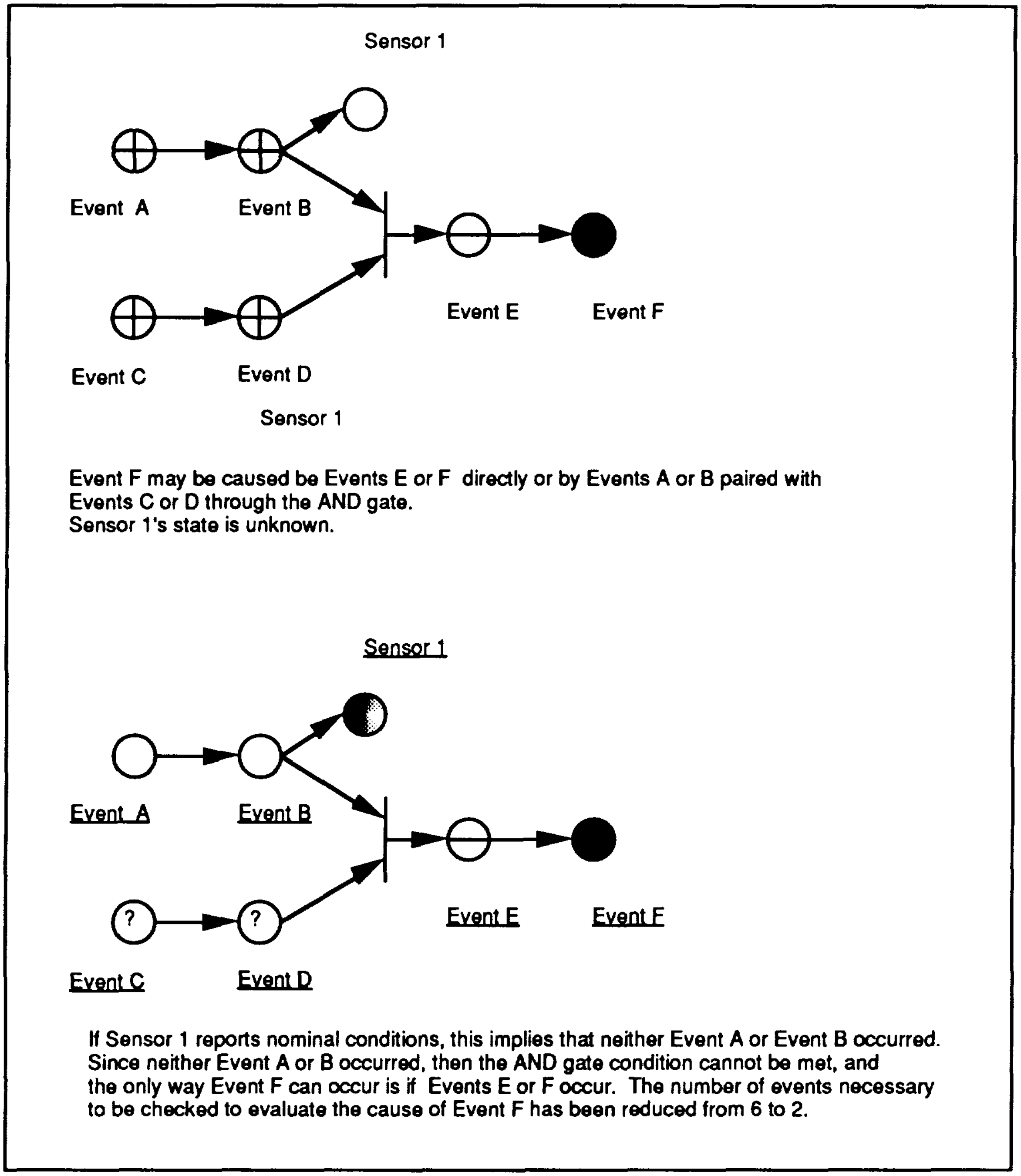
Space Shuttle FEAT is scheduled to fly on STS-52 as a Detailed Test Objective (DTO). A FEAT model of the S-band Communications System has been installed on an Apple™ Powerbook™, which will be flown aboard the shuttle. Astronauts will use the model to perform on-board fault isolation for the S-band Communication System. They will be able to configure the model to match the actual S-band system configuration, and then will use FEAT to identify possible causes of failures of the S-band system.

FUTURE APPLICATIONS OF DIGRAPHS

Digraphs are gaining acceptance, within the NASA community, as a viable method for conducting many kinds of analysis. Space Station Freedom Program and Operations, has mandated the use of digraph analysis for the Space Station Level II Integration effort; and many others are beginning to take up the banner. Some of the potential areas of application include the following:

Fault Isolation/Testability

FEAT's ability to model and analyze system failures make it a natural candidate for fault isolation efforts. If a failure event occurs, FEAT can display all of the possible single and paired causes for that event. However, in a large system, potential causes can be enormous in number. A method of pruning the list of possible causes is then necessary. Sensor information associated with the system can be used to remove candidate causes which occur downstream of a known nominal condition. Incorporation of sensor data, into the analysis, can help to reduce the number of candidate failures to a manageable sum. Then using traditional techniques, further isolation can be accomplished. Figure 8 shows an example of such a case.



Event F may be caused be Events E or F directly or by Events A or B paired with Events C or D through the AND gate.
Sensor 1's state is unknown.

If Sensor 1 reports nominal conditions, this implies that neither Event A or Event B occurred. Since neither Event A or B occurred, then the AND gate condition cannot be met, and the only way Event F can occur is if Events E or F occur. The number of events necessary to be checked to evaluate the cause of Event F has been reduced from 6 to 2.

Figure 8

Sensor data may also be combined with FEAT to identify the potential for cascading alarms. For instance, if a fault occurs downstream from a sensor, the sensors upstream will eventually alarm as a result of the fault. FEAT can show the effects of a fault on the downstream sensors.

This solution is being implemented by NASA, in an extension of FEAT, called Extended Real-time FEAT (ERF). ERF automatically prunes the list of possible faults, according to sensor information. ERF is being developed as a part of the FDIR system for the On-orbit Control Center Complex. Mission Controllers will use ERF to resolve off-nominal system behavior, by reducing the potential number of failure causes.

FEAT developers are pursing the possibility of incorporating, or interfacing with, a testability analysis tool which will help to evaluate sensor coverage in systems, and make recommendations regarding appropriate sensor locations. ERF is dependent upon adequate sensor information and proper placement of the sensors. Properly placed sensors provide information to quickly and accurately locate faults. The combination of FEAT, ERF and testability tools will make a very powerful fault isolation system.

Temporal Analysis

Not every event immediately affects the next downstream event. There may be appreciable delays within an event and between events. For example, an inappropriately shut valve, may not for some time, cause the pressure in the system to rise to an unacceptable level. In such a situation, time delay is an important aspect of calculating the potential failure space.

This issue will be addressed in FEAT when a modification is made to Digraph Editor to allow modelers to include time delays within events, and delays between events. FEAT will then compute the maximum and minimum time delay between selected events. This capability will be supplied in a future version of FEAT.

Software Modeling

Physical systems are not the only candidates for digraph analysis. Software functions and data flow can be modeled as well. Particularly, the flow and effect of invalid/improper data can be modeled. This can provide insight to the designer in determining mission critical software functions. Additionally, the effect of invalid data on other system functions (both software and hardware) may be shown. For instance, a software functional component that generates invalid data as an event; may then provide that data to other software and hardware as an invalid data input event. FEAT can be used to model these behaviors too.

Design Evaluation and Redundancy Management

Digraph models can be used to determine whether or not a system design provides sufficient redundancy. Maintenance and configuration effects on the system, can be evaluated by selectively removing (setting) components from the system. The reconfigured system can then be evaluated for induced single and paired events. This can be particularly useful in determining new vulnerabilities after a system has encountered failures and/or has portions of the system secured for maintenance.

FEAT contributes to design evaluation by rapidly displaying all single events caused by the event of interest, and all pairs of events that will result in that event. Unexpected single point common cause events are also quickly identified. As the design is modified to provide additional redundancy, the digraph model can be updated to reflect the changes, and the new set of single events and pairs of events can be evaluated.

Logistics Analysis

Logistics analysis addresses corrective and preventive maintenance tasks, and determines the kinds and numbers of repair parts needed for a system. This type of analysis is associated with the reliability and availability (reference 6), of systems. Reliability is defined as the measure of the mean time between failure (MTBF) and, concerns the probability that a system will operate over a specified period of time. No provision is made for repair when calculating reliability. Availability varies from reliability, in that it is a measure of the mean time to repair (MTTR), or, the probability that the system will operate over a period of time considering that something can be done to restore functionality lost as a result of a failure. How system repairs can be supported, or supportability, is important to determining availability. If repairs can be made instantaneously, availability is increased. However, long delays between failure and repairs makes the system proportionally less available.

FEAT models can help to identify critical components and the effect of their failure upon the system. Digraph models of the system can, along with specific part reliability, help to determine priorities for inventory stocks, and schedules for maintenance. Spare parts inventories are a major factor in determining supportability. For example, spares for parts that cause single point common cause events should have higher priority for stocking than parts that contribute to pairs of events.

Maintainability concerns the time it takes to remove and replace a component. Digraph models can identify components prone to low reliability, and single common cause failure. Designers can then either improve the reliability of the component or ensure that such items are accessible and easily replaced.

SUMMARY

As NASA continues to search for better and innovative approaches to new and old problems, directed graph analysis has emerged as an attractive expansion of the methods applied to Risks Assessment. Directed graphs are a well established area of mathematical study and analysis. Digraphs provide an easily comprehendible visual representation of cause and effect relationships. Conversion of the digraph to an equivalent matrix is straightforward, and allows analysis of digraphs to be mathematically calculated and verified. The nature of matrices also makes them ideally suited for computerized calculations, which in turn provides a vehicle for automating the task of risk assessment and failure analysis.

FEAT uses directed graph theory to provide engineers and analysts with a powerful and flexible automated analytic helper. FEAT can provide end to end analysis of cause and effect events. Very large systems can be modeled in modules, then connected to form the entire system. This feature also allows digraphs to be arranged in mix and match fashion. FEAT can detect and return information about single point failure vulnerability, failure event pairs, common cause events, and reduced capability analysis. FEAT shows the results of event propagation on system schematics and on the associated digraph. Digraph Editor provides a helpful way for the analyst to create digraphs.

The FEAT Project is funded by the NASA Space Station Freedom (SSF) Advanced Programs Development Office (Code MT) and the SSF Program Office (Code MS).

REFERENCES

1. L. Levy, *Discrete Structures of Computer Science*. John Wiley & Sons, 1980.
2. R. Stevenson, J. Miller and M. Austin. Failure Environment Analysis Tool (FEAT) Development Status. *AIAA Computing in Aerospace VIII, AIAA-91-3803*. October 1991.
3. I. Sacks. Digraph Matrix Analysis. *IEEE Transactions on Reliability, Vol. R-34, No. 5*. December 1985.
4. I. Sacks, G. Keller, and R. Rauch. Application of Digraph Matrix Analysis to the Space Station. *RDA , Logicon, R & D Associates, RDS-TR-148400-001*. September 1987.
5. J. Schier. Integrated Risk Assessment (IRA): Defining the Level II Safety & Reliability Job and Implementation Plan using Digraphs. *Unpublished Grumman presentation*. September, 1992.
6. B. Blanchard. *Logistics Engineering and Management*. 4th Edition. Prentis-Hall, Inc. Englewood Cliff, NJ. 1992.

BIBLIOGRAPHY

D. Haasl, N. Roberts, W. Vesely, F. Goldberg. *Fault Tree Handbook*. GPO Sales Program, U.S. Nuclear Regulatory Commission, Washington, DC. 1981.

J.Pearl, *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kauffman, 1988.

INFORMATION AND COMMUNICATIONS PART 5: COMPUTER SIMULATION, VIDEO, AND IMAGING TECHNOLOGY

Development of Interactive Multimedia Applications

Albert Leigh[1]
Lui Wang
Software Technology Branch/PT4
NASA Lyndon B. Johnson Space Center
Houston, TX 77058

ABSTRACT

Multimedia is making an increasingly significant contribution to our informational society. The usefulness of this technology is already evident in education, business presentations, informational kiosks (e.g. in museums), training and the entertainment environment. Institutions, from grade schools to medical schools, are exploring the use of multifaceted electronic text books and teaching aids to enhance course materials. Through multimedia, teachers and students can take full advantage of the cognitive value of animation, audio, video and other data types in a seamless application. The Software Technology Branch at NASA Johnson Space Center (NASA/JSC) is taking similar approaches to apply the state-of-the-art technology to space training, mission operations and other applications. This paper discusses the characteristics and development of multimedia applications at the NASA/JSC.

1.0 INTRODUCTION

Multimedia embraces many technologies and disciplines including videography, music, signal and image processing, artificial intelligence, computer graphics, database and data communication. It is the fastest growing segment in the computer industry today. With incorporation of animation, audio, video and interactive navigational links, digital multimedia technology is changing the way computers are applied. For instance, computers are emerging as successful supplements to formal classroom instruction and as viable alternatives to expensive hands-on simulators and trainers. In education and aerospace training environments it has also become necessary to maximize resources. Whether these resources are in the form of instructors, materials, or time, all must be prudently allocated in a cost effective manner. Computerized instruction utilizing existing tools and developing technologies is being substantiated with a growing list of applications and increasing return on investment. Such applications can better stimulate human senses and draw closer attention than traditional systems.

All computer manufacturers support multimedia capabilities in one form or another. Companies like Apple Computers and Microsoft Corporation have released multimedia capable operating systems and Graphical User Interface (GUI) environments. Many other software and hardware vendors have also launched products for these environments that comply to the standards defined by international and professional organizations. Through

[1] Albert Leigh is with McDonnell Douglas Space Systems Company, 16055 Space Center Blvd, Houston, TX 77062.

these non-proprietary standards, products have become more homogeneous and can be utitized over a large sector.

In this paper, the characteristics and components of multimedia environments will be discussed. A summary of multi-platform development activities of interactive multimedia applications at the NASA/JSC's Software Technology Branch (STB) will be reported. The final section outlines the experiences gained from developing these projects.

2.0 MULTIMEDIA ENVIRONMENT

As the term implies, multimedia is the integration of several media in an interactive computing environment (Figure 2.1). Conventional media include audio, video, still-photographs and printed documents, whereas computerized media consist of test, graphics, animation, spreadsheets and navigational links. When combined with the computer's interactive capability, audiences need no longer be passive in a multimedia system because they can navigate through a maze of information based on an intuitive structure. They decide when to start, when to stop, which piece of information to retrieve, and where to go for pertinent data. Although the term multimedia is vaguely defined, it is obvious that multimedia is used to enhance the way messages are conveyed in a multi-sensory form. As a picture is worth a thousand words, time-sequenced pictures like video, combined with audio, transmit information more effectively. It is one of the most exciting things to happen with computers in many years.

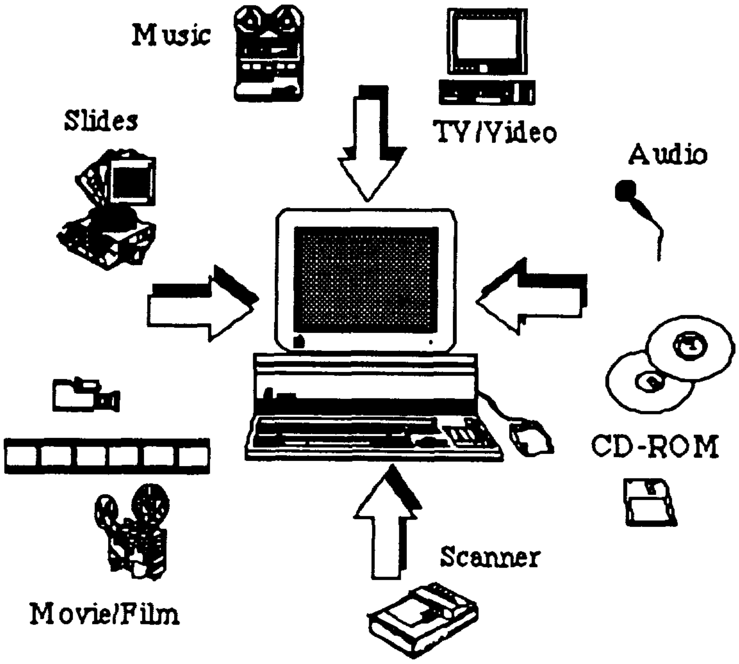


Figure 2.1 A Multimedia Environment

A multimedia software environment generally consists of two levels: viewing and authoring. Viewing refers to accessing a pre-composed system interactively without the capability to make modifications. Authoring involves organizing, mixing, and matching different forms of media; generating links for navigation; designing user-interfaces; and, creating error handling routines. Scanning devices are used to scan in photographs and flat documents, while appropriate digitizing devices are used to convert audio and video data from analog to digital forms. Of all the included media, video is the most elaborate, complicated, and intriguing. It is composed of sequences of video frames or images which are stored and replayed in rapid succession. The National Television and Standard Committee (NTSC) format used in the United States displays thirty frames per second. Thus, video documents containing a wealth of information, such as sound and still and motion images, demand greater system resources. Many computer hardware and software companies, including consumer electronic companies, have released desktop multimedia products with the necessary components and capabilities.

2.1 Trends

Available components include single chip solutions for video encoding (i.e. frame grabbing), display, and compression/decompression (CODEC), respectively. With these components, high-resolution displays are common to today's computers. They can now display up to 1280 by 1024 resolution and show 24-bit color photo-realistic images, at a fraction of the cost of earlier high-priced graphics workstations. Also, a single video overlay board (e.g. NTSC to VGA) can display a live television window on a computer, and can capture and compress video data on a hard disk. As a result, these capabilities allow software vendors to develop and market authoring tools (e.g. Authorware's *Authorware*), non-linear video editing software packages (e.g. Adobe System's *Premiere*), and video special effects systems (e.g. NewTek's *Video Toaster*). In addition, mass media companies are trying to establish their presence in the new media frontier. For instance, the Wall Street Journal reported that Turner Broadcasting System, Inc. is creating interactive news documentaries using the company's Cable News Network (CNN) footage and interactive games based on characters from its Hanna-Barbera film library to be used on CD-ROMs, i.e. compact disk-read only memory. As shown in Table 2.1, many similar *titles* ranging from children books to instructional materials are already on the market.

Title	Description	System
Amanda Stories	Children adventures and stories	Mac/PC
Beethoven: Symphony No. 9	Interactive classical music entertainment and lesson	Mac/PC
I Photograph to Remember	A still image essay the works of Pedro Meyer, a renowned photographer	Mac/PC
Columbus: Encounter, Discovery and Beyond	An interactive multimedia lesson of Columbus' voyage to the new world	PC
Microsoft Bookshelf for Windows	A reference library includes an encyclopedia, a dictionary, a thesaurus, a world alamanac, an atlas and two books of quotations	PC
The Virtual Museum	An interactive electronic museum where users can move from room to room and select any exhibit for more detailed examination	Mac
The Madness of Roland	The world's first hypermedia novel with digital video	Mac

Table 2.1: Commercial multimedia publications

As these data types demand increasing storage space, manufacturers have responded with winchester (hard) disk drives that are smaller in physical size, but greater in capacity. It is not uncommon for new PCs to have 2 billion bytes (GB) of storage. Distribution medium for developed applications also needs to be more efficient and economical. CD-ROM, for example, can store up to about 650 million bytes (MB), and yet, each disk costs as little as a box of floppy disks to produce. The greater the number of disks produced, the less expensive per disk is the production cost. In comparison, a box of ten diskettes can store approximately 15 MB. Then, there are magneto-optical drives (~1 GB), 3.5 inch *floptical* drives (~120 MB), 8 mm backup tapes (2.3 GB), and digital audio tape (DAT) (8 GB).

It is now obvious that the multimedia industry has enjoyed an exponential growth within the last few years. In fact, it is considered to be one of the fastest growing industries. However, the growth could have been much greater if there were open standards for this technology. A lack of standards has resulted in proprietary systems that are costly and incompatible. This is the main reason that multimedia technology did not reach the mainstreams until recently.

So, what standards exist today, and what are needed in the future? The raster display technology of the past decade is one of the forerunners to the current multimedia realm. Digital audio has also become a common data type on computer systems from Apple Computer, Commodore Computers, Sun Microsystems and PC vendors. Most importantly, as of 1992, video data is becoming more common on systems like Macintosh and PCs. This is primarily due to the fact that there are currently three non-proprietary digital video CODEC standards: Joint Photograph Experts Group (JPEG) standard for still image compression; the Consultative Committee on International Telephony and Telegraphy (CCITT) Recommendation H.261 for video tele-conferencing; and, the Moving Pictures Experts Group (MPEG) for full-motion video compression on digital storage media (DSM). Table 2.2 lists the experimental compression ratios that JPEG variant coders can achieve and the resulting image quality [2].

Bits/Pixel	Image Quality	Ratio
0.1	Recognizable image	160:1
0.25	Useful image	64:1
0.75	Excellent quality	22:1
1.5	Near original	11:1
8	Lossless JPEG	2:1

Table 2.2: JPEG Compression Ratios

Intel Corporation is one of the pioneers in the digital video compression arena. Intel's Digital Video Interactive (DVI) provides powerful CODEC capabilities with a programmable processor. DVI allows application developers to choose CODEC algorithms for better image quality; more simultaneous video operations, such as scaling motion video to a re-sized window or increased flexibility to support special needs of embedded applications; or any combination of the above. A key new feature of the components is their ability to compress or decompress JPEG images in near realtime. IBM Corporation is aggressively supporting this technology in its ActionMedia product line. NewVideo Company is another vendor actively developing DVI boards for Apple Macintosh systems.

Finally, Apple Computers has released a Macintosh system extension called QuickTime that allows temporal data types, such as audio and video, to be integrated into applications without special hardware. This extension can be used along with the above CODEC standards as long as the software interface conforms to QuickTime protocol. With it, video messages can be embedded in electronic mail, spreadsheets and presentations.

3.0 DEVELOPMENT ACTIVITIES

The use of various forms of media in computer applications has been in place at the Software Technology Branch (STB)for several years. However, the applications developed today utilize more multimedia capabilities by incorporating audio and video which was not feasible in the past. The following sections describe the to-date activities of developing multimedia and related applications at NASA/JSC.

3.1 Multimedia

The Automated Information Center (AIC) project is an interactive multimedia tour of the NASA/JSC Information Systems Directorate Products Center (IPC). The IPC is a facility dedicated to providing computer-related hardware and software products for sale and loan, information searches, product demonstrations, monthly newsletters, and many other services to the JSC community. The AIC system was developed to provide IPC customers with on-demand access to information about the IPC and its products and services. Also, to lessen the IPC staff's burden of continually being asked questions by customers about general operational topics. By automating this task, the IPC staff will be able to provide more individualized assistance to customers requiring detailed information, as well as being allowed greater freedom to accomplish day-to-day activities. This system was built as a fully automated multimedia environment utilizing state-of-the-art personal computer components and software in a kiosk-style booth at the IPC facility (Figure 3.1).

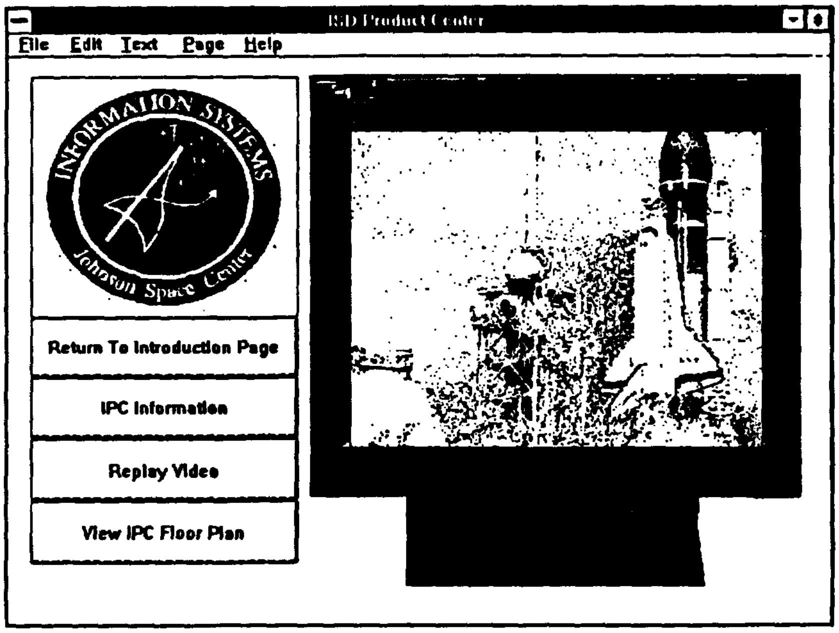


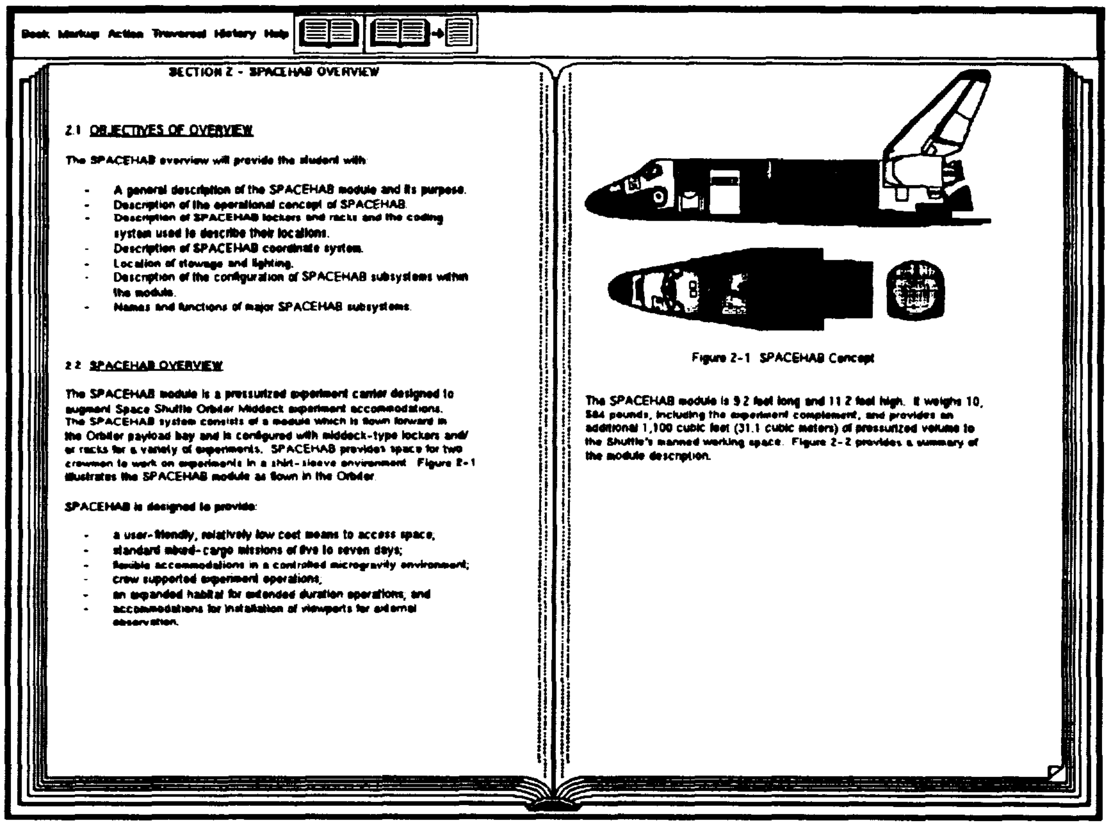
Figure 3.1: Automated Information Center's digital video screen

The general system configuration includes a generic 486 PC with Microsoft Windows, digital CODEC boards, an audio board and a CD-ROM drive. In the application, the user is presented with a touch screen interface and menus to invoke general information about the IPC and detailed descriptions about the IPC services, products and activities. The user could choose either of these paths from the initial screen. The upper portion of the screen contains general navigational buttons. Upon selecting certain buttons, the user receives audio and video explanations, schedules of events, software and hardware catalogs, and floor plan information. The user can also try out some on-line software or can view software demonstration videos.

Based on this system, STB is exploring potential development of a very large scale multimedia information retrieval system for the NASA/JSC's Public Affair Office and the U.S. Navy's Informational Survivability Management System (ISMS) for advanced damage control.

3.2 Hypermedia

A hypermedia product, called Hyperman, was designed and developed by STB. Hyperman is a software tool which enables the users of technical manuals to have rapid on-line access to documentation with a full range of hypermedia capabilities. With Hyperman users can parse documents in their native word processing format and display these documents on a UNIX platform employing X-Windows with the Motif Toolkit as shown in Figure 3.2. Hyperman is able to display a broad range of media, including text, equations, tables, graphics and audio.



SECTION 2 - SPACEHAB OVERVIEW

2.1 OBJECTIVES OF OVERVIEW

The SPACEHAB overview will provide the student with:

- A general description of the SPACEHAB module and its purpose.
- Description of the operational concept of SPACEHAB.
- Description of SPACEHAB lockers and racks and the coding system used to describe their locations.
- Description of SPACEHAB coordinate system.
- Location of stowage and lighting.
- Description of the configuration of SPACEHAB subsystems within the module.
- Names and functions of major SPACEHAB subsystems.

2.2 SPACEHAB OVERVIEW

The SPACEHAB module is a pressurized experiment carrier designed to augment Space Shuttle Orbiter Middeck experiment accommodations. The SPACEHAB system consists of a module which is flown forward in the Orbiter payload bay and is configured with middeck-type lockers and/or racks for a variety of experiments. SPACEHAB provides space for two crewmen to work on experiments in a shirt-sleeve environment. Figure 2-1 illustrates the SPACEHAB module as flown in the Orbiter.

SPACEHAB is designed to provide:

- a user-friendly, relatively low cost means to access space;
- standard mixed-cargo missions of five to seven days;
- flexible accommodations in a controlled microgravity environment;
- crew supported experiment operations;
- an expanded habitat for extended duration operations; and
- accommodations for installation of viewports for external observation.

Figure 2-1 SPACEHAB Concept

The SPACEHAB module is 9.2 feet long and 11.2 feet high. It weighs 10,584 pounds, including the experiment complement, and provides an additional 1,100 cubic feet (31.1 cubic meters) of pressurized volume to the Shuttle's manned working space. Figure 2-2 provides a summary of the module description.

Figure 3.2 Hyperman's on-line documentation

Hyperman was created with two separate configuration options. The first option creates a stand-alone hypermedia tool. The second option creates a "help-system" version of Hyperman. In the second configuration, Hyperman can exist within another host software tool. When the user has a need for on-line help, Hyperman can provide context sensitive help based on the users current location in the software.

Hyperman employs a number of techniques to provide for easy transition between the user's paper and electronic document domains. These techniques include; electronic highlights, electronic notepads, full text search, hyper-links, preservation of the look of the book from paper to screen , and easy access to the UNIX environment. This access to the UNIX environment means that applications and processes, including the attachment of video, animation and audio to Hyperman text and graphic objects, can be started from within Hyperman which will augment the users comprehension. Hyperman's parsing preserves not only the text and page formats but all of the formats to the text including bold-facing, underlining, scripting, subscription, and italics. These capabilities are encased in a Graphical User Interface (GUI) which attempts to make the task of document management, information retrieval and diverse data integration easy.

3.3 Intelligent Computer-Aided Training (ICAT)

Expert system technology has been used to develop autonomous training systems for use by astronauts, flight controllers and NASA engineers in learning to perform a wide range of procedural tasks. STB has developed many ICAT systems based on this technology. The Payload Deployment ICAT (PD/ICAT) and Center Information Systems/Computer Operations ICAT (CISCO/ICAT) are examples of such applications developed on UNIX/X-Windows and Microsoft Windows environments, respectively. The Active Thermal Control System (ATCS) ICAT for Space Station Freedom training was developed on the Apple Macintosh environment. It incorporates extensive multimedia in the form of scanned color photographs, animated graphics, digitized audio, and QuickTime digital video to engage the trainee's interest and to present a variety of concepts and system functions in a simulated real-world environment.

Audio and video are significant elements in many subject areas. Flight controllers rely heavily on audio information received via their headphones, and astronauts must train for many tasks requiring visual inspection. Yet, audio and video elements have been omitted from most ICAT systems because they are system dependent. Integrating Apple's QuickTime software into the ICAT development environment offers three major advantages: it eliminates the cost of specialized hardware, it speeds development time and reduces maintenance costs, and it enables the cross-platform portability of the ICAT development environment.

Using interactive graphics of display screens, high resolution photographs of equipment and control panels, and full-motion video of operation and maintenance procedures, ICAT systems are being developed to supplement and potentially replace large, expensive part task trainers and provide "on-demand" training wherever and whenever it is needed.

3.4 Virtual Environments

Virtual Environment technology, which is generally referred to as *virtual reality* or *artificial reality*, has the potential to provide an intuitive human-computer interface of unprecedented power (Figure 3.3). The STB is exploring the use of this technology for training as an adjunct to its ongoing work in ICAT. Virtual Environments (VE) can *place* an individual into any scenario that can be copied or imagined. The use of VE for training is obvious. Crew members can experience a virtual cockpit or an Extravehicular Activity (EVA) and both develop and be trained in new procedures. Payload specialists can learn to operate virtual instruments before they are built or flown. The key to VE is the *immersion* of the user in another world. The ability of VE to cause the *suspension of disbelief* is a power argument for the infusion of VE technology into NASA's training program.

Figure 3.3: A VR system in use at NASA/JSC

The STB is actively pursuing a number of applications projects that will utilize virtual environment technology. Among these projects are training for Space Shuttle EVA, Space Station Freedom operations especially those that are cupola based, and Hubble Space Telescope repair and maintenance. In addition, STB is working jointly with Marshall Space Flight Center to enable personnel at the two centers to simultaneously share the same VE, eliminating the usual constraints of location from joint training activities. Finally, prototypical science laboratories are being developed to enable students to observe physical phenomena not available in typical student laboratories.

4.0 CONCLUSION

Multimedia technology is in a very dynamic growth period with industry standards being defined and technical advancements with hardware occurring in many cases faster than new products can be released. Applications are being developed to exercise these available technologies. Software is reaching cross-platform capability status with peripheral component interface utilities and more extensive capabilities and functions incorporated into products. As multimedia system and component costs decrease, these systems will become attainable by a wider group for varied implementations.

Although multimedia is here today, the technology is still in its infancy. Mixing and matching peripherals with existing peripherals bring up conflicts and incompatibilities. On a distributed network, transferring multimedia data can decrease the network throughput and efficiency dramatically. The STB is actively pursuing many avenues to overcome these deficiencies. Many multimedia data types are quite portable now. Animation, audio, graphics and images, spreadsheets and text data can easily be transported between hetrogeneous platforms. What is needed most at this time is non-proprietary software video compression and decompression capabilities such that the applications developed can be delivered without requiring costly hardware add-ons. More research and development needs to be done to incorporate search and indexing, knowledge capture, and higher bandwidth networking for distributed capabilities.

As important as the hardware and software technology is the skill level of an integrated development team. This team of creative professional technical and artistic people from various disciplines is necessary for a worthwhile product. With a good design and creative skills, a multidisciplinary effort of computer engineers, graphics artists, a producer or manager and other relevant personnel can conquer the risks and barriers. It will also result in a high-tech *masterpiece* application and yield high pay-offs. In conclusion, to meet the demands of our informational society, the full potential of this revolutionary digital technology must be continue to be realized and utilized.

5.0 REFERENCES

1. Ang, P. et al., "Video Compression Makes Big Gains", IEEE Spectrum, pp. 16-19, October 1991

2. Cavigioli, C., "JPEG Compression: Spelling Out the Options", Advanced Imaging, pp. 44-47, March 1991

3. Fox, E., "Advances in Interactive Digital Multimedia Systems", IEEE Computer, pp. 9-21, October 1991

4. Kenney, P. et al., "New Media Information and Training Applications in Aerospace and Education", NASA/JSC's ICAT Conference Proceedings, November 1991

5. Korenjak, A., "Video Compression: Higher Performance for Increased Flexibility", Intel InteraActivities, Volume 2, No. 2, Winter 1990-91

6. Ripley, G., "DVI - A Digital Multimedia Technology", Communications of the ACM, pp. 811-822, July 1989

VISUAL COMMUNICATION IN MULTIMEDIA CYBERSPACES

This paper was withdrawn from presentation

N93-22172

S32-60

150227

P-5

MICRO-VIDEO DISPLAY WITH OCULAR TRACKING AND INTERACTIVE VOICE CONTROL

James E. Miller
Naval Undersea Warfare Center Division, Newport
Missiles Division, Code 831
Newport, RI 02841

ABSTRACT

In certain space-restricted environments, many of the benefits resulting from computer technology have been foregone because of the size, weight, inconvenience, and lack of mobility associated with existing computer interface devices. Accordingly, an effort to develop a highly miniaturized and "wearable" computer display and control interface device, referred to as the Sensory Integrated Data Interface (SIDI), is underway. The system incorporates a micro-video display that provides data display and ocular tracking on a lightweight headset. Software commands are implemented by conjunctive eye movement and voice commands of the operator. In this initial prototyping effort, various "off-the-shelf" components have been integrated into a desktop computer and with a customized menu-tree software application to demonstrate feasibility and conceptual capabilities. When fully developed as a customized system, the interface device will allow mobile, "hands-free" operation of portable computer equipment. It will thus allow integration of information technology applications into those restrictive environments, both military and industrial, that have not yet taken advantage of the computer revolution. This effort is Phase I of Small Business Innovative Research (SBIR) Topic #N90-331 sponsored by the Naval Undersea Warfare Center Division, Newport. The prime contractor is Foster-Miller, Inc. of Waltham, MA.

INTRODUCTION

During the last decade, there has been a rapid development of technologies that support computer miniaturization. These technologies have significantly reduced computer workstation size and, at the same time, improved processor speed, memory capacity, and software utilization. It is interesting to note, however, that functional improvements in the interface devices for computers have been relatively few. With the exception of the computer mouse/trackball and new "pen-based" input devices, virtually all computer systems still rely on large and cumbersome interfaces, namely CRT/LCD screen displays and keyboards. There has been relatively little ergogenic improvement with these interfaces since their 1950s predecessors were introduced.

Unfortunately, many of the benefits associated with computer technology have been foregone because of this lack of appropriate interface devices. This is particularly true for certain space-restricted environments such as aircraft, submarines, and spacecraft. The disadvantages imposed by size, weight, and inconvenience of existing interfaces as well as the lack of mobility for the operators using them has resulted in the avoidance of "computer solutions."

The development effort described below seeks to solve this problem by integrating several new "off-the-shelf" technology products to produce a highly miniaturized, micro-video computer display and control system which incorporates both interactive ocular tracking and voice control. The goal is to create a "wearable" interface mounted on a portable headset. Further development will lead to the implementation of highly miniaturized, portable alternatives to currently used cathode-ray tube/liquid crystal display and keyboard control devices. The advantages of computer technology will thus become more accessible for space restricted and operator-mobile applications and environments.

DISCUSSION

A number of emerging technologies are now available which can provide significant improvement for the human-to-computer interface. Singularly, each one of these technologies has the ability to provide improvement, but significant synergy is gained by combining them together into a single, integrated system. The technologies applied in this effort are those of micro-video displays, ocular tracking systems, and computer voice recognition. Based upon a capability verses cost analysis of commercially available products, as described in reference (1), the component subsystems described below were selected for use in this project. Since the interface concept relies on human sensory interaction, the system is referred to as the Sensory Integrated Data Interface (SIDI)

Once the component subsystems were selected, the effort consisted of assembling them into a functional interface. On an adjustable headset were mounted a micro-video display, an associated Infra-Red (IR) illumination source and video camera, and a microphone. The micro-video display presents computer data to the operator. Using eye-tracking hardware installed in a PC expansion slot, screen cursor movement is caused to follow eye movement. Software menu selections are then implemented by gaze-fixation timing and voice command. An analogy to the "point and click" functionality of a computer mouse can be drawn. The line-of-sight gaze is used to "point" to a menu option and the voice command "clicks on it" or activates the selected item.

Figure 1 illustrates how the components were assembled on an operator headset. Note that the display projector/eye-tracker is worn in a monocular arrangement in front of one eye. This allows the operator to access computer data while operating other equipment or performing other tasks. Note also that the use of the headset allows "hands-free" operation of the computer. Figure 2 illustrates a high level component integration scheme for the desktop computer. A customized software application was developed to demonstrate conceptual capabilities. As shown in Figure 3, it consists of a menu-tree listing of topics which the operator activates by gazing and, when ocular lock is achieved, initiates with a verbal command. Running underneath this application are the customized software device drivers for the system components.

Micro-Video Display

Data display for SIDI is provided by the "Private Eye" developed by Reflection Technology, Inc. The "Private Eye" uses a new proprietary technology which uniquely combines conventional semiconductor and optical techniques to create an image of a 12 inch monitor in a miniature package measuring only 1.1 inches X 1.2 inches X 3.2 inches. It weighs less than two ounces. The display uses emissive elements to provide a high definition, high contrast, and fast responding image viewable in bright daylight.

Worn in a monocular position in front of one eye, the unit interfaces with a PC computer via an interface card in an expansion slot. A virtual image of the computer CRT display is projected approximately 18 inches in front of the viewer's eye. The unit displays 720 X 280 pixels which can be formatted as 25 lines with 80 characters per line or can be used to show bit-mapped graphics. The monochrome image appears to float in space in front of the viewer with a quality and resolution matching that of a standard display. A lens system allows image focusing to accommodate viewing with or without eye glasses.

Ocular Tracking System

The subsystem chosen to provide line-of-sight ocular tracking for the SIDI was the Eye Slaved Pointing (ESP) System from ISCAN, Inc. It is available as a PC compatible, turnkey eye movement monitoring system that serves as a hybrid interface device. Terminate and Stay Resident (TSR) software developed for the ESP is designed to substitute an operator's line-of-sight gaze for standard computer pointing devices. The system provides cursor positional data to software applications which permit or require a pointing device.

The eye-tracker functions by using a custom set of integrated circuits and tracking algorithms to lock on and track a point on the operator's eye in real-time. This is accomplished by measuring and tracking a low-level Infra-Red (IR) light beam reflected from the eye surface. A miniature IR video camera is used to acquire the pupil and corneal images needed by the tracking algorithms. The operator views a series of positional data points on a display to calibrate the tracking function for the display. From that point on, the tracking function converts eye movement into screen cursor movement. Since the display and eye-tracking camera are positionally coupled by the headset, movement of the head does not affect the tracking function. Both the ESP and Private Eye are mounted so as to be adjustable on the headset and allow variable positioning of the components for different operators.

Voice Recognition Unit

The Voice Recognition Unit (VRU) chosen to provide voice control for the SIDI was the Voice Master from Covox, Inc. The system consists of a control unit, speaker, microphone, and software which can provide both speech recognition and generation capability for a computer using an available RS-232 port. The operator trains the system to accept specific voice commands which implement menu options once the ESP has achieved ocular lock on a menu option.

VRUs allow a computer to translate any spoken language into accurate, intelligent commands. It does so by breaking down a spoken command into its frequency components and then comparing those components with those of pre-stored commands. Most language requirements for a computer can be carried out with a relatively small vocabulary when combined with gaze-directed screen interaction. Within 10 - 15 minutes, an experienced operator can fully train the voice unit to recognize the required sets of alpha-numeric commands/inputs.

While voice input alone can be used to select items from a display menu, there are many situations where it does not perform well. For example, voice commands are poor for most interactions with graphic images, like pointing to a specific wire in a schematic or selecting a location to insert information in a text display. Even with text applications, voice interaction is difficult when selecting one item from a display which has many similar items, like a stock inventory sheet or a page of text. Eye-directed interaction in conjunction with voice commands handles these situations quickly and naturally.

Software

A customized software application was developed to demonstrate functional and conceptual capabilities of the SIDI. The software consists of both the user application and the underlying TSR that handles hardware interrupts. As shown in Figure 3, the user application consists of a menu tree that allows the operator to access various types of documentation for submarine system and subsystem components. The operator makes a selection by gazing at a major topic item. When the SIDI achieves ocular lock-on, the selection flashes at the operator signifying that verbal commands can be initiated. Lock-on is achieved quickly enough so that the operator perceives that the system is operating concurrently with his thought processes. By giving various commands such as "ACCEPT", "UP", "DOWN", and "BACK" the operator can move around the menu to view desired data. The software was written in the C programming language with some in-line assembly code.

SIDI APPLICATIONS

There is a wide spectrum of applications for the SIDI or a SIDI derivative. They range from various consumer entertainment products to business applications to military documentation applications. Any current computer application that could benefit from operator mobility is a candidate SIDI application. Figure 4 illustrates an application for a technician performing a maintenance routine on a cruise missile using procedural documentation displayed by a SIDI system. A few of the many possible implementations that could be envisioned are as follows:

- Paperless technical manuals
- Equipment diagnostic/repair systems
- Personalized training systems
- Entertainment electronics
- Portable inventory management systems
- Procedural documentation in space-restricted areas
- Computer interface for the physically handicapped

FUTURE WORK

The goal of this project was to prototype existing commercial off-the-shelf hardware to demonstrate the feasibility of the SIDI concept. That goal was achieved. A functional prototype system was developed that successfully allows hands-free operation of a computer. The next step, as described in reference (2), is to refine the concept by developing a more compact and optimized interface device that provides operator mobility, requires less hardware adjustment to accommodate different operators, and is compatible with commercial software applications.

To maximize operator mobility, it is proposed that the interface be developed to serve as a remote terminal served by a computer base station. The user module will contain the minimum amount of electronics required to drive the display, eye tracking camera, and microphone. The base station will provide all the intelligence for these components via a data link. The data link will be preferably an un-tethered technology, such as an Infra-Red (IR) spacial optical link, thereby providing the operator unrestricted movement.

To reduce the amount of component adjustment for each individual who operates the interface, the headset hardware will be miniaturized and integrated into a single, easy-to-use, composite unit. The hardware will be optimized for

- Ease of alignment to obtain the eye image
- Ease of calibration to the user's eye
- Better tolerance to room lighting variations
- Better display resolution
- A separation of component functions to allow base station interaction

Finally, to preclude the need for customized software applications, the output of the eye tracking system will be modified to reflect standard computer mouse and trackball conventions. This will allow the use of standard software interface environments.

REFERENCES

1. Kendrick, W. K. "Component Survey and System Definition Report," NAV-9371, Contract N66604-91-C-0936, Foster-Miller, Inc. March 25, 1992.

2. Nappi, B. "Sensory Integrated Data Interface (SIDI)," Small Business Innovative Research (SBIR) Phase II Proposal, Foster-Miller, Inc. August 21, 1992.

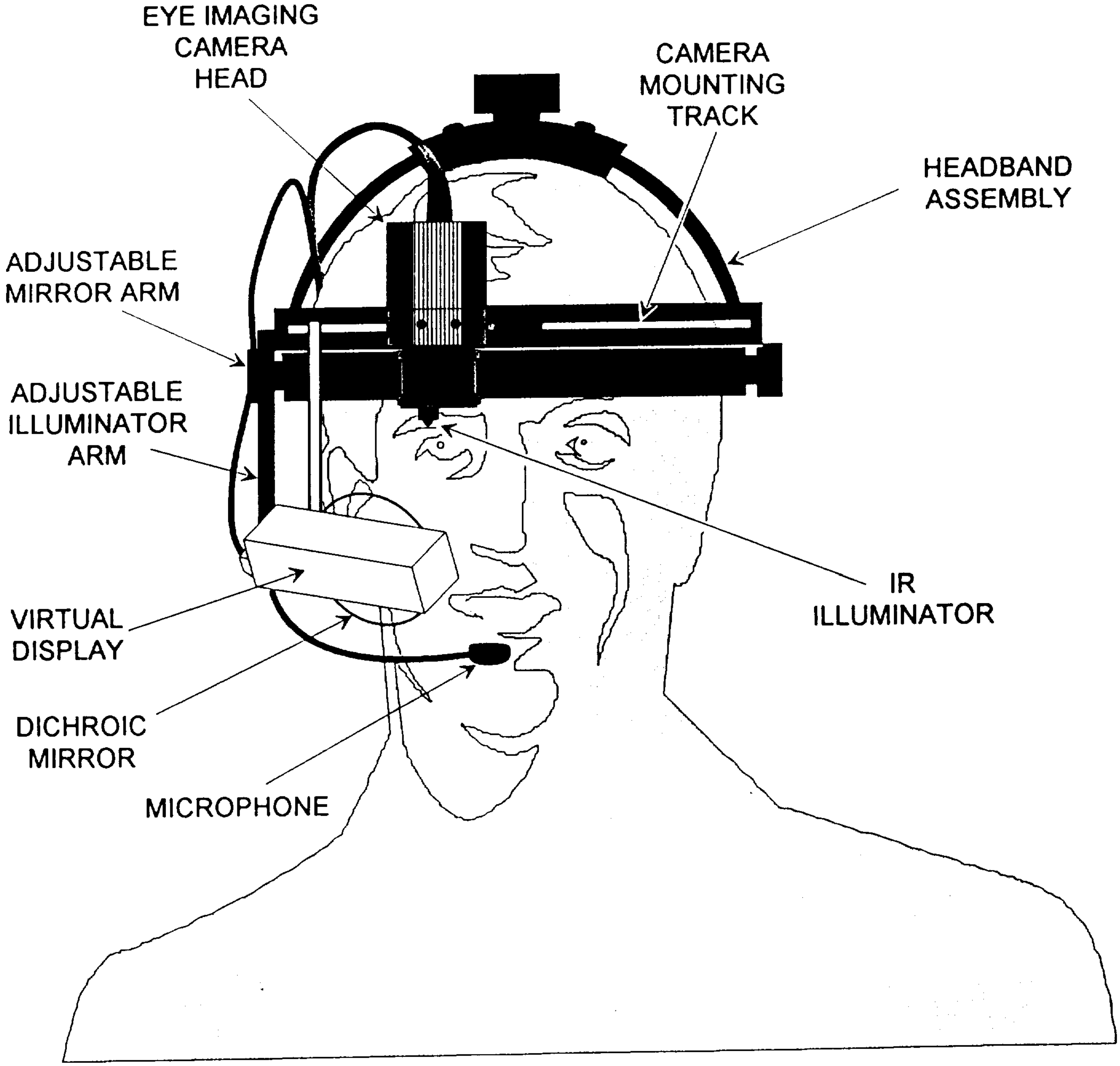

EYE IMAGING CAMERA HEAD
CAMERA MOUNTING TRACK
HEADBAND ASSEMBLY
ADJUSTABLE MIRROR ARM
ADJUSTABLE ILLUMINATOR ARM
IR ILLUMINATOR
VIRTUAL DISPLAY
DICHROIC MIRROR
MICROPHONE

FIGURE 1

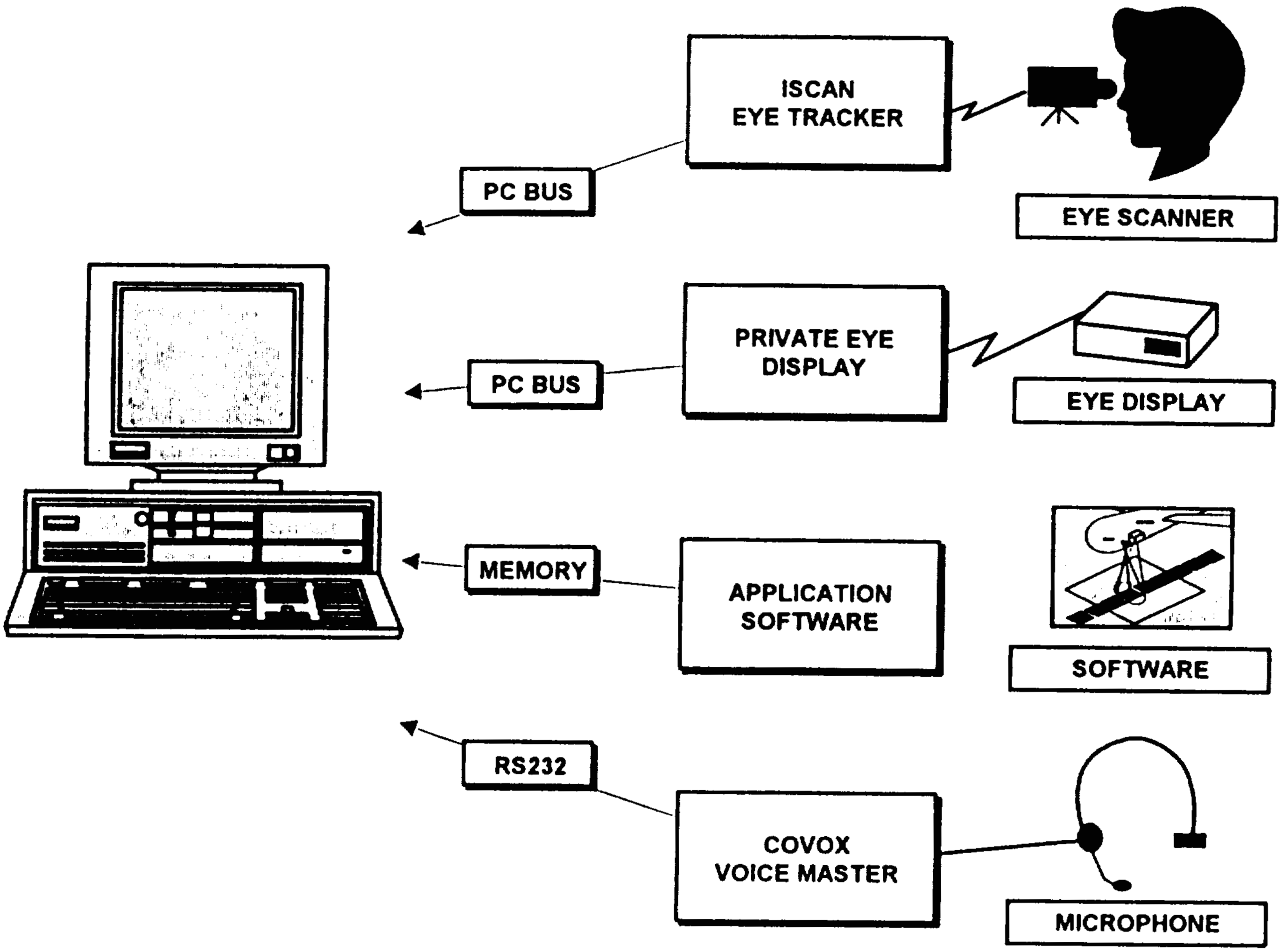

ISCAN EYE TRACKER
PC BUS
EYE SCANNER
PRIVATE EYE DISPLAY
PC BUS
EYE DISPLAY
MEMORY
APPLICATION SOFTWARE
SOFTWARE
RS232
COVOX VOICE MASTER
MICROPHONE

FIGURE 2

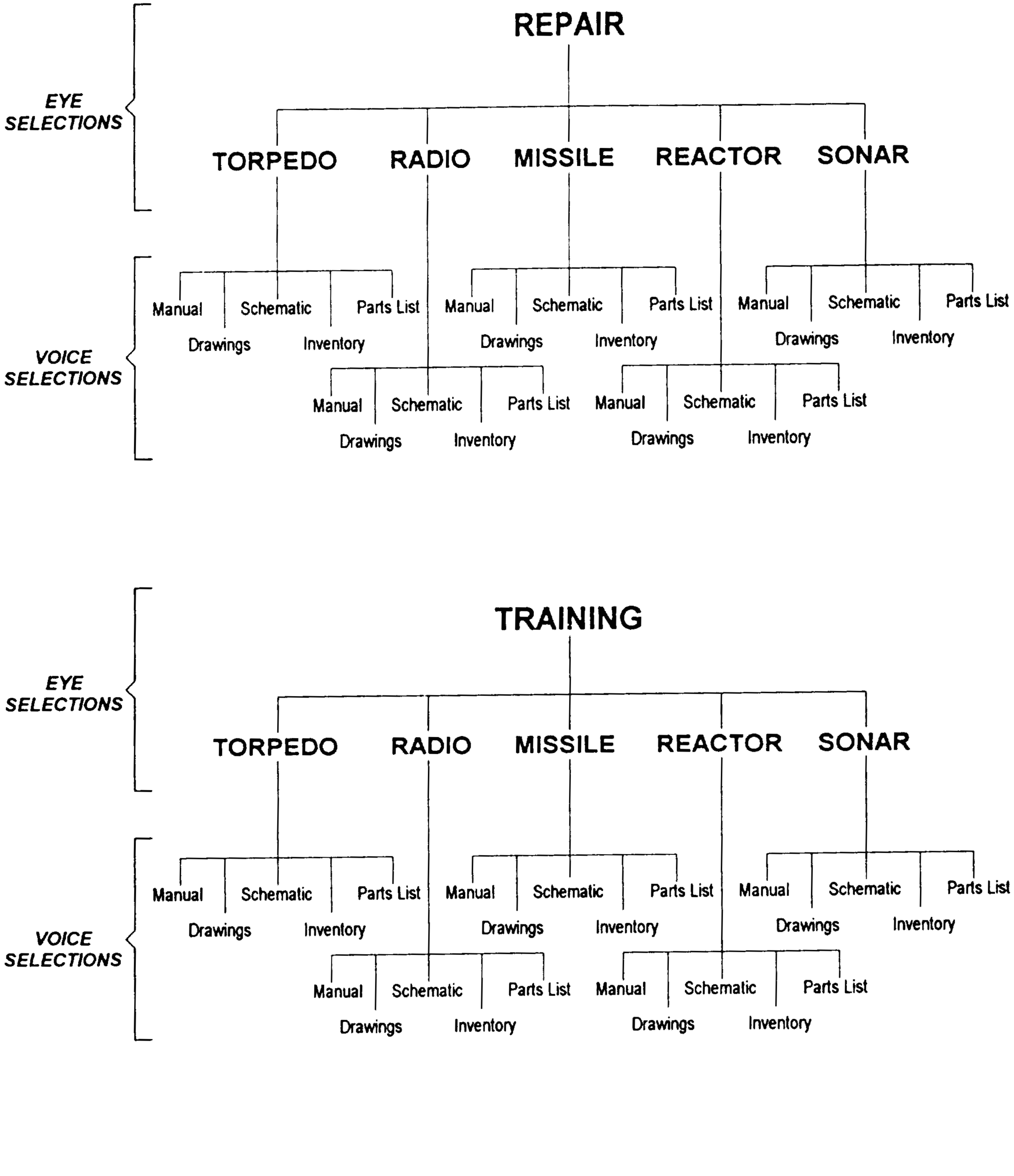

REPAIR
EYE SELECTIONS
TORPEDO
RADIO
MISSILE
REACTOR
SONAR
VOICE SELECTIONS
Manual
Drawings
Schematic
Inventory
Parts List
Manual
Drawings
Schematic
Inventory
Parts List
Manual
Drawings
Schematic
Inventory
Parts List
Manual
Drawings
Schematic
Inventory
Parts List
Manual
Drawings
Schematic
Inventory
Parts List
TRAINING
EYE SELECTIONS
TORPEDO
RADIO
MISSILE
REACTOR
SONAR
VOICE SELECTIONS
Manual
Drawings
Schematic
Inventory
Parts List
Manual
Drawings
Schematic
Inventory
Parts List
Manual
Drawings
Schematic
Inventory
Parts List
Manual
Drawings
Schematic
Inventory
Parts List
Manual
Drawings
Schematic
Inventory
Parts List

FIGURE 3

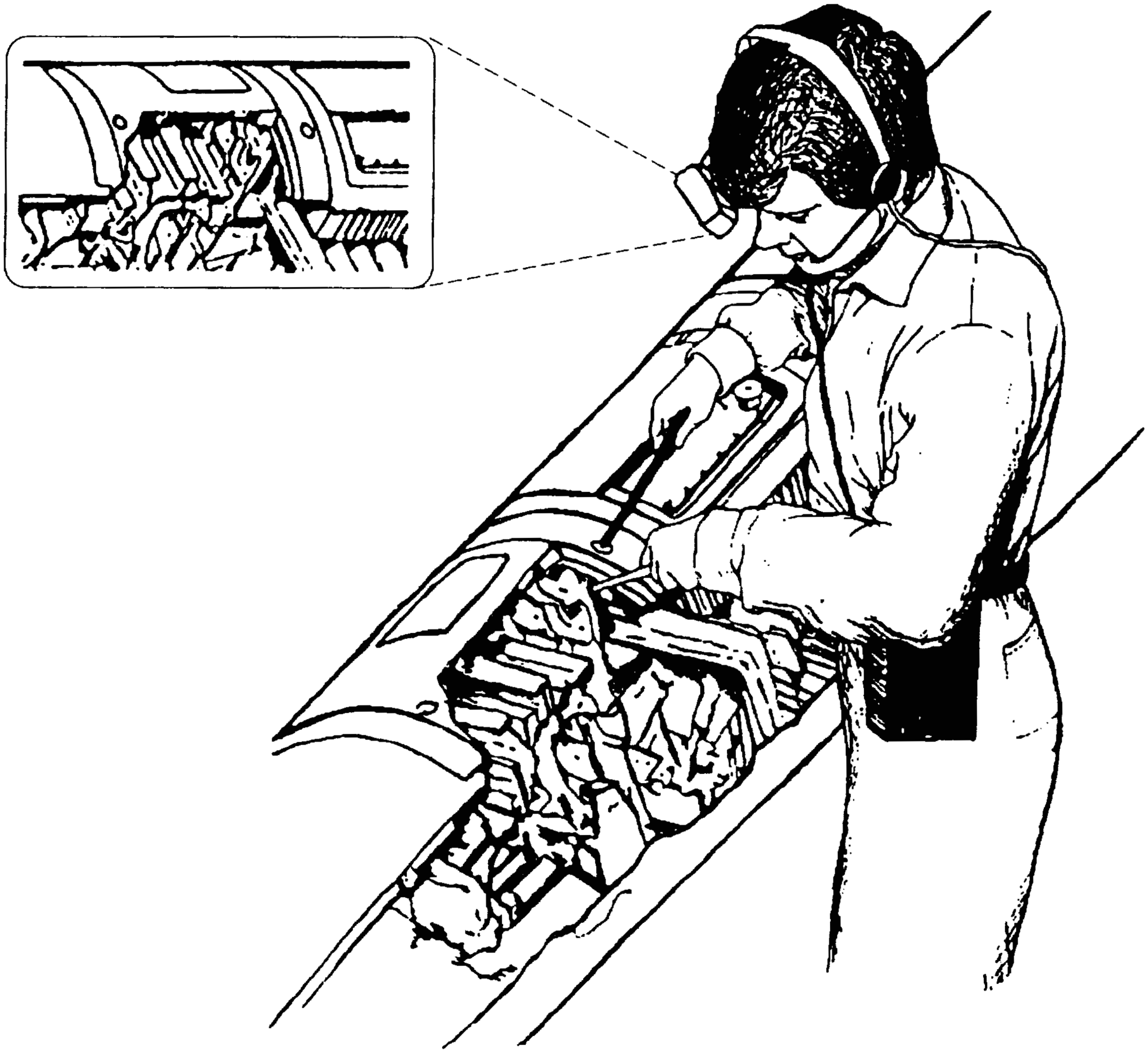


FIGURE 4

VIDEO CONFERENCING MADE EASY

D. Gail Larsen
INEL/EG&G Idaho, Inc.
P.O. Box 1625
Idaho Falls, ID 83415-1500

Paul R. Schwieder
INEL/EG&G Idaho, Inc.
P.O. Box 1625
Idaho Falls, ID 83415-1500

Prepared for the
U.S. Department of Energy
Office of New Production Reactors
Under DOE Contract No. DE-AC07-76ID01570

ABSTRACT

Network video conferencing is advancing rapidly throughout the nation, and the Idaho National Engineering Laboratory (INEL), a Department of Energy (DOE) facility, is at the forefront of the development. Engineers at INEL/EG&G designed and installed a very unique DOE video conferencing system, offering many outstanding features, that include true multipoint conferencing, user-friendly design and operation with no full-time operators required, and the potential for cost effective expansion of the system.

One area where INEL/EG&G engineers made a significant contribution to video conferencing was in the development of effective, user-friendly, end station driven scheduling software. A PC at each user site is used to schedule conferences via a windows package. This software interface provides information to the users concerning conference availability, scheduling, initiation, and termination. The menus are "mouse" controlled. Once a conference is scheduled, a workstation at the hubs monitors the network to initiate all scheduled conferences. No active operator participation is required once a user schedules a conference through the local PC; the workstation automatically initiates and terminates the conference as scheduled. As each conference is scheduled, hard copy notification is also printed at each participating site.

Video conferencing is the wave of the future. The use of these user-friendly systems will save millions in lost productivity and travel cost throughout the nation. The ease of operation and conference scheduling will play a key role on the extent industry uses this new technology. The INEL/EG&G has developed a prototype scheduling system for both commercial and federal government use.

INTRODUCTION

Network video communications is advancing rapidly, and the Idaho National Engineering Laboratory (INEL), a Department of Energy (DOE) facility operated by EG&G Idaho, Inc., is at the forefront of this development. Engineers at EG&G have been involved in the design and development of several video conferencing installations for the Department of Defense (DoD), the INEL, and DOE. These installations have included; point-to-point and multipoint NSA approved secure and nonsecure video conferencing and data networks; local and long haul communication networks; multilevel security systems; remote digital video imaging; mobile video and audio systems;and multinodal, multimedia information networks. Typically these systems incorporate multiplexed digital data traffic that can be carried over land lines, microwave, fiber cable or satellite communications media. INEL engineers have developed expertise and implementation tools in all the above areas over the past eight years, that has placed our design and installation teams in a unique position to help government and private groups develop video conferencing and data transmission capabilities. INEL engineers have, in all cases, carried these jobs from start to finish. Using industry standard equipment and have interfaced with government and private agencies to work both the technical and operational challenges associated with video conferencing and data networks.

The principle subject of this paper is a video conferencing and data network developed for the DOE that embodies many of the features that are both common and unique to high level video conferencing and data transmission needs, including true multipoint conferencing, user-friendly design and operation, no full-time operators required, and the potential for cost-effective expansion of the system. This development is an excellent example of an information system that serves todays needs, using current technology, and at the same time looks forward to the future needs of video conferencing and data transmission. Although the conferencing system developed for the DOE was designed for a limited user community, the work performed by INEL engineers forms the groundwork for video conferencing and data communications networks of any size.

The DOE identified the need for a nation-wide video and data network and contracted the INEL/EG&G Idaho to design it. This network was to link remotely located sites through an effective use of both video and data communications capabilities. Eight sites were selected for the initial system. A dual hub topology (one in the eastern US and one in western US) was selected to minimize the number and length of communication lines and lower costs (see Figure 1). The system configuration at each user site consists of video conference equipment (monitors, cameras, etc.), a compressor/decompressor (CODEC) for bandwidth reduction, a multiplexer for multiple data input, a gateway to a nationwide communications network, and a personal computer (PC) linked to centrally located workstations for local systems and video

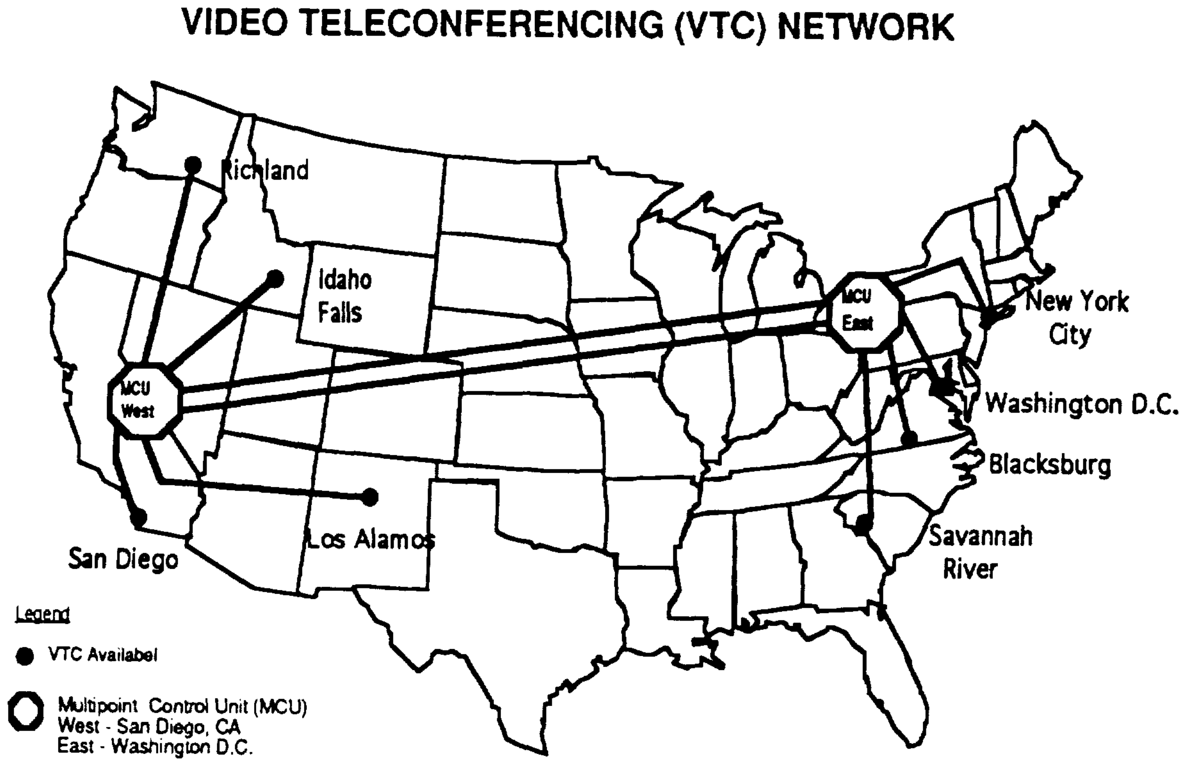


FIGURE 1

BASIC MULTIPOINT VIDEO CONFERENCING NETWORK

network scheduling. A workstation and a Multipoint Control Unit (MCU) used for video switching are located at each hub. The workstation provides control of the network-wide video scheduling database. Each site is equipped with a bandwidth multiplexer (mux) used to multiplex the video conferencing signal as well as other types of digital data transmissions on a T1 communication line leased from AT&T. The multiplexers, connected to the communications network, create a communications backbone capable of serving not only video conferencing, but many other forms of data transmission between the involved sites. Data transmission and network control interface directly to the multiplexers; only video conferencing signals pass through the CODEC units and the MCU.

THE HARDWARE

The hardware is all off-the-shelf and readily available. Rooms or studios are set up using commercial video and audio equipment, with either customized configurations or modular consoles. The communications backbone is a multiplexer based, multipath nodal system, capable of carrying many different types of data traffic, including the video conferencing. The multipath capability supports reliability, by providing more than one path for data flow, as well as supporting both point-to-point and star/hub type communications links, simultaneously. Each node's multiplexer allows for the support of computer traffic, voice, fax, high resolution graphics or imagery, or any form of digital information exchange, up to the limit of available bandwidth. Riding along on the same communications lines to perform housekeeping chores is the network wide control and data routing information provided by an easy to use control and scheduling system. Developing the network architecture into either, or both a star/hub, or point-to-point configuration is simply a matter of choice in a well designed nodal communications network. The star/hub configuration is necessary for multipoint video conferencing applications. The point-to-point applications support point-to-point video conferencing, as well as all manner of digital data traffic, and all can be served simultaneously, again up to the limit of the available bandwidth. Fractional T1, T1, or even T3 communication services can be employed to support the users needs and budgets.

The Video Conferencing CODEC is located on the equipment side of the nodes multiplexer and supports only video conferencing applications. Its primary role is to reduce the full motion video bandwidth (typically 92 mb/s) down to near full motion (1.5 mb/s or less) to allow for affordable transmission between sites. The CODEC also provides data formatting, encryption, bandwidth multiplexing and other services.

The MCU located at each hub is an audio-activated digital video switcher. This video input to output switching capability is key to multipoint conferencing. To accomplish this switching, the MCU separates each site's incoming audio signals from the video, determines which signal is the loudest (dominant) at any given time, and switches the transmission of the associated video to all participants. This results in "video-follow-voice"conferencing, in which multipoint participants see the video image of the dominant speaker. The audio is party line; all participants hear each other at all times. As another participant becomes the dominant speaker, the MCU automatically switches to broadcast their video image. The DOE configuration uses two dedicated MCUs, one in the eastern U.S. and one in the western US, linked together via redundant communication lines in a cascade configuration, with a combined potential for five separate simultaneous conferences involving up to twelve separate participant sites. A follow-on to this existing design will use the MCUs as free floating independently assigned conference servers, allowing several MCUs at each hub to serve a much larger conferencing community, all under the control of the automatic conferencing control and scheduling system.

CONFERENCE CONTROL AND SCHEDULING

One area where INEL engineers have made a significant contribution to video conferencing is in the development of an effective, user-friendly video conferencing scheduling system. A PC at each site is used to schedule current and future conferences via a windows menu package. Users are provided information on current and future conferencing schedules, conference initiation and termination, and video system status. All inputs to the system are via a "mouse" or keyboard. Following the prompts and selecting the desired options enables the user to schedule current and future conferences on the system as well as do conference status inquiry.

Once a conference is in the workstation's scheduling database, the workstation monitors the common network time (based on east coast time, but displayed to each user in local time) and activates and terminates each scheduled conference automatically by sending commands to the MCUs at the hubs. Conference scheduling time slots are broken into 15-minute intervals. Start immediate and end immediate commands are also available for spontaneous unscheduled conferences allowing initiation and termination of a conference at any time.

No active operator participation is required once a user schedules a conference through the local PC. Under the control of the scheduling database, the workstation automatically brings up the conference as scheduled. Hard copy notification of the scheduled conference is provided at each participating site via a dedicated printer. There can be up to 5 separate and independent video conferences on the network simultaneously, however, users can participate in only one conference at a time.

The scheduling system is designed so that a central manned operation center is not needed. Individual users of the system schedule and operate the system from their own facility. Typically, any user could operate and schedule a conference after just a few minutes of training. The following example demonstrates how easy scheduling a conference can be:

> "A user at Site 1 needs to schedule a video conference meeting for Monday at 10:00 a.m. He opens the scheduling log on his PC screen and discovers that the necessary conferencing sites are free on Monday at 10:00 a.m. With a few clicks of the mouse, identifying the sites involved and the conference date and time, the conference is entered into the time slot from 10:00 to 10:30 a.m. The user selects a closed conference to ensure privacy. All sites in the conference receive a hard copy notification of the scheduled conference.
>
> A few minutes before 10:00 a.m., on the day of the conference, one of the participants involved walks into their conference room turns on a single switch to activate the local system. The system comes up in a loop-back (to the MCU and back) configuration wherein the local sites initially see themselves. Cameras at the sites are adjusted by the conference attendees until everyone in the room can be seen on the systems monitors. Microphones are arranged to make sure that all voices can be comfortably heard. At 10:00 a.m., the hub MCUs receives a signal from the workstation to complete the required conference interconnections and the conference begins. An ascending three-tone signal is automatically sent to all participating sites announcing the start of the conference. The originator makes the necessary introductions and initiates the conference.
>
> As the conference progresses, the sites' main video monitors switch between the sites of the dominant speakers, and graphics are sent and displayed on each sites' auxiliary monitors. Discussions center on the problems at hand and after twenty minutes, issues are cleared up and direction given. Since the conference is completed early, the originator brings up the scheduling menu at the local PC and selects the "End Conference Immediately" option. A descending three-tone signal is sent to all three sites, and each sites monitors return to a view of their own conference room. The conference is complete."

DATA TRAFFIC

The multiplexers and the dedicated full-time T1 (1.544 Mb/s) communication lines form the backbone of the communication network. Independent data transmission occurring simultaneous with a video conference is possible because the multiplexers share the use of the T1 bandwidth through division of the 1.544 Mb/s between data, video, and control traffic. During off-shift times or on weekends, when video conferencing is not being conducted, all but the control portion (19.2 Kb/s) of the T1 bandwidth can be manually allocated to data transmission tasks. Work is progressing to include automatic, dynamic network bandwidth control as part of the workstation functions.

PRIVACY

Scheduling PCs at each site as well as the workstations at the hubs are password protected. Video conferencing signal encryption is optional. Encryption can be performed in the CODEC and in the MCU through the use of encryption keys. Each site receiving encrypted transmission must use the same key.

Conference privacy can also be assured by making a conference closed to all but invited participants. Open conferences can be joined at any time. However, when an additional participant joins an open conference, his presence is announced by a brief series of audio tones. Closed conferences allow only those sites scheduled by the conference originator to participate. The conference is designated open or closed when it is scheduled on the originator's PC.

SECURE SYSTEM APPLICATIONS

NSA approved security is an easy extension of the above described network. DoD networks using the MCUs in secure conferencing configurations have been established by INEL engineers at the Strategic Air Command (SAC), at Offutt AFB; the Tactical Air Command (TAC), at Langley AFB; the Pentagon; Norfolk Naval Center; Cheyenne Mountain/Peterson AFB in Colorado; and the Material Air Command (MAC) at Scott AFB in Illinois. These networks typically employ KG-94/194 encryption devices at the Red/Black boundaries of the secure facilities. It should be noted that these systems use a combination of fiber links, microwave links, commercial telephone links, and satellite links as communication media, all working through the KG devices. These secure systems have worked reliably and well.

In secure applications, the multiplexers and nodal communication equipment exists on the black side of the secure boundary and receive black (encrypted) data from the secure areas, or non secure information from non secure areas. For video conferencing the MCU and CODECs must operate on unencrypted, or red data, thus requiring the MCU and CODECs at both end station and hubs to be located in secure (Red) areas. As long as Red/Black boundaries are observed, both encrypted secure and non secure data traffic can be handled simultaneously by the network.

GOVERNMENT AND COMMERCIAL BENEFITS FROM VIDEO CONFERENCING

Although these systems are new, benefits have already been realized. Among them are:

*** **Managers and key technical people are more reachable.**
*** **Lost productivity due to travel is reduced.**
*** **Less time is lost in clearing calendars and making travel plans.**
*** **Meetings can be held that normally would be cost and schedule prohibitive.**
*** **Meetings can be expanded to include additional personnel who would normally not travel.**
*** **A greater number of people can benefit from special expertise.**
*** **Meetings are better structured and more task-oriented.**
*** **People make firmer commitments to a person they can see rather than just hear.**
*** **Territorial issues do not have to be addressed, so the focus is more likely to be on the original purpose of the meeting.**
*** **Decisions can be reached faster.**
*** **Sessions can be recorded and replayed as needed.**

Although travel reduction is the most common reason for businesses and other agencies to consider teleconferencing, it isn't the biggest factor in savings. Our experience has shown that although travel decreases, the cost of design, installation, and maintenance of a video teleconferencing center offsets the saving in travel significantly. What really improves (thereby justifying the expense of a facility) is the response time to problems in development or delivery of a product. All the best minds of an organization, not just the ones who are free to travel, can be put on the problem at once, with immediate feedback.

It is also our experience that teleconferencing enhances group decision making, and that in situations involving conflict, it facilitates negotiation. Teleconferencing participants tend to be less dogmatic, and more compromising, allowing opinions to be changed more easily, resulting in formation of fewer coalitions.

WHERE VIDEO CONFERENCING IS HEADED

The pressing need in the video conferencing world is to reduce the number of "islands" of video conferencing which cannot intercommunicate, in favor of single networks that serve larger user communities. The various architectures must look forward to emerging standards and related technical developments that will support system interconnectivity. However, a dominating adherence to these standards is still a few years in the future. Pseudo standards tend to be set up by systems that have the most equipment in place. Growth and expansion, of necessity, perpetuates the acquisition and installation of compatible, usually single source equipment. Until equipment is truly standardized, networks large or small, must acquire certain critical equipment from a single manufacturer. This is more critically true for digital multipoint conferencing hardware than for the simpler point-to-point digital links. The development of a large nationwide network requires settling on a single architecture for certain critical components, including the CODEC, the MCU, and separately the multiplexers and communication backbone equipment prior to finalizing design.

INEL Engineering propose the development of new, or the expansion of existing systems into larger networks, to be designed to satisfy the following concepts, and comply with the configuration shown in Figure 2.

*** **Use of a large area network multiplexer backbone capable of serving a large number of sites.**

*** **Banks of MCUs used as free-floating conference servers assigned to users on a conferencing basis and not permanently assigned to any given user, thus significantly expanding the potential numbers of users.**

*** **Untended network accessible through any end station conference facility.**

*** **A large, user-friendly conference control database.**

*** **Dynamic network bandwidth control and allocation of data and video through the workstation database.**

*** **Simultaneous use of the dedicated bandwidth for video conferencing and other user information exchange needs to better utilize expensive communications costs.**

*** **Incorporation into the network of dial-up and switched services.**

INEL engineers contend that the network developed for DOE and the experience gained from the DoD installations provide a viable base from which new larger systems can be developed and to which older systems can migrate to establish nationwide video conferencing networks. Hardware compatibility will continue to be a problem for the next few years, however, despite these problems large multiuser, multipoint and multiservice conferencing and data networks are possible using today's technology.

This larger network would be similar in operation and application to the limited DOE networks discussed in the previous section. The primary differences would be (1) larger multinode, multiplexed communications network forming the backbone, (2) the use of MCUs as free floating servers assigned as needed to video conferencing users, and (3) the expansion of the control database to support a larger user community with multimedia, multidata, and multiuser services. The end station video equipment and network access and control equipment will remain much as it is in the above DOE system. Our existing database and network control will be expanded, along lines already established, to serve a larger user community.

Using an innovative structure such as that shown in Figure 2 assures compatible, low-maintenance systems accessible to a large user community at a minimum cost. Video conferencing is the wave of the future. Swift action to make it accessible nationwide should be a critical priority of both private industry and government agencies.

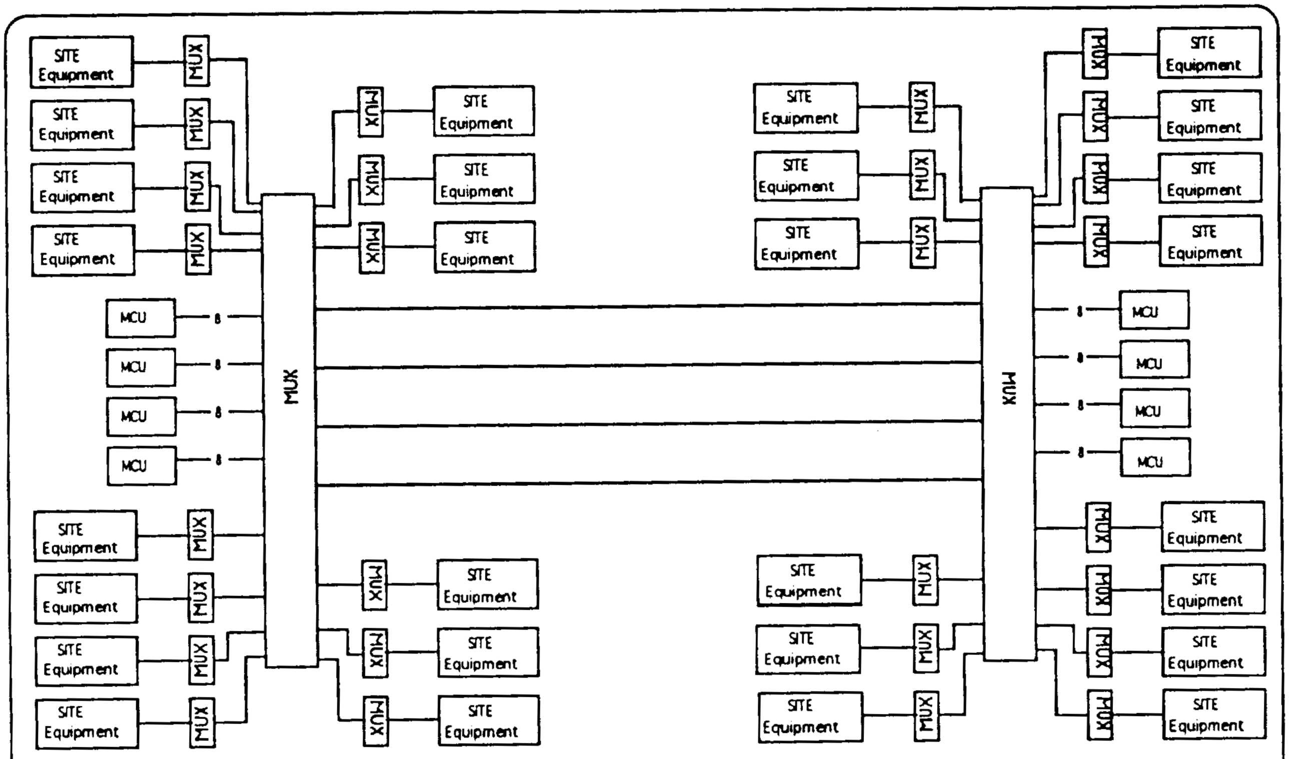


FIGURE 2

LARGE USER COMMUNITY VIDEO CONFERENCING AND DATA NETWORK

Omit

MANUFACTURING TECHNOLOGY PART 4: INTELLIGENT TOOLS

PRECEDING PAGE BLANK NOT FILMED

242 INTENTIONALLY BLANK

INTEGRATED FLEXIBLE MANUFACTURING PROGRAM FOR MANUFACTURING AUTOMATION AND RAPID PROTOTYPING

S.L. Brooks
C.W. Brown
M.S. King
W.R. Simons
J.J. Zimmerman
Allied Signal, Inc. - Kansas City Division [1]
Kansas City, MO 64141

ABSTRACT

The Kansas City Division of Allied Signal Inc., as part of the Integrated Flexible Manufacturing Program (IFMP), is developing an integrated manufacturing environment. Several systems are being developed to produce standards and automation tools for specific activities within the manufacturing environment. The Advanced Manufacturing Development System (AMDS) is concentrating on information standards (STEP) and product data transfer; the Expert Cut Planner system (XCUT) is concentrating on machining operation process planning standards and automation capabilities; the Advanced Numerical Control system (ANC) is concentrating on NC data preparation standards and NC data generation tools; the Inspection Planning and Programming Expert system (IPPEX) is concentrating on inspection process planning, coordinate measuring machine (CMM) inspection standards and CMM part program generation tools; and the Intelligent Scheduling and Planning System (ISAPS) is concentrating on planning and scheduling tools for a flexible manufacturing system environment. All of these projects are working together to address information exchange, standardization, and information sharing to support rapid prototyping in a Flexible Manufacturing System (FMS) environment.

INTRODUCTION

As industry strives towards a Computer Integrated Manufacturing (CIM) environment, many technological advances, such as computer aided design systems and computer controlled production systems, have been accomplished. However, these advances have created many islands of automation in which integration between these areas is still a labor intensive effort. As a result, an automated link must be established between the product definition and numerical control machines. The IFMP systems are knowledge based and reflect an open systems architecture. The projects are working together to develop an object-oriented database that houses a persistent representation of products, process plans, resource, NC, inspection, and other manufacturing support information. This integrated database will allow manufacturing personnel to share all essential manufacturing information in a common database, thus providing quick information access to the manufacturing community. The database will also allow concurrent processes to utilize the same information. The data structure within the database is an implementation of the Standard for Exchange of Product Data (STEP), an international standard. These data structures provide a standard format for information exchange inside and outside of the manufacturing facility. The IFMP systems utilize leading edge solid modeling technology and feature-base tolerancing to automate tasks that were impossible to automate in the wireframe world. Solid modeling and other leading edge technologies utilized within the systems open the door to meeting product requirements with a smaller fraction of money and resources and provide the capability to reprogram for quick production turn around of new product designs.

This paper presents a brief overview of the IFMP automation projects, how they are attempting to achieve interoperability and concludes with a description of the manufacturing environment these projects are being developed to support.

AUTOMATION PROJECTS

Advanced Manufacturing Development System

The AMDS system is a next generation product data translation and data management environment that is driven directly from international product data standards. AMDS embodies the technologies of distributed

[1] Operated for the United States Department of Energy under Contract Number: DE-AC04-76-DP00613

workstation database management, standard product data models, and high productivity object-oriented programming that will prepare KCD to adapt to the evolution toward open systems integration. Using AMDS , KCD has demonstrated that it can move files formatted in the emerging international Standard for the Exchange of Product Model Data (STEP) between commercial solid modelers and into a vendor independent distributed solid model product definition database. Using AMDS, KCD has demonstrated that it can transfer prismatic parts between three commercially available solid modelers. AMDS is used within the IFMP system as a transfer mechanism to move product definition from a client into the IFMP manufacturing database so that XCUT, IPPEX, and ANC can have concurrent access to product definition.

The AMDS architecture (see Figure 1) consists of two major pieces: a generator and a repository. The generator reads in an electronic form of STEP and automatically generates the repository which consists of an object-oriented database and in/out file translators for importing and exporting STEP ASCII files in standard format. The AMDS can be used in file exchange mode or applications can be written directly against the database. The AMDS generator guarantees that the database and its in/out file translators are consistent with each other. The repository is based on the ITASCA object-oriented database management system. ITASCA is a true object-oriented system and treats programs and data uniformly as objects. This uniform treatment of data and programs allows economies of programming that have greatly accelerated the development of AMDS software.

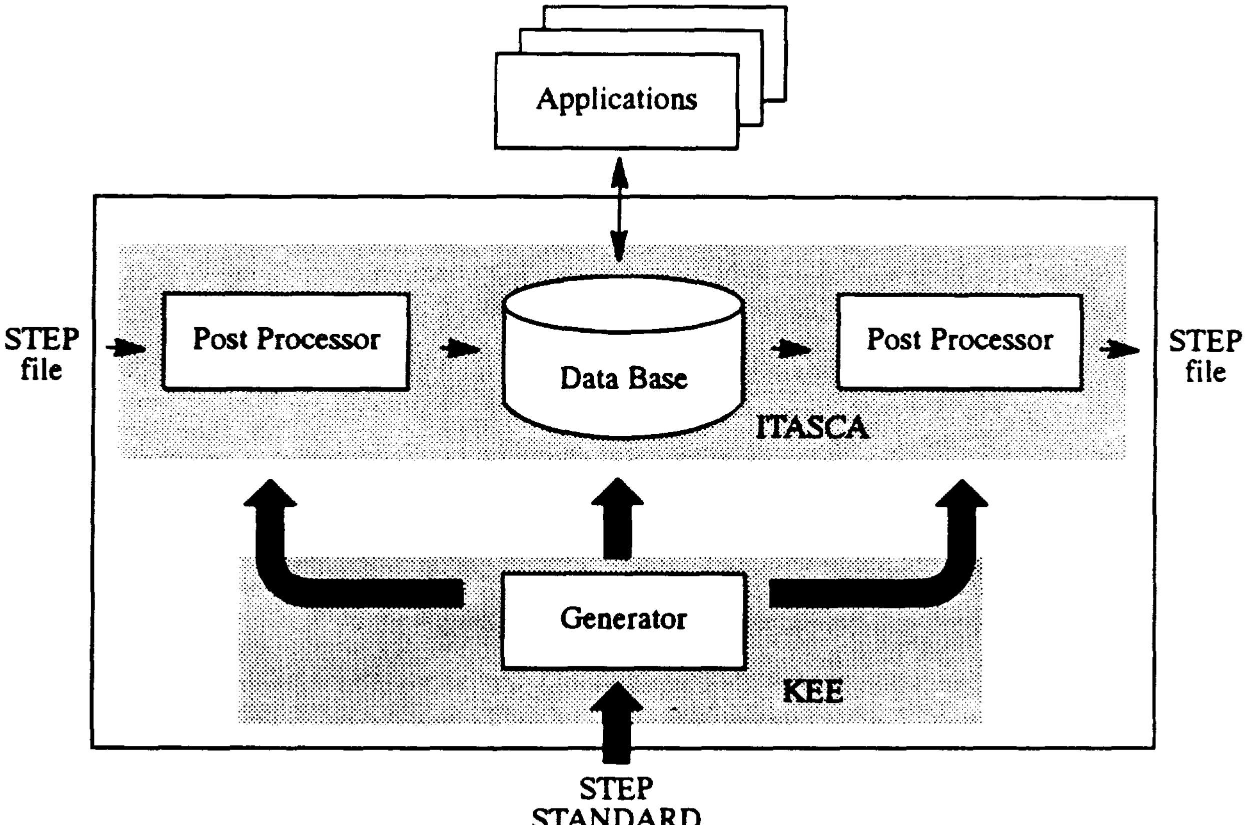


Figure 1. AMDS System Architecture

Expert Cut Planner System

A process plan is the sequence of operations necessary to transform raw material into a finished part. XCUT[1] maintains a persistent definition of products, and process plans in an object-oriented database. The information stored in the database is an implementation of the Product Data Exchange Specification (PDES), an international standard. XCUT incorporates the PDES models for product definition, geometric shape, form features, and process plans, among others. PDES data models specify the definitions of objects as well as an

ASCII exchange file format for transferring instances of those objects. Each object in the XCUT database has methods for storing and retrieving itself from the database and for reading and writing itself to an exchange file.

XCUT (see Figure 2) shares its object-oriented database with two other advanced manufacturing projects at the Kansas City Division, the advanced numerical control planning system ANC, and the inspection planning system IPPEX. All three systems will share the same information, providing seamless integration between process planning, NC, and inspection. Data modeling in EXPRESS-G and the EXPRESS language is used for defining objects and their relationships and attributes. The implementation of the database has been automated by a program that parses the EXPRESS files and generates C++ code and the database schema. The code generated by the parser produces all methods necessary for accessing objects from the database and for importing and exporting objects to data exchange files.

XCUT is linked with a solid modeling system to provide the spatial reasoning capabilities needed in process planning. The solid modeler provides visualization graphics and is used to identify the set of manufacturing features removed from the raw material to make the finished part. The definitions of all solid models used by XCUT are stored in its object-oriented database and are recreated in the solid modeler at run time.

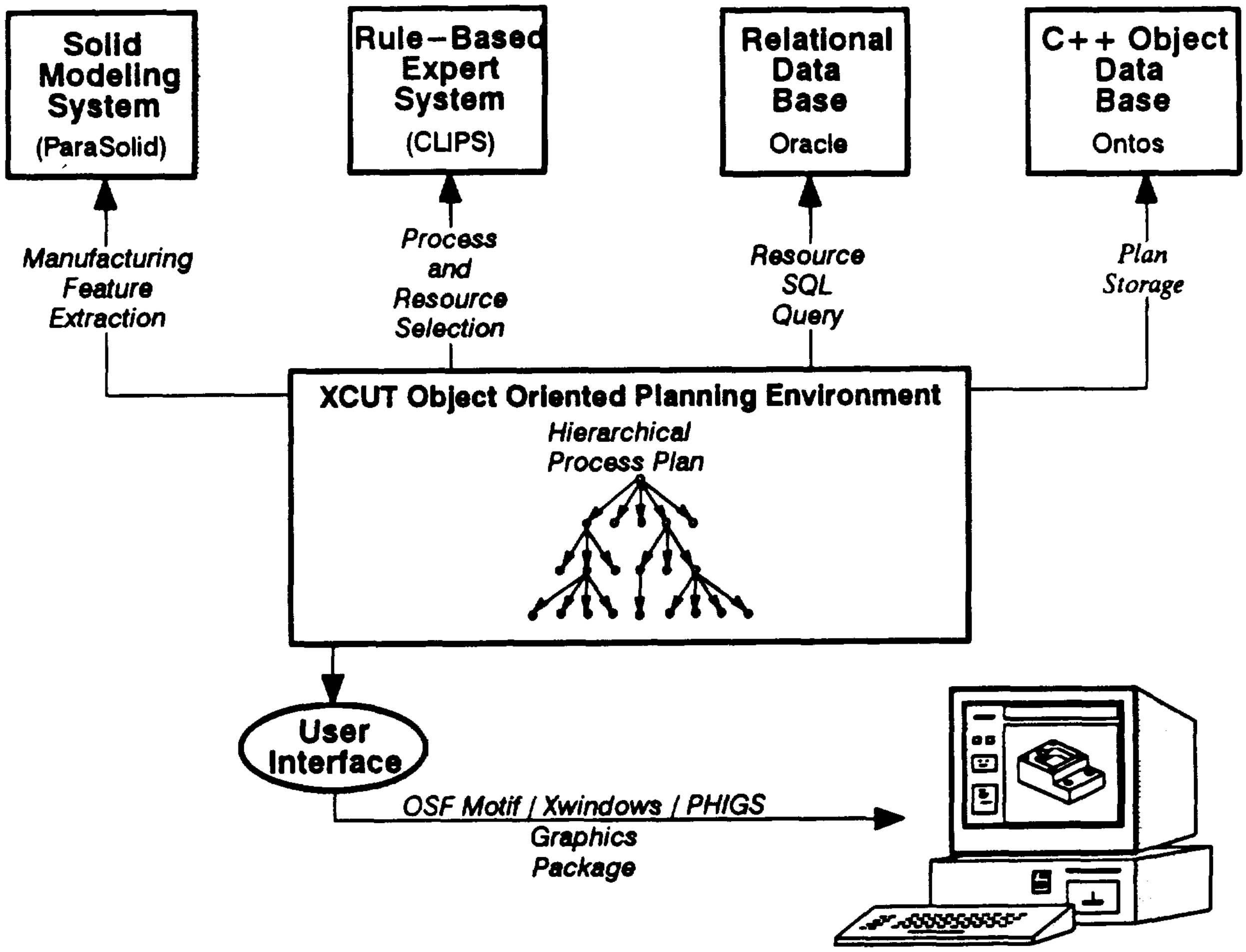


Figure 2. XCUT System Architecture

Advanced Numerical Control System

The numerical control (N/C) analyst's instruction set is embodied by a process plan. A process plan includes a sequence of the operations and processes necessary to transform raw material into a finished part[2]. The ANC system uses a process plan in an electronic form. The process plan is broken down into "bite sized" pieces that represent singular cutting tool operations or "single tool uses". A single tool use references one or many manufacturing features which provide the ANC system with design information describing the material to be removed. A sub-component of the manufacturing feature is a "delta volume". A delta volume is a solid model that provides a geometric representation of the volume of material to be removed. The manufacturing feature also includes non-geometric design attributes such as surface finish, tolerance, threads and edge conditions. Other information necessary for automation of N/C data preparation activities include machine tool characteristics, cutting tool characteristics, work holding devices and associated manufacturing resource information.

The ANC system (see Figure 3) uses an object-oriented database as a persistent repository for all of its information. An object-oriented kernel solid modeler provides the capability to answer arbitrary geometric questions algorithmically. The main component within the system is a solid model based manufacturing package. This software supplies solid model based toolpath generation capability. It also provides a user interface for viewing, manipulating and creating solid models.

"Knowledge based sub-systems are extremely important in the capture, reduction, packaging, expression and dissemination of the knowledge utilized in operating a manufacturing enterprise" [3]. The ANC system incorporates a knowledge based system to provide these capabilities. Manufacturing rules for this "system of experts" are being developed by a group of resident experts in the field of numerical control and process planning. An inference engine uses these manufacturing rules, along with the related part, machine and tool information to determine the appropriate feed rates, spindle speeds, depths of cut, step over distances and motion parameters required to generate an N/C toolpath.

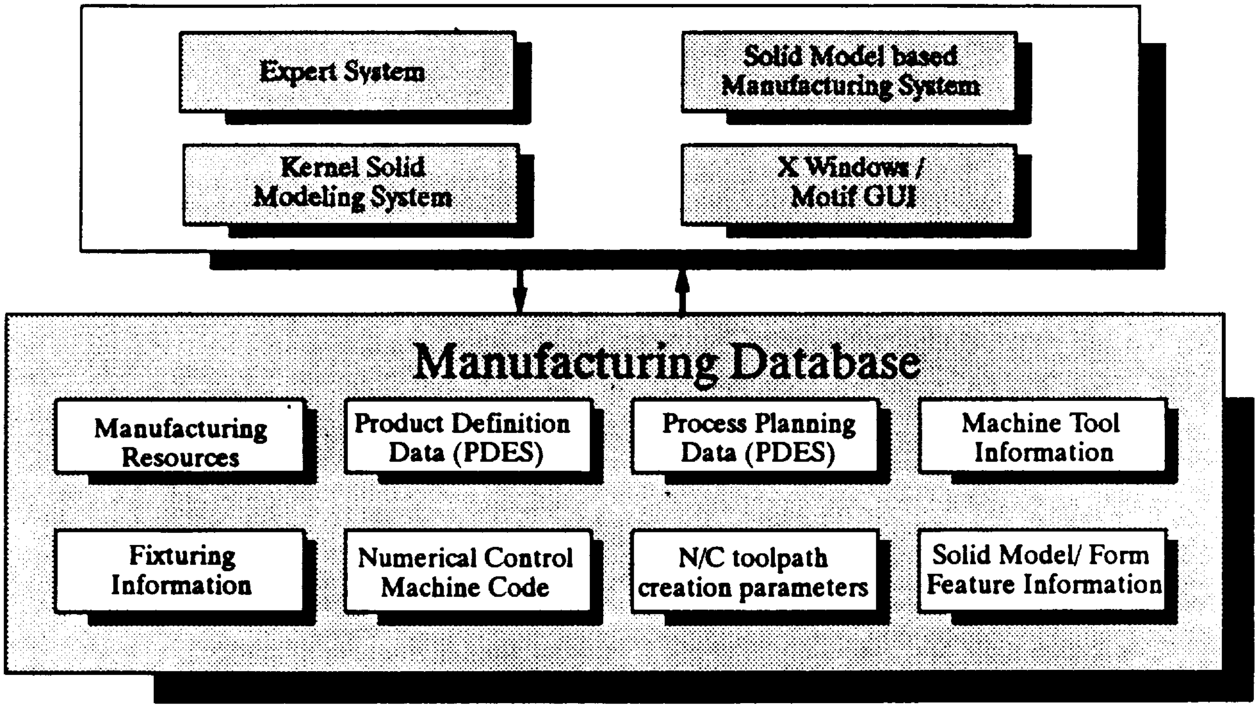

Expert System
Solid Model based Manufacturing System
Kernel Solid Modeling System
X Windows / Motif GUI
Manufacturing Database
Manufacturing Resources
Product Definition Data (PDES)
Process Planning Data (PDES)
Machine Tool Information
Fixturing Information
Numerical Control Machine Code
N/C toolpath creation parameters
Solid Model/ Form Feature Information

Figure 3. ANC System Architecture

The ANC system interprets process plans and analyzes supporting manufacturing, product and solid model feature information to determine appropriate motion controls for machining a part. The machinable volumes generated by XCUT are used as geometric input to the ANC system. The ANC system also utilizes process plans that are generated by XCUT. Solid models that represent the volume of space through which the cutting tool travels are generated for each single machining operation. These solid models are compared to solid models representing the fixture assembly to detect collision. They are also used to generate solid models that represent the shape of the part after a single machining operation is performed. These solid models (classified as In Process States) are used to verify the accuracy of the NC data, and to check for collision on entry motion for subsequent machining operations. The system also includes the capability to automatically regenerate NC data based on changes to the part design.

Inspection Planning and Programming Expert System

IPPEX (Inspection Process Planning EXpert)[4], is a knowledge-based system currently being developed for the dimensional inspection of piece parts at the KCD. The objective of IPPEX is to make CMMs more effective production support inspection tools by creating consistent and standard inspections, enhancing their productivity, and capturing inspection expertise. This is accomplished through applying product modeling, incorporating an explicit tolerances representation, establishing dimensional inspection techniques and embedding an inference mechanism.

The IPPEX system automates the generation of inspection process plans and part programs for measuring piece parts with coordinate measuring machines (CMMs). While the XCUT and ANC create plans and instructions for machining, IPPEX concurrently will create the appropriate inspection planning and generate the part programming code necessary for sample-point dimensional measurement[5]. The IPPEX inspection activities will integrate with the XCUT activity plan to create a final product process plan. Given the inspection scope, defined by feature-based tolerances in the product model, IPPEX will plan the sequence of operations necessary to verify that the manufactured part conforms to requirements. These operations will contain activity objects that will identify resources such as measuring machines, part set-ups, and probe configurations, and tasks such as establish datum reference frame, verify tolerance, and measure feature. The process plans will also contain inspection techniques based upon the feature's current measurement criteria. The inspection techniques determine the number of sample points, the distribution of these sample points, and the selection of the appropriate substitute geometry algorithm. Based upon this inspection process plan, a Dimensional Measurement Interface Standard (DMIS)[6] formatted CMM part program will be created along with the appropriate part set-up and probe configuration support documents.

As illustrated in Figure 4, the IPPEX system consists of five major components: a user interface, a product modelling system, a relational database management system, an object-oriented database system, and an expert system environment which involves an inferencing mechanism and multiple knowledge-bases. The user interface provides the user access to IPPEX's functions. The product modeler supplies the product definition information. The expert system environment controls the inspection knowledge bases and the inference mechanism. The databases contain resource data and plan storage

The current IPPEX prototype system runs on an HP/Apollo engineering workstation. The user interface is an icon-based menu-driven module which interfaces the user to the product modeler and the IPPEX planning system. The current product modeling system involves the Parasolid Solid Modeler[7] complimented by CAM-I's Dimension and Tolerance (D&T) Modeler[8]. The connection to the geometric solid model and D&T information is acquired through the modeler's application programmers interface subroutine library. The data in the relational database are programmatically accessed through embedding SQL constructs in the C language. Finally, the current inferencing environment is NASA's C Language Integrated Production System (CLIPS)[9]. CLIPS is a production rule-base system. It is activated by the IPPEX system through C functions to CLIPS. IPPEX's application routines can initiate CLIPS, assert new facts into the CLIPS fact base, retract facts from the fact base, execute a set of rules, and transfer inferenced decisions.

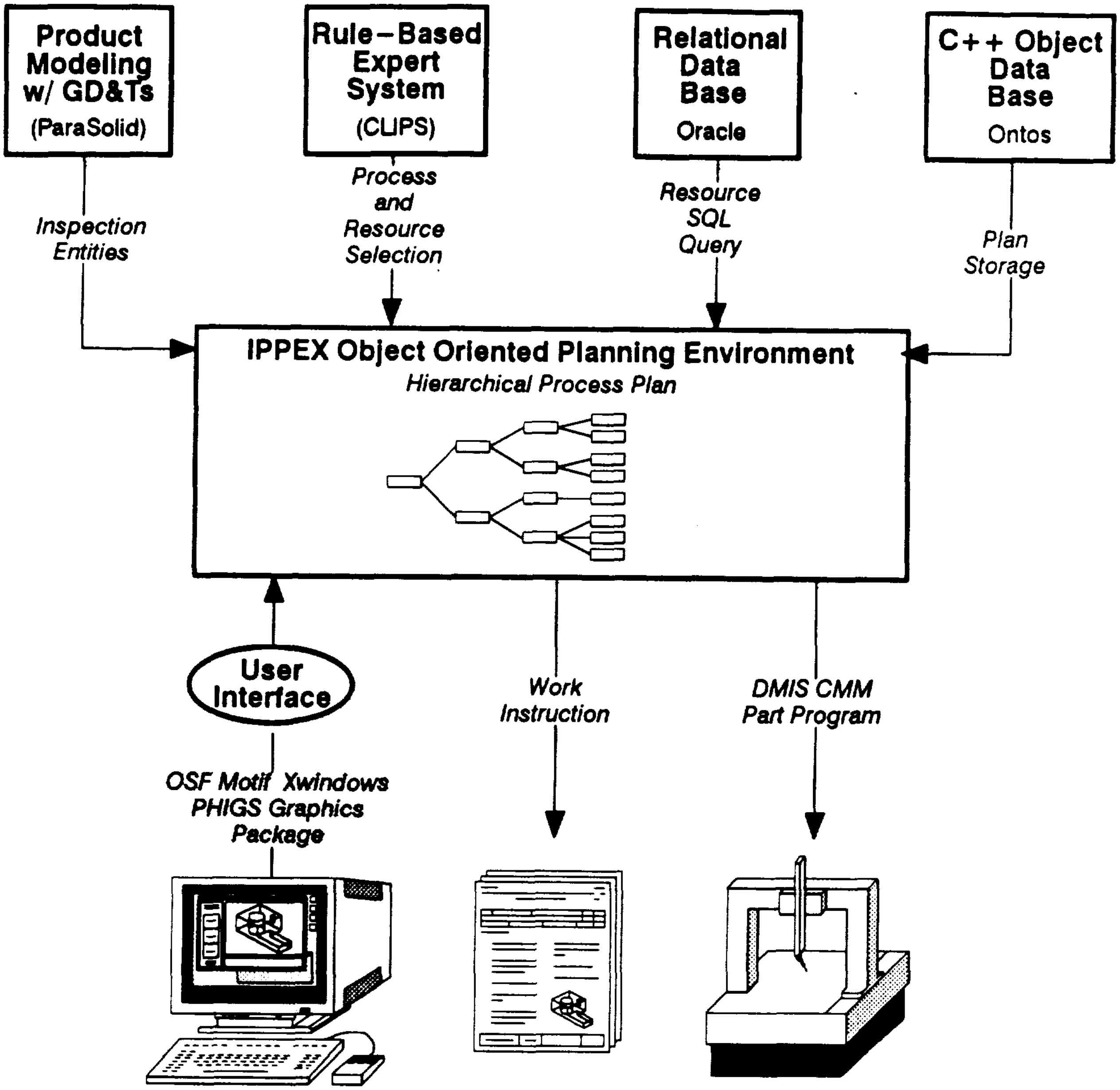


Figure 4. IPPEX System Architecture

Intelligent Scheduling and Planning System

ISAPS[10] is a scheduling and planning tool for shop floor personnel with the responsibility of producing discrete, machined electrical component housings in a Flexible Manufacturing System (FMS) environment. The ISAP system (see Figure 5) has two integrated components: the Predictive Scheduler (PS) and the Reactive Scheduler (RS). These components work cooperatively to satisfy the four goals of the ISAP system, which are: 1) meet production due dates, 2) maximize machining center utilization, 3) minimize cutting tool migration, and 4) minimize product flow time.

The PS is used to establish schedules for new production requirements on a variable planning horizon. It provides finite capacity scheduling for six machining centers, two coordinate measuring machines (CMMs), one automatic wash station, and five manual stations for tooling, part, and fixture preparation. The RS is used to adjust the schedules produced by the PS for unforeseen events that occur during production operations, such as equipment failures, changing priorities, and product mix.

A common model of the FMS is employed by the PS and RS which defines the basic system configuration and availability of resources to be considered for scheduling. The PS and RS subsystem prototypes have been developed using KEE[11], an expert system shell from IntelliCorp, and Common Lisp. The schedules developed by the ISAP system have been validated using a discrete event simulation model of the FMS. The prototypes are currently being converted to the object-oriented C environment, ProKappa[12], also from IntelliCorp.

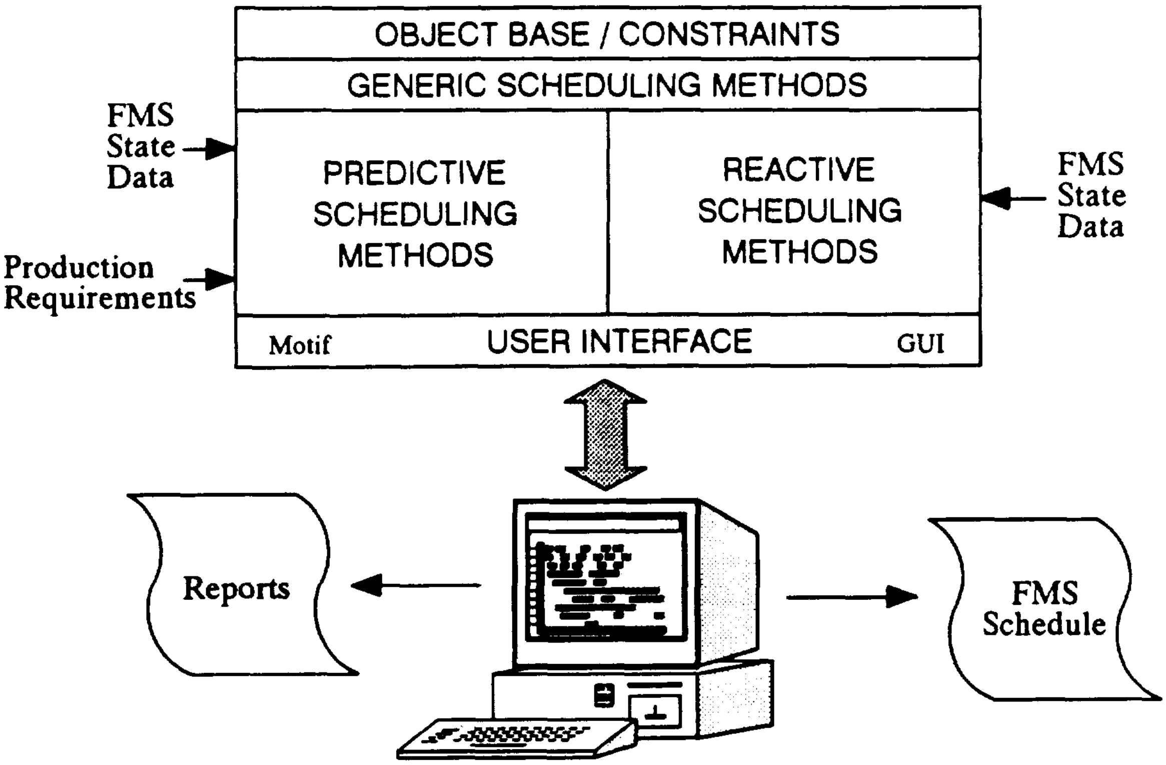

OBJECT BASE / CONSTRAINTS
GENERIC SCHEDULING METHODS
FMS State Data
Production Requirements
PREDICTIVE SCHEDULING METHODS
REACTIVE SCHEDULING METHODS
FMS State Data
Motif
USER INTERFACE
GUI
Reports
FMS Schedule

Figure 5. ISAPS System Architecture

IFMP SYSTEM INTEGRATION

The IFMP is in the process of integrating all of the previously described systems into one seamless manufacturing environment. Figure 6 illustrates the flow of information that is required to rapidly produce products targeted for the Flexible Manufacturing System at KCD.

Product definitions will be obtained in electronic form from an outside client in STEP ASCII standard format. The AMDS system will automatically translate this information into product definitions to feed XCUT, IPPEX and ANC. XCUT will use the definition to generate the process plans for the machining operations. IPPEX generates similar plans for the inspection process. These plans are joined together to produce the production plan for the product. ANC takes the production plan, along with the product definition from AMDS, and generates the required N/C programs to machine the product. The second pass of IPPEX generates the DMIS part inspection program. The N/C and DMIS programs are post-processed to their respective native machine codes and electronically shipped to the FMS. ISAPS uses the information embedded in these programs and master production schedule requirements to determine the appropriate schedule to manufacture the product.

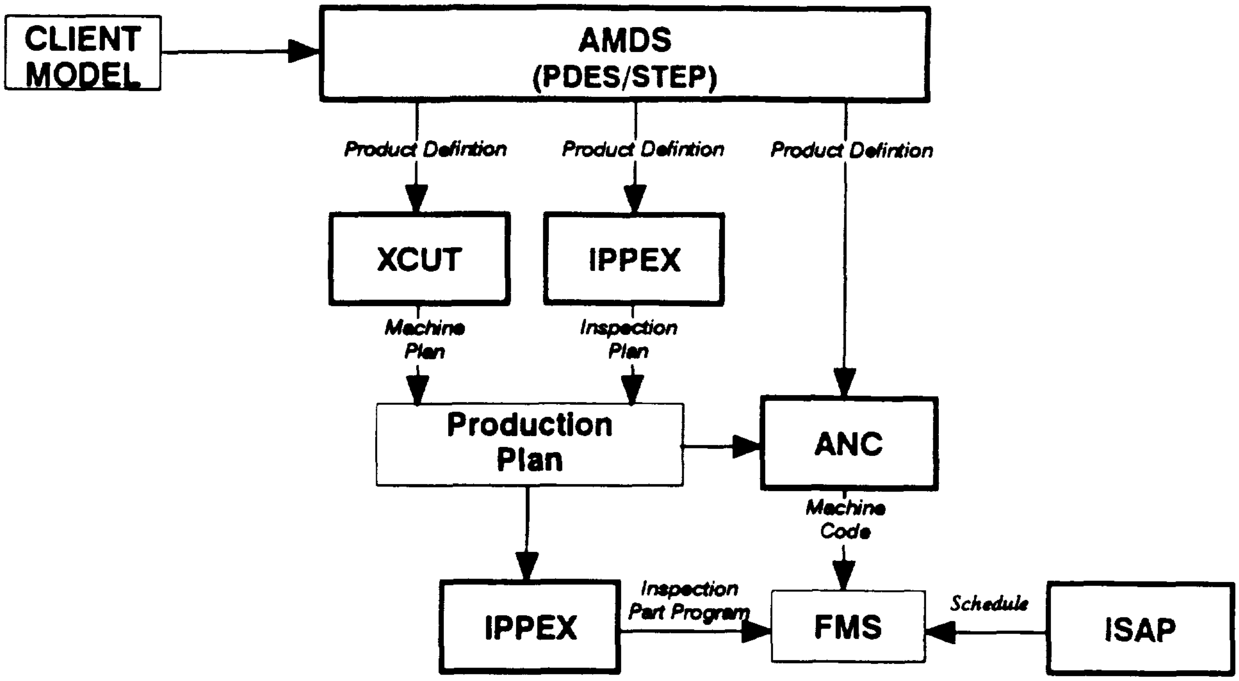


Figure 6. IFMP Information Flow

REFERENCES

[1] Hummel, K.E., Brooks, S.L., and Wolf, M.L., "XCUT: An Expert System for Generative Process Planning," *International Industrial Engineering Conference*, Toronto, Ontario, May 12–18, 1989.

[2] Wolf, M.L., "AIS in Automated Process Planning", *CAM–I* May 17–18, 1988.

[3] Preiss, K., *21st Century Manufacturing Enterprise Strategy*, Iacocca Institute, Lehigh University, Vol. 2 pg. 27, November 1991.

[4] Brown, C.W., "IPPEX: An Automated Planning System for Dimensional Inspection", *Proceedings of the 22nd CIRP International Seminar on Manufacturing Systems: Computer Aided Process Planning*, Vol., 20 No. 2, 1991.

[5] Brown, C.W., "Dimensional Inspection Techniques for Sample-Point Measurement Technology", *Precision Engineering Journal*, American Society for Precision Engineering, Vol. 14, No. 2, April 1992, pp. 110-112.

[6] ANSI/CAM–I 101–1990 "Dimensional Measuring Interface Specification", *Computer Aided Manufacturing – International*, Arlington, TX., 1990.

[7] "Parasolid v3.0 Reference Manual", June 1990., Shape Data Ltd.

[8] Ranyak, P. S.; Fridshal R. "Features for Tolerancing a Solid Model", *Proceedings of 1988 ASME International Computers in Engineering Conference and Exhibition at San Francisco, CA.*, July, 1988, pp. 275–280.

[9] Culbert, C. J. "CLIPS Reference Manual", *National Aeronautics and Space Administration*, Version 4.0, March 1987.

[10] King, M.S., et al, "ISAPS – Intelligent Scheduling And Planning System", *Proceedings of the Design Productivity International Conference*, Honolulu, Hawaii, Vol. 2, Feb. 6–9, 1991, pp. 661–666.

[11] "KEE Software Development System Core Reference Manual", Version 3.1, IntelliCorp, Inc. Mountain View, CA, July, 1990.

[12] "PROKAPPA Programmer's Reference Manual", Version 2.0, IntelliCorp, Inc. Mountain View, CA, October, 1991.

ALLIED SIGNAL, INC.
TECHNICAL CONTACTS FOR IFMP PROJECTS

- **AMDS:**
  John Zimmerman
  Staff Engineer
  (816) 997-2932

- **XCUT:**
  Steve Brooks
  Staff Engineer
  (816) 997-4329

- **IPPEX:**
  Curtis Brown
  Staff Engineer
  (816) 997-3548

- **ANC:**
  Bill Simons
  Staff Numerical Control Analyst
  (816) 997-4739

- **FMS / ISAPS:**
  Mike King
  Staff Engineer
  (816) 997-5175

- **TECHNOLOGY TRANSFER:**
  Dennis Stittsworth
  Manager Technology Transfer
  (816) 997-4596

AN EXPERT SYSTEM FOR SUPERPLASTIC FORMING IN CONCURRENT ENGINEERING ENVIRONMENTS

This paper was withdrawn from presentation

AUTOMATED FIBER PLACEMENT COMPOSITE MANUFACTURING: THE MISSION AT MSFC'S PRODUCTIVITY ENHANCEMENT COMPLEX

John H. Vickers
NASA Marshall Space Flight Center
MSFC, AL 35812

Larry I. Pelham
Thiokol Corporation, Space Operations
MSFC, AL 35812

ABSTRACT

Automated fiber placement is a manufacturing process used for producing complex composite structures. It is a notable leap to the state-of-the-art in technology for automated composite manufacturing. The fiber placement capability was established at the Marshall Space Flight Center's (MSFC) Productivity Enhancement Complex in 1992 in collaboration with Thiokol Corporation to provide materials and processes research and development, and to fabricate components for many of the Center's Programs. The Fiber Placement System (FPX) was developed as a distinct solution to problems inherent to other automated composite manufacturing systems. This equipment provides unique capabilities to build composite parts in complex 3-D shapes with concave and other asymmetrical configurations. Components with complex geometries and localized reinforcements usually require labor intensive efforts resulting in expensive, less reproducible components; the fiber placement system has the features necessary to overcome these conditions. The mechanical systems of the equipment have the motion characteristics of a filament winder and the fiber lay-up attributes of a tape laying machine, with the additional capabilities of differential tow payout speeds, compaction and cut-restart to selectively place the correct number of fibers where the design dictates. This capability will produce a repeatable process resulting in lower cost and improved quality and reliability.

INTRODUCTION

A rather unique situation exists at the Productivity Enhancement Complex (PEC) in that NASA engineers and scientists work together with industry and university experts to solve material and process problems and perform advanced research and development. Each of the research cells is designed to concentrate on a specific materials need, a specific process or set of related processing activities. This arrangement allowed us to combine the new fiber placement capability with existing filament winding, tape laying, tape wrapping, pultrusion and ancillary facilities.

Developing manufacturing technology is one of the most important challenges in the field of composites. The need exists to establish manufacturing methods that meet the requirements of both superior material properties and program economics. Fabrication of components using composite materials prompts two primary catalysts for our research and development initiatives: the characterization/optimization of the composite processes and fabrication of components for the many projects throughout the Center. To characterize and optimize the process, we must first understand and define the critical parameters involved. To do this, we use a design of experiments statistical methods that will evaluate the available data and provide an understanding of the interactive effects of these parameters. From this understanding, we are able to sustain a repeatable process resulting in lower cost, improved quality and reliability and enhanced performance. We will achieve a process that we can control and provide ourselves and industry alternative solutions for materials and processing selection. At the MSFC we also have an advantage over industry in that we are not under the gun to meet production schedules with this equipment. We can develop and implement improvements without interruption and not just hold the status-quo. The second impetus mentioned was to fabricate components for the Center's programs. Working with the Program offices, the Science and Engineering Organizations, the Technology Utilization Office, and others, we provide very sophisticated flight components, lightweight robotic end effectors, or simpler secondary structures, brackets and fixtures.

THE FIBER PLACEMENT SYSTEM MECHANICS

The fiber placement system is shown as it is located at the PEC in Figure 1. Fiber placement represents one of the key elements of composite manufacturing technology and is a result of efforts to improve and merge capabilities from filament winding, tape laying, robotics and machine tool rigidity. The fiber placement process utilizes unidirectional prepreg tow material that is applied by compressing the tow material between a roller and part mandrel during movement of the machine and/or mandrel. The fiber placement machine is robotically programmed to maintain the compaction roller on the mandrel. With linear and rotational motions and the capability to vary the number of tows it can precisely place material onto complex surfaces. Repeated application of the composite material onto the mandrel and underlying layers is continually compacted by the roller, thus forming a consolidated laminate.

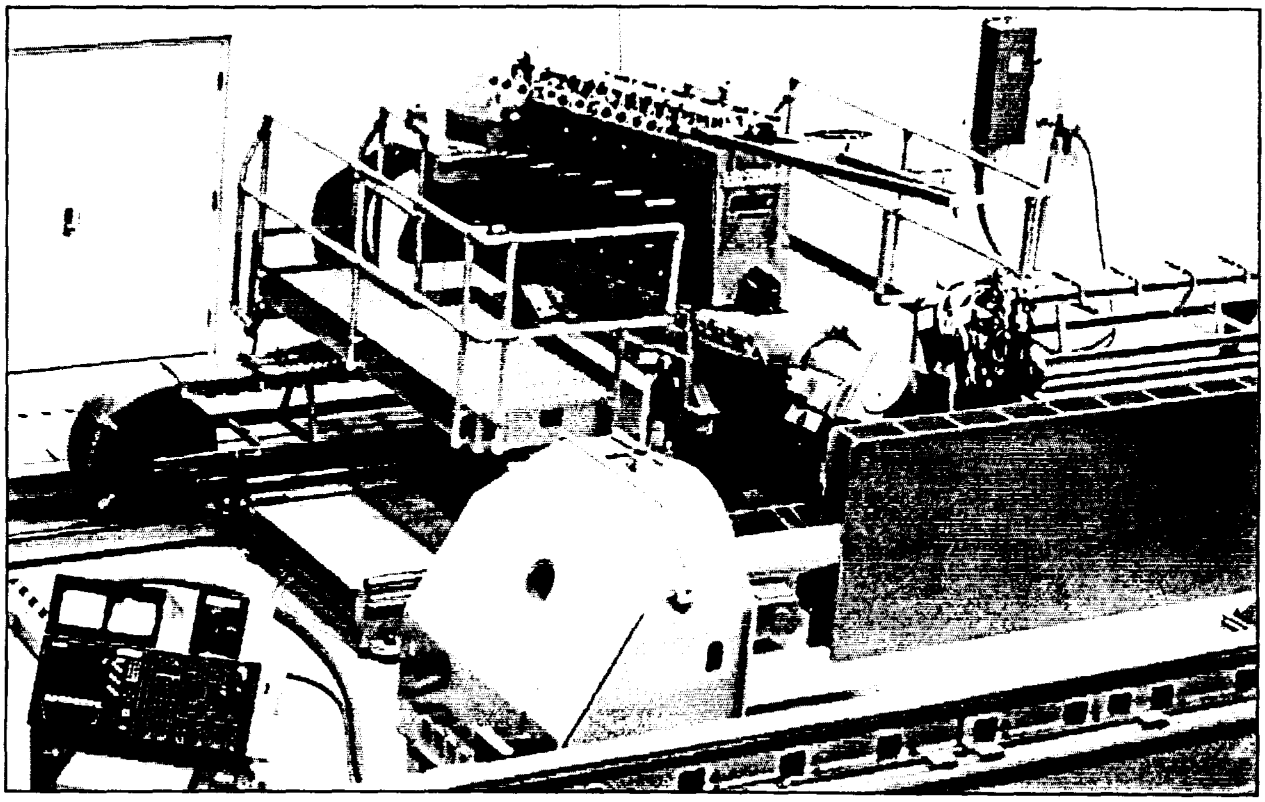


FIGURE 1. VIEW OF THE FPX

The FPX machine has seven major axes of motion as shown in figure (2). It is computer numerically controlled (CNC). The carriage and bed have two linear axes (X and Z) and a tilt axis (Y). The mandrel is located parallel to the carriage and has rotational motion (C axis). The head is attached to the robotic wrist that is capable of three axes of motion: yaw (I), pitch (J), and roll (K). All of the machine axes are employed to keep the compaction roller normal to the mandrel surface.

The machine is capable of applying up to 24 individual tows at a time producing 3-inch wide collimated fiber array. The tow material is delivered from a creel located on the crossfeed bed and equipped with bi-directional tensioners capable of retracting material to avoid slack as the machine moves through its motions. The tows course through guide rollers to a servo-controlled redirect roller located on the head to maintain fiber alignment throughout the machines motions. Certainly, the most sophisticated component of the machine is the delivery head. The computer controlled delivery head precisely dispenses, cuts, clamps, and restarts tow material automatically. The fiber array travels from the redirect roller to the cut/clamp/restart mechanism (CCR). The CCR's are where individual tows are cut to drop off segments of the array, and restarted to resume that segment of the array. To maintain the proper adhesive characteristics (tack), heating and cooling sources are strategically located within the head. The machine is capable of holding parts which are 84 inches in diameter, 342 inches long, and 20,000 pounds.

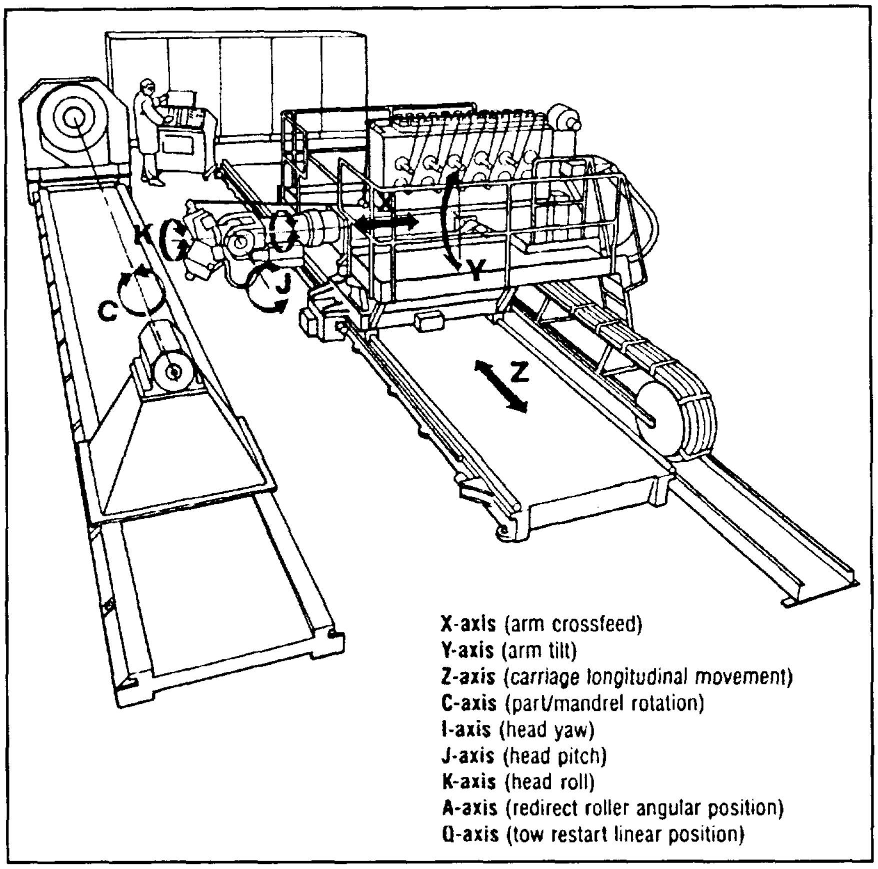

K
X
Y
C
J
Z
X-axis (arm crossfeed)
Y-axis (arm tilt)
Z-axis (carriage longitudinal movement)
C-axis (part/mandrel rotation)
I-axis (head yaw)
J-axis (head pitch)
K-axis (head roll)
A-axis (redirect roller angular position)
Q-axis (tow restart linear position)

FIGURE 2. SEVEN MAJOR AXIS OF MOTION

Machine control is achieved with Milacron's Acramatic 975F CNC. The control system uses 32-bit buss type open architecture incorporating two 386 processors, two 387 processors, and nine 186 processors to provide high speed instantaneous control to 86 distinct mechanical and pneumatic devices. SDRC's *I-DEAS*™ software is used to provide a complete offline programming system to implement a comprehensive mechanical design automation system. The system software is integrated for design, analysis, and manufacturing. A solid model is generated to provide a complete 3-dimensional product definition that can be easily analyzed to determine component performance and manufacturing parameters. Engineering analysis is performed directly from this model as is a manufacturing simulation. Using this approach, alternative designs are optimized to meet established engineering and manufacturing criteria. Productivity is substantially enhanced using this topdown integrated manufacturing strategy. Design, analysis, simulation, machine programming, and data management are all administered from within one software environment.

THE FIBER PLACEMENT PROCESS

The prevailing method for producing composite components with highly complex geometries is by manual layup. Manual layup of these components is simply not productive and until now, the automated machines available for manufacturing were also less than adequate. Filament winding and tape laying machines are the most widely used methods for manufacturing composite components. The FPX will not render either of these processes obsolete, what it will do is fill the role where filament winders and tape layers machine limitations fail specific geometries. The FPX expands the boundaries of composite processing erasing previous impediments to manufacture complex structures from advanced composite materials both efficiently and reliably.

With a wide variety of integrated machine technology, the FPX is able to provide one-step fabrication of symmetrical or asymmetrical, simple or complicated composite structures. Figure 3 illustrates the method by which precise placement of the tow material is achieved to comprise the laminate component. A mechanical compaction roller laminates the tow onto the mandrel or part surface. By mechanically pressing the tows onto the surface, entrapped air and inner band gaps are eliminated. Uniform compaction reduces debulking requirements, processes concave and asymmetrical surfaces, and supports the fiber steering capability. Fiber steering is achieved by differential tow payout speed and compaction, allowing continuous fibers to be directed around openings to eliminate machining or unnecessary buildups. Fiber orientation that is precisely controlled will counter shear stresses heretofore necessarily considered in the design. Fiber steering capabilities are shown in Figure 4. One or more tows are delivered by a method of cut/clamp/restart, programmably controlled individual tows can be started and stopped to precisely place material. The ability to add and drop tows can maintain part boundaries and uniform part thickness by eliminating overlap or increase thickness where the design dictates. This saves valuable material and eliminates manual insertion of material. The combined capability to cut/restart tows with in-process compaction is the first automated method to produce a constant zero degree wind angle on concave or convex structures. Machine tool quality and rigidity and high torque brushless servodrives maintain linear axis repeatability within .002 inches. The FPX uses advanced composite prepreg tow material that in all cases does not exhibit the same tack properties. To conform to different material properties, provisions for variable heating and cooling are incorporated into the delivery head.

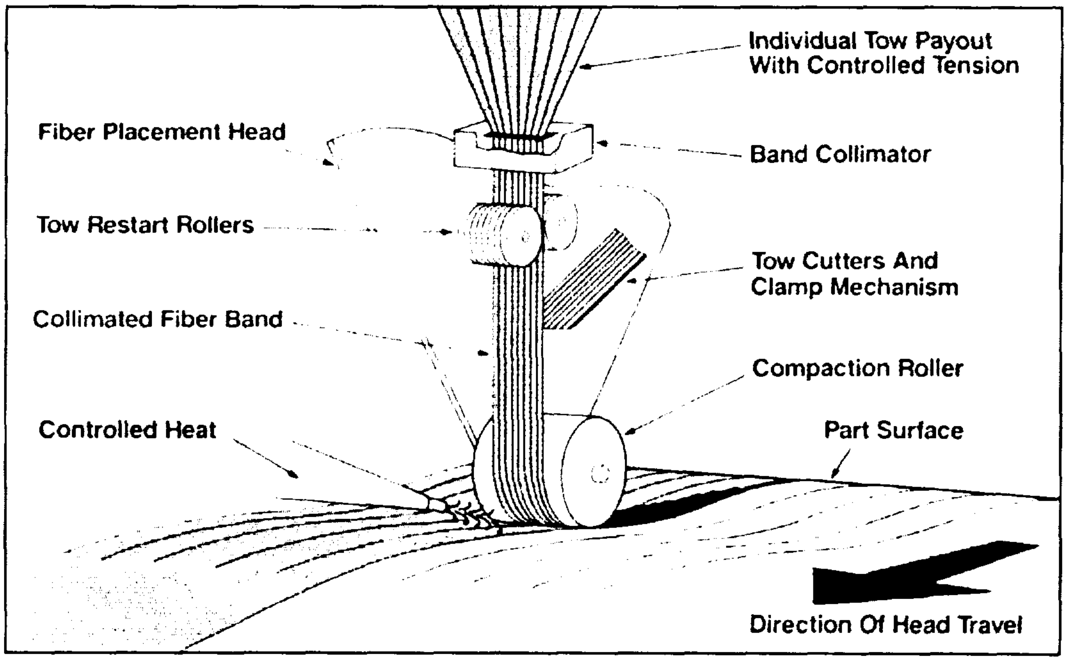


FIGURE 3. FIBER PLACEMENT APPLICATION

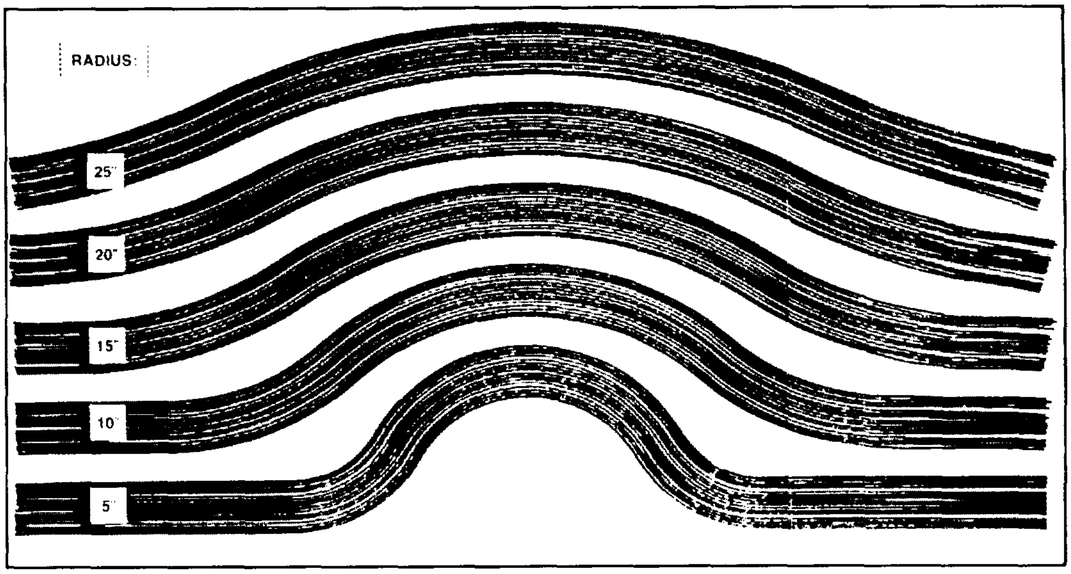


FIGURE 4. FIBER STEERING

COST CONSIDERATIONS

Performance benefits have been the incentive for using composite materials for aerospace applications in that performance has traditionally outweighed cost. Composite structures have a reputation for high cost that is not necessarily deserved. Generally, this misconception stems from comparison of raw material costs with other structural material choices like aluminum; typically prepreg graphite-epoxy can cost seven to ten times more per pound than aluminum. Recent studies however, have integrated total costs for design, manufacturing, assembly and support and in many cases, composites have provided a cost savings as well as the performance benefits. Of the many considerations that affect the cost of composite structures, one of the more important is design. Designers must be educated so that they no longer simply utilize composite materials in aluminum designs. There is a diverse, flexible family of systems are available for selection. Automation usually corresponds to lower cost and in the right circumstances this is appropriate. The FPX has feedrates of up to 2400 inches per minute and coupled with in-process compaction, can provide material application rates of over 6 times that of hand layup. Since products are produced with near net shape, scrap rates of up to 40% can be negated. However, problems and discrepancies arise when sophisticated, expensive automated systems are specified for components that could be efficiently manufactured without automation. The choice of fabrication processes for specific composite components should always be thoroughly investigated and with the appropriate strategy, the costs can be significantly less than for equivalent metal components. Concurrent engineering within the design and manufacturing disciplines will ultimately most influence the cost of composite structures.

MSFC'S FPX MISSION OBJECTIVES
PROGRAM PLANS

The multiyear plan for fiber placement consists of a set of programs and activities that will retain and extend our leadership in aerospace manufacturing. The MSFC Fiber Placement System (FPS) is unique and has not been used to fabricate flight quality structures to date. A significant level of confidence will be gained in the manufacturability of these structures. Specific optimal design and manufacturing parameters will be instituted and a database will be established so that other research and flight programs could withdraw information as well as make contributions. This work will transfer the knowledge of technology to an Industry that is avidly awaiting published study results to provide the basis for proposal of composite materials using these manufacturing methods. This study will provide MSFC and Industry manufacturers a baseline understanding of the manufacturing effects on product performance.

Optical Structures Programs

The use of composite materials for optical bench structures has increased significantly over the last few years. The primary justification for selection of these materials is the ability to maintain precise focal length without thermal control or active focusing systems; coefficients of thermal expansion of zero can be obtained. While structures are successfully fabricated for this application, they are manufactured as one-of-a-kind and primarily by hand. This considerable manual element of the fabrication process reduces repeatability thereby inhibiting analysis and correlation of resultant data. The objective of this research project is to utilize a fully automated processing method to fabricate an optimally designed optical bench. Research associated with optical structures has to date been limited to design and analysis and has not considered the inherent deficiencies of the processing equipment and operator variability. Automation of the fabrication process will reduce processing and operator variability. The optical benches for this study will be designed, manufactured and verified to obtain precise values of axial and radial thermal vacuum expansion. Using current analytical methods, we will develop a model of an optimally designed optical bench that is similar in functional characteristics to an existing design so that its performance characteristics can be compared to existing data for flight hardware. A structure approximately 8.0 inches in diameter, 30 inches in length and .20 inches in wall thickness would correspond to typical structures fabricated in house.

We will use a statistical design of experiments approach to determine the critical processing parameters and interactions that affect the performance of the optical bench. These relationships can be used to develop a prediction equation for determining end performance mean response values. The fiber placement equipment will be used to fabricate test samples and optical bench structures to verify predicted performance characteristics. Mechanical properties and thermal-dimensional stability testing will be conducted to establish performance characteristics.

RSRM Composite Stiffener Ring

The Redesigned Solid Rocket Motor existing stiffener ring design is inadequate for chance splashdown load conditions that occur unpredictably. A damage condition known a "cavity collapse" may result due to these insufficient structural margins. An investigation is underway to demonstrate manufacturability of a composite replacement stiffener ring to alleviate this problem by increasing these margins with a creative design. The design can be fabricated only manually or with fiber placement. Fabrication of a subscale composite stiffener ring will provides a processing analog to confirm the FPX manufacturing capability.

Other Programs

The Materials and Process Laboratory manages the Productivity Enhancement Complex at the MSFC, where the composite manufacturing capability resides. Affiliated personnel are continually involved in manufacturing, evaluation and review of existing and planned programs both in house and for contracted efforts. The availability of multiple process methods at the PEC will permit us to perform process comparison studies on specific projects as well as pure investigative studies. Previously we have fabricated the optical benches for the Space Science Laboratory's (SSL) Mission "Multi-Spectral Solar Telescope Array" that was successfully flown May 13, 1991. Also for the SSL optical bench structures were fabricated for the Water Window Imaging X-Ray Microscope. An optical bench was fabricated for the University of Alabama Huntsville's "Newton Telescope" to be used by the Students for the Exploration and Development of Space (SEDS). The Solar X-Ray Imager (SXI) is an MSFC in house project that will require design, development and verification of a flight telescope using a composite optical bench that is deliverable in 1995. Many of the DD&V activities for the Solar X-Ray Imager will directly benefit from these study results. AXAF-S is an earth orbiting X-ray spectrometer that will use a composite optical bench structure also to be fabricated at the PEC. Past MSFC programs that could have benefitted include Hubble Space Telescope and the Soft X-Ray Telescope. We are currently assisting AXAF-I with their evaluation of contracted efforts to fabricate composite optical bench structures. Surveys have indicated that many more programs could benefit from the fiber placement technology: joint IR&D programs, Technology Transfer, and other facility usage agreements have been proposed by government and aerospace contractors. Each of these programs will be evaluated so that appropriate priorities can be determined.

CONCLUSION

The FPX is a prominent addition to the MSFC in house advanced technology facilities. To augment this capability, the proposed work and other research programs will strengthen our in-house expertise and showcase our can-do-ability. With continued emphasis on the MSFC's composite manufacturing capability, we will confirm and retain a world class status. Since technology is a vital ingredient in the Nation's economic competitiveness, it is clear that it will continue to be one of NASA's principal goals to achieve technology transfer, and do it largely through direct interactions between researchers and engineers. Fundamental research and employment of innovative concepts like fiber placement will provide a continuum of technological development for America's space program. By committing ourselves and our resources to restoring our Nation's technology base, we can assume and maintain the leadership in many key industries.

REFERENCES

1. "Engineered Materials Handbook." Volume 1. Composites, ASM international, May 1988.

2. Strong, Dr. A. Brent: "Fundamentals of Composite Manufacturing: Materials, Methods, and Applications." Society of Manufacturing Engineers, 1989.

3. Schwartz, Mel M.: "Composite Materials Handbook." McGraw-Hill Book Company, 1984.

4. Enders, Mark L.: "The Fiber Placement Process." Thiokol Corporation, 1992.

5. Evans, Don O. and Others: "Fiber Placement Process Study." SAMPE 34th Symposium, May 1989.

6. Robinson, M. J.: "A QUALITATIVE ANALYSIS OF SOME OF THE ISSUES AFFECTING THE COST OF COMPOSITE STRUCTURES." 23rd International SAMPE Technical Conference, October 1991.

Learning Characteristics of a Space-Time Neural Network As a Tether "Skiprope Observer"

Robert N. Lea and James A. Villarreal
NASA / Lyndon B. Johnson Space Center
Houston, Texas 77058

Yashvant Jani
Togai InfraLogic Inc.
Houston, Texas 77058

Charles Copeland
Loral Space Information Systems
Houston, Texas 77058

Abstract :

The Software Technology Laboratory at the Johnson Space Center is testing a Space Time Neural Network (STNN) for observing tether oscillations present during retrieval of a tethered satellite. Proper identification of tether oscillations, known as "skiprope" motion, is vital to safe retrieval of the tethered satellite. Our studies indicate that STNN has certain learning characteristics that must be understood properly to utilize this type of neural network for the tethered satellite problem. We present our findings on the learning characteristics including a learning rate versus momentum performance table.

1.0 Introduction

NASA and the Italian Space Agency plan to fly the Tethered Satellite System (TSS) aboard the Space Shuttle in July, 1992. The mission, lasting approximately 40 hours, will deploy a 500 kg satellite upward (away from the earth) [1, 2] to a length of 20 km, perform scientific experiments while on-station, and retrieve the satellite safely. Throughout the deployment, experimentation, and retrieval, the satellite will remain attached to the Orbiter by a thin tether through which current passes, providing power to experiments on-board the satellite. In addition to the scientific experiments on-board the satellite, the dynamics of the TSS itself will be studied. The TSS dynamics are complex and non-linear due to the mass as well as spring-like characteristics of the tether. When the tether is modeled as a massless spring, it typically exhibits longitudinal and librational oscillations [2]. However, when the tether is modeled as beads connected via springs as shown in fig. 1, the dynamics of TSS includes longitudinal, librational and transverse circular oscillations or so-called "skiprope" phenomenon. These circular oscillations are generally induced when current pulsing through the tether interacts with the Earth's magnetic field [3, 4]. The center bead typically displaces the most from the center line. Thus, the "skiprope" can be viewed (fig. 2) by plotting a trajectory of the mid-point of the tether as it is retrieved slowly from the Onstation-2 phase in a high fidelity simulation test case. Detection and control of the various tether modes, including the 'skiprope' effect, is essential for a successful mission. Since there are no sensors that can directly provide a measure of skiprope oscillations, indirect methods like Time Domain Skiprope Observer [4] and Frequency Domain Skiprope Observer [3] are being developed for the mission. We are investigating a Space Time Neural Network (STNN) based skiprope observer.

The STNN is basically an extension to a standard backpropagation network [5,6,7] in which the single interconnection weight between two processing elements is replaced with a number of Finite Impulse Response (FIR) filters [8]. The use of adaptable, adjustable filters as interconnection weights provides a distributed temporal memory that facilitates the recognition of temporal sequences inherent in a complex dynamic system such as the TSS. We have performed experiments in detecting various parameters of skiprope motion using an STNN.

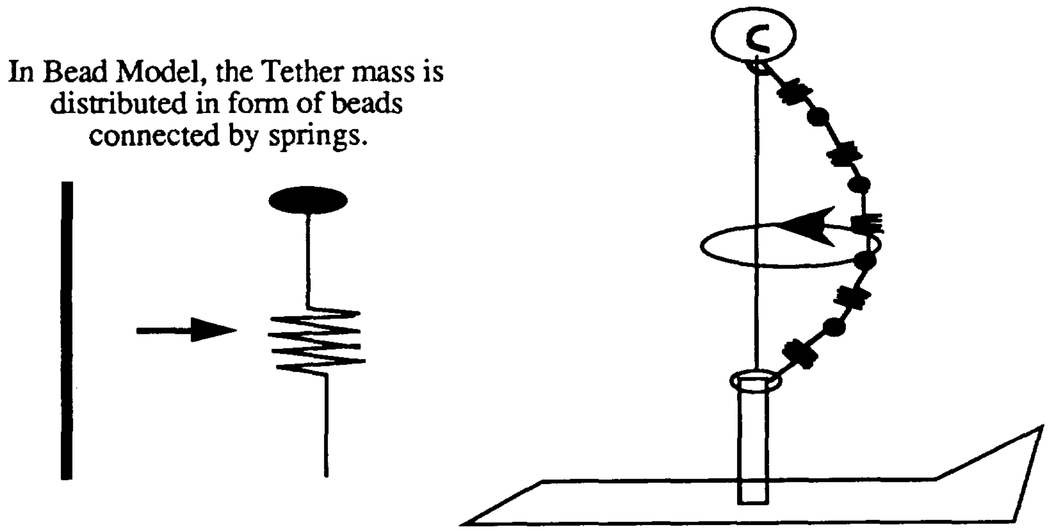

In Bead Model, the Tether mass is distributed in form of beads connected by springs.

Fig. 1 When tether is modelled as beads, the transverse circular oscillations known as "skiprope" are induced during retrieval.

Extensive studies using high fidelity simulations have shown that the tethered satellite exhibits characteristic rate oscillations in the presence of skiprope motion as shown in figure 3. Since these rate oscillations are measured by the satellite's on-board rate gyros, the measured rates can be used as inputs to a skiprope detection system along with other measured parameters such as tension and length [9]. We have trained an STNN using data logged from a high fidelity Orbital Operations Simulator (OOS) [10] which models the behavior of the TSS. The parameters used in network training include satellite roll, pitch, and yaw rates, sensed tension and length of the tether, and the position of the mid-point of the tether during skiprope motion. In this paper, we first describe the STNN architecture in section 2. The STNN configuration used for skiprope observation is described in section 3 along with training and test data generated by the simulation test cases. Learning characteristics are discussed in section 4, and conclusions are summarized in section 5.

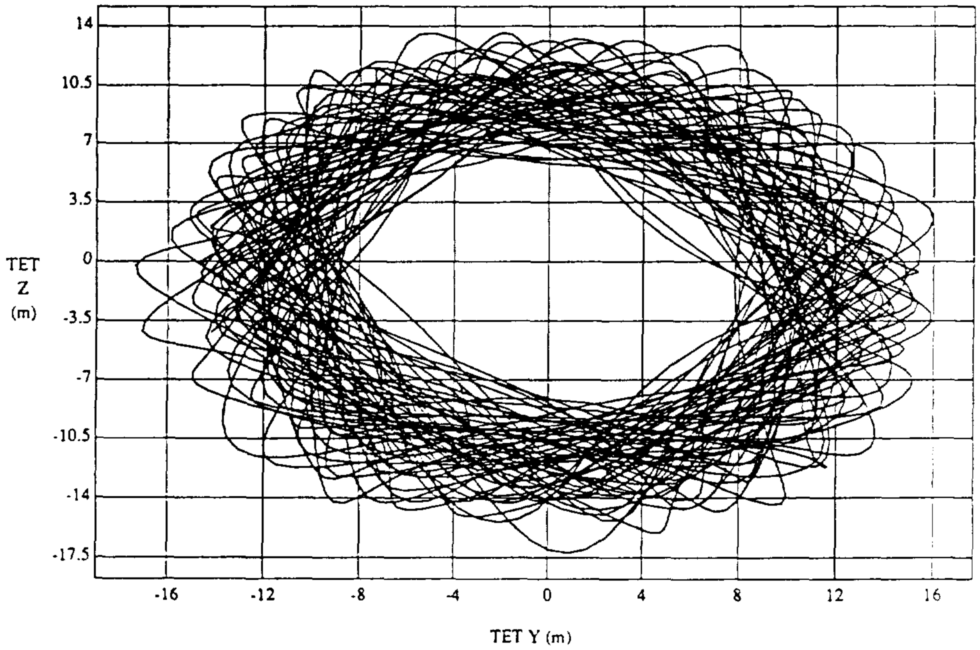


Figure 2 - Trajectory of tether mid-point during "skiprope".

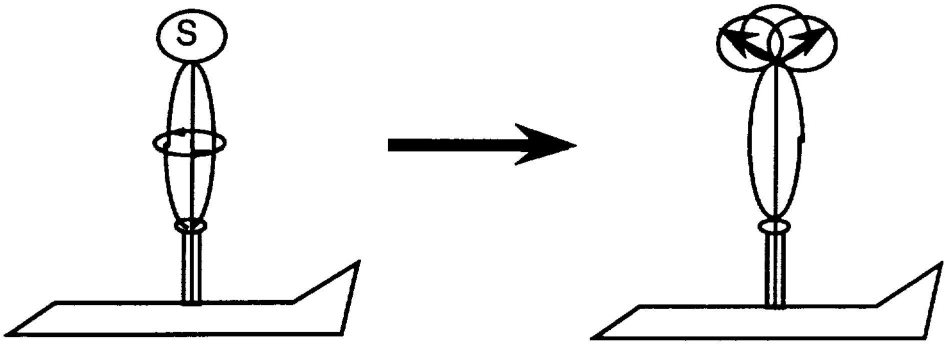


Figure 3 - Tether "skip rope" effect leads to highly characteristic satellite attitude oscillations which can be used to detect the magnitude and phase of the skiprope

2.0 STNN Architecture

The STNN architecture [8] allows the dimension of time to be added to the strong spatial modelling capabilities found in neural networks. The time dimension can be added to the standard processing element used in conventional neural networks by replacing the synaptic weights between two processing elements with an adaptable-adjustable filter as shown in figure 4.

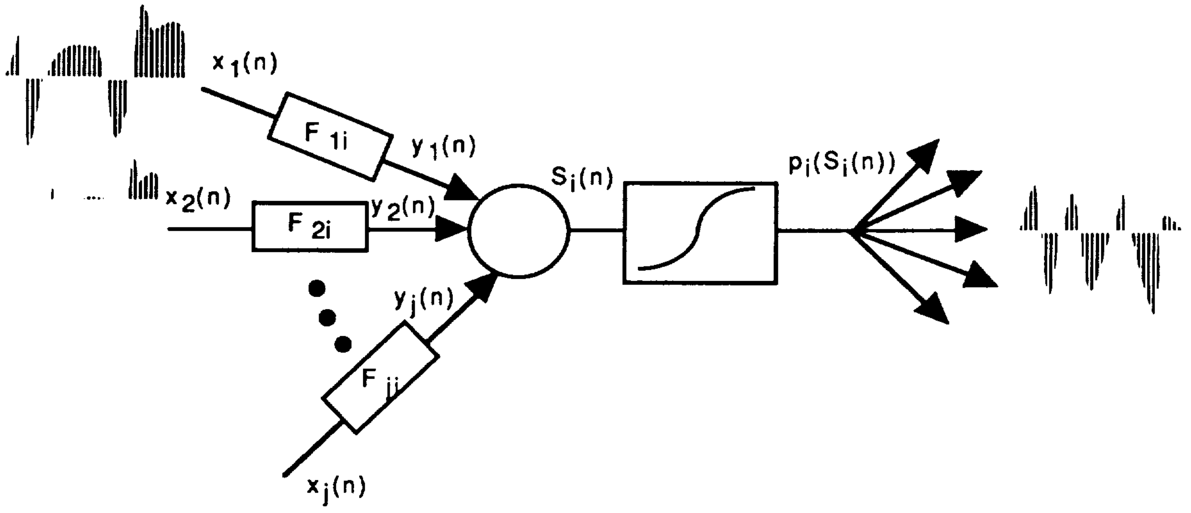


Figure 4 - A pictorial representation of the Space-Time processing element.

Instead of a single synaptic weight with which the standard backpropagation neural network represented the association between two individual processing elements, there are now several weights representing not only spatial association, but also temporal dependencies. In this case, the synaptic weights are the coefficients to the adaptable digital filters:

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

Here the x and y time sampled sequences are the input and output respectively of the filter and a_m's and b_k's are the coefficients of the filter. Thus, if there are j parameters going into a neuron, the y_j

are input into the neuron, where each y_j is a filtered value of the x_j using n time series samples as shown in fig. 4. The x_j's are the real input from an external source. Thus, the STNN is learning a temporal dependency of the input parameters.

A space-time neural network includes at least two layers of filter elements fully interconnected and buffered by sigmoid transfer nodes at the intermediate and output layers as shown in figure 5. A sigmoid transfer function is not used at the input. Forward propagation involves presenting a separate sequence dependent vector to each input, propagating those signals throughout the intermediate layers until the signal reaches the output processing elements. In adjusting the weighting structure to minimize the error for static networks, such as the standard backpropagation, the solution is straightforward. However, adjusting the weighting structure in a space-time network is more complex because not only must present contributions be accounted for but contributions from past history must also be considered. Therefore, the problem is that of specifying the appropriate error signal at each time and thereby the appropriate weight adjustment of each coefficient governing past histories to influence the present set of responses. A detailed discussion of the algorithm can be found in the reference [8].

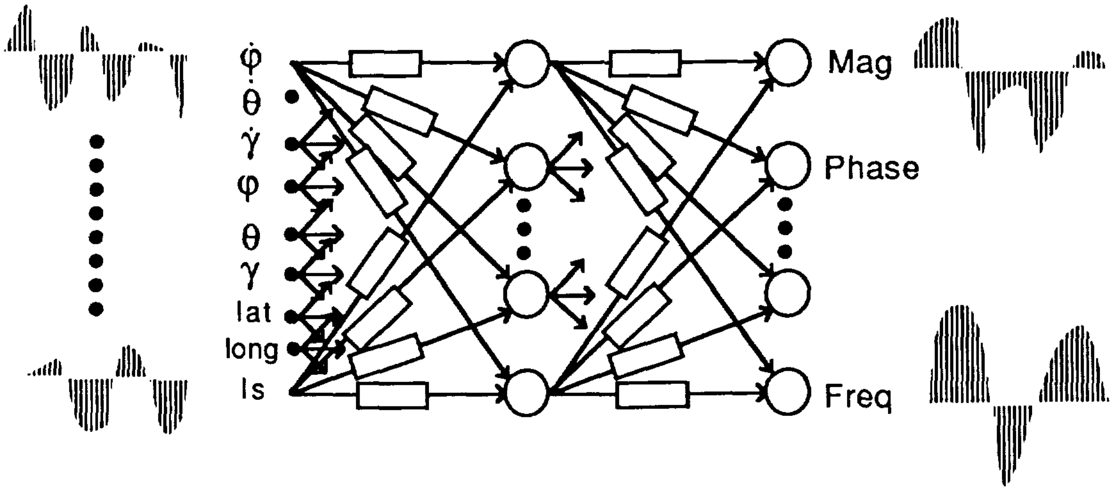


Figure 5 - A depiction of a STNN architecture showing the distribution of complex signals in the input space.

3.0 STNN Configuration and Test/Training Data

Several different simulation runs were used to gather data for STNN training. The simulation runs are consistent with the requirement that the skiprope observer must be capable of performing during various combinations of current flow through the tether and satellite spin. For example, one simulation represents a case in which current flows through the tether continuously, and the satellite is in yaw hold. Another simulation represents the case in which current flows through the tether only during the on-station phase, and the satellite is in yaw hold. A third simulation represents continuous current flow, and satellite spin at 4.2 degrees/second. These three scenarios will form the basis for STNN skiprope observer training and testing, and are consistent with simulations that are used for testing the Time-Domain Skiprope Observer (TDSO) [4] which will be flown on TSS-1.

Ultimately, the network should utilize only roll rate, pitch rate, yaw rate, sensed tension and sensed length since these are the only directly measurable parameters. However, we have conducted experiments using derived parameters such as roll, pitch, and yaw position in addition to rates with no significant improvement. The biggest challenge to network training so far has been to learn the phase mapping. Several different network configurations have yielded good results in predicting skiprope amplitude, but we have not been as lucky with skiprope phase. Since the ultimate goal is to provide the crew with accurate measurements of skiprope amplitude and phase to support the yaw maneuver, the skiprope observer should learn to predict amplitude and phase based on the available inputs. However, decisions concerning the yaw maneuver can be based on the x and y coordinates of the mid-point of the skiprope motion as well. Therefore, the basic network configuration consists of 6 inputs (roll rate, pitch rate, yaw rate, sensed tension, x(t), and y(t)) and 2 outputs (x (t+1) and y(t+1)). Notice that we are training on the current x and y position and predicting x and y position for the next time step. In previous experiments we focussed on finding the optimum network configuration in terms of numbers of hidden units and numbers of zeros from layer to layer. Through experimentation, we settled on 30 hidden units and 30 zeros from the input layer to the hidden layer, and 30 zeros from the hidden layer to the output layer, although slight deviations in these parameters have little or no effect in network performance. In this paper we concentrate primarily on the effects of learning rate and momentum on the overall generalization of the Space-Time Neural Network.

4.0 Learning Characteristics

A well known characteristic of backpropagation networks, or networks derived from backpropagation, is that in order to achieve reasonable generalization, the network must learn the training data. Experiments have indicated that, like standard backpropagation, the learning characteristics of STNN are such that if the training data is not learned, generalization will not occur. These and other learning characteristics dictate that a particular sequence of steps be followed in the training and testing of STNN. The following general steps were used as guidelines throughout the STNN testing.Please note that the use of the word "momentum" in this report refers to a term in the learning algorithm that represents a fraction of the previous weight change rather than any physical properties of the TSS.

1. Train and test - evaluate learnability of training data.
2. Adjust network as necessary (set learning rate and momentum in updating of interconnection weights).
3. If network is unable to obtain sufficient convergence on training data, test individual parameters one at a time. Eliminate un-mappable parameters and start over.
4. If reasonable convergence is realized on training data, divide the data set into a training set and a separate test set.
5. When reasonable performance is achieved on the separate test data, then go for multi-test case generalization.

Step 2 above generally involves trying different combinations of learning rate and momentum in the interconnection weight update formulas. Table 1 illustrates the test case matrix we have identified in order to test the effects of different combinations of learning rate and momentum.

The results that follow are from training and testing using data from the simulation which includes current pulsing and satellite spin, which is considered the most difficult case. Following our general training and testing steps listed above, we verified that the STNN was able to learn the training data using a learning rate of 0.05, and momentum set to 0.9. We trained and tested on all 3500 Input/Output pairs and achieved a MAX error of 0.08 and RMS error of 0.02 at 140 cycles. Since the network will be trained off-line before being placed in the operational environment, we must determine how well the network will perform when presented with data that it has not previously seen. Therefore, to test the generalization ability of STNN, we train on only the first and last 400 input/output pairs from the full 3500, and test separately on the middle 2700 input/output pairs while trying various combinations of learning rate and momentum with the following results. First, with a momentum of 0.9, we tried learning rates of 0.05, 0.2, and 0.7 (test cases #4-6 in Table 1). Test case #4 resulted in MAX error = 0.43, and RMS error = 0.04 at cycle 100. Figure 6 shows the error plot for test case #4 up to 500 cycles. Figures 7a and 7b show a portion of the x and y predictions from test case #4. Test case #5 resulted in MAX error = 0.43 and RMS error = 0.04 at cycle 480. Figure 8 shows that the network prediction of y in test case 5 is similar to that of test case #4. Increasing the learning rate to 0.7 in test case #6 results in the network never reaching errors as low as in the previous two test cases (at least not within 500 cycles) and overall performance is similarly degraded as is seen in figures 9a and 9b. Next we set momentum to 0.2 and try learning rates of 0.05, 0.2, and 0.7 (test cases #1-3 in Table 1). Test case #1 yielded MAX error = 0.44, and RMS error = 0.05 at 100 cycles, as is shown in figure 10a. Figure 10b shows that the network's prediction of x in this test case is not quite as accurate as test cases #4 and #5. As we increase learning rate from 0.05 to 0.2, performance degrades significantly as is shown in figure 11a. The error graph in figure 11b shows that no learning occurred in test case #2, as RMS error never dropped significantly below 0.5, and MAX error remained near 0.8. Similar results occurred in test case #3 as we increased the learning rate from 0.2 to 0.7. The overall test errors are summarized in Table II.

Table 1 - Learning Rate Versus Momentum in STNN Weight Update Formulas

Test Case	Momentum in weight update	Learning Rate
1	0.2	0.05
2	0.2	0.2
3	0.2	0.7
4	0.9	0.05
5	0.9	0.2
6	0.9	0.7
7	0.95	0.05
8	0.98	0.05

Table II - Number of Training Cycles to Reach Lowest Test Errors.

Test Case	Max Error	RMS Error	Number of Cycles
1	0.44	0.05	100
2	0.78	0.49	280
3	0.8	0.5	480
4	0.43	0.04	100
5	0.43	0.04	480
6	0.5	0.09	400
7	0.41	0.04	480
8	0.41	0.04	480

5.0 Conclusions

Through experimentation, we have gained insight into the learning characteristics of STNN in terms of learning rate and momentum parameters. In particular, we find that the skiprope observer problem requires high momentum and very low learning rate. In test case 4 we have seen that the RMS error drops to 4 % within only 100 cycles of learning. We further verified this by performing two test cases (#7 and #8) with high momentum and low learning rate. It should be noted that the max error is reduced in both cases.

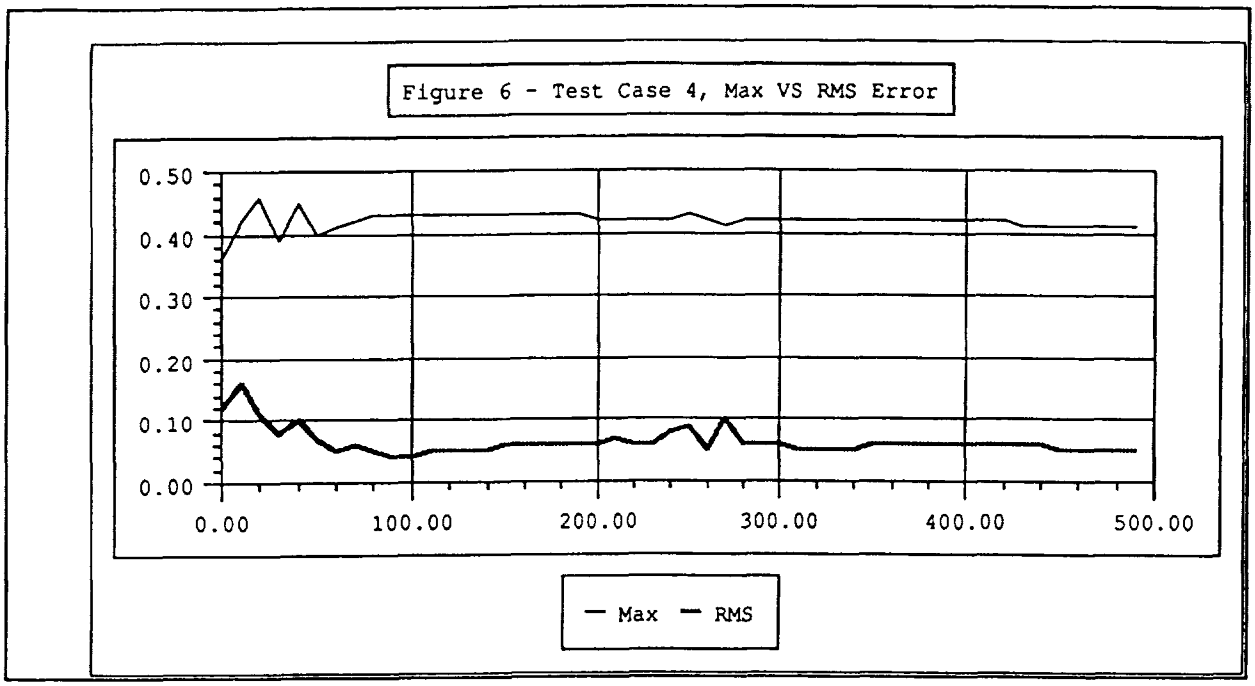


Figure 6 - Test Case 4, Max VS RMS Error

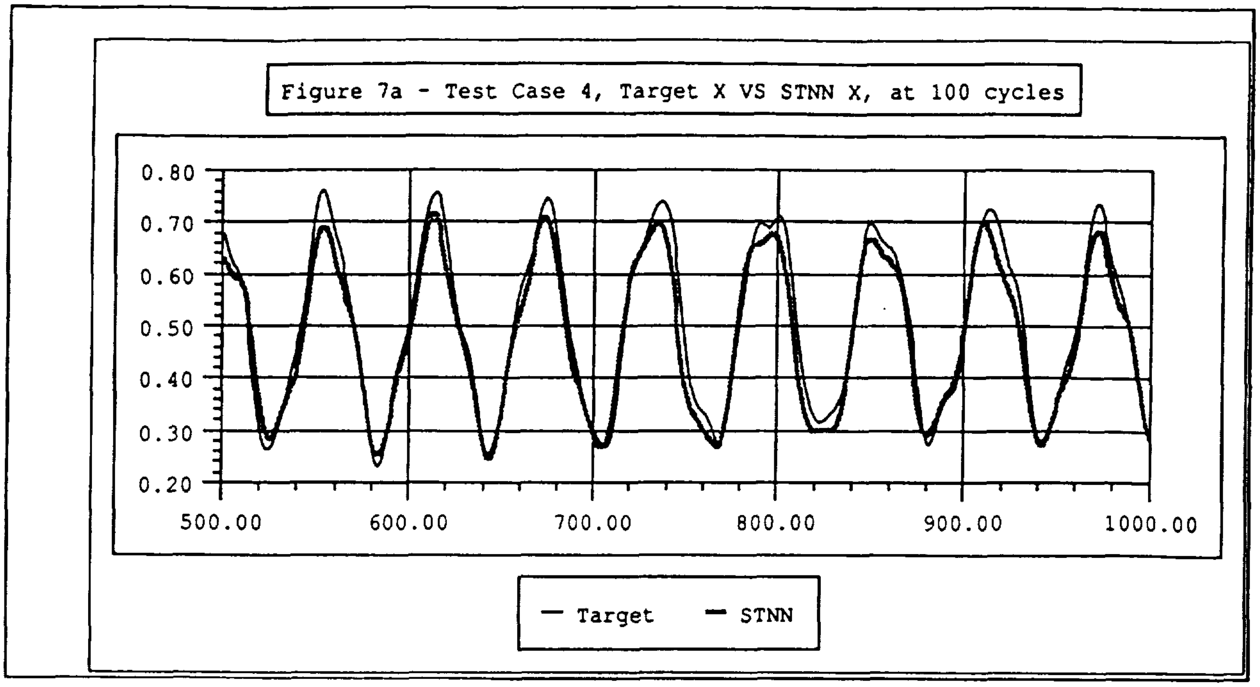


Figure 7a - Test Case 4, Target X VS STNN X, at 100 cycles

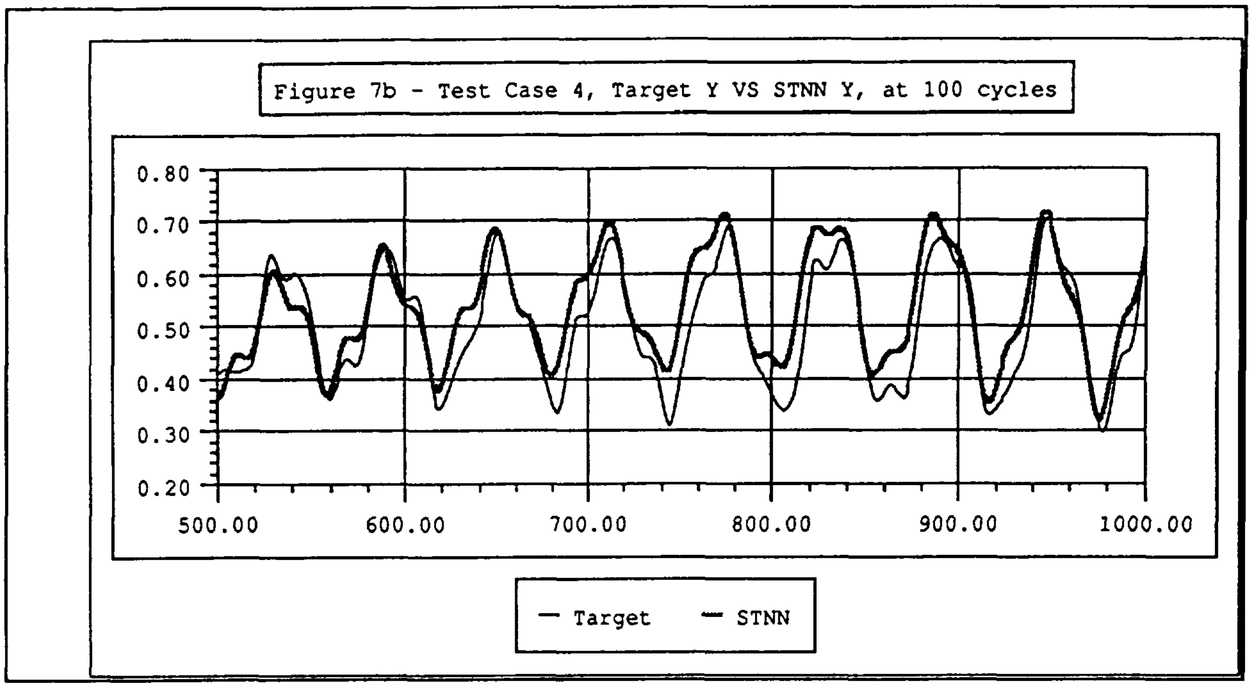


Figure 7b - Test Case 4, Target Y VS STNN Y, at 100 cycles

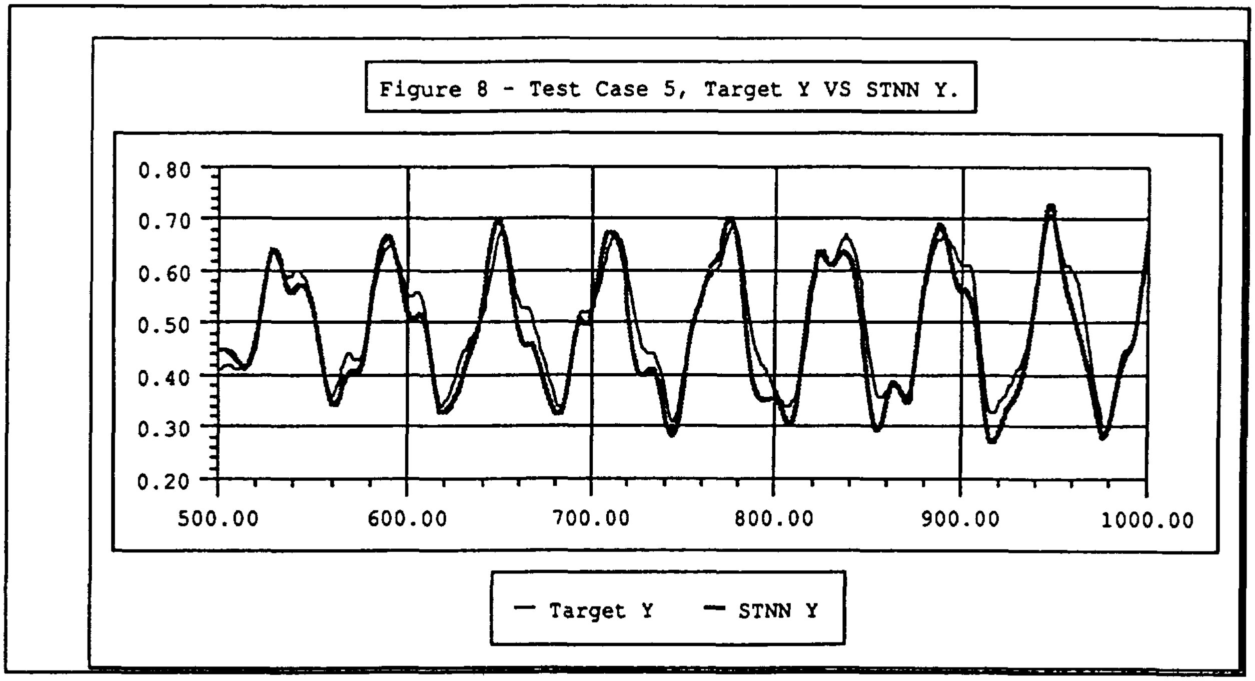


Figure 8 - Test Case 5, Target Y VS STNN Y.

Figure 9a - Test Case 6, Max VS RMS Error

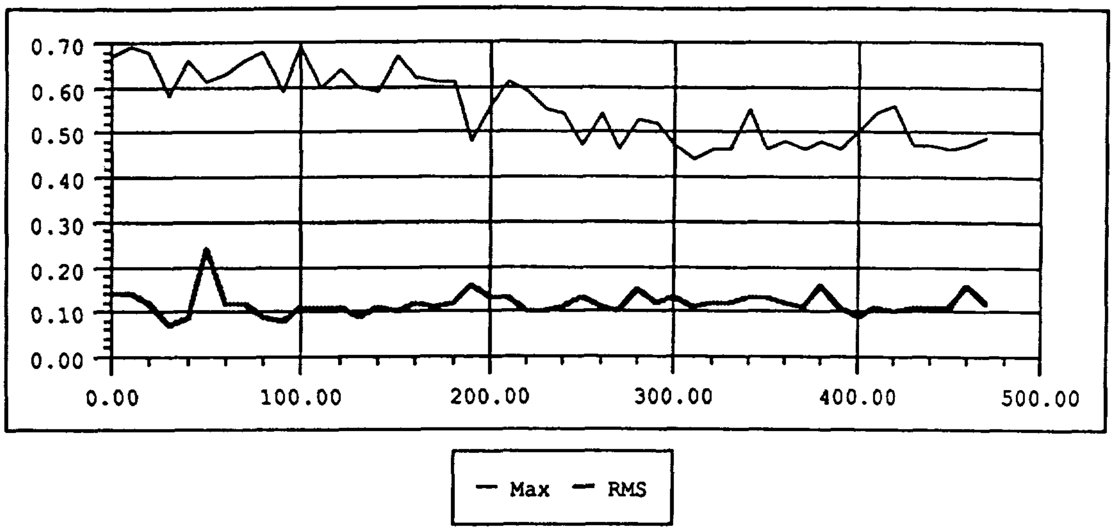
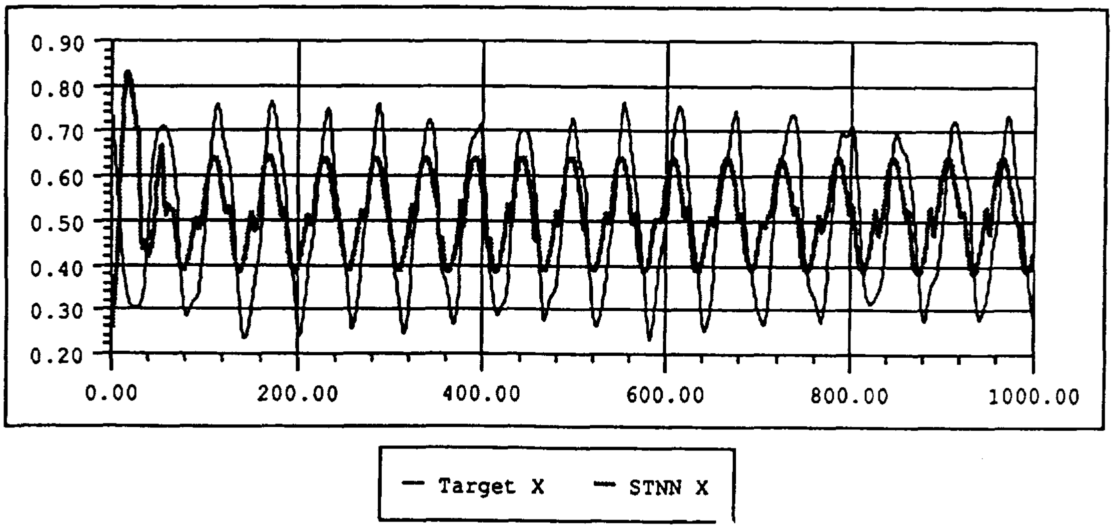


Figure 9b - Test Case 6, Target X VS STNN X, 400 cycles



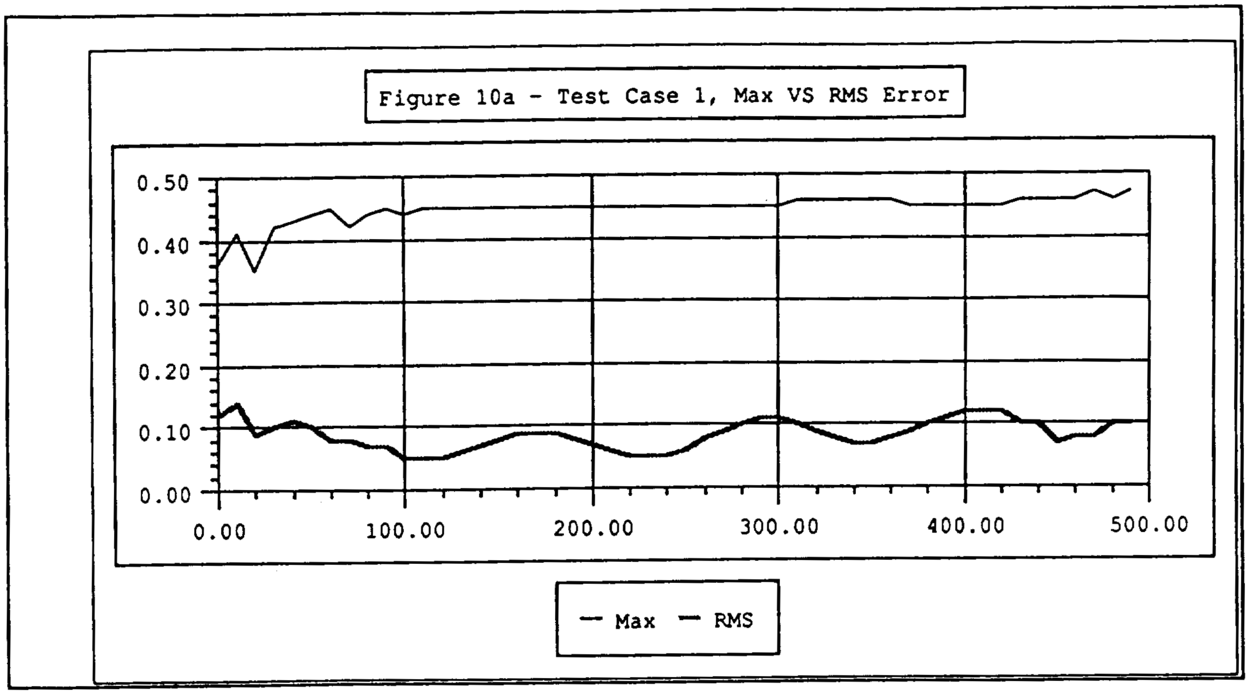


Figure 10a - Test Case 1, Max VS RMS Error

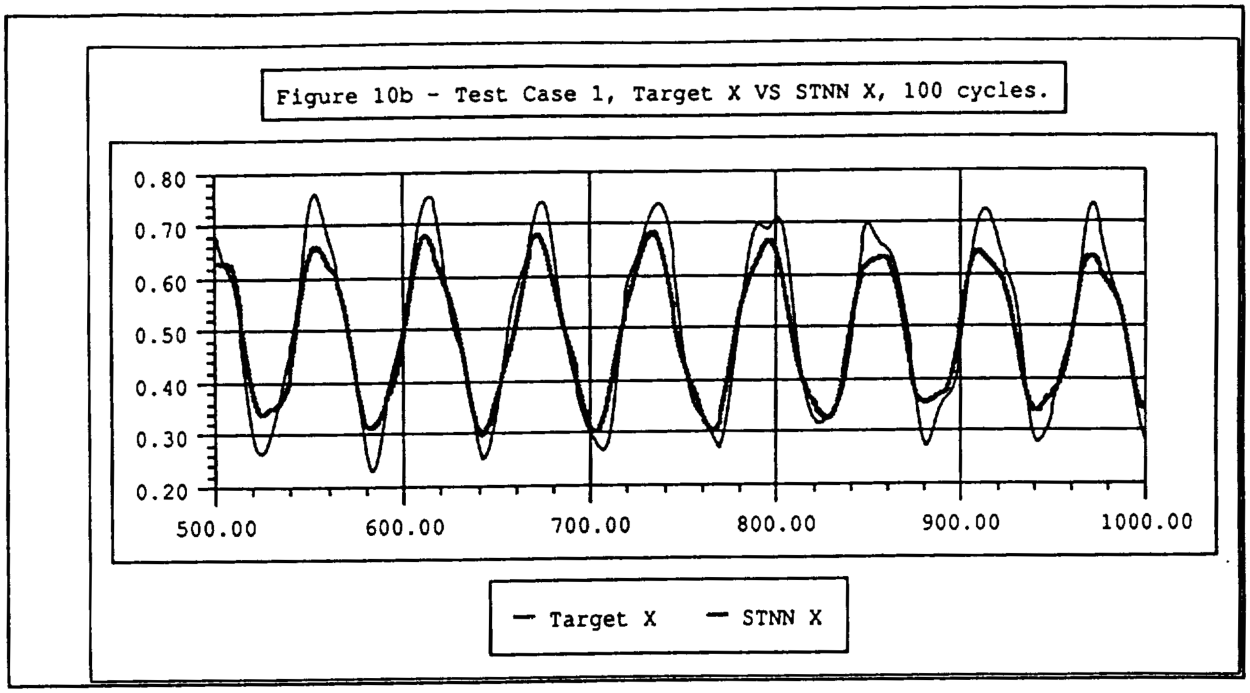


Figure 10b - Test Case 1, Target X VS STNN X, 100 cycles.

Figure 11a - Test Case 2, Target Y VS STNN Y, 280 cycles.

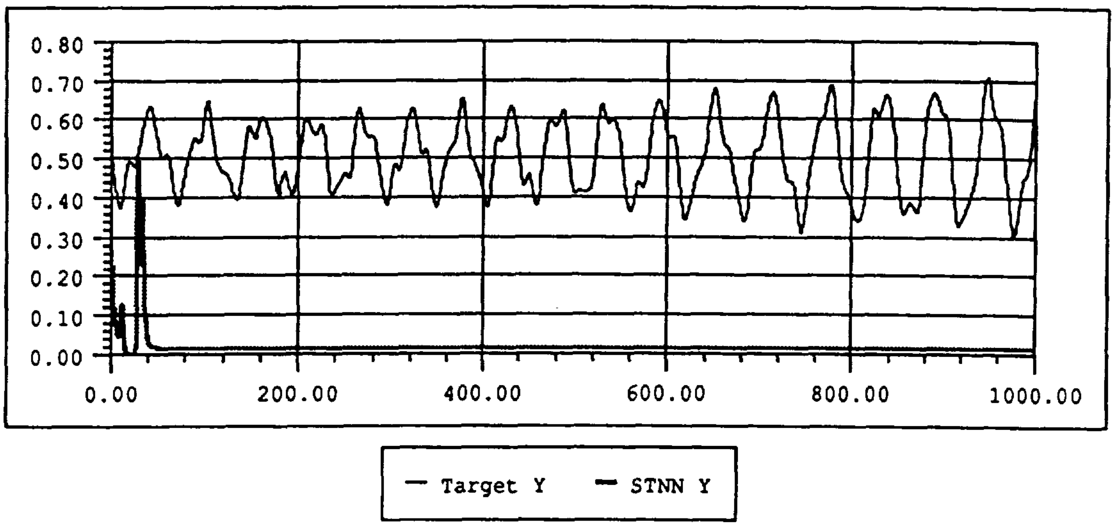
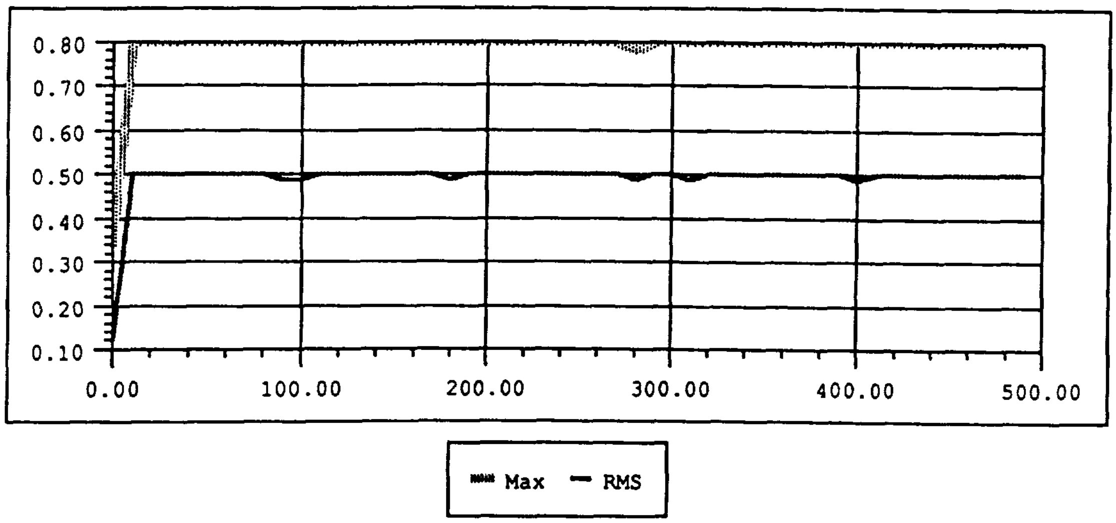


Figure 11b - Test Case 2, Max VS RMS Error



Based on our earlier results, we conclude that the STNN is slow in learning sharp discontinuities like those encountered in phase behavior. The value of the phase goes from 180 to -180 abruptly when the circle is complete. When we changed to the x- and y- component form (rather than amplitude and phase), the STNN based skiprope observer performed much better in predicting x and y coordinates of the mid-point of the tether.

We will have an opportunity to perform a side-by-side comparison of the STNN based skiprope observer and the TDSO using simulation data. Next, we will test the STNN based skiprope observer with the post mission data after the TSS-1 flight.

References :

1. Coledan, S. : "Tether Satellite Advances", Space News, vol. 2, no. 15, p. 8, 1991.
2. Powers, C.B., Shea, C., and McMahan, T. : "The First Mission of the Tethered Satellite System", A special brochure developed by the Tethered Satellite System Project Office, NASA/Marshall Space Flight Center, Huntsville, Alabama, U.S. GPO 1992-324-999, 1992.
3. Ioup, G.E., Ioup, J.W., Rodrigue, S.M., Amini, A.M., Rayborn, G.H., and Carroll, S. : "Frequency Domain Skiprope Observer", Skiprope Containment Status Meeting held at Denver, Sep. 10-11, 1991. (Research supported by NASA Contract NA8-38841)
4. Glowczwski, R. : " Time Domain Skiprope Observer Overview", Skiprope Containment Status Meeting held at Martin Marietta, Denver, Sept. 10-11, 1991.
5. Kosko, B. : "Neural Networks and Fuzzy Systems", Prentice-Hall, New Jersey, 1992.
6. J.A.Freeman and D.M.Skapura : "Neural Networks Algorithms, Applications and Programming Techniques", Addison-Wesley Publishing Company, Reading, MA. 1991.
7. Wasserman, P.D. : "Neural Computing Theory and Practice" Van Nostrand Reinhold, New York, 1989.
8. Villarreal, J.A., and Shelton, R.O. : "A Space-Time Neural Network", International Journal of Approximate Reasoning , 6(2), 133-149, 1992.
9. Lea, R.N., Villarreal, J.A., Jani, Y., and Copeland, C. : "Application of Space Time Neural Networks to detect Tether Skiprope Phenomenon in Space Operations", Proceedings of the AIAA GN&C Conference held at Hilton Head, South Carolina, August 10-12, 1992.
10. Edwards, H. C., and Bailey, R. : "The Orbital Operations Simulator User's Guide", LinCom corporation, ref. LM85-1001-01, June 87.

omit

MICROELECTRONICS/OPTOELECTRONICS
PART 2

WIRELESS INFRARED COMMUNICATIONS FOR SPACE AND TERRESTRIAL APPLICATIONS

James W. Crimmins
Wilton Division
K&M Electronics, Inc.
Ridgefield, CT 06877

ABSTRACT

Voice and data communications via wireless (and fiberless) optical means has been commonplace for many years. However, continuous advances in optoelectronics and microelectronics have resulted in significant advances in wireless optical communications over the last decade. Wilton has specialized in diffuse infrared voice and data communications since 1979.

In 1986, NASA Johnson Space Center invited Wilton to apply its wireless telecommunications and factory floor technology to astronaut voice communications aboard the Shuttle. In September, 1988 a special infrared voice communications system flew aboard a "Discovery" Shuttle mission as a flight experiment. Since then the technology has been further developed, resulting in a general purpose 2 Mbs wireless voice/data LAN which has being tested for a variety of applications including use aboard Spacelab.

Funds for Wilton's wireless IR development were provided in part by NASA's Technology Utilization Office and by the NASA Small Business Innovative Research Program. As a consequence, Wilton's commercial product capability has been significantly enhanced to include diffuse infrared wireless LAN's as well as wireless infrared telecommunication systems for voice and data.

The technology and resulting commercial products are reviewed.

INTRODUCTION

Diffuse infrared communications is a powerful, well established wireless technology which has greatly benefited by recent advances. Since 1979, Wilton's sole specialty has been diffuse infrared communications. Former limitations of infrared communication such as the need for a direct line-of-sight have been eliminated by new techniques. Currently, increasing numbers of system designers wishing to eliminate cables are discovering this unfamiliar technology.

Several wireless communication systems have been developed in a series of projects completed by Wilton for NASA Johnson Space Center. Included are infrared wireless systems that provide multi-channel, multi-user communications for both voice and data. The high speed, digital design of these systems provides the flexibility to configure wireless networks to handle voice, data or a combination of both at the same time.

Wilton's diffuse infrared technology has been significantly enhanced through its NASA related activity and as a result, new wireless tools are being made available to industry.

This paper describes the nature of diffuse infrared communications, the developments related to the Wilton/NASA activity and the resulting commercial products and technology that Wilton is offering to industry.

DIFFUSE INFRARED COMMUNICATIONS DEVELOPMENT

Early Development Work

For many years, wireless communication using infrared light has been implemented using modulated infrared light beams which linked the transmitter to the receiver. If the beam was broken by an obstruction, communication was interrupted. With the commercial availability of infrared light emitting diodes and low noise photo diodes, it became feasible to construct very sensitive IR receivers which were sensitive enough to detect indirect infrared light - light which after many reflections from surfaces within a room, reaches the receiver greatly attenuated. IR communication over indirect paths is termed "diffuse" IR communications.

This development greatly enhanced the utility of infrared as a communication medium. IR transceivers were no longer required to be stationary and pointing of the optics was no longer necessary. Communication systems could be built which were "RF like" but did not suffer from FCC regulation, electrical interference, neighboring system interference, compromised security and RF health hazards. Applications such as infrared cordless telephones and wireless handheld computers became feasible.

New problems however came with the highly sensitive IR receivers:

Σ Sources of randomly modulated infrared light (such as fluorescent lights) that once may have been considered too low in level to be of concern, were now formidable sources of IR interference. Means were needed to reject this interference.

Σ For some two way systems in which a transmitter and receiver are co-located, the local transmitter may interfere with the highly sensitive receiver. Accordingly, means were needed to reject the transmitter's interference.

Σ Portable diffuse IR transmitters required excessive battery power to reach even the most sensitive receiver. An approach was required to cover large areas or multiple rooms using low powered diffuse IR transmitters.

Wilton's work in the early eighties generated solutions to problems such as these and provided a foundation for the system development that followed.

Diffuse IR Systems

By the mid 80's, Diffuse IR systems were constructed at Wilton which permitted high quality full duplex voice or asynchronous data communications between portable battery operated transceivers.

Some of these early systems were used by Intel's Systems Group for their voice recognition quality audit systems. Many such systems were installed at Ford Motor Company by Intel. Paint inspectors wearing IR voice transceivers verbally enter inspection data directly into Ford's database. The inspectors walk freely within the 40 by 160 foot inspection area. Full duplex, noise free voice communication is provided in spite of banks of high intensity fluorescent lights that are needed for visual inspection. The inspector's belt mounted IR transceivers also provide a channel for a handheld bar code scanner. Up to four such systems could be used simultaneously in the same area.

In 1987, Engineers at NASA Johnson Space Center concluded that Wilton's technology had promise for applications aboard spacecraft given additional development. An enhanced version of Wilton's voice system was developed and flown on Discovery, September 1988. NASA. At that point it was decided to move the technology to the next level.

NASA JSC/Wilton Activity

The Discovery experiment had proved the utility of Infrared communications in spacecraft. Wilton proposed a new system which would provide a means to wirelessly interconnect people and devices by a general purpose, multi-user, multi-channel communication system. Wilton won Phase I and Phase II contracts with NASA JSC through the Small Business Innovative Research Program, which resulted in the delivery of equipment June 30, 1992.

The new multi-user wireless interconnect system was based on a diffuse infrared wireless local area network (IRplex 7000), which can accommodate up to 64 wireless nodes. The network data rate is 2 Mbs with a throughput of 1.8 Mbs. Because of its deterministic design, the IRplex 7000 wireless LAN is able to handle digitized voice as well as other digitized data. As a result, special wireless nodes have been configured for voice, asynchronous data, and digitized physiological data at 160 Kbs. Interfaces to standard LANs such as Ethernet are being explored.

Wilton's previous technology for wireless asynchronous data was enhanced and prepared for commercial packaging with the help of funding from the Technology Utilization Office at NASA JSC. Exhibits show advance product sheets for the IRplex 1000 and IRplex 3100 product lines which describe the results of this effort.

Also aided by this program was the conversion of IRplex 7000 technology into a general purpose computer LAN (Wilton's IRplex 6000). This diffuse, wireless LAN operates at 2.5 Mbs and can readily be employed by any computer having a standard ARCNET adapter installed. This activity has resulted in IRplex 7000 technology being converted for use in the commercial market place.

Wilton's Commercial Activity

Wilton is moving forward in several commercial areas related to wireless infrared technology:

WILTON'S INFRARED CORDLESS TELEPHONE

Σ SINGLE ROOM OR MULTIPLE ROOM COVERAGE

Σ HIGH QUALITY VOICE TRANSMISSION DUE TO GENEROUS BANDWIDTH.

Σ INTERFERENCE FREE.

Σ SECURE.

Σ NO DANGER OF MUTUAL INTERFERENCE IN HIGH DENSITY MULTI-OFFICE APPLICATIONS.

Σ SUITABLE FOR BUSINESS USE

FACTORY FLOOR WIRELESS VOICE COMMUNICATION SYSTEM

Σ INTEL COMBINED WILTON'S FULL DUPLEX IR VOICE SYSTEM WITH INTEL'S VOICE RECOGNITION SYSTEM.

Σ PAINT INSPECTORS AT FORD ENTER DATA VERBALLY INTO FORD'S DATABASE.

Σ IR COVERS A 40 BY 160 FOOT INSPECTION AREA. FOUR CHANNELS AVAILABLE.

Σ ROBUST OPERATION IN PRESENCE OF INTENSE FLORESCENT LIGHTING.

Σ BAR CODE SCANNER CHANNEL

WILTON'S PRESENT COMMERCIAL ACTIVITY

Σ IRplex 6000 WIRELESS ARCNET (PRE-INTRODUCTION).

Σ IRplex 1000, IRplex 3100 ASYNCHRONOUS WIRELESS PORTS WITH MULTI-ROOM COVERAGE (PRE-INTRODUCTION).

Σ IRplex 2500 CENTRAL OFFICE TALK/TEST SYSTEM .

Σ INFRARED BADGES FOR PERSONNEL LOCATION.

Σ IR WATTHOUR METER/LOGGER FOR REMOTE READING (BEING SUPPLIED TO A POWER UTILITY)

Conclusion

Wireless communication is proving to be the technology of choice for wireless application ranging from space to commercial telecommunications. With the current trend toward miniature cordless telephone and computers, the need for wireless communications has exceeded the available radio spectrum.

Wireless infrared communications, by merit of its cellular nature, can offer gigabauds of simultaneous wireless communications within a single building. The NASA/Wilton activity has made available attractive wireless solutions for both space and terrestrial use and for both government and commercial markets.

It is likely that by 2002, standard IR ports will be commonplace on most telephone and computer related devices.

FLEXIBLE HIGH SPEED CODEC*

JAMES V WERNLUND
HARRIS GCSD
PO BOX 91000
PALM BAY, FLA. 32901

ABSTRACT

HARRIS, under contract with NASA Lewis, has developed a hard decision BCH (Bose-Chaudhuri-Hocquenghem (ref. 1)) triple error correcting block CODEC ASIC, that can be used in either a bursted or continuous mode. The ASIC contains both encoder and decoder functions, programmable lock thresholds, and PSK related functions. The CODEC provides up to 4 dB of coding gain for data rates up to 300 Mbps. The overhead is selectable from 7/8 to 15/16 resulting in minimal band spreading, for a given BER. Many of the internal calculations are brought out enabling the CODEC to be incorporated in more complex designs. The ASIC has been tested in BPSK, QPSK and 16-ary PSK link simulators and found to perform to with in 0.1 dB of theory. for BER's of 10^{-2} to 10^{-9}. The ASIC itself, being a hard decision CODEC, is not limited to PSK modulation formats. Unlike most hard decision CODEC's, the HARRIS CODEC doesn't degrade BER performance significantly at high BER's but rather becomes transparent.

INTRODUCTION

This paper details the development of the BCH ASIC and its features. Control of the ASIC through a single control line (Block Mark) is discussed. Operation in both bursted and continuous modes of synchronization are detailed. Many of the special features, enabling use of the ASIC in more complex coding schemes, is discussed. In particular an architecture enabling use of the CODEC as a soft decision CODEC is detailed and performance is evaluated. Theoretical performance is predicted, through simulation, and compared to the performance data taken using a digital noise test set and a commercially available bit error rate test set. Details of the test set are presented and programming is discussed. The test set developed under this program is interesting in that any modulation format between 2-ary and 16-ary can be evaluated.

The NASA contract that funded the development of the ASIC also funded the development of the soft decision architecture, discussed in this paper, and the digital noise test set. Ten of the CODEC's were manufactured and supplied to NASA. A preliminary data sheet has been generated for the CODEC and is available.

The ASIC

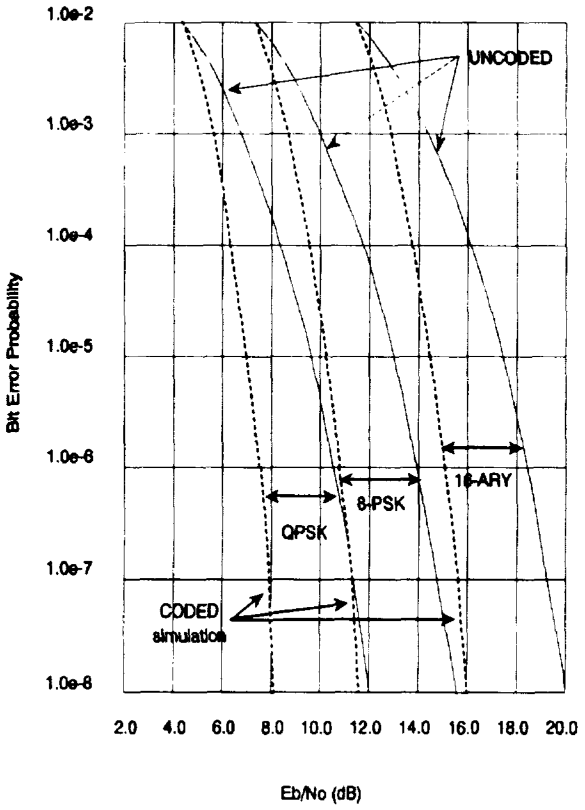
The ASIC is a high speed low power CMOS design. It was developed using VLSI design tools and was fully simulated. It is packaged in a 132 pin PGA and consumes 1.5 Watts of power when clocked at 45 Mhz. The ASIC contains both the encoder and decoder functions, input/output formatting functions, block mark generation circuits, lock detection circuits and PSK carrier phase ambiguity circuits. As a hard decision CODEC operating in the continuous mode the ASIC can provide all the functions necessary for stand alone operation.

The CODEC can provide up to 4 dB of hard decision coding gain at 10^{-8} BER for bit rates to 300 Mbps (see Fig 1.). The ASIC supports interface widths of 1, 2, 4 or 8 bits. This interface will operate up to a rate of 43 Mhz providing a 38 Mbps, 75 Mbps, 150 Mbps or 300 Mbps data rate for the specified interface width. The data format can be either continuous or bursted.

*This work is funded by NASA Lewis Research Center under Contract #NAS3-25087;
Contract Manager: Robert Jones

In the continuous mode the ASIC generates all dynamic control signals internally. Static control signals, such as interface width and lock threshold, must still be supplied. In this mode the ASIC generates a gated clock for clocking data to and from the user. Coded data out of the encoder and into the decoder is continuous. Code words are formed by appending 32 parity bits to every 256 bits of data, resulting in a code rate of (256)/(32+256) or 7/8. The resulting gated clock, supplied to the user, is therefore on for 224 bit times and off for 32. The decoder is self synchronizing in this mode. Circuits within the decoder search for the code word boundaries and a lock detect signal indicates when the decoder has locked. One of the special features of this ASIC is it can resolve carrier phase ambiguities for BPSK, QPSK and 16-ary PSK modulation formats. Static control signals are used to set the modulation mode when this feature is enabled. Note: the acquisition time increases when this mode is enabled.

Hard Decision Performance of a (256, 224) Code



Hard Decision Coding Gain (in dB) (256, 224) Code				Hard Decision Coding Gain (in dB) (512, 480) Code			
BER	QPSK	8-PSK	16-PSK	BER	QPSK	8-PSK	16-PSK
10^{-4}	2.0	2.2	2.4	10^{-4}	1.8	2.0	2.2
10^{-6}	3.0	3.2	3.4	10^{-6}	2.8	3.0	3.2
10^{-8}	3.8	3.9	4.0	10^{-8}	3.6	3.7	3.8

Fig. 1 Hard Decision Coding Gains

The parity bits appended to the data are made up of two parts. The first 28 bits are true code word parity. The last 4 bits are not important to the decoder operation and are therefore supplied to the user (i.e. USER BITS). It is important to note the bits are not protected by the coding process. The decoder removes these bits and supplies them to the user on the receive side. These bits can be used to support network protocol or order wire functions.

Operation of the CODEC in the bursted mode is very similar to the continuous mode. The exception is the user supplies the code word boundaries to both the encoder and the decoder. These boundaries are determined by the control signal Block Mark. This control signal is a TTL level signal which is high for 224 - 480 bit times and low for 32 bit times. The one constraint is once a burst begins transitions in Block Mark must occur on 16 bit boundaries. Bursts longer than 480 bits are formed by concatenating blocks (see Fig 2). This allows very long bursts to be formed. If a the Block Mark signal is left low for more than 32 bit times then the encoder and decoder assume a new burst and a reset is initiated, for the next burst. Note the code rate in the bursted mode is determined by the code word boundaries supplied by the USER and can be as high as (480/(32+480)) 15/16. The coding gain for this over head is only 0.2 dB less than that of the shortest code word. As in the continuous mode the four USER BITS at the end of every code word are made available to the user.

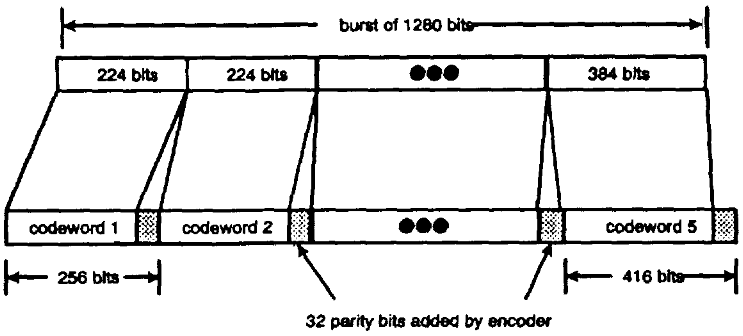


Fig. 2 Concatenated blocks

In addition to providing the decoded data to the user, the decoder can tell the user which bits within a code word were changed by the decoder. This feature was incorporated to enable the ASIC to take advantage of soft decisions when they are available. The application, considered in the NASA funded program, incorporates a Chase Algorithm. (ref. 2) to increase coding gain.This approach is referred to as the FHSC CODEC.

FHSC CODEC

In the FHSC CODEC the decoding function incorporates fourBCH Decoder ASIC's to perform the decoding (fig. 3). The soft decisions are used to generate four code words, one to each of the four decoders. A likelihood term for each of the four code words is also generated. The four code words generated differ in only three bit locations. The bit locations changed are determined by the soft decisions and are the bits with the poorest statistics. Strictly speaking the approach would require 8 decoders to consider all possible combinations of the three bit locations. But, by taking into account the code word parity and the fact that the decoder can only decode code words with 3 errors or less the number of decoders can be reduced to four. The four decoders perform theirdecoding and output the decoded data. The Chase Post Processor circuits read the changed bit locations from each of the 4 decoders. The soft decisions for each of the changed bit locations and the pre-likelihood's calculated previously are then used to calculate a final likelihood for each of the four decoders. The most likely decoder is finally selected and it's decoded output is chosen as the decoded data for that received code word.

Simulations of the FHSC CODEC indicate this technique can provide as much as 1.5 dB additional coding gain (fig 4). Paper designs were generated that preserved all of the features of the ASIC. These designs were ECL circuits which handled interface widths of 3 or 4 bits. With these widths the FHSC can handle 2-ary, 4-ary, 8-ary and 16-ary modulation formats. The design was a four card design using 300K ECL and FCT logic.

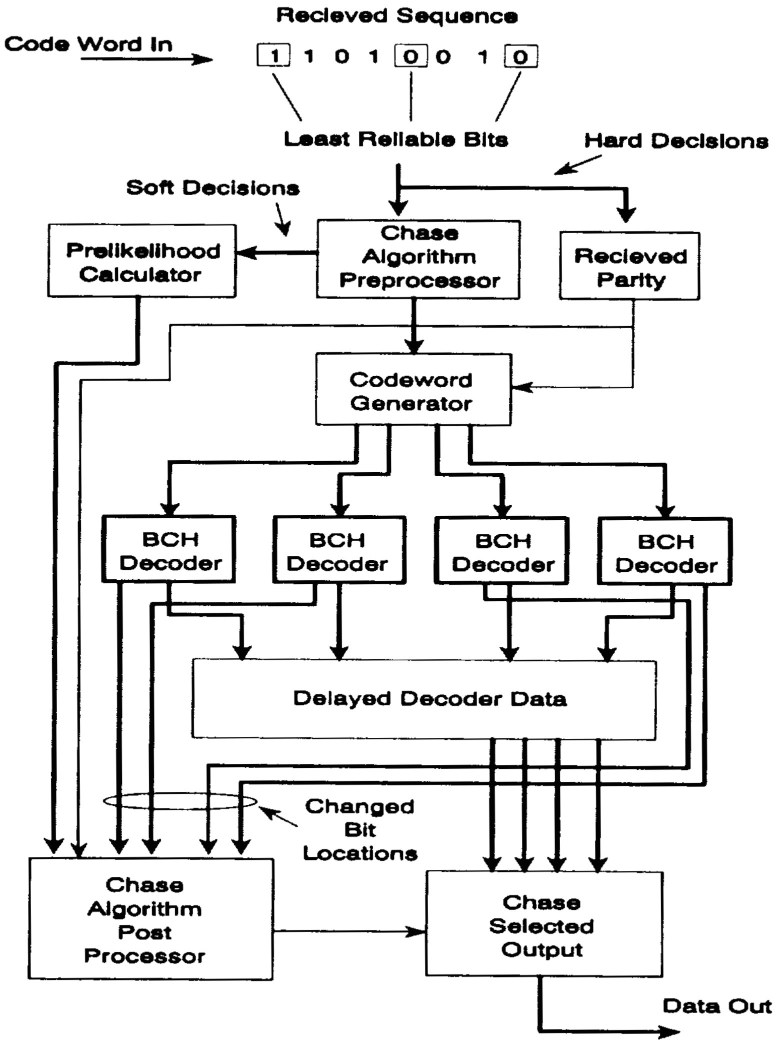
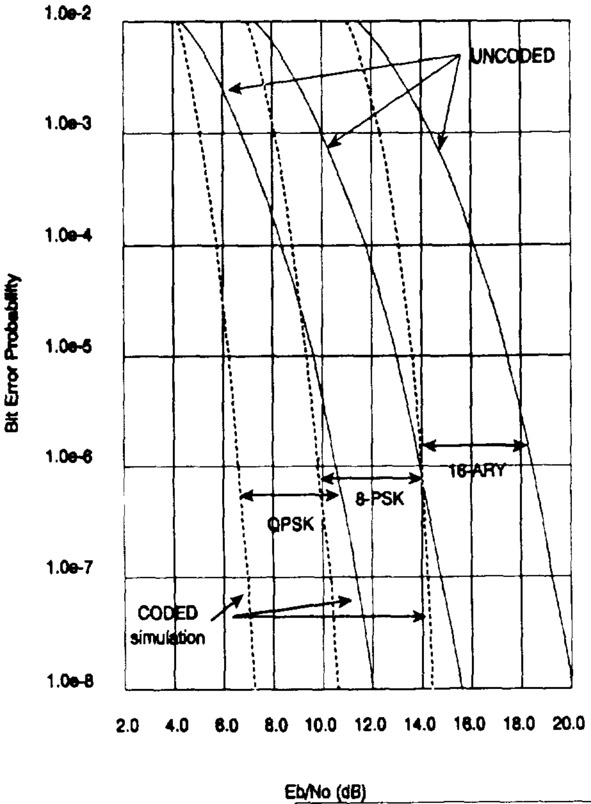


Fig. 3 The Chase Appliqué

Soft Decision Performance of a (256, 224) Code



Soft Decision Coding Gain (in dB) (256, 244) Code			
BER	QPSK	8-PSK	16-PSK
10 $^{-4}$	2.7	2.9	3.1
10 $^{-6}$	4.0	4.3	4.5
10 $^{-8}$	4.9	5.2	5.4

Soft Decision Coding Gain (in dB) (512, 480) Code			
BER	QPSK	8-PSK	16-PSK
10 $^{-4}$	2.5	2.7	2.9
10 $^{-6}$	3.8	4.1	4.3
10 $^{-8}$	4.7	5.0	5.2

Fig. 4 Soft Decision Coding Gains

ASIC testing was done using test and development equipment (TDE) developed under the NASA contract and a commercially available BERT. The TDE equipment generated all of the control signals needed for control of the ASIC in both the bursted and continuous modes and provides a digital link simulator for BER testing. Testing was performed for BPSK, QPSK and 16-ary PSK signals, at Eb/No's ranging from 0 dB to + 20 dB. In all cases the BCH CODEC performed to within 0.1 dB of theory. Calibration of the TDE equipment was accomplished by turning the CODEC off and measuring the resulting BER. The coding gain was then measured by turning the CODEC back on and measuring the BER. After compensating for the appropriate band spreading the gain was then determined.

Although the link simulator supports data rates up to 250 Mbps, signals within the TDE equipment indicate the ASIC did support encoder/decoder functions to 300 Mbps. This includes the lock indicator signal of the ASIC.

The digital link simulator is interesting in that it is completely under PC control. In order to handle all the modulation modes it generates quantized I and Q samples based on the modulation mode and the desired Eb/No (fig. 5.). These samples are then mapped into the desired hard decision bits biased on the modulation mode. The modes tested were all PSK signals but there is nothing stopping the RAM memory from being loaded with QAM profiles or FSK profiles. The one limit is that only 2-ary, 4-ary, 8-ary and 16-ary signals can be tested. The TDE equipment was designed to handle both the ASIC and the FHSC. It therefore can also generate the soft decisions needed for the FHSC CODEC designed but never built.

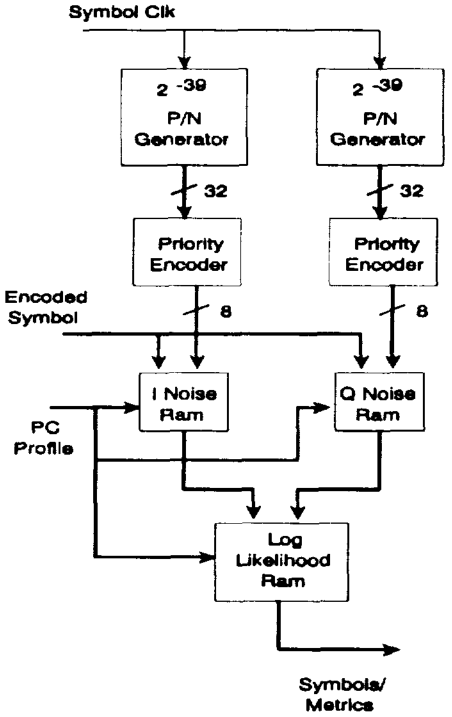


Fig. 5 Noise Generator / Log Likelihood

POTENTIAL APPLICATIONS

The BCH CODEC was developed to deal with the problems of satellite communication systems, in particularly the poor signal quality typically associated with satellites. Many of the video bandwidth compression algorithms require BER' s better than can be maintained on satellites. The BCH CODEC can support the high data rates needed for video signals while providing considerable improvement to the signal quality for a very small overhead. The networking of computers and machines over long distances or to remote locations also may be best served by satellites. Once again the BCH CODEC can provide the coding gain needed to guarantee link integrity. The ability to resolve carrier phase ambiguities, in PSK systems, is an attractive alternative to differential encoding. The self synchronization feature minimizes the circuitry needed to integrate the ASIC into a continuous link.

Though the ASIC's design was tailored to the parameters of a satellite link the basic BCH encoder/decoder can provide coding gain to any link. Cellular phones could pick up considerable coding gain to combat the problems of fading and multi-path associated with travel around a city. High speed digital transcontinental links like those being developed for the phone system could benefit. Digital audio links, such as those being developed for the Cable TV industries, where as many as 30 digital audio channels are multiplexed together forming a high speed digital link, could also use the CODEC.

Considerable interest has already been generated by the chip's development. For many the low overhead rate of the code coupled with magnitude of the coding gain is the most desirable feature. Speed is not an issue. The primary concern of these users is the power consumption and cost. The power consumption of the BCH

CODEC ASIC, as with all CMOS designs, is a function of the clock speed. For clock speeds less than 100 Khz the power consumed is essentially the DC power of the chip and is less than 50 mWatts (Fig. 6).

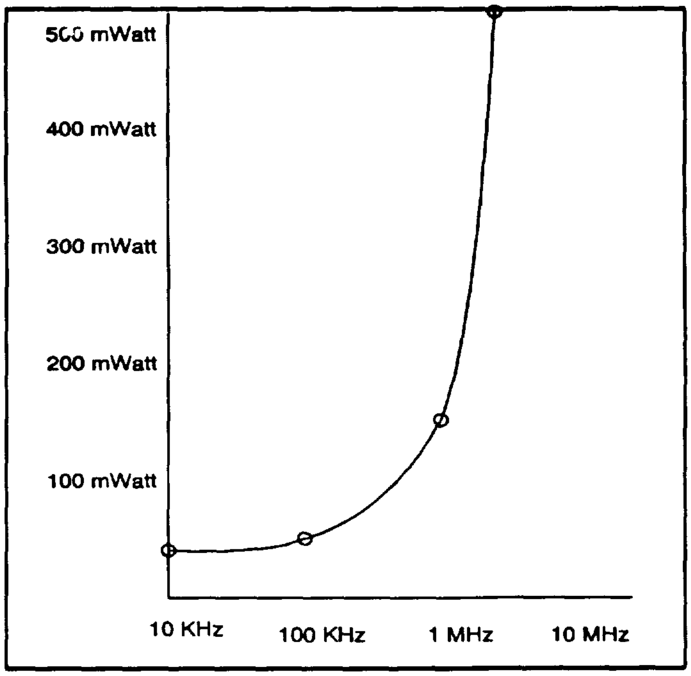


Fig. 6 Power Consumption vs Speed

Cost and Availability

Currently the only chips in existence are those belonging to NASA. Harris is currently evaluating the market. If a market can be identified, several things could be done to drive the costs down. Currently the ASIC's die size is unnecessarily large, due to the large package size needed to accommodate the pin count needed to support all the features of the ASIC. Elimination of the carrier phase ROM's, the bit location circuitry and optimization of the lock detection circuits would result in a sizable reduction in the die size. In addition, the die size could be further reduced by implementing the resulting design in the smaller gate technologies now available. The smaller die would result in an increase in yield and a reduction in cost. The resulting die could be packaged in a 64 pin PLCC rather than the 132 pin PGA, further reducing the price.

CONCLUSIONS

A high speed, high rate CODEC suitable for both burst and continuous modes of operation has been developed by NASA and Harris. It can operate as a single chip hard decision CODEC or, with a decoding appliqué, it can utilize soft decision information in the decoding process. Coding gains up to 4 dB are obtained by the BCH CODEC ASIC, increasing to up to 5.5 dB with soft decisions.

Error correction coding has long been considered a good means to lower the required EIRP in communication systems having unlimited bandwidth. However, high-rate codes such as the one described are also well suited for bandwidth efficient systems. The CODEC rate and interface are matched to the larger signaling alphabets used for constrained bandwidth communications. Performance data indicates that coding gain improves slightly with increasing modulation alphabet size and is a weak function of code word length. Even with the overhead required to insert parity bits, the net result is less power required to communicate a given data rate over a fixed bandwidth channel.

Performance testing indicates the BCH CODEC performs very close to the theory. Using the TDE equipment, coding gains for many other modulation formats can be evaluated. The approach is extremely flexible by design. The BCH CODEC supports several different modulation formats and interface modes, at data rates up to 300 Mbps/s.

It is believed that the approach and hardware resulting from this project will prove useful to a variety systems. Tailoring the design to a specific application would reduce the size cost and power consumption of the CODEC, required by many potential applications.

References :

1) Polkinghorn, F. Jr., "Decoding of Double and Triple Error Correcting Bose-Chaudhuri Codes," IEEE Transactions of Information Theory, October 1966. pp. 480-481.

2) Chase, D., "A Class of Algorithms for Decoding Block Codes With Channel Measurement Information," IEEE Transactions on Information Theory, January 1972, pp. 170-182.

3) Boyd, R. and Hartman, W., "Flexible High Speed CODEC," Advanced Modulation and Coding Technology Conference Publication, June 1989, Session 3, article 6.

4) Segallis G. and Wernlund J. , "Flexible High Speed CODEC," The 14th International Communication Satellite Systems Conference and Exhibit, Conference Publication Part 2, March 22-26, 1992, pp.820-825

ULTRA-STABLE, LOW PHASE NOISE DIELECTRIC RESONATOR STABILIZED OSCILLATORS FOR MILITARY AND COMMERCIAL SYSTEMS

Muhammad Mizan, Thomas Higgins, Dana Sturzebecher
Army Research Laboratory, E&PS Directorate,
AMSRL-EP-MA, Fort Monmouth, N.J. 07703-5601.

ABSTRACT

EPSD has designed, fabricated and tested, ultra-stable, low phase noise microwave dielectric resonator oscillators (DROs) at S, X, Ku, and K-bands, for potential application to high dynamic range and low radar cross section target detection radar systems. The phase noise and the temperature stability surpass commercially available DROs. Low phase noise signals are critical for CW doppler radars, at both very close-in and large offset frequencies from the carrier. The oscillators were built without any temperature compensation techniques and exhibited a temperature stability of 25 parts per million (ppm) over an extended temperature range. The oscillators are lightweight, small and low cost compared to BAW & SAW oscillators, and can impact commercial systems such as telecommunications, built-in-test equipment, cellular phone and satellite communications systems. The key to obtaining this performance was a high Q factor resonant structure (RS) and careful circuit design techniques. The high Q RS consists of a dielectric resonator (DR) supported by a low loss spacer inside a metal cavity. The S and the X-band resonant structures demonstrated loaded Q values of 20,300 and 12,700, respectively.

INTRODUCTION

Systems with stringent performance requirements can benefit from the ultra-stable, low phase noise microwave dielectric resonator oscillators (DROs), at S, X, Ku, and K-bands, reported in this paper. The oscillators, which exhibited excellent temperature stability over extended temperature ranges, were built without any temperature compensation techniques. System designers now have the option of selecting fundamentally operating high frequency DROs without forfeiting critical performance. The key to obtaining this performance was a high Q factor resonant structure (RS) and careful RF circuit design. DROs will play an important role in future military and commercial systems because of their reliability, simple construction, small size, high efficiency, low cost and spurious-free RF output spectrum.

OSCILLATOR DESIGN

The analysis of basic feedback type of circuitry was first given by Leeson.[1] The feedback oscillator configuration allows the circuit designer to isolate low quality or faulty components by measuring the residual noise of the oscillator's components before they are employed in an oscillator circuit. Knowing the magnitude of residual noise of individual components, such as the loop amplifier, resonator and power divider, the absolute phase noise of an oscillator utilizing these components can be estimated.[2] Because of this advantage, the feedback loop (parallel feedback) oscillator configuration, shown in Figure 1, was chosen. Because of similarities in design, the X-band oscillator design is described in detail with only the performance of the S, Ku and K-band DROs reported, all being summarized in Table 1.

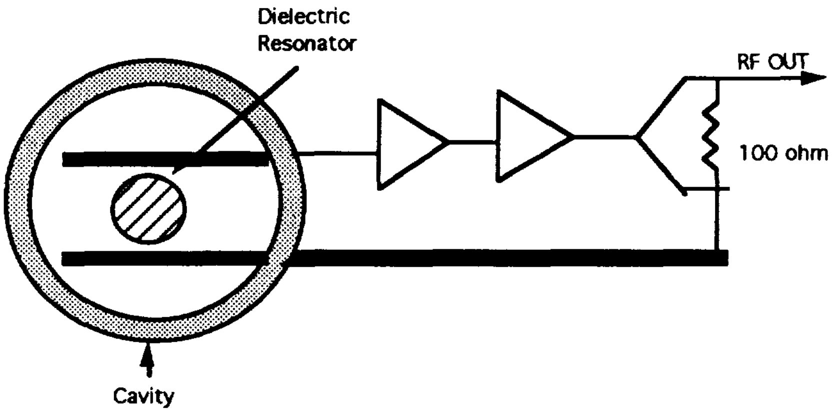


Figure 1. Parallel feedback configuration

DIELECTRIC RESONATOR LOADED CAVITY DESIGN

The cavity dimensions were chosen such that the TE$_{01\partial}$ mode of the resonator was well separated from the cavity modes. Also, the amount of coupling, and the positioning of the dielectric resonator (DR) in the cavity are very critical in obtaining optimum performance. A spacer made of a material with a low dielectric constant was used for mounting the DR inside the cavity, and is shown in Figure 2. Improper mounting will degrade the Q and increase vibration sensitivity.

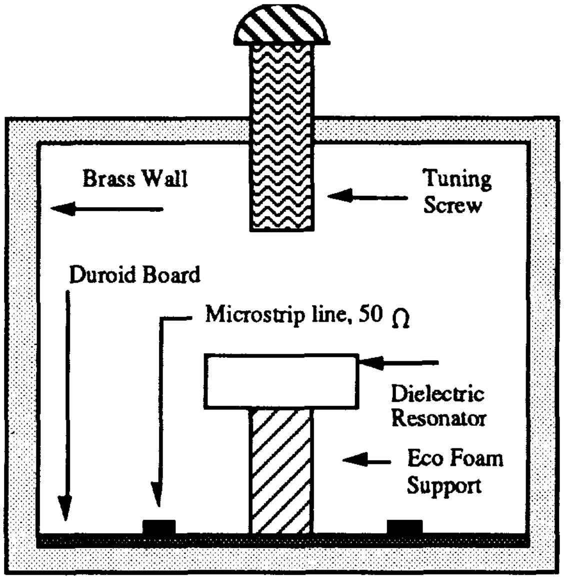


Figure 2. Cavity Configuration

The modes of the cylindrical brass cavity, in the absence of the DR, were analyzed using the cylindrical cavity resonant frequency formulas for TE$_{nml}$ and TM$_{nml}$, which are given by[3]:

$$f_{nm\ell} = \frac{c}{2\pi\sqrt{\mu_r \varepsilon_r}}\sqrt{\left(\frac{p'_{nm}}{p}\right)^2 + \left(\frac{\ell\pi}{d}\right)^2} \quad \text{and} \quad f_{nm\ell} = \frac{c}{2\pi\sqrt{\mu_r \varepsilon_r}}\sqrt{\left(\frac{p_{nm}}{p}\right)^2 + \left(\frac{\ell\pi}{d}\right)^2} \qquad (1)$$

The calculated modes were then verified by network analysis measurements. Figure 3 shows the air filled cavity mode resonances, which were identified as TE_{111}, TM_{010} and TM_{011}, at 8.765 GHz, 8.985 GHz and 10.415 GHz, respectively. These modes were excited by the 50 ohm microstrip transmission line located at the bottom of the cavity. Shown in Figure 4, is the response of the cavity in the presence of the DR and as expected, the cavity mode resonances shifted lower in frequency. Figure 4 also shows the separation between the $TE_{01\partial}$ mode and the cavity modes. The resonant frequency of the DR was very sensitive to movement of a metal screw, which tunes via fringing field perturbations. Figure 5 shows the behavior of the cavity modes as the tuning screw was plunged into the cavity from the top. The TM_{011} mode was also extremely sensitive to the tuning screw and completely overlaps the $TE_{01\partial}$ mode at some tuning positions. Figure 5 suggests that one has to be very careful with the tuning screw, because carelessness could result in operation on an undesired cavity mode.

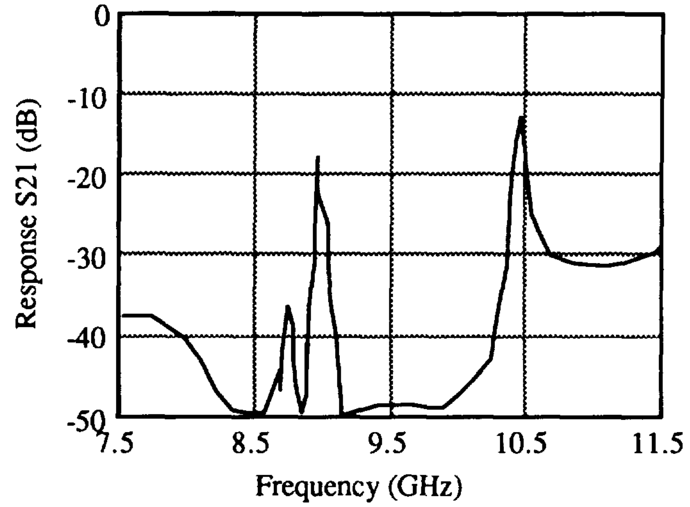


Figure 3. Air Filled Cavity Response

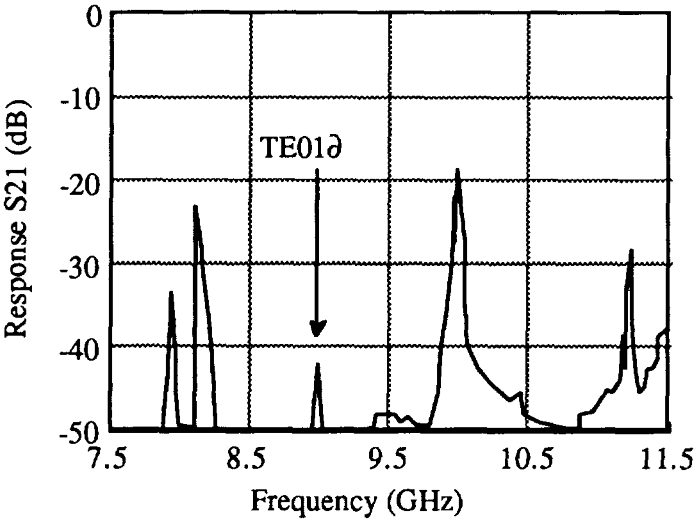


Figure 4. Response of Cavity with DR in Place

The loaded Q measurement for the X-band dielectric resonator is shown in Figure 6, showing a 3 dB bandwidth of 709 KHz at a center frequency of 9.043 GHz, which corresponds to a Q of 12,700. The insertion loss was about 12 dB. Higher loaded Q values were easily attainable by varying the position of the DR, however, at the cost of higher insertion loss. A Q as high as 18,000 was attained at X-band with 20 dB insertion loss, which if used in an oscillator configuration, would have required two additional gain stages to be overcome.

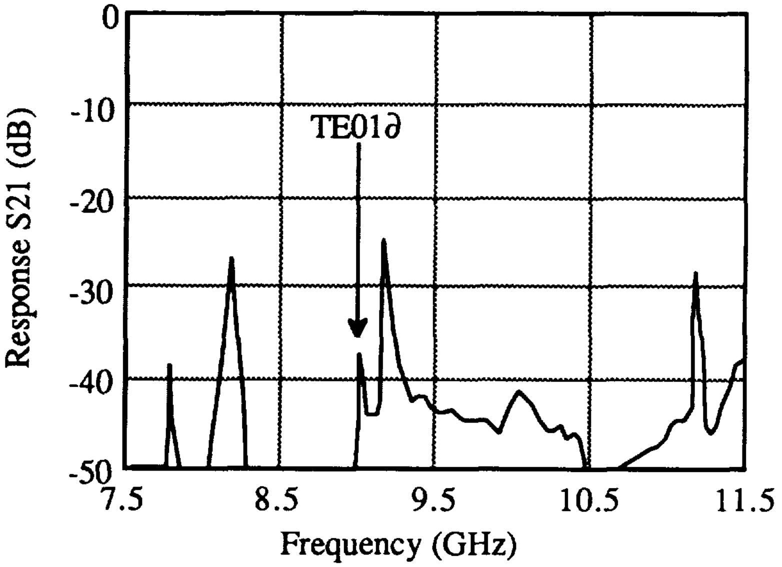


Figure 5. Cavity Response with Tuning Screw Movement

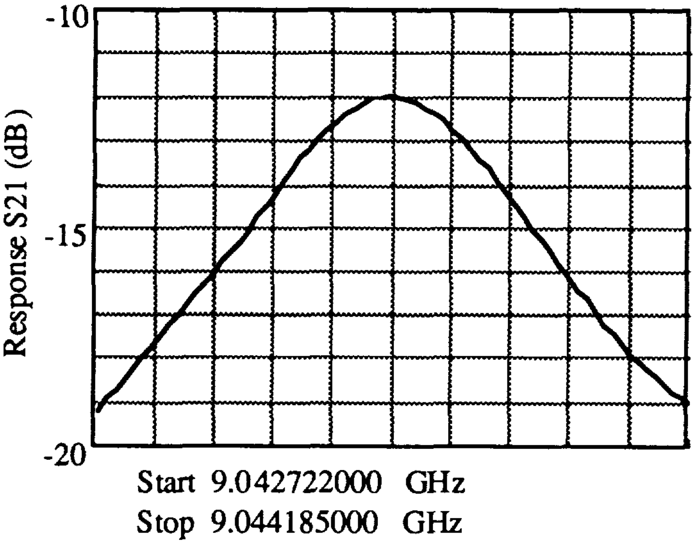


Figure 6. Response of the X-Band Resonant Structure

AMPLIFIER DESIGN

It is well known that bipolar junction transistor (BJT) amplifiers have much lower 1/f noise than GaAs MESFET amplifiers, and for this reason the S-band DRO utilized BJTs. At X-band there existed several different choices of active devices. The BJT was eliminated because of the much reduced gain at the higher frequencies, with the other alternatives being HEMTs, HBTs and MESFETs. Several amplifiers utilizing these devices were built and measured to determine which offered the lowest 1/f noise. An amplifier employing a Fujitsu FSX52WF MESFET was found have the lowest 1/f noise at X-band, however the same device was unable to be used at Ku-band because of its low gain, so another Fujitsu device was identified for use.

The absolute phase noise of a DRO can be improved either by increasing the loaded Q factor of the RS and/or by lowering the 1/f noise of the oscillator's loop amplifier. Since the loaded Q is related to the insertion loss of the RS, additional gain would need to be designed into the loop amplifier in order to achieve higher Q factors, however too many gain stages in the loop amplifier would degrade the overall phase noise due to the addition of 1/f noise. A two stage amplifier was found to be the best compromise between a high loaded Q and the number of gain stages. With only one amplifier, the highest Q achievable was approximately 4,000. Whereas with two stages, the achievable Q was approximately 13,000. This is a factor of 3.25 improvement. It should be noted that a 6 dB improvement in overall phase noise is realized by doubling the loaded[2] Q, while there is only a 3 dB degradation in phase noise because of the added gain stage.

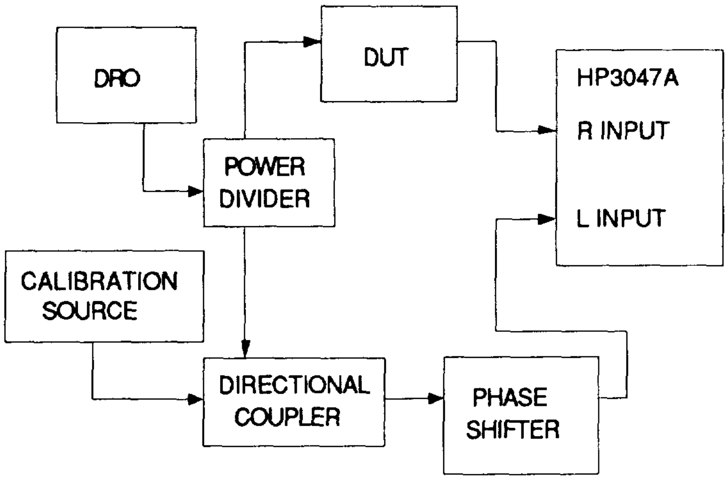


Figure 7. System Used to Measure 1/f Noise

The test setup used for measuring the amplifier 1/f noise is shown in Figure 7. The system is driven by a low noise DRO in order to get the best system noise floor. The system is capable of detecting a noise level of -140 dBc/Hz at 100 Hz offset from a 9 GHz carrier frequency. The residual phase noise of a single stage Fujitsu MESFET amplifier was measured, and found to be at or below the noise floor of -140 dBc/Hz. The drain bias had a significant effect on the 1/f noise, and it was found that the noise level improved as the drain current increased from its normal operating point. The X-band oscillator incorporated two of these amplifiers to overcome the insertion loss of the high loaded Q RS, which thereby offset the noise degradation of the additional gain stage.

The loop amplifiers for the other frequency bands were evaluated in the same manner as for X-band. The residual phase noise (1/f noise) of the S-band BJT amplifier was and found to be at or below a noise floor level of -145 dBc/Hz.

RESULTS

The single side-band absolute phase noise of the 9 GHz two-stage MESFET DRO was measured by downconverting the test DRO to 153 MHz. The down conversion was achieved by mixing the 9 GHz DRO with a high-overtone bulk acoustic resonator (HBAR) oscillator. Then the measurement was made by phase locking the 153 MHz signal to a HP 8662A frequency synthesizer driven from an external 10 MHz VCXO. The measurement configuration is shown in Figure 8. The 9 GHz DRO exhibited a SSB phase noise level of -65 dBc/Hz at a 100 Hz carrier offset frequency,which is an improvement of 6-10 dBc/Hz over previously published data.

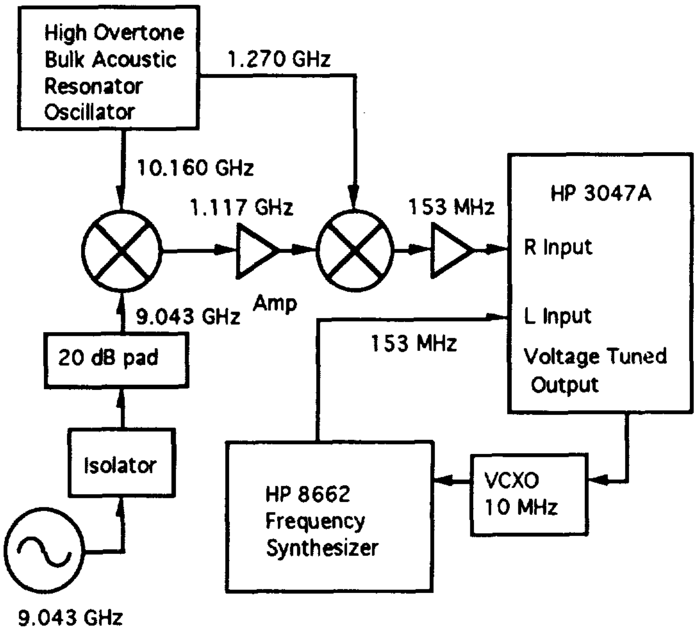


Figure 8. System Used to Measure Absolute Phase Noise

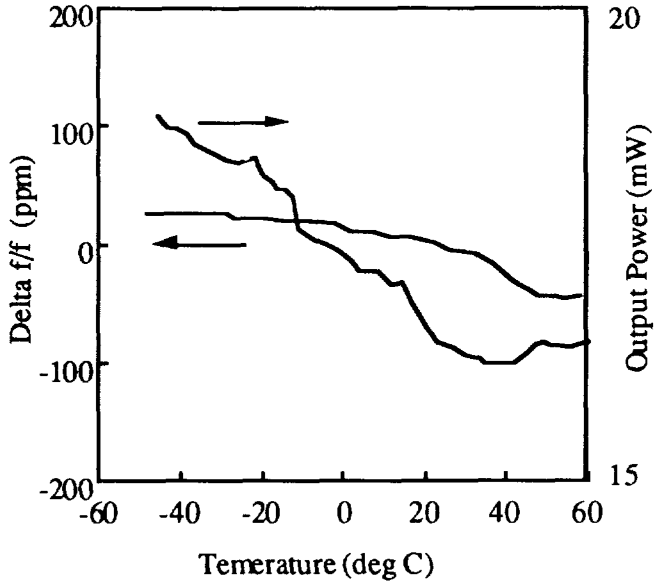


Figure 9. Power & Frequency Variation vs. Temperature of X-Band DRO

Frequency stability and RF output power vs. temperature was measured by using a computer controlled temperature chamber. The DRO was subjected to a temperature profile that began with a 15 minute soak at +55°C and then proceeded to drop at a one degree C per minute from +55°C to -45°C. The total frequency drift of the X-band DRO from -50°C to +20°C was only 25 ppm, and 65 ppm from -50°C to +50°C. Typical RF power output at room temperature was 16.5 mW, and the maximum variation over the full temperature range was 3 mW. The frequency and power vs temperature is shown in Figure 9. The frequency variation with bias voltage (voltage pushing) was found to less than 25 KHz/V over a 4 volt range. The DROs at the other operating frequencies were tested in the same manner, and all the results are summarized in Table 1.

CONCLUSION

Signals with low phase noise are critical for CW doppler radars, at both very close-in and large offset frequencies from the carrier. Design procedures have been presented here which show an improvement over previously published data for temperature stability (for uncompensated DROs) and phase noise.

Band	Frequency (GHz)	Insertion. Loss dB	Q_L	Phase Noise @100 Hz dBc/Hz	Phase Noise @1 KHz dBc/Hz	Temp Stability ppm
S	2.091	21	20,.300	-97	-127	150
X	9.043	12	12,700	-65	-93	65
Ku	15.900	12	5,800	-40	-70	99
K	24.973	To be determined				45

Table 1. Summary of Measured Results

REFERENCES

[1] D. B. Leeson, "A Simple Model of Feedback Oscillator Noise Spectrum", Proceedings of the IEEE, Vol. 54, No. 2, pp. 329-330, February 1966

[2] G. K. Montress and T. E. Parker, "Design Techniques for Achieving State-of-the-Art Oscillator Performance," Proceeding of 44th Annual Symposium on Frequency Control 1990, pp. 522-535. IEEE Catalog No. 90CH2818-3.

[3] D. M. Pozar, Microwave Engineering Addison Wesley, 1990

EXCIMER LASER PROCESSING OF BACKSIDE-ILLUMINATED CCDS

S. D. Russell
Naval Command, Control and Ocean Surveillance Center,
RDT&E Division (NRaD), Code 553, San Diego, CA 92152-5000

ABSTRACT

An excimer laser is used to activate previously implanted dopants on the backside of a backside-illuminated CCD. The controlled ion implantation of the backside and subsequent thin layer heating and recrystallization by the short wavelength pulsed excimer laser simultaneously activates the dopant and anneals out implant damage. This improves the dark current response, repairs defective pixels and improves spectral response. This process heats a very thin layer of the material to high temperatures on a nanosecond time scale while the bulk of the delicate CCD substrate remains at low temperature. Excimer laser processing of backside-illuminated CCDs enables salvage and utilization of otherwise nonfunctional components by bringing their dark current response to within an acceptable range. This process is particularly useful for solid state imaging detectors used in commercial, scientific and government applications requiring a wide spectral response and low light level detection.

BACKGROUND

A number of image-gathering detectors use charge coupled devices (CCDs) with varying degrees of sensitivity and resolution. CCDs are solid state electronic imaging devices which read out image charges from wells in an array of pixels. CCDs designed for solid-state cameras, such as camcorders, are in great demand and are widely available. They have been designed to provide adequate performance when viewing brightly illuminated scenes. However, in astronomical, scientific and military applications their spectral response, dark current, and other characteristics are not satisfactory. Excimer laser processing of backside-illuminated CCDs will be shown to overcome these deficiencies.

Spectral Response:

To overcome the limitations of imaging through the polysilicon gates that necessarily cover all of the sensitive pixel array, it would be desirable to illuminate the CCD from the backside if the silicon substrate were thin enough. In other words, a solution to obtaining better light sensitivity would be the thinning of the backside of the CCD to a total thickness of roughly 10 microns and illumination from the backside. When the silicon substrate upon which the array resides is made thin enough to permit short-wavelength light (blue and ultraviolet) to penetrate into the active regions of the device, improved spectral response has been obtained. However, for a backside-illuminated CCD, the electrical characteristics of the shallow region near the back surface dominate the CCD response to short wavelength photons. Silicon develops a thin native oxide ($<$ 30 Å thick) that can contain enough trapped positive charge to deplete a region several thousand Angstroms deep into the CCD. The absorption depth for high energy (UV or blue) photons in silicon is very short (about 30 Å for 250 nm light and about 900 Å for 400 nm light). Therefore, photogenerated electrons created in this region can drift toward the Si/SiO_2 interface and become trapped or recombine thereby drastically reducing the quantum efficiency in the UV and blue.

One method of accumulating the backside of the CCD is by ion implantation of the backside and subsequent heating to activate the dopant. Initially, only a fraction of the implanted dopant atoms reside in locations in the crystal which are electrically active. Thermal energy is provided to permit the migration of dopant atoms into active sites. The obstacle that must be overcome by fabricators when this approach is relied on is that the backside doping process (and heating) occurs after all frontside device fabrication. A large temperature elevation of the frontside circuitry at this point in the process can cause deleterious effects. For example, backside doping of a silicon substrate with boron has been attempted to enhance the spectral response and suppress the dark current of CCD detectors. Boron implantation is normally followed by a thermal anneal at 1000°C for thirty minutes. But temperatures above about 600°C can cause damaged contacts (spiking) and damage to metal layers in a device. Temperatures exceeding about 800°C can cause diffusion of dopants affecting transistor threshold and leakage values. Since the final implant occurs after all frontside device fabrication, the anneal temperature is restricted to 400°C. At this temperature, boron doses of approximately 10^{13} ions/cm^2 have only 10 to 20% of the dopants activated [1]. ·As the implant dosage increases, the silicon crystal becomes more damaged and the percentage activation decreases. The consequence is that frequency response is affected and dark current can rise to objectionable levels (see discussion below). Therefore, with the standard processing scheme, there is a tradeoff between improving spectral response and improving dark current.

Dark Current:

High performance low light detecting CCDs are also susceptible to dark current, i.e. the thermally generated charge carriers under zero illumination. Excessive dark current will destroy the dynamic range of the imager thereby masking low light level signals. In addition, variations across the array will degrade image quality and can be misinterpreted by subsequent signal processing circuitry. Dark current effects in CCDs range from individual pixels with excessive dark current to high average dark current and variations in dark current across the imaging array. While dark current is normally associated with front side circuitry, crystalline damage arising from unannealed implanted dopants can lead to generation sites for dark current. Therefore, fabrication techniques to improve both the spectral response and dark current of CCDs is highly desirable.

EXPERIMENTAL

CCD Test Vehicle:

Backside-illuminated CCDs containing a 90 pixel x 90 pixel array were used as test vehicles for the laser process. The CCD was a conventional 4-phase buried channel device. In addition to the dark current and spectral response reported in this paper, the CCDs were fully tested for functionality before and after laser processing to ensure no damage to the either the imaging area or the associated electronics.

Excimer Laser Processing Apparatus:

Figure 1 schematically shows the laser processing system. The excimer laser beam is directed into an optical path which homogenizes and shapes the intensity profile to provide uniform illumination across the active area of the CCD without scanning the beam. This is required since the intensity profile emitted by the excimer laser exhibits both spatial and temporal (pulse-to-pulse) nonuniformities and must be correctly shaped for the specific CCD array geometry. A typical intensity profile emitted by an excimer laser is shown in Figure 2 (A). A subsequent shaped and homogenized intensity profile which is directed into a processing chamber and then onto the CCD to be processed is shown in Figure 2 (B). The laser processing system in Figure 1 includes in-situ characterization of the laser process using a reflectivity monitor to measure the melt duration of the silicon material which is an important process control parameter. Additional process controls may include mass flow controllers for process and purge gases, evacuation systems and alignment systems. Details of these subsystems have been described elsewhere [2].

Excimer Laser Process Recipe:

The device physics dictates that high concentration of p^+ dopants be used to prevent the trapping of photogenerated charges near the back surface of the CCD as described above. Therefore, the CCDs were ion implanted with boron to doses of 5 x 10^{13} cm^{-2}. Rather than receiving a 30 minute anneal at 400°C in a furnace as prescribed in the conventional fabrication for the backside implant, the devices were subsequently transferred to the laser processing chamber and the ambient evacuated and backfilled with an inert gas. Processing was conducted using the excimer laser operating at 248 nm with a KrF gain medium. Pulse repetition rates up to 100 Hz were attainable with pulse energies up to 750 mJ. The laser intensity profile was homogenized, shaped and directed normal to the sample surface. Upon illumination with sufficient laser fluence ($\phi_{melt} \geq 0.7$ J/cm^2), the silicon melts allowing redistribution of the dopants within tens of nanoseconds. The silicon then recrystallizes, resulting in dopants in electrically active sites and annealing of crystalline damage caused by the ion implantation. Processing parameters along with typical and ranges of values are given in Table I.

TABLE I. Processing Parameters

PARAMETER	TYPICAL VALUE	RANGE
implant dose	5x10^{13} ions/cm^2	1x10^{13} - 1x10^{15} ions/cm^2
implant species	boron	B, BF_2 or none (see alternate process in text)
implant depth	120 nm	5 - 150 nm
furnace anneal	none	≤ 400°C, 30 min
process ambient	helium	inert or dopant gases (see alternate process in text)
sample temperature	23°C	< 400 °C (restricted by the device not the laser process)
laser fluence	1.0 J/cm^2	0.7 - 2.0 J/cm^2
laser wavelength	248 nm	157 - 351 nm
laser intensity profile	tophat, < 5% nonuniformity	< 10% nonuniformity
laser temporal profile	23 ns	10 - 30 ns
number of laser pulses	10	1 - 10

Note, an alternative process involves the elimination of the ion implantation step altogether. In such a case, the process ambient is a dopant gas such as boron trifluoride (BF_3), which can be photolytically or pyrolytically decomposed by the laser and incorporated into the molten silicon. Subsequent laser annealing in an inert ambient follows, as above. This in-situ laser doping process is described in more detail elsewhere [3].

RESULTS

Figure 3 shows the electrically active charge carrier profiles of dopant concentration vs. depth obtained using the spreading resistance profiling technique for three individual samples which were all treated in the manner listed in Table I, except for variations in laser fluence. The silicon samples were identical to those used in the fabrication of the CCD arrays. The samples were irradiated with ten pulses with laser fluences of 0.7, 0.8 and 0.9 J/cm^2, respectively. Samples undergoing conventional furnace annealing used for the backside implant showed approximately 10 to 20% boron activation while the laser activated samples shown here exhibit approximately 100% activation. As shown in Figure 3, the dopant profile may be controlled by changes in laser fluence since the depth of active dopant distribution increases with increasing laser pulse energy. Similarly, varying the number of laser pulses can change the profile from a graded profile with peak concentration near the surface to that of a uniform dopant distribution.

Responsivity Improvements:

Responsivity improvements occur in two areas. First, the responsivity becomes more uniform across the array. This is a product of the uniform illumination and subsequent uniform recrystallization of the backside of the CCD. All samples show an improvement in response uniformity originally degraded by the spatial nonuniformities in the implant. Secondly, the response of the CCD to blue and shorter wavelengths of light is improved by providing for a peak dopant concentration at the back surface to allow for the photons absorbed near the back surface to be collected by the pixel electrodes as described in the background. Tests were performed to detect the observed improvement in responsivity with the laser process. Test CCDs were laser annealed on one half of the array with 10 pulses at 1.5 J/cm^2. The devices were subsequently flood illuminated with blue light (400 nm) and a line scan across the device was measured. Figure 4 shows the spectral response line scan from a representative CCD. Note that on the half of the CCD which received laser processing (the left hand side in Figure 4), the response to the blue light improved over that half which received no laser processing (43.4 nA/μW vs. 37.8 nA/μW). The 20% increase demonstrated here is not optimized, but is representative of the improvement obtained using this technique. Additional improvements can be obtained using lower laser fluences and the in-situ laser doping process mentioned above to create a more shallow anneal thereby keeping the peak dopant concentration closer to the surface.

Dark Current Improvements:

CCDs were fabricated and tested. Devices were selected which exhibited dark current defects, i.e. excessive average dark current, nonuniformities, and/or individual pixels with excessive dark current. The test devices were then laser processed in accordance with the recipe in Table I and retested. Results of a typical (not best case) laser annealed sample will be discussed and is shown in Figure 5. This sample was chosen due to the unique spiral defect structure which was "repaired" by laser processing. Figure 5A shows a map of the dark current for the 90 x 90 pixel array prior to laser processing. Numerous defective pixels, with dark currents exceeding 11 nA/cm^2 are present. Figure 5B shows the dark current map following the laser process. Note that all defective pixels were improved. The mean dark current of the 8100 pixels in the array decreased from 9.958 nA/cm^2 to 9.658 nA/cm^2. The standard deviation of the dark current, which is representative of the uniformity in the array, decreased from 4.826 nA/cm^2 to 0.812 nA/cm^2. Furthermore, individual pixels with excessive dark current, attributed to generation at crystalline defects were eliminated or reduced. Table II shows a summary of a correlation of the pixel data before and after laser processing demonstrating that defective pixels with a severe dark current level (> 50 nA/cm^2) are reduced to low or moderate dark current levels. Those defective pixels with low or moderately high dark current levels (< 50 nA/cm^2) are removed completely.

TABLE II. CCD Pixel Dark Current

DARK CURRENT	DEFECT TYPE	NUMBER OF PIXELS	
		PRE-LASER PROCESSING	POST-LASER PROCESSING
0 - 5 nA/cm^2 above array mean	none	7974	8098
5 - 10 nA/cm^2 above array mean	low	66	1
10 - 50 nA/cm^2 above array mean	moderate	46	1
≥ 50 nA/cm^2 above array mean	severe	14	0

CONCLUSION

An excimer laser has been used to activate the final boron implant on the backside of backside-illuminated CCD arrays. Results indicate that the laser fully activates the dopant resulting in a significant reduction in the mean dark current, improved dark current uniformity and repair of defective pixels with excessive dark current. Furthermore, responsivity improvements occur both in the uniformity and the sensitivity (quantum efficiency) to short wavelength (blue and UV) light. Apparent minor modifications in processing techniques in semiconductor fabrication can lead to major cost savings, yield and reliability improvements due to the large volume production and repetitive nature of processing. Therefore, extensive effort is placed on eliminating even one step from a process flow since each step has an associated yield. This is more important involving backside processing of CCDs since substantial fabrication costs and time are invested in the device by this step in the fabrication. Therefore, it is apparent that the laser process described for repair of defected pixels in addition to improving dark current and enhancing blue response simultaneously in one process step is highly desirable alternative to conventional processing. In addition, the laser process described herein is compatible with other laser techniques, e.g. backside thinning (etching) or sidewall texturing which may be implemented in prior processing steps [2,4,5].

Acknowledgements:

This work was done in collaboration with D. Sexton, E. P. Kelley & R. E. Reedy at the Naval Ocean Systems Center (now NCCOSC, RDT&E Division).

REFERENCES

1. H. Ryssel, I. Ruge, *Ion Implantation*, (New York, John Wiley & Sons, 1986), page 248.

2. S. D. Russell, D. A. Sexton, "Excimer Laser Thinning of Backside Illuminated CCDs", NOSC-TD-1697, November 1989.

3. D. A. Sexton, S. D. Russell, R. E. Reedy, E. P. Kelley, "Excimer Laser Dopant Activation of Backside Illuminated CCDs", Navy Case No. 72,219 (patent pending).

4. S. D. Russell, D. A. Sexton, "Responsivity Uniformity Enhancements for Backside-Illuminated Charge-Coupled Devices (BICCDs) by Excimer Laser-Assisted Etching", NOSC-TD-2103, May 1991.

5. These additional laser processes applicable to backside-illuminated CCDs are described in the following pending patents: Navy Case No. 71,978, 72,726, 73,014, 74,142, 74,182, and 74,183. Information may be obtained by writing to the Patent Counsel, Code 0012 at NCCOSC, RDT&E Division , San Diego, CA, 92152-5000.

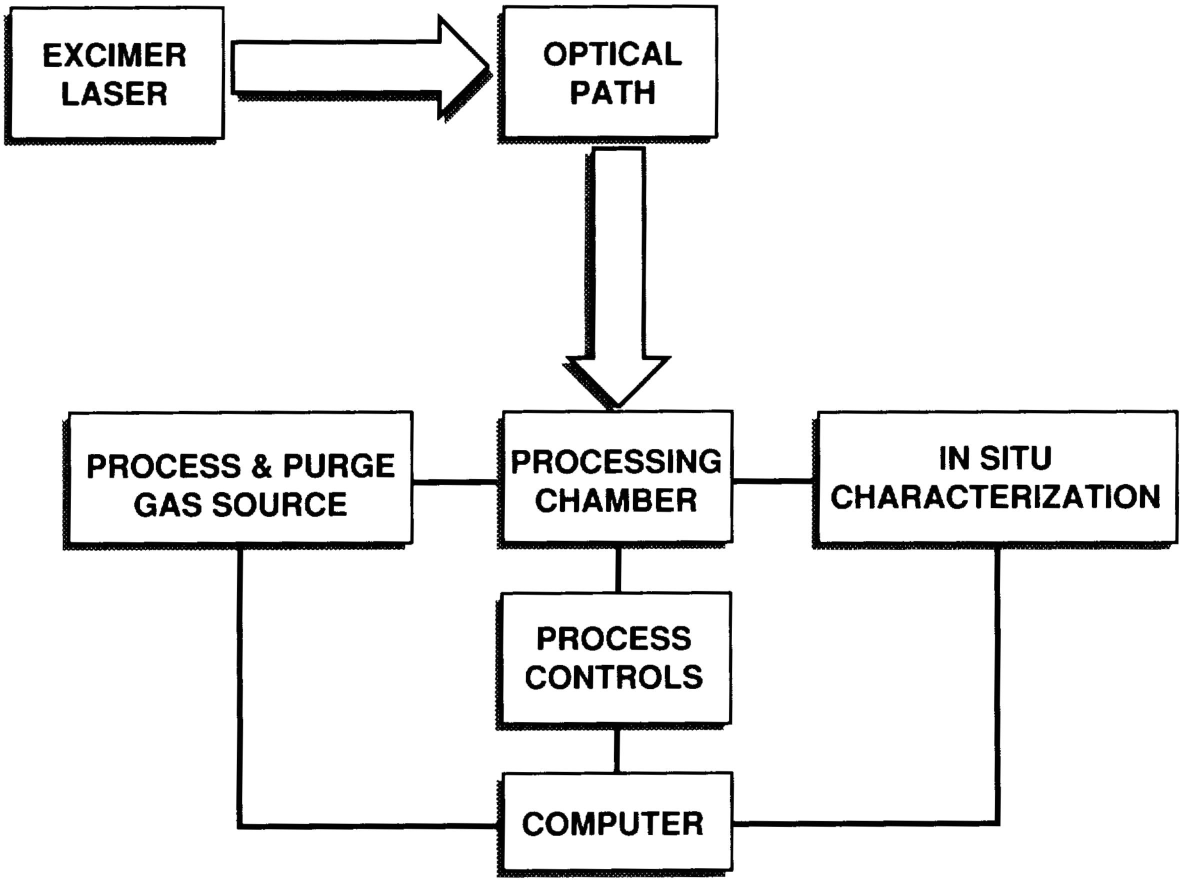
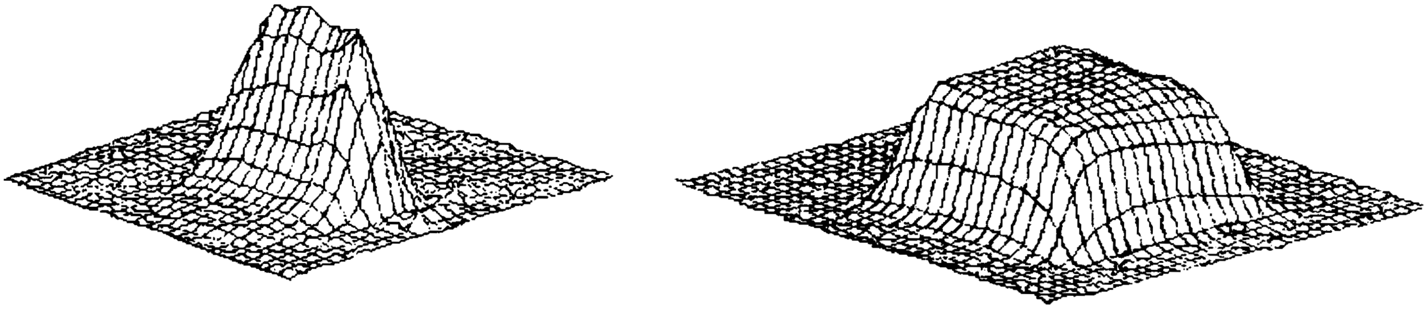


Figure 1. Schematic of excimer laser processing system.



(a) Typical excimer laser intensity profile.

(b) Homogenized intensity profile.

Figure 2.

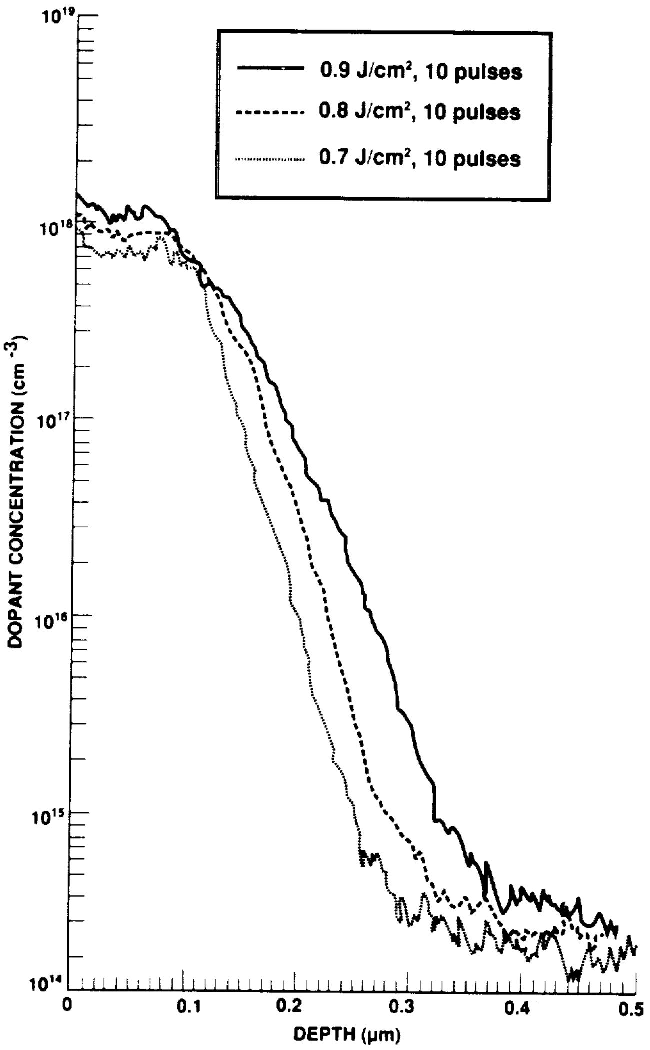


Figure 3. Dopant concentration vs. laser fluence.

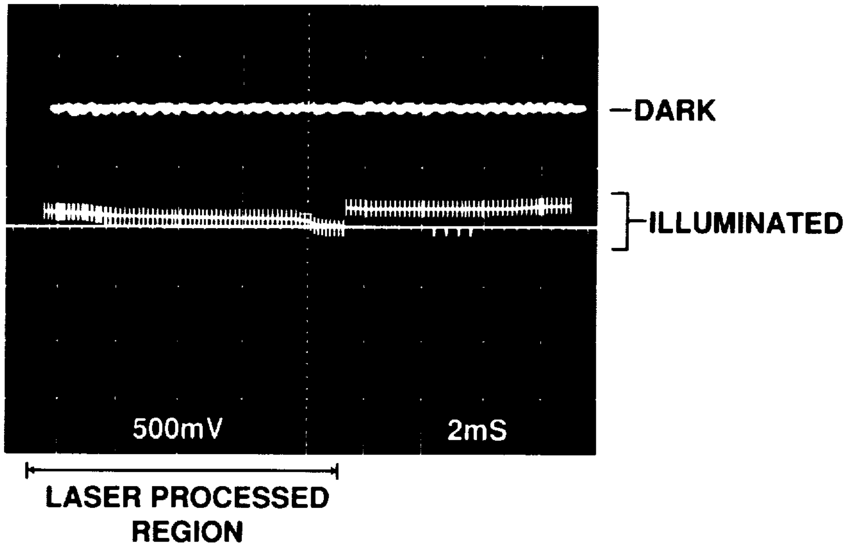
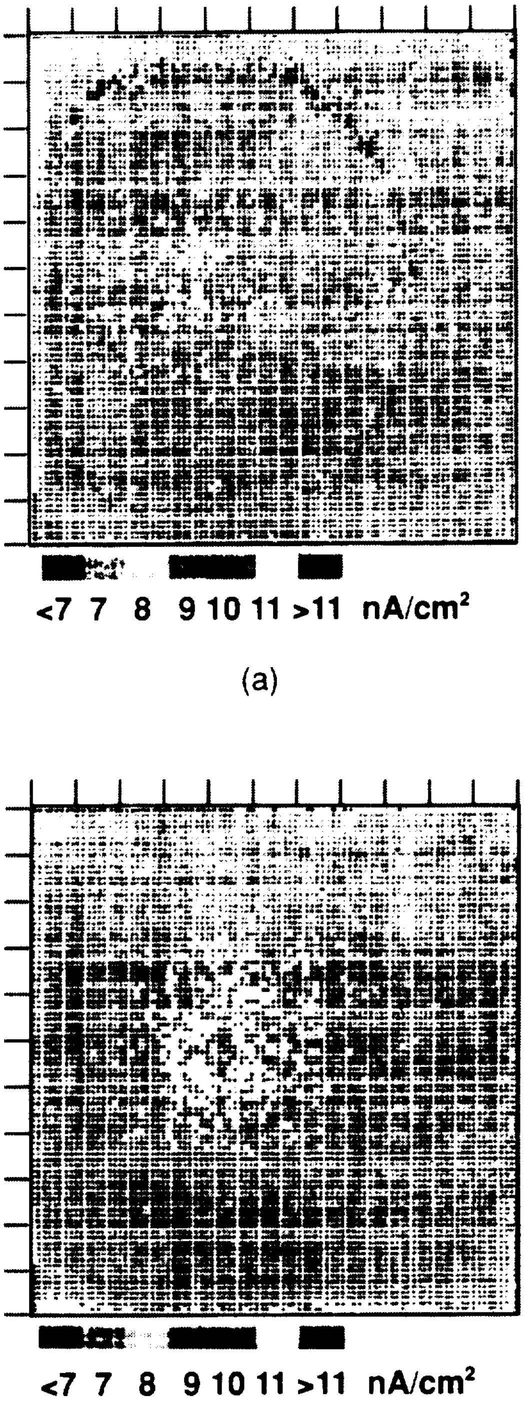


Figure 4. Spectral response line scan.



(b)

Figure 5. Dark current map (a) before, and (b) after laser processing.

ADVANCED MATERIALS
PART 4

p. 10

APPLICATIONS OF FIBROUS SUBSTRATES CONTAINING INSOLUBILIZED PHASE CHANGE POLYMERS

Tyrone L. Vigo and Joseph S. Bruno
USDA, ARS, SRRC
1100 R. E. Lee Blvd.
New Orleans, LA. 70124

ABSTRACT

Incorporation of polyethylene glycols into fibrous substrates produces several improved functional properties when they are insolubilized by crosslinking with a methylolamide resin or by polyacetal formation by their reaction with glyoxal. The range of molecular weights of polyols that may be insolubilized is broad (M_n of 600-20,000) as are the curing conditions (0.25-10 min at 80-200°C). Most representative fiber types and blends (natural and synthetic) and all types of fabric constructions (woven, nonwoven and knit) have been modified by incorporation of the bound polyols. The most novel property is the thermal adaptability of the modified substrates to many climatic conditions. This adaptability is due to the high latent heat of the crosslinked polyols that function as phase change materials, the hydrophilic nature of the crosslinked polymer and its enhanced thermal conductivity. Other enhanced properties imparted to fabrics include flex and flat abrasion, antimicrobial activity, reduced static charge, resistance to oily soils, resiliency, wind resistance and reduced lint loss. Applications commercialized in the U. S. and Japan include sportswear and skiwear. Several examples of eclectic sets of properties useful for specific end uses are given. In addition, other uses are biomedical, horticultural, aerospace, indoor insulation, automotive interiors and components and packaging material.

INTRODUCTION

Unique sets of properties imparted to fibrous substrates or materials containing crosslinked polyethylene glycols has been described in our publications on the insolubilization of polyols by crosslinking with DMDHEU or dimethyloldihydroxyethyleneurea with acid catalysts [1,2,3,4] and more recently by their reaction with glyoxal to form polyacetals by tosylate intermediates [5]. Multifunctional property improvements vary to some extent with the nature, construction and porosity of the fibrous substrate, the curing conditions employed, the type of crosslinking agent employed, molecular weight of the polyol, and various additives in the solutions used to effect polymerization or insolubilization. Moreover, certain applications require an improved set of product attributes different from those required in other applications. Commercial products and applications in which there are current industrial interest are evaluated with regard to desirable property improvements imparted by this process. Structural aspects of the crosslinked or insolubilize polyols that are responsible for property enhancement are discussed. The interactive nature of processing and polymerization conditions with the nature of the fiber assembly and geometry are also explored. Future developments and improvements based on these concepts and processes are presented.

MULTIFUNCTIONAL PROPERTIES IMPARTED

Water-soluble polyols may be insolubilized onto fibrous materials by reacting with the tetrafunctional resin DMDHEU or with glyoxal by a sulfonate intermediate. Representative structural units of the modified polyols are shown in Figure 1. The existence of a fiber-polymer matrix results in many improved functional properties compared to the untreated substrate or fabric. Each of these attributes is due to one or more aspects of the modified polyols bound to the fiber and are noted in Table I. Thermal adaptability is the most novel of all the properties imparted to fibers and is probably due to three simultaneous phenomena. First, the insolublized polyethylene glycols (as a network polymer in either reaction system) function as PCM's (phase change materials) with high latent heats of fusion and crystallization, thus buffering temperature changes in hot and cold weather. Secondly, these modified polymers are also very hydrophilic and provide enhanced thermal comfort due to their water sorption properties and the amount of energy stored during the evaporation and condensation of excess water. Thirdly, there is also evidence from our infrared thermography studies that the crosslinked polymer in the fiber matrix changes the thermal conductivity of the fiber surface. This latter effect has also been observed by scientists engaged in aerospace research [6]. The increased moisture content, sorption and capacity of the modified fabrics are also due to the hydrophilic nature of the crosslinked polymer. The marked increase in wear life (usually observed by increases in flex abrasion) and in sizable reduction of lint loss or particle release and fiber entanglement (pilling) on the surface is attributable to the elastomeric nature of the insoluble polyols. Such an elastomeric nature gives the coated fibers the ability to absorb mechanical stresses and deformations and prolongs failure that would more readily occur in the unmodified fibers. Enhancement of the soil release and antistatic behavior of fabrics containing the bound polyols was expected, since the incorporation of various types of polyethylene oxides to fibers is well-documented and was observed as early as 1957. Durable press properties have been observed (conditioned wrinkle recovery angles as high as 325° in all cotton fabrics). The mechanism of resiliency probably differs conventional crosslinking of cellulosic fibers with resins in that this effect can be achieved at very low curing temperatures and involves grafting and homopolymerization of the polyol to provide such resiliency. Although the rationale for enhanced wind resistance was not explained, it may be due to removal of moisture from the air as it passes through the fabric; further study of this effect is in progress. The antimicrobial activity may be due one or two factors: (a) the slow release of an active antimicrobial agent (such as formaldehyde) under certain conditions and/or (b) the hydrophilic nature of the coated fiber that dessicates the microorganisms and thus deprives them of moisture needed to sustain growth. The mechanism of antimicrobial activity is also being studied in more detail.

TABLE I. Multifunctional Properties Imparted to Fibrous Substrates Containing Insolubilized Polyols

Attributes of modified substrates	Cause(s) of attributes
Thermal adaptability and thermal comfort	Latent heat of polyethylene glycols Moisture content and water transport properties Thermal conductivity of polymeric coating
Water sorption, capacity and content	Hydrophilic nature of crosslinked polymer
Resistance to wear, lint loss and pilling [7]	Flexibility of polymeric gel or elastomer
Reduction of static charge and improved soil release [8]	Nonionic and hydrophilic nature of polymer
Resiliency or anti-wrinkling [9]	Grafting to cellulosics and elastomeric nature of polymer
Enhanced wind resistance [10]	Possible removal of moisture from air by hydrophilic polymer
Antimicrobial activity [11]	Hydrophilicity of polymer or slow release of antimicrobial agent

INFLUENCE OF FIBER TYPE AND FABRIC CONSTRUCTION

Due to the nature of the reactions used for insolubilizing the polyols (condensation reaction of hydroxyl end groups of the polymer with a tetrafunctional N-methylol cyclic urea and use of the dialdehyde glyoxal), certain types of functional groups on fibers will also react with these resins or functional groups. Grafting of the polymer onto cellulosic fibers as well as some reaction of resin with the cellulosic fibers occurs even at low curing temperatures (120°C or less). This has been verified by using scanning electron microscopy to detect the insolubility of modified cellulose fabrics in solvents such as cupriethylenediamine [12]. The reaction of cellulosic fibers with glyoxal under a variety of conditions is also well documented. When the curing temperature is higher, the cellulose reacts as readily as the polyol does with the crosslinking resin and glyoxal, and the resultant fabric has much poorer mechanical properties than fabrics cured under milder conditions. Wool, polyamide and polyurethane fibers contain some primary amino (-NH_2) groups on the surface of the fibers. Thus, there is also the possibility of grafting the polymer onto these fibers as well as reaction of the amino groups with the crosslinking resin and reaction of these groups with the dialdehyde glyoxal to form imine bonds. However, durability of the polymer coating or the polymer in the fiber matrix is not as good for wool as it is for polyamide and polyurethane substrates. This is due to the poor physical bonding of the polymer to the scales of the wool fiber.

Other types of fibers, such as polyester and acrylic, have fewer functional groups available (such as -OH) that make grafting and/or direct reaction with the resin or glyoxal less likely than with cellulosic, proteinaceous and fibers containing amino- groups. However, durability of the polymer on these types of fibers is good. Thus, bonding may occur in these types of fibers by some ionic interactions and possibly even hydrophobic type bonds between the polymer and fiber surfaces. Chemically inert fibers such as polypropylene and glass have some durability of the crosslinked polymer in the fiber matrix. This may be due to the presence of a few hydroxyl groups on the fiber surface and/or hydrophobic bonding between the polymer and the fiber surface. In addition to the nature of the reactive groups of the fiber, the nature of the fiber surface will also affect the adhesion, type of bonding and durability of the crosslinked polymer in the fiber/polymer matrix.

The physical nature of the fabric or fibrous assembly also has some influence on the amount of polymer incorporated into the matrix, its distribution in the matrix and some of the physical properties of the modified fabric or surface. These effects are primarily independent of fiber type and related to the porosity, air permeability and construction of the fabric. For example, the amount of polymer bound in a loosely woven cotton and loosely woven polypropylene would be about the same. Moreover, the distribution of the polymer in each type of fabric would be similar. In this instance, the open fabric structure would result in more polymer being bound between fibers and very little polymer on the fabric surface. For tightly woven fabrics, polymer attachment is more difficult, and more polymer tends to be present on the fabric surface than between fibers. More surface deposition of polymer also results in a stiffer fabric or material than in fabrics where surface deposition is minimal. For knits, these effects are the same as those observed in wovens. Since knit constructions tend to be more open and porous, modified knits usually have a good hand after treatment. Nonwovens are very porous and open structures (relative to most wovens and knits) and thus have the best surface aesthetics of all three major types of fabric constructions when they contain the crosslinked and bound polymers.

Since the modified fabrics contain a crosslinked polymer, their dimensional stability to laundering, that is, their ability to retain their original dimensions, is excellent. This is particularly useful for knit fabrics, since these types of constructions have poor dimensional stability in the unmodified state. In contrast, the ability of the modified fabrics to retain their dimensions in the wet state varies markedly with the fabric construction. Nonwoven fabrics containing the crosslinked polymer have very good wet dimensional stability. With most nonwovens, there is 0-2% shrinkage in the wet state. With wovens, this wet shrinkage can be between 7-15% in area. With certain types of knits, the wet shrinkage is dramatic (ranges of 25-40% in area), while with other knits, wet shrinkage may be 15% in one direction, but expand 15% in another direction, giving an overall area change of zero. Again, this phenomenon is independent of fiber type. All types of fabric constructions, even those with high wet shrinkage, return to their original dimensions on drying, and have outstanding dry dimensional stability. The wet shrinkage in certain constructions has been viewed by some as a negative attribute in early stages of the development of textile products. However, some scientists and engineers are now exploring how this dramatic difference in fabric dimensions in the wet and dry state can be used to produce novel mechanical effects useful in irrigation and other applications as self-adaptive structures with a shape memory that responds to changes in humidity and water content.

EFFECT OF SOLUTION COMPOSITION

The amount of polyols bound to fibrous surfaces and the resultant properties imparted are not only dependent on the fiber type and fabric construction but are also dependent on the molecular weight of the polyethylene glycol (M_n 600 to 20,000), the composition of resin and acid catalyst employed for the DMDHEU reaction. For the insolubilized polyacetal derived from the polyol, it is dependent on both the concentration of glyoxal and the concentration and type of sulfonic acid used to form sulfonate intermediates as well as the molecular weight of the polyol. Unreacted, lower molecular weight polyols, such as those with M_n of 600 and 1,000, are fairly amorphous and have low melting (T_m) and crystallization (T_c) temperatures and moderate enthalpies of fusion (H_f) and crystallization (H_c). As the molecular weight increases, the polyols become more crystalline and have higher H_f and H_c, and correspondingly higher T_m and T_c. Enthalpies are optimum at M_n of 6,000-10,000, and the melting and crystallization temperatures tend to level off at these molecular weights. The same trend in the thermal properties is observed when these polyols are either crosslinked with the tetrafunctional resin DMDHEU or react with glycols to form insoluble polyacetals. However, their latent heat values and melting and crystallization points are lower than those of the unmodified polymers. The magnitude of their heat absorption (H_f) and heat release (H_c) generally increases with increasing molecular weight. This magnitude is more dependent on the curing conditions than on the fiber type and fabric construction. When these polymers are crosslinked, the amorphous regions are the first to react, and thus higher molecular weight polyols have more crystalline material remaining to provide better latent heat properties than lower molecular weight, amorphous p polyols. Figure 2 shows thermal scans of modified fabrics containing insolubilized polyols derived from reaction with PEG-1,000/DMDHEU and from PEG-3,350/glyoxal/methanesulfonic acid. In each instance, the area under the curve is the latent heat of fusion absorbed with increasing temperature. The maximum heat absorption occurs at T_m. When such modified fabrics are cooled, comparable heat release occurs at lower temperatures with maximum heat release occurring at T_c.

Polyethylene glycols may be reacted with unalkylated DMDHEU resins (e.g., the structure shown in Figure 1 (a) 0in which all four reactive groups are hydroxyl or may be reacted with DMDHEU resins in which the two $-N-CH_2OH$ groups are partially alkylated and/or glycolated. Several catalysts are effective (such as p-toluenesulfonic acid, a mixed system with $MgCl_2.6H_2O$/citric acid or $NaHSO_4$) for network polymerization of both types of resins at the same level of reactivity, but there are at least two major advantages to using the alkylated DMDHEU resin. The first advantage is that there are much lower levels of free formaldehyde in the solution containing the alkylated DMDHEU than those containing the unalkylated DMDHEU. Nevertheless, most fabrics treated with either resin system have negligible amounts of formaldehyde release (0-70 ppm) after washing and drying. The second major advantage is that the resultant fabrics are usually much softer when the alkylated DMDHEU is used as a crosslinking agent than when the unalkylated DMDHEU is used. This phenomenon is probably due to less crosslinking (slower reaction rate) of alkylated groups relative to unalkylated groups in the resin. However, there are some instances where certain types of fabrics are equally supple when either resin may be used to crosslink a polyol of the same molecular weight. A good example is the treatment of knit 90/10 cotton/Lycra (elastomeric polyurethane) fabrics with PEG-1,000/DMDHEU. Scanning electron microscopy indicates that there is little surface deposition of polymer for both resins used to treat this fabric. Each of the resultant fabrics was softer than the corresponding untreated control fabric.

For reaction of the polyols with glyoxal to form insolubilized polyol bound to the fibers, stoichiometric amounts of sulfonic acids are required. Since methanesulfonic acid is half the molecular weight of p-toluenesulfonic acid, it may be effectively used at concentrations as low as 6% to form sulfonate end groups of the polyols that subsequently react with glyoxal. This system has the advantage of being totally formaldehyde-free, but adhesion an of the polyacetal to the fibrous surface and prolonged durability to laundering appears not to be as good as that of polymer derived from reaction with DMDHEU resins. Studies are in progress to modify fiber surfaces and use other techniques to improve such durability.

EFFECT OF CURING CONDITIONS

The polyethylene glycols were initially insolubilized on various types of fibers by a conventional pad-dry-cure process [1]. This process usually consisted of immersing the fabrics in solutions containing approximately 60% solids (polyol + resin), removing excess solution through squeeze rolls, drying 5-7 minutes at 85°C, then curing for 2-3 minutes at 140-160°C. Although modified fabrics prepared by this process had acceptable thermal properties, cellulosic fabrics such as cotton had substantial strength losses that were similar to those usually obtained when cotton was treated with any durable press agent. Moreover, all types of treated fabrics were much stiffer than the corresponding untreated fabrics. It was later determined that using such conventional process conditions resulted in overcuring that reduced desirable thermal properties by reaction of the resin with crystalline regions of the polyols. Moreover, at high cure temperatures the hydroxyl groups in the cotton fabric react at the same rate with the DMDHEU as the hydroxyl end groups of the polyol react with DMDHEU. This leads to a rigid system in which the cellulosic fibers lose at least half their tensile strength.

When the mildest curing conditions (time/temperature) were employed in a single step to bind and insolubilize polyols to fibers, most functional improvements were superior to those obtained by the above two-step conventional dry-cure process. The most comprehensive change and improvement were in the heats of fusion and crystallization of the crosslinked polymer and corresponding higher temperatures of T_m and T_c of the bound polymer. In many instances, fabrics cured by a single step method had superior thermal properties to those cured by the conventional two-step method even when the former fabrics had less polymer incorporated than the latter fabrics. A representative example is curing of 55/45 pulp/polyester nonwoven fabric treated with PEG-1,000/DMDHEU. The fabric cured in a single step for 1.5 min. at 90°C had a weight gain of only 46%, but had a T_m of 35°C, a T_c of 10°C and corresponding H_f of 22 J/g and H_c of 21 J/g. The same fabric cured in a single step for 3 min. at 90°C had a weight gain of 87%. However, even under these conditions it was overcured, since it had a T_m of 12°C, a T_c of -10°C and corresponding H_f of 16 J/g and H_c of 15 J/g.

Cotton fabric treated with PEG-1,000/DMDHEU, then cured by a conventional two-step process, had breaking strength and flex abrasion losses of about 50% compared to the control. This treated fabric was also about three times stiffer than untreated fabric. In contrast, cotton fabric treated with an identical solution, then cured for 5 min/85°C lost only 20% of its breaking strength and had an 800% increase in its flex life compared to untreated cotton fabric. Moreover, the treated cotton fabric was about 40% softer than the untreated fabric. Thus, it is usually beneficial to use the minimum curing conditions to insolubilize the polymer inside the fiber matrix.

Insolubilization of the polyols via formation of polyacetals also is sensitive to optimum curing conditions for a particular formulation and molecular weight of polyol. Preliminary results [5] indicate that somewhat higher curing temperatures (usually above 125°C) are required to insolubilize the polyols by this route than are required to insolubilize them by reaction with DMDHEU resins in the presence of acid catalysts.

APPLICATIONS SUITABLE FOR MULTIPROPERTY IMPROVEMENTS

The diversity of improved properties imparted to fabrics and fibrous substrates make the modified materials suitable for many applications. Nevertheless, there are certain sets of properties that are more useful for each application. Table II illustrates a few examples of applications and the sets of functional properties appropriate for that application.

TABLE II. Relationship between Application and Functional Properties Imparted to Fibrous Substrates Containing Crosslinked Polyols

Application	Group of Desirable Functional Properties
Sportswear and skiwear	Thermal adaptability, wind resistance, water sorption softness and dimensional stability
Garments for biomedical and computer clean rooms	Thermal adaptability, resistance to static charge, antimicrobial activity, lint loss, water sorption
Shoe components and socks	Thermal adaptability, durability to wear, water sorption and antimicrobial activity
Automotive interiors	Thermal adaptability, resistance to static charge, oily soil release, wear resistance
Industrial and consumer wipes	Sorption of water and oil, liquid capacity, antimicrobial activity, wet dimensional stability
Work uniforms	Thermal adaptability, resistance to static charge, pilling and oily soil release, anti-wrinkling, durability to wear and prolonged laundering
Indoor insulation	Thermal adaptability, ability to remove humidity from air, dimensional stability, antimicrobial effects

Pulp/polyester nonwoven fabrics used for surgical scrub suits afford a representative example of property improvements that are specific and relevant to this end use. When these fabrics are treated with PEG-1,000/DMDHEU and subsequently cured, thermal adaptability is imparted (heat absorption in the range of 25-35°C). The modified fabric absorbs at least twice the amount of water in the liquid and gaseous state, has a 100 fold decrease in static charge, 300% improvement in its flex life, and has significantly less particles released than from the corresponding untreated nonwoven fabric. Numerous other examples of eclectic sets of properties suitable for a specific application could be cited, and currently form the basis for many commercial products from current and prospective licensees. In addition to the applications or end uses in Table II, other possible end uses include blankets and bedding materials, geotextiles, horticultural and agricultural applications and defense/aerospace applications such as space suits and military uniforms and apparel. Some applications are to replace existing materials while other applications are more novel and may lead to the development of products that did not previously exist.

FUTURE TRENDS AND OPPORTUNITIES

Predicting future commercial developments from present scientific trends and interest in any area of science is quite difficult. However, there appear to be several fundamental concepts and practical opportunities in the application of polymers to fibers to improve many functional properties by a single process. It would be worthwhile to investigate further the structural aspects of polymers that cause them to function as phase change materials. A fundamental comparison of the few compounds, polymers and other compositions with high latent heats of fusion and crystallization should also be interesting and rewarding.

Another area of opportunity is to synthesize new polymers or modify existing polymers that impart multifunctional property improvement when they are bound to a fibrous substrate. Combinations of properties should lead to the development and commercialization of materials that could function in physically and chemically hazardous environments for both civilian and military uses. Opportunities exist for use of these new materials in the biomedical and health care areas as well as for pollution control and improvement of the environment.

A third area of opportunity is modification of fiber surfaces by high energy sources such as plasma, glow discharge and excimer lasers. This research has intensified in the past decade [11], and has led to improvement in the adhesion, wettability and dyeability of fibrous surfaces. In conjunction with the attachment of tailor-made polymers to fibrous surfaces, these studies should produce new types of fibrous composites, protective clothing and materials suitable for many new and existing applications.

A final area of opportunity is the emerging science and technology of intelligent materials and self-adaptive structures. Application of polymers to fibers that impart a thermal, mechanical or other type of memory should lead to modified materials that retain or change their shape or structure when exposed to external stimuli such as heat, light, electric current and/or moisture. These concepts, in conjunction with existing knowledge of modified fibrous materials, should lead to new products and technologies in electronic, health care, aerospace and consumer applications.

REFERENCES

[1] T. L. Vigo and J. S. Bruno, "Improvement of Various Properties of Fiber Surfaces Containing Crosslinked Polyethylene Glycols," J. Appl. Poly. Sci., Vol. 37, pp. 371-379, Jan, 1989.

[2] T. L. Vigo and J. S. Bruno, "Fibers with Multifunctional Properties: A Holistic Approach," in M. Lewin and J. Preston, Eds., "High Performance Fibers," Marcel Dekker, New York, in press, 1992.

[3] T. L. Vigo, J. S. Bruno and W. R. Goynes, "Enhanced Wear and Surface Characteristics of Polyol-modified Fibrous Substrates," in L. Rebenfeld, Ed., "Science and Technology of Fibers and Related Materials," Wiley-Interscience, New York, J. Appl. Poly. Sci. Symposia Series, Vol. 47, pp. 417-435, 1991.

[4] T. L. Vigo, C. M. Frost, J. S. Bruno and G. F. Danna, "Temperature-adaptable fibers and method of producing same," U. S. Pat. 4,851,291, July 25, 1989.

[5] T. L. Vigo, G. F. Danna and J. S. Bruno, Proc. Poly. Sci. & Eng., Am. Chem. Soc., Vol. 67, pp. 503-504, Apr. 1992.

[6] L. F. Kuznetz, Ph.D., NASA Ames Center, private communication, Jan. 23, 1991.

[7] T. L. Vigo, G. E. R. Lamb, S. Kepka and B. Miller, "Abrasion and Lint Loss Properties of Fabrics Containing Crosslinked Polyethylene Glycol," Textile Res. J., Vol. 60, pp. 160-171, May, 1991.

[8] E. Kissa, " Soil-Release Finishes," and S. B. Sello and C. V. Stevens, "Antistatic Treatments," in M. Lewin and S. B. Sello, Eds., "Handbook of Fiber Science and Technology, Vol. II, Chemical Processing of Fibers and Fabrics: Functional Finishes, Pt. B.," Chapters 3 and 4, Marcel Dekker, New York, 1984.

[9] J. S. Bruno and T. L. Vigo, "Temperature-adaptable Fabrics with Multifunctional Properties," Proc. Natl. Assoc. Am. Text. Chem. Color. Conf., pp. 258-364, 1987.

[10] S. L. Harlan, "A New Concept in Temperature-adaptable Fabrics Containing Polyethylene Glycols for Skiing and Skiing-like Activities," pp. 248-259, in T. L. Vigo and A. F. Turbak, Eds., "High-tech Fibrous Materials," Am. Chem. Soc. Symp. Series No. 457, Washington, D.C., March, 1991.

[11] T. L. Vigo, "Antimicrobial Activity of Cotton Fabrics Containing Crosslinked Polyols," for presentation at 1993 Beltwide Cotton Conference, New Orleans, LA., Jan. 1993.

[12] W. R. Goynes, T. L. Vigo and J. S. Bruno, "Microstructure of fabrics chemically finished for thermal adaptability," Textile Res. J., Vol. 59, pp. 277-284, May, 1990.

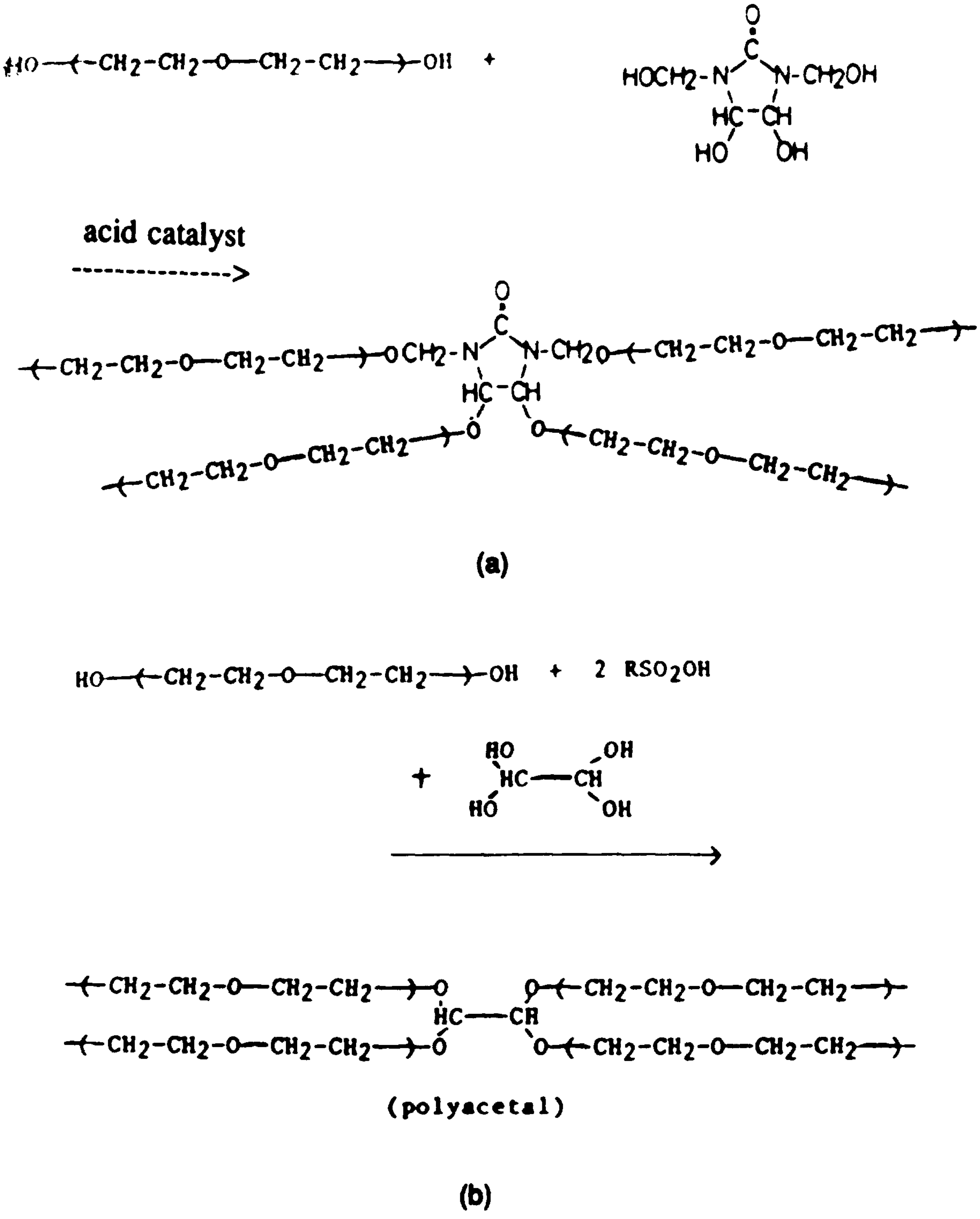


Figure 1. Reaction of end groups of difunctional polyols with (a) tetrafunctional resin DMDHEU to form network or crosslinked structure and (b) dihydrate of glyoxal + sulfonic acid to form polyacetals.

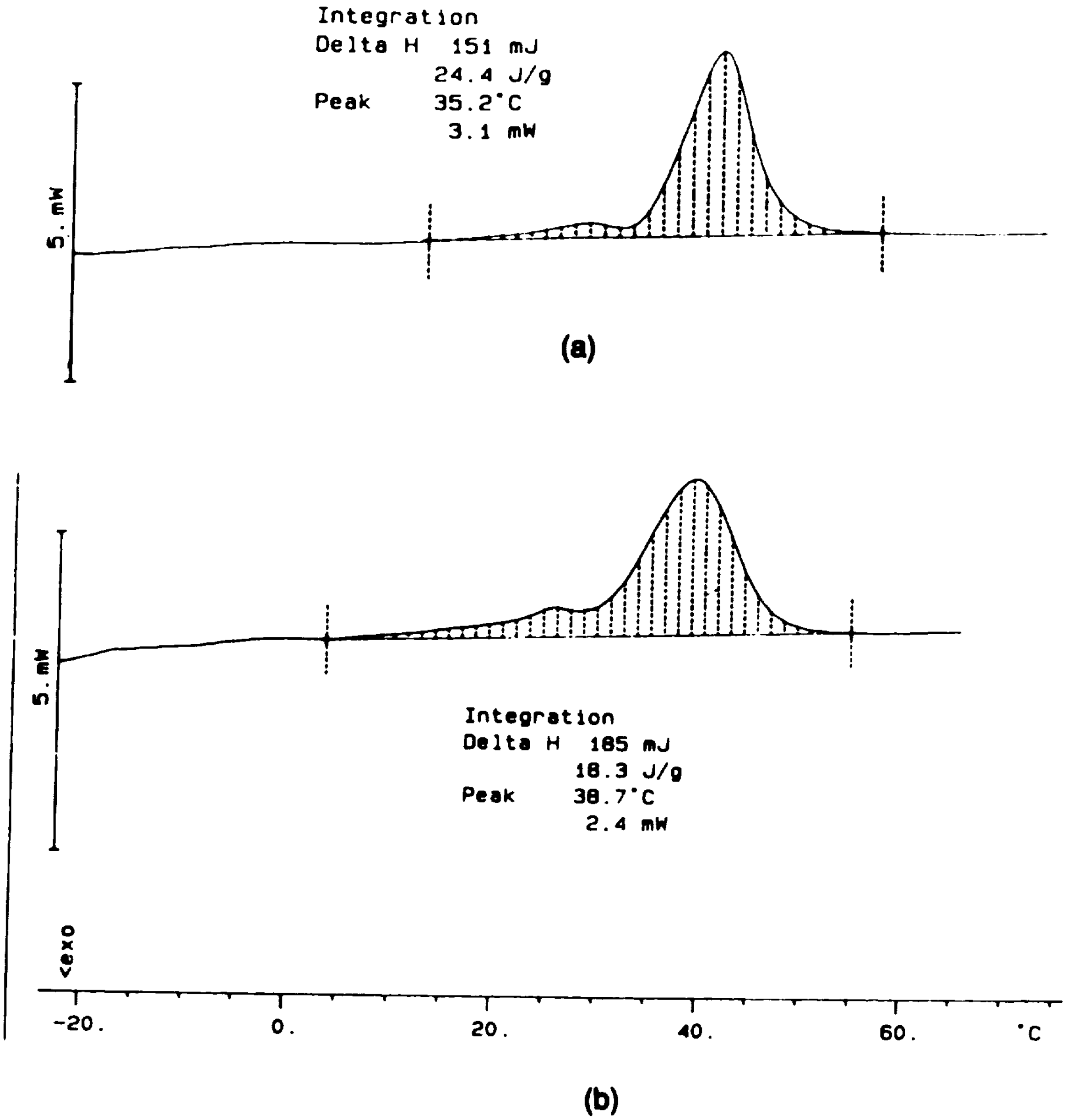


Figure 2. Thermal scans (heating mode) of (a) 55/45 pulp/polyester nonwoven fabric treated with PEG-1000/DMDHEU, cured to wt. gain of 50% and (b) 100% cotton woven fabric treated with PEG-3350/glyoxal methanesulfonic acid to wt. gain of 32%.

RUST TRANSFORMERS/RUST COMPATIBLE PRIMERS

Dario A. Emeric and Christopher E. Miller
U.S. Army Belvoir Research, Development
& Engineering Center
Fort Belvoir, VA 22060-5606

ABSTRACT

Proper surface preparation has been the key to obtain good performance by a surface coating. The major obstacle in preparing a corroded or rusted surface is the complete removal of the contaminants and the corrosion products. Sandblasting has been traditionally used to remove the corrosion products before painting. However, sandblasting can be expensive, may be prohibited by local health regulations and is not applicable in every situation. To get around these obstacles, Industry developed rust converters/rust transformers and rust compatible primers (high solids epoxies).

The potential use of these products for military equipment led personnel of the Belvoir Research, Development and Engineering Center (BRDEC) to evaluate the commercially available rust transformers and rust compatible primers. Prior laboratory experience with commercially available rust converters, as well as field studies in Hawaii and Puerto Rico, revealed poor per-formance, several inherent limitations, and lack of reliability. It was obvious from our studies that the performance of rust converting products was more dependent on the amount and type of rust present, as well as the degree of permeability of the coating, than on the product's ability to form an organometallic complex with the rust. Based on these results, it was decided that the Military should develop their own rust converter formulation and specification. The compound described in the specification is for use on a rusted surface before the application of an organic coating (bituminous compounds, primer or topcoat). These coatings should end the need for sandblasting or the removing of the adherent corrosion products. They also will prepare the surface for the application of the organic coating.

Several commercially available rust compatible primers (RCP) were also tested using corroded surfaces. All of the evaluated RCP failed our laboratory tests for primers.

INTRODUCTION

Proper surface preparation has been the key to obtain good performance by a coating. The major obstacle in preparing a corroded or rusted surface, before the application of an organic coating, has been the complete removal of the contaminants and the corrosion products. Sandblasting has been traditionally used to remove the corrosion products before painting. However, sand-blasting can be expensive, it may be prohibited by local health regulations, and it may not apply in every situation. To get around these obstacles, Industry developed the rust converters /rust transformers, and the rust compatible primers.

The potential use of these products in fielded equipment by the Military, led personnel of the Belvoir Research, Development and Engineering Center (BRDEC) to test the commercially available rust transformers[1] and rust compatible primers[2]. Prior field and laboratory experience with commercially available rust transformers, as well as field studies in Hawaii and Puerto Rico, revealed poor performance and lack of reliability. It was obvious from our studies, that the performance of rust converting products was more dependent on the amount and type of rust present, and on the degree of permeability of the coating, than on the product's ability to form an organometallic complex with the rust. Based on these results, it was decided that the Military should develop their own rust converter formulation (U.S. Patent 4,880,478)[3] and specification (MIL-R-53086)[4]. The compound described in the specification is for use on a rusted surface before the application of an organic coating (bituminous compounds, primer or topcoat). The proper application of the product should end the need for sandblasting or the removing of the adherent corrosion products. It also will produce the proper surface for the application of an organic coating.

Commercially available rust compatible primers (RCP) that were tested[2] failed our laboratory tests for primers. The Chemical Agent Resistant Coating (CARC)[5] system used by the Department of Defense on tactical equipment performed better on corroded surfaces than the so called rust compatible primers.

EXPERIMENTAL PROCEDURE

Cold rolled steel panels[1] cleaned in an aqueous non-ionic detergent solution (0.1%Triton-X100) were exposed for 5 minutes to a 5 % salt fog solution according to ASTM B117. These panels were placed outdoors for three weeks[2]. At this point, the panels were ready to be treated and painted. When the corroded surface was treated with a rust converter the following procedure was used:

Step 1. Wet panel with either distilled or demineralized water
Step 2. Allow excess water to run off by tilting panel
Step 3. Apply on first coat (sprayed or brushed) of rust converter
Step 4. Allow panel to dry for not less than 12 hours
Step 5. Rinse in flowing water and repeat steps 2 through 4
Step 6. Spray 5% sodium bicarbonate solution on panels
Step 7. After 3 minutes, rinse solution off panels
Step 8. Allow panels to dry at room temperature for not more than 24 hours

The treated panels were painted with the CARC primer (MIL-P-53022)[6] to a 1.5 mils (38 μm) dry film thickness (DFT) or with the (MIL-C-62218)[8] bituminous compound to a DFT of 6 mils (152 μm).

Air assisted spray[3], airless spray, and drawdowns were tried to apply the rust compatible primers in one coat to a DFT of 5-6 mils (127-152 μm). Two methods for the application of the primers were developed. These methods should prevent the formation of pinholes that could develop during application of the primers. The first method was to spray the initial coat as a mist or flash coat onto the surface of the panels. The second method used the 8 step rust converter procedure described above. Both processes produced a leveled profile (hills and valleys) of the corroded surface. The panels were then coated to a DFT of 3.5 mils (89 μm) with the rust compatible primers. The paint was allowed to dry for 24 hours, after which, the panels were recoated with 3.5 mils (89 μm) DFT of the respective products to achieve a final DFT of 7.0 mils. Before having the panels tested, they were allowed to dry at room temperature for one week.[4]

[1] Q-Panels, R-Type, 0.023" thickness, dull matte finish, 3" x 6" size, Q-Panel Company, 26200 First Street, Cleveland, OH 44145.

[2] After approximately 12 days of exposure, the panels were powerwashed and sprayed with a 5% salt solution. This procedure was followed to obtain uniform, adherent corrosion on the surface of the panel. After the pre-corrosion process, the loose corrosion was removed using a high pressure power-washer.

[3] Binks Spray Gun, Model 2001, Binks Manufacturing Company 4.5 in. (11.4 cm.) Rubber Coated.

[4] The CARC panels used as controls were wash primed with DOD-P-15328[7] to a 0.4 mils (10 um) DFT prior to the application of the primer MIL-P-53022[6].

WET ADHESION TESTING

The treated panels were tested according to Federal Test Standard No. 141 Method 6301.2[9]. The panels were evaluated according to ASTM D3359[10] (measuring adhesion by tape test).

SALT SPRAY/WET ADHESION TESTING

After the preparation and curing processes were complete, the panels were placed in the salt spray chamber for periods of 336 hours and 500 hours. The test was run according to ASTM B117[11]. Once removed, the panels were rinsed in tapwater, dried, and tested according to Federal Test Method Standard 6301.2[9]. The panels were then tested according to either ASTM D3359[10] (measuring adhesion by tape test) or ASTM D610[12] (non-impinged area) and ASTM D1654[13] (impinged area).

SALT SPRAY/GRAVELOMETER TESTING

After the preparation and curing processes were completed, the panels, before testing, were placed in the cold temperature chamber and conditioned for two hours. The panels were tested according to ASTM D3170[14]. After the gravelometer testing was completed, the panels were placed in the salt spray chamber[5] for 336 and 500 hours according to ASTM B117[12]. Upon removal of the samples from the salt spray cabinet, they were rinsed with tapwater, dried, and evaluated according to ASTM D610[12] (non-impinged area) and ASTM D1654[13] (impinged area).

OUTDOOR EXPOSURE TESTING

Once scribed with a diamond pattern, the panels, treated with the rust converters, were placed outdoors at Fort Belvoir. They were oriented at an angle of 45° from the horizontal facing south for one year. After one year, the panels were removed and tested according to ASTM D610[12] (non-impinged area) and ASTM D1654[13] (impinged area). The rust compatible primers were not tested outdoors because of their failure in the Laboratory.

RESULTS

The CARC topcoat blistered and delaminated from the panels that had been treated with the phosphoric acid base and tannic acid base rust converters. No signs of blistering or delamination were observed on the panels treated with the BRDEC formulation (Tables 1, 2 & 3).

The MIL-C-62218[8] bituminous coating failed on the edges of the panels when treated with the commercially available rust converters. No signs of failure were observed on the panels treated with the BRDEC formulation.

No problems were found, when the RCP was applied to uncorroded surfaces, but as expected problems were found when the RCP was applied to the corroded test panels. The problems were due to the high profile of the corroded surface that prevented the high viscosity paint to fully wet the surface. Several different application methods were tried but the results were the same. None of these paints could be successfully applied to the proper thickness in one coat. It was decided that once the loose corrosion was removed, that either a mist coat of the RCP or rust converter should be applied to lower the corrosion profile and to improve surface wetting. Final film thickness will be achieved by using a two coat process with 24 hours drying period between first and second coats. No differences were found between the products in this test. All the products performed well in the salt spray tests when there were no chips or cuts in the coating (Tables 5, 6 & 7). There was one rust compatible primer that out-performed the others when a damaged paint film was exposed to 500 hours of salt spray. However, when the same product was applied to

[5] Harshaw Salt Fog Cabinet, Model #22, Harshaw Chemical Company

an uncorroded surface and treated in the same manner, it delaminated three days after testing was completed. The reason for the delamination is not known and it may necessitate further work.

CONCLUSIONS

The tests and evaluations discussed in this report were under-taken to find out if the commercially available rust converters or rust compatible primers could be used by the Military on corroded surfaces of military equipment without first having to sandblast the surfaces. The decision was that any rust converter to be used by the Military will have to meet the requirements of MIL-R-53086[4] and not to use any rust compatible primer if the CARC[5] system is going to be used.

REFERENCES

1. D.A. Emeric, B. Westich, R.C. McNeil, Rust Converters, BRDEC Report 2457, November 1987.

2. C. Miller, B. Westich, Evaluation of High Solids Paint, BRDEC Report 2492, May 1990.

3. D.A. Emeric et al, Protective Coatings for Steel Surfaces and Method Application, U.S. Patent No. 4,880,478 November 14, 1989.

4. MIL-R-53086, Rust Converter, Metric

5. MIL-C-46168, Coating, Aliphatic Polyurethane, Chemical Agent Resistant

6. MIL-P-53022, Primer, Epoxy Coating, Corrosion Inhibiting, Lead & Chromate Free

7. DOD-P-15328, Primer (Wash), Pretreatment (Formula No. 117 for Metals)

8. MIL-C-62218, Corrosion Preventive Compounds, Cold Application (For New & Fielded Motor Vehicles & Trailers)

9. Federal Test Method Standard (FTMS) 141, method 6301.2, "Adhesion (wet) tape test"

10. ASTM D3359, "Measuring Adhesion by Tape Test"

11. ASTM B117, "Standard method of salt spray (fog) testing"

12. ASTM D610, "Standard method of evaluating degree of rusting of painted steel surfaces"

13. ASTM D1654, "Standard method for evaluation of painted or coated specimens subjected to corrosive environments"

14. ASTM D3170, "Chip Resistance of Coatings"

TABLES

Table 1
Wet Adhesion
of Corroded Panels
ASTM D3359

Panel	In the Laboratory	In the Field
BRDEC formulation	5B	4B
Phosphoric Acid	4B	2B
Tannic Acid	4B	1B
RCP	3B	3B

Key: RCP - rust compatible primer - high solids epoxy

Table 2
Salt Spray/Wet Adhesion
of Corroded Panels

Panel	ASTM D610	ASTM D3359
	(Rusting)	(Adhesion)
BRDEC formulation	7	4B
Phosphoric Acid	7	4B
Tannic Acid	9	1B
RCP	7	2B

Key: RCP - rust compatible primer - high solids epoxy

Table 3
Outdoor Exposure Testing
of Corroded Panels

Panel	ASTM D610	ASTM D1654
	(Rusting)	(Scribe)
BRDEC formulation	6	3
Phosphoric Acid	5	3
Tannic Acid	2	2
RCP	4	3

Key: RCP - rust compatible primer - high solids epoxy

Table 4

Visual Analysis before Testing

Panel	No Rust Converter Comment	With Rust Converter Comment
Control	ok	
53022	ok	ok
Tannic acid		ok
Phosphoric Acid		ok
BRDEC Formulation		ok
RCP	pinholes throughout	ok

Key: Control -- Cold rolled steel panels + 0.4 mils (10.2 μm) of DOD-P-15328 wash primer + 1.0 mil (25.4 μm) MIL-P-53022.
RCP - rust compatible primer - high solids epoxy
53022 - epoxy primer as per MIL-P-53022

Table 5
Wet Adhesion Results

Panel	No Rust Converter ASTM D3359 Rating (Adhesion)	With Rust Converter ASTM D3359 Rating (Adhesion)
Control	5B	
53022	3B	4B
Tannic Acid		4B
Phosphoric Acid		4B
BRDEC Formulation		3B
RCP	3B	4B

Key: Control -- Cold rolled steel panels + 0.4 mils (10.2 μm) of DOD-P-15328 wash primer + 1.0 mil (25.4 μm) MIL-P-53022

Table 6
Salt Spray/Gravelometer Results (336 hours)

	No rust converter		With Rust Converter	
Panel	ASTM D610	ASTM D1654	ASTM D610	ASTM D1654
	(Rusting)	(Scribe)	(Rusting)	(Scribe)
Control	10	7		
53022	10	8	10	9
Tannic Acid			10	9
Phosphoric Acid			10	9
BRDEC Formulation			10	8
RCP	10	7	10	9

Key: Control -- Cold rolled steel panels + 0.4 mils (10.2 μm) of DOD-P-15328 wash primer + 1.0 mil (25.4 μm) MIL-P-53022
RCP - rust compatible primer - high solids epoxy
53022 - epoxy primer as per MIL-P-53022

Table 7
Salt Spray/Gravelometer Results (500 hours)

Panel	No Rust Converter		With Rust Converter	
	ASTM D610 (Rusting)	ASTM D1654 (Scribe)	ASTM D610 (Rusting)	ASTM D1654 (Scribe)
Control	10	10		
53022	10	9	10	9
Tannic Acid			10	9
Phosphoric Acid			10	7
BRDEC Formulation			10	7
RCP	10	6	10	8

Key: Control -- Cold rolled steel panels + 0.4 mils (10.2 μm) of DOD-P-15328 wash primer + 1.0 mil (25.4 μm) MIL-P-53022
RCP - rust compatible primer - high solids epoxy
53022 - epoxy primer as per MIL-P-53022

S34-25

1[illegible]3

N93-22183

METHODS FOR PREDICTING PROPERTIES AND TAILORING SALT SOLUTIONS FOR INDUSTRIAL PROCESSES

0

p.

Moonis R. Ally
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6044

ABSTRACT

An algorithm developed at Oak Ridge National Laboratory (1) accurately and quickly predicts thermodynamic properties of concentrated aqueous salt solutions. This algorithm is much simpler and much faster than other modeling schemes and is unique because it can predict solution behavior at very high concentrations and under varying conditions. Typical industrial applications of this algorithm would be in manufacture of inorganic chemicals by crystallization, thermal storage, refrigeration and cooling, extraction of metals, emissions control, etc.

INTRODUCTION

The use of electrolytes is ubquitious in commerce and industry, and much work has been done to describe the properties of electrolytes, especially aqueous electrolytes. Description of the properties of such electrolytes in terms of electrostatic interactions has made the treatment cumbersome and difficult to apply in a practical sense because of the requirement for a large number of experimentally determined parameters.

Until this algorithm was developed at the Oak Ridge National Laboratory, no modeling scheme could quickly predict the relative performance of salt solutions in the concentration range from 2 to 98 mole %. Earlier modeling schemes address many parameters to evaluate a solution's performance. Such complex modeling schemes are slow, are limited to predicting the performance of low (~6 molal) concentrations, and do not permit extrapolation from one set of conditions to another. The algorithm runs on a personal computer and can easily generate in a day the equivalent of one year experimental work. Only two or three parameters are required by the algorithm. For many salts, these parameters are included in a data base that can be augmented and edited by the user. The algorithm is user friendly with pull-down menus. Results can be obtained in either tabulated or graphical form, depending upon user preference.

COMPARISON OF COMPUTED VS. MEASURED DATA

Comparisons of the computed and empirical data on vapor pressure-composition-temperature, molar volume-composition-temperature, enthalpy-concentration-temperature, and solid phase behavior for electrolytes are shown in Figures 1-3 and Tables 1 and 2.

USER FRIENDLY SOFTWARE

A user friendly software for use on IBM personal computers is available for licensing.

REFERENCES

1. Ally, M.R., and Braunstein, J., (1991). Process for Preparing Salt Mixtures with Predicted Crystalline Phases. Patent Pending.

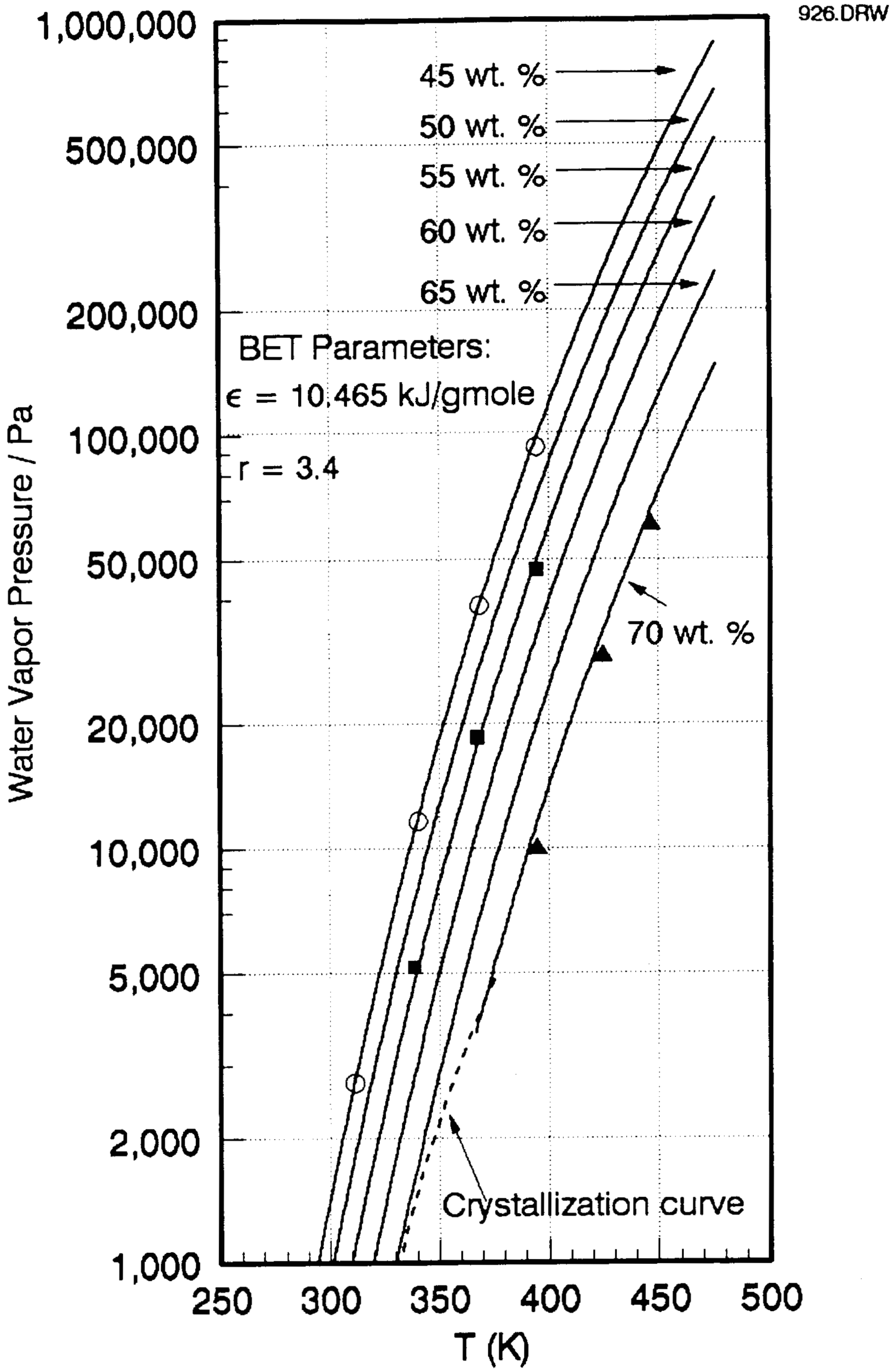


Fig. 1. Vapor pressure-composition-temperature for aqueous LiBr solutions. Solid lines indicate fitted experimental data. Symbols represent predicted values from ORNL algorithm.

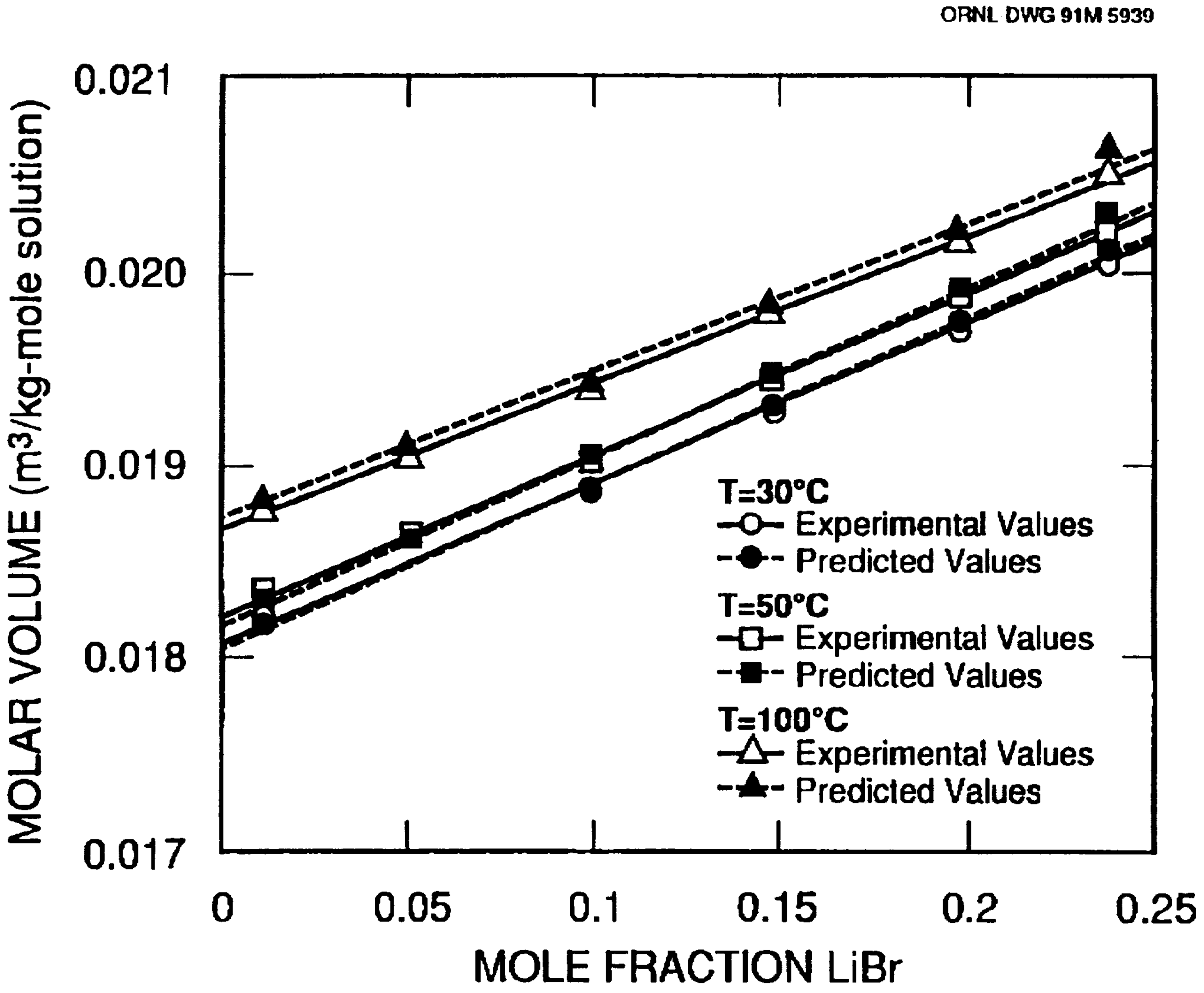


Fig. 2. Comparison of predicted and experimental molar volume-composition-temperature data for aqueous LiBr solutions.

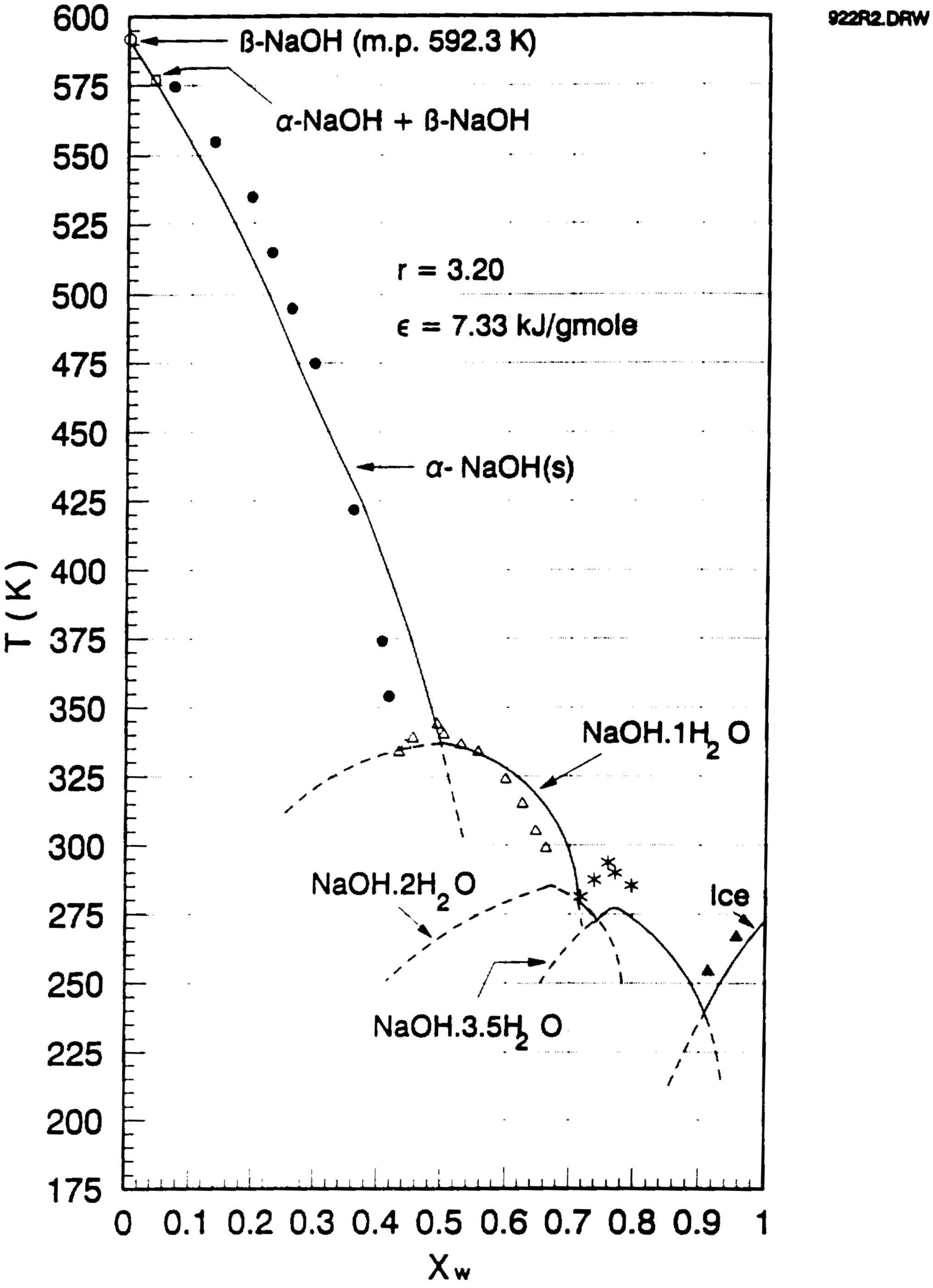


Fig. 3. Some crystalline phases in aqueous NaOH solutions. ___ Prediction of stable phases from ORNL algorithm. - - - - Prediction of metastable phases from ORNL algorithm. Symbols represent experimental data on crystalline phases.

Table 1. Comparison of experimental and predicted molar volumes of aqueous (Li,K,Na)NO_3 solutions

Parameters: $r = 2.0$, $\epsilon = 3.8$ kJ/g-mole
$\epsilon' = 20.3\text{x}10^{-4}$ kJ/kg-mole-Pa ($20.3\text{x}10^{-1}$ m^3/kg-mole)

Wt. % (Li,K,Na)NO_3	Mole Fraction (Li,K,Na)NO_3, x_s	Correlated Experimental Molar volume * m^3/kg-mole	Predicted Molar volume $\bar{V}$, m^3/kg-mole
t = 50° C (323.15 K)			
4.231	0.0100	0.01833	0.01840
18.71	0.0500	0.01905	0.01907
32.70	0.1000	0.01993	0.01993
43.56	0.1500	0.02082	0.02080
52.23	0.2000	0.02173	0.02170
59.31	0.250•	0.02266	0.02263
65.21	0.300	0.02364	0.02359
70.19	0.350	0.02468	0.02459
74.46	0.400	0.02579	0.02562
78.16	0.450	0.02699	0.02667
t = 100° C (373.15K)			
4.231	0.010	0.01906	0.01898
18.71	0.050	0.01979	0.01973
32.70	0.100	0.02071	0.02067
43.56	0.150	0.02165	0.02164
52.23	0.200	0.02259	0.02263
59.31	0.250	0.02355	0.02365
70.19	0.300	0.02452	0.02470
74.46	0.400	0.02651	0.02690

* Goodness of fit defined by (a) average absolute deviation = $(\sum | d_i |)$/Nx100%, d_i = (ρ(observed) - ρ(predicted))/ρ(observed); (b) Residual sum of squares = $\sum(\rho(\text{observed}) - \rho(\text{predicted}))^2$. Number of data points, N, between 440 and 524 for each wt. % solution (Ally et al., 1991). ϵ evaluated from mixing rules at 120°C.

Table 2. Comparison of correlated and predicted molar enthalpies of aqueous LiBr solutions

Parameters: r = 3.4, ϵ = 10.465 kJ/g-mole

t = 100 F (311 K)			
Wt. % LiBr(dry) in Solution	Mole Fraction LiBr, x_s	Correlated Molar Enthalpy kJ/kg-mole sol'n $\hat{H}$	Predicted Molar Enthalpy kJ/kg-mole sol'n $\hat{H}$
10	.02253	2.85	2.75
30	0.08165	2.13	2.51
40	0.12150	2.04	2.38
50	0.1718	2.31	2.31
60	0.2373	3.82	3.01
70	0.3262	7.07	6.13
t = 220 F (378 K)			
10	.02253	7.88	7.68
30	0.08165	6.54	7.18
40	0.12150	6.40	6.90
50	0.1718	6.70	6.71
60	0.2373	8.26	7.26
70	0.3262	11.64	9.78
t = 280 F (411 K)			
10	.02253	10.44	10.18
30	0.08165	8.78	9.54
40	0.12150	8.59	9.17
50	0.1718	8.89	8.90
60	0.2373	10.44	9.35
70	0.3262	13.89	11.57
t = 360 F (456 K)			
10	.02253	13.92	13.56
30	0.08165	11.82	12.67
40	0.12150	11.55	12.17
50	0.1718	11.84	11.75
60	0.2373	13.35	11.98
70	0.3262	16.79	13.73

AN X-RAY SCATTER APPROACH FOR NON-DESTRUCTIVE CHEMICAL ANALYSIS OF LOW ATOMIC NUMBERED ELEMENTS

H. Richard Ross
Sverdrup Technology, Inc.
NASA John C. Stennis Space Center, MS 39529

ABSTRACT

A non-destructive x-ray scatter (XRS) approach has been developed, along with a rapid atomic scatter algorithm for the detection and analysis of low atomic-numbered elements in solids, powders and liquids. The present method of energy dispersive x-ray fluorescence spectroscopy (EDXRF) makes the analysis of light elements (i.e. less than sodium; < 11) extremely difficult. Detection and measurement become progressively worse as atomic numbers become smaller, due to a competing process called "Auger Emission", which reduces fluorescent intensity, coupled with the high mass absorption coefficients exhibited by low energy x-rays, the detection and determination of low atomic-numbered elements by x-ray spectrometry is limited. However, an indirect approach based on the intensity ratio of Compton and Rayleigh scatter has been used to define light element components in alloys, plastics and other materials. This XRS technique provides qualitative and quantitative information about the overall constituents of a variety of samples.

INTRODUCTION

Qualitative and quantitative determination of the elemental content of a wide variety of samples is vital in many fields of science, technology, and industry. Energy dispersive x-ray fluorescence spectroscopy (EDXRF) is the best known rapid, sensitive, and non-destructive technique for identifying unknown elements in solids, powders, and liquids. Samples that contain high levels of light elements are not amenable to EDXRF detection. Therefore, this method is not capable of providing reliable estimates of elemental concentrations for low atomic number based materials. Corrosion and contamination products are often compounds containing low atomic numbered elements (i.e. oxides, carbonates, nitrates and hydrates), which are costly to detect and time consuming. EDXRF analysis reveals only heavy elements, therefore failing to represent the bulk composition of the contaminants.

Recently, a non-destructive x-ray scatter (XRS) approach has been developed at the NASA John C. Stennis Space Center. XRS can provide qualitative and quantitative information about the overall constituents of a variety of samples. The experimental results presented in this paper, show the versatility and importance of XRS in analyzing complex low atomic numbered materials.

EXPERIMENTAL CONDITIONS

System

A schematic of the Spectrace Corporation (Mountain View, California) EDXRF system geometry and components is shown in Figure 1. Developments in the system's hardware, the technique itself, the theory, and the performance characteristics are mentioned in the literature [1-2].

Operating Parameters

A complete list of the experimental conditions is provided in Table 1. The spectrometer used in this study was a Spectrace 440 EDXRF Analyzer equipped with a molybdenum target x-ray tube, which operates at 50W of power. All scatter measurements were made with a 0.0625 inch diameter collimator and were taken for 90 seconds per sample. In all tests a minimum of 20,000 counts were acquired for each sample, and the counting procedure was repeated four times. The percent RSD was less than 3%.

Spectrometer	Spectrace 440 Energy Dispersive Analyzer
X-ray Tube	Molybdenum Anode Target
Excitation Conditions	50 kV, 0.1 mA, Direct Excitation
Count Cycle	90 Seconds Real Time
Dead Time	48%
Collimator Diameter	0.0625 inches
Pulse Processor	Time Constant 12.5 μs
X-ray Path	Air

Table 1: Experimental Conditions

THEORY

The presence of Compton (incoherent) and Rayleigh (coherent) scatter peaks are routinely observed in x-ray fluorescence spectra and are considered a nuisance. In fact, scatter is a major source of background which limits analytical sensitivity and may cause spectral interference [3]. Figure 2 displays the output of a Rhodium x-ray tube operated at 30kV scattered from a polyester plug. The intense continuum output of the tube scattered from the polyester plug is apparent. Rhodium characteristic lines appear in the form of the Compton and Rayleigh scatter peaks. Rayleigh scatter is indicated by characteristic x-rays from the anode of the x-ray tube scattered to the detector without a change in energy, and Compton scatter is characterized by an energy loss that occurs in the sample.

When x-ray photons strike a collection of atoms, the photons may interact with electrons of the target atoms resulting in the scatter of the x-ray photons as illustrated in Figure 3. The scatter of the x-ray photons is caused mainly by outer, weakly-held electrons. If the collisions are elastic, scatter occurs with no loss of photon energy and is known as Rayleigh scatter. If the photon loses energy, causing the ejection of an electron, the scatter is inelastic and results in Compton scatter [1,4].

The energy loss associated with Compton scatter results in a predictable change in wavelength of the radiation given by Eq. 1:

$$\Delta\lambda = 0.243\ (1 - \cos\phi), \tag{1}$$

because most x-ray spectrometers have a primary beam-sample-detector angle of 90°, $\phi = 90$ and $\cos\phi = 0$. Therefore, the Compton wavelength shift, which is shown in Eq. 2, is known as the Compton Wavelength;

$$\Delta\lambda = 0.0243. \tag{2}$$

In energy dispersive systems the Compton shift may be more effectively expressed as shown in Eq. 3:

$$E = 12.396 \div 0.0243, \text{ when } \Delta E = 0.510, \tag{3}$$

where E is in kV and λ is in Angstroms. For example, a molybdenum Compton scatter signal has an observed energy peak at 16.968 kV, as shown in Eq. 4:

$$17.478\text{ kV} - 0.510\text{ kV} = 16.968\text{ kV}. \tag{4}$$

There are two important points to recognize concerning the application of x-ray scatter intensities for measuring light element contributions of a sample; a larger observed Compton scatter is seen from samples with low atomic number matrices because there is less absorption by the sample; and the ratio of Compton-to-Rayleigh scatter intensity increases as the average atomic number of the sample decreases, as shown in Figure 4 [6]. Therefore, x-ray fluorescence scatter can provide qualitative and quantitative information about a variety of samples. Most significantly, the x-ray scatter spectrum signal yields information about the overall sample constituents and exploits the sensitivity of scatter parameters by combining the estimates of light elements with EDXRF measurement of the remaining elements. The x-ray scatter system helps discriminate light elements in metal alloys, and also determines the percentage of each component by using the XRS signal to compute the Atomic Scatter Factor (ASF) as shown in Eq. 5:

$$ASF=\sum_{i}^{i=N}(wi)(Zi). \tag{5}$$

The ASF includes the following properties: the ASF of an element is the same as the atomic (Z) number of the element, and the ASF of a compound or a mixture of elements is the mass sum fraction(s) of element i(wi) multiplied by the Z number of element i, where N=number of elements. The XRS algorithms contain their own calibration curves, relating the Compton/Rayleigh scatter signal ratio to atomic number and for calculating a given chemical formula.

EXPERIMENTAL RESULTS

Corrosion and Contamination

Corrosion and contamination often represent critical problems for government agencies as well as industry. XRS analysis can effectively solve these problems, since compounds containing light elements are often associated with corrosion and contamination products. EDXRF analysis of contaminant spots on a low-alloy steel housing reveals iron as the only detectable heavy element. XRS analysis indicate that these spots have an ASF of 20.6. This information was input into the XRS program which generated the formula Fe_2O_3 as the only possibility, thus revealing the contamination to be simple rust spots and not a more complex contaminant.

The XRS identification algorithm is based upon comparison of net intensities of the selected elements in the unknown sample with those of the known reference samples. First a "Library" of references must be created using a set of known samples as shown in Table 2.

Substance	Atomic Scatter Factor	C/R
Fe	26.0	0.58
Fe_3O_4	21.0	0.89
Fe_2O_3	20.6	1.18
$FeCO_3$	16.5	1.65
$FeN_3O_9 \cdot 9H_2O$	10.1	3.71

Table 2: "Library" of Known References

After choosing the elements to be employed in the identification, each reference sample is measured long enough (usually 90-120 sec) to make the statistical counting error negligible. The net intensities of the selected elements (Z > 12) and the Compton-Rayleigh backscatter ratios are calculated and stored in the "Scatter Library" along with their standard deviations and a formula label of the reference sample.

After the library of references is complete, a measurement of the unknown sample can be performed. The net intensities (I_i) of the unknown sample are calculated and a statistic (t_s) is created for each of the possible K pairs, where K is the number of references in the library. The statistic (t_s) is shown in Eq. 6:

$$t_s=\sum_{i=1}^{N}\frac{(I_{ix}-I_{ik})^2}{\sigma^2(I_{ix})+\sigma^2(I_{ik})}, \tag{6}$$

where I_x and I_k are the net intensities of the ith element of the unknown and the kth reference, respectively; σI_{ix} and σI_{ik} are the standard deviations of these intensities; and N is the number of intensities measured. The statistic (t_s) is a squared Euclidean distance coefficient (dxk), between the unknown sample and the kth reference, weighted with the variances of measured intensities. During the actual identification, the program sequentially compares intensities of the unknown with those of the references, (See Figure 5 & 6) calculating for each pair (unknowns) it's t_s value, and selecting the smallest of them. If the smallest t_s value is greater than 90% threshold, then the second smallest t_s value is also retrieved and names of both references are displayed along with the message "possible fit". However, both t_s's must also be less than the 99.9% confidence level threshold, or the algorithm will determine that none of the references match the sample and will display "no match found".

Solvent Mixtures

The analysis of hydrocarbon components is of critical importance in the assessment of the integrity of clean systems used in the National Aeronautics and Space Administration (NASA) Space Shuttle Main Engine (SSME) Testing Program. These analyses are based on the removal of residues from hardware critical surfaces by means of flushing the surfaces with approved halogenated solvents. Infrared spectroscopy has been widely used in the detection of these impurities. However, there are limitations with this technique caused by the strong infrared absorbance band that are characteristic of these halogenated solvents. Consequently, when analyzed by infrared spectroscopy, the solvent bands are totally absorbing, which makes the detection of several hydrocarbons in the infrared spectral region impossible.

XRS spectroscopy has proven to be a significant improvement in the speed and accuracy of infrared measurements. This procedure is based on the application of hydrocarbons expressed as isopropyl alcohol (IPA) in chlorofluorocarbon (CFC) 113. Scatter rate ratios were obtained from the calibration mixtures and are tabulated in Table 3.

% CFC 113	% IPA	Molecular Wt.	% Hydrogen	C/R
100	0	187.37	0.0	0.6673
80	20	161.92	0.51	0.8584
60	40	136.47	1.2	1.1835
50	50	123.74	1.9	1.2740
40	60	111.01	2.5	1.5656
20	80	85.55	5.3	2.6341
0	100	60.10	13.42	5.1118

Table 3: Calibration data and results for hydrogen x-ray scatter analysis of IPA in CFC 113.

The calibration curve for hydrogen (IPA) determination is given in Figure 7. A low atomic number element such as hydrogen produces very strong Compton scattering. The calibration curve was linear for the range of hydrogen as hydrogen produces very strong Compton scattering. The calibration curve was linear for the range of hydrogen concentrations studied. The relative standard deviation was found to be 2%, which yields a relative error in the hydrogen determination of ± 0.15%.

Alloy Identification

XRS can also be effective in the identification of alloys containing low Z elements. An aluminum alloy was analyzed by EDXRF and XRS. The EDXRF analysis shows aluminum as the only detectable heavy element. The XRS routine measured an atomic scatter factor of 12.46, and it is known without using the ASF identification algorithm that the ASF of pure aluminum is 13.0. The measured ASF of 12.46 is obviously lower than that of pure aluminum which indicates the presence of a light element. The XRS analysis for low atomic numbered compounds revealed that the ASF of 12.46 corresponded to the material having a composition of $AlLi_{0.012}$, which corresponds to approximately 5% by weight of lithium in the aluminum.

The sample was presented as a flat sheet, therefore no sample preparation was required. The XRS analysis time was approximately 2 minutes.

DISCUSSION

The determination of light elements by EDXRF analysis is hindered by a number of difficulties. One of which is the fluorescence yield, which for atomic numbers below 16, does not exceed 0.05. When an electron transition fills a vacancy in an inner shell there is a certain probability that the emitted x-ray photon will be absorbed in the atom. That is, the emitted x-ray photon ejects an Auger (secondary) electron in one of the outer shells. Figure 8 shows a plot of fluorescence yield (ω) versus atomic number for K, L, and M x-ray lines. Auger electron production is a process that is competitive with x-ray photon emission. The Auger yield ($1-\omega$) increases with decreasing atomic number. Therefore, Auger analysis seems more promising than x-ray analysis, since the low energy photons can not escape from the sample. Typically, the Auger information comes from a surface layer approximately 10-20 Å in depth. By using the Compton-Rayleigh scatter method with molybdenum k-alpha as primary x-ray radiation, the critical depth can be as high as in 300 microns in an aluminum matrix. This critical depth is more representative of the whole sample, and smaller errors will be found when samples are not perfectly homogenous.

CONCLUSION

X-ray scatter spectroscopy has considerable usage for the detection and analysis of low atomic numbered elements from hydrogen to sodium. This technique can provide results that are comparable to conventional infrared and inductively coupled plasma analysis methods.

Covering a wide range of applications, XRS spectroscopy can be used in microelectronics and other contamination control industries; in disciplines involving natural compounds, such as geology, mineralogy, archaeology, and biology; in industries using microstructure composites; in art and artifacts fields for authentication purposes; in material verification; and in environmental problem solving and failure analysis. Obtaining both, XRS and EDXRF information from a single sample will allow the instrument to provide simultaneous chemical and mineralogical data that could be used to characterize unknown materials. Utilization of this information would be indispensable for applications in space exploration such as future missions to the moon and to Mars. This technique can be used for on-stream analysis, continuous control, improved product quality, raw materials savings, and automation of industrial processes.

ACKNOWLEDGEMENTS

The author wishes to extend appreciation to Lore' M. Gandy for her editorial assistance and critical review of this paper. Appreciation is also extended to Chris M. Bourgeois for preparing the samples and for data analysis support.

REFERENCES

(1) D. E. Leyden, *Fundamentals of X-Ray Spectrometry as Applied to Energy Dispersive Techniques* (Tracor X-ray, Mountain View, California, 1984).

(2) T. B. Johansson, R. E. Van Greiken, J. W. Nelson, and J. W. Winchester, *Anal. Chem.* **47**, 854-860 (1975).

(3) E. P. Bertin, *Principles and Practice of X-Ray Spectrometric Analysis* (2nd Ed., Plenum Publishing Corp., N. Y., New York, 1975).

(4) D. E. Leyden, "Energy Dispersive X-Ray Spectrometry", *Spectroscopy* **2**, 28-36, 1988).

(5) K. K. Nielson, "Application of Direct Peak Analysis to Energy Dispersive X-Ray Fluorescence Spectra", *X-Ray Spectrometry* **7**, 15-22 (1978).

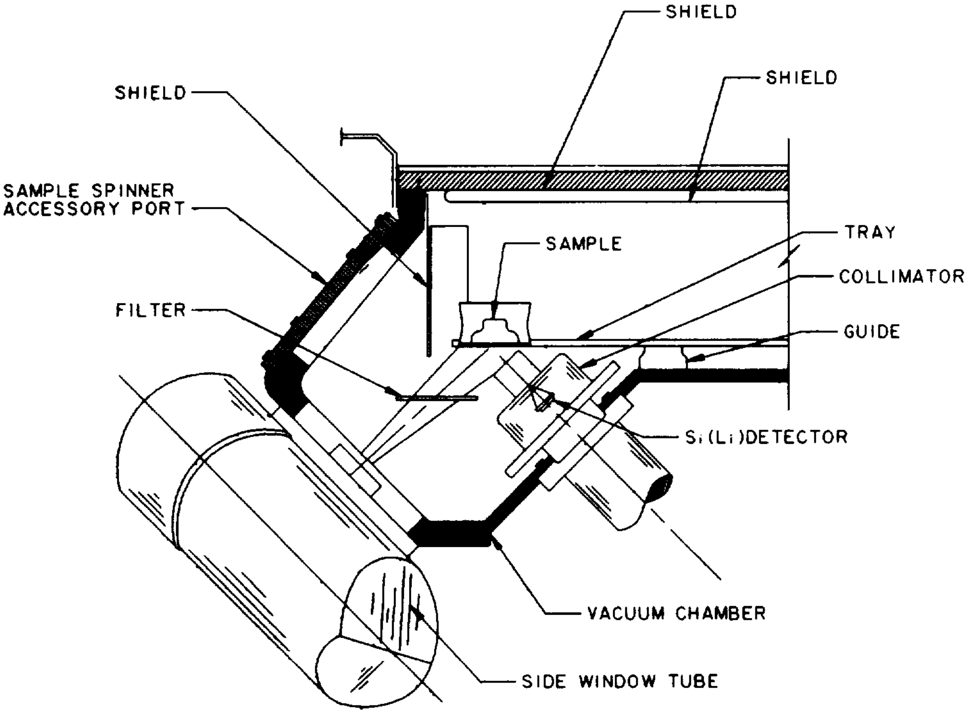


Figure 1: EDXRF System

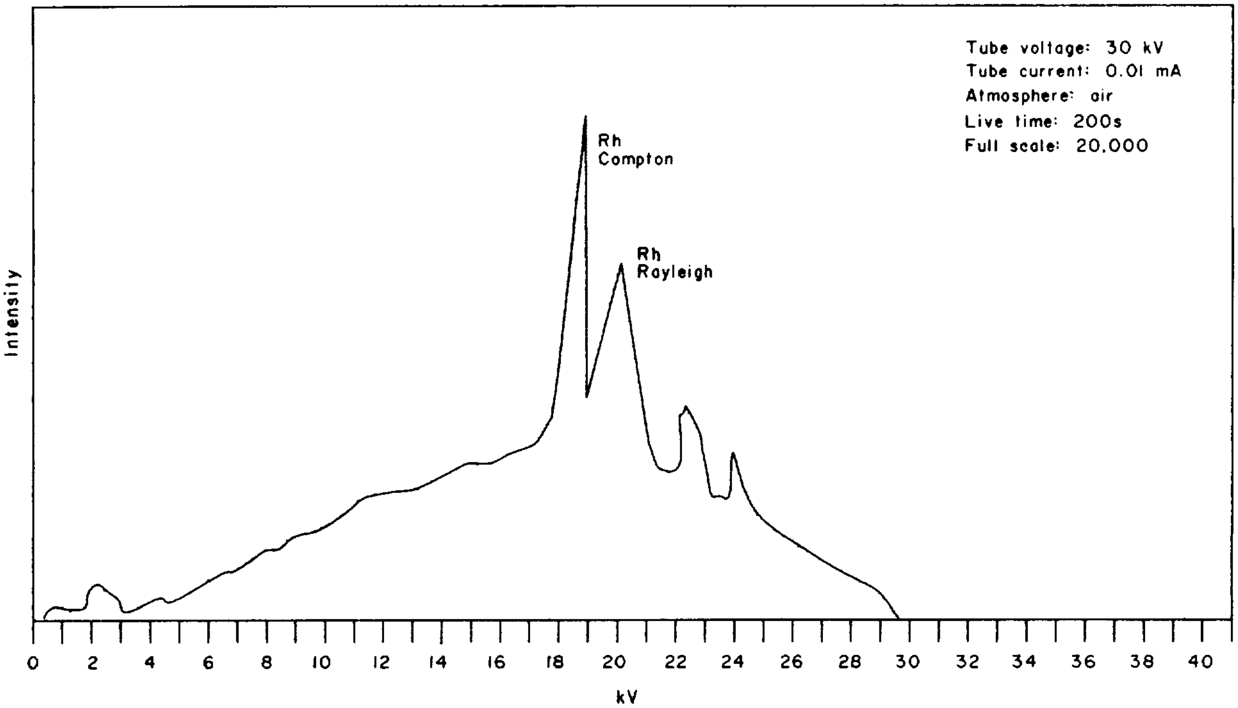


Figure 2: Output of a Rhodium x-ray tube operated at 30 kV scattered from a polyester plug.

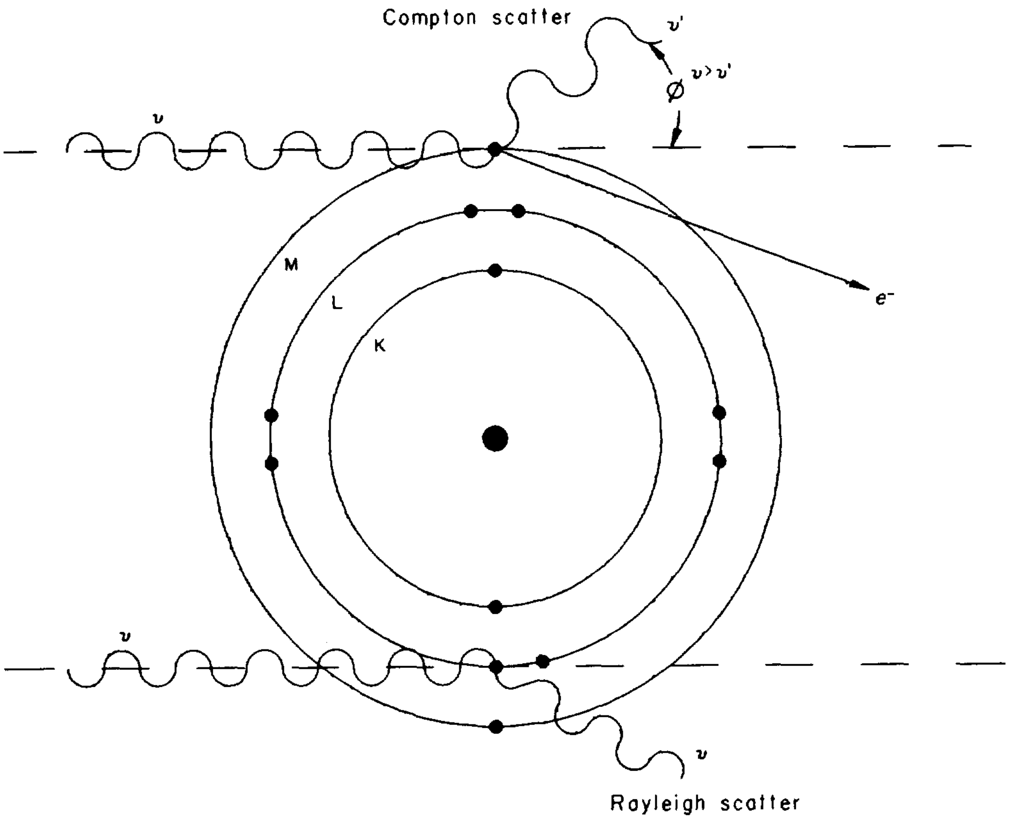


Figure 3: Illustration of Compton and Rayleigh scatter.

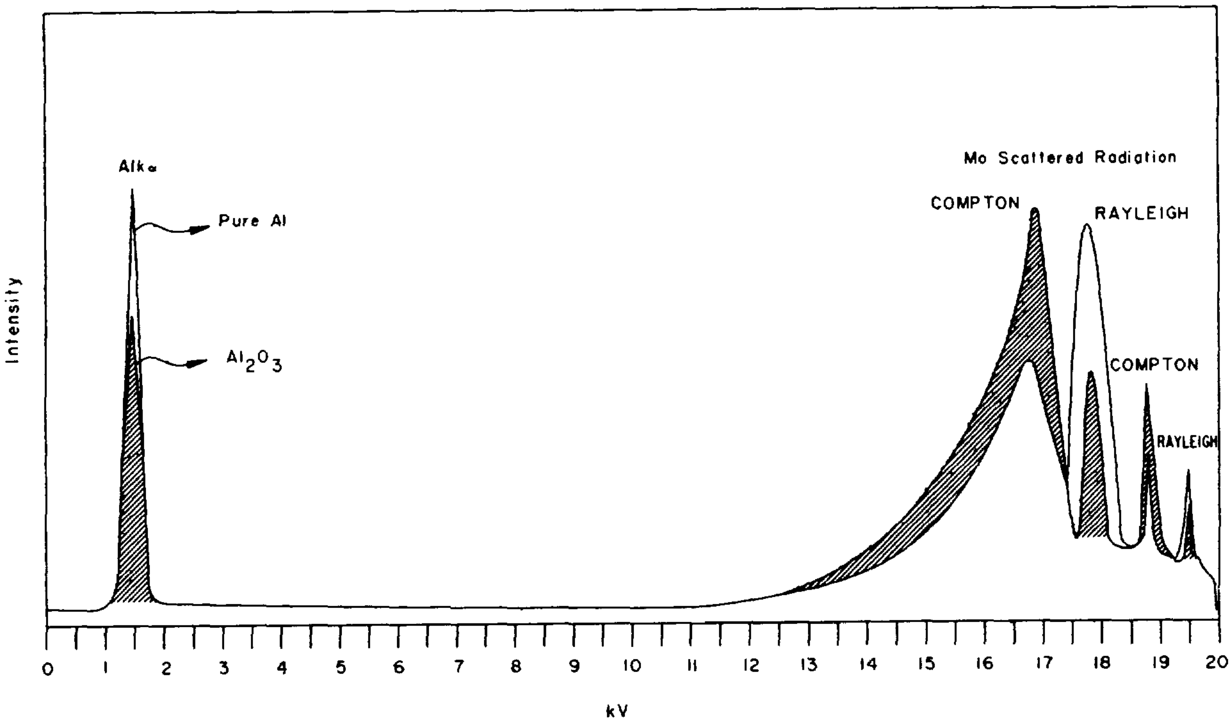
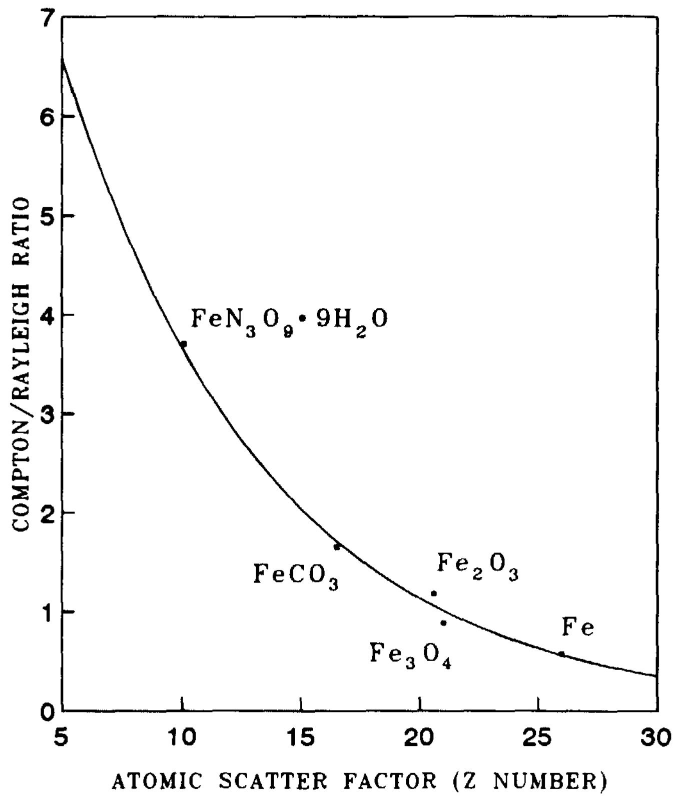
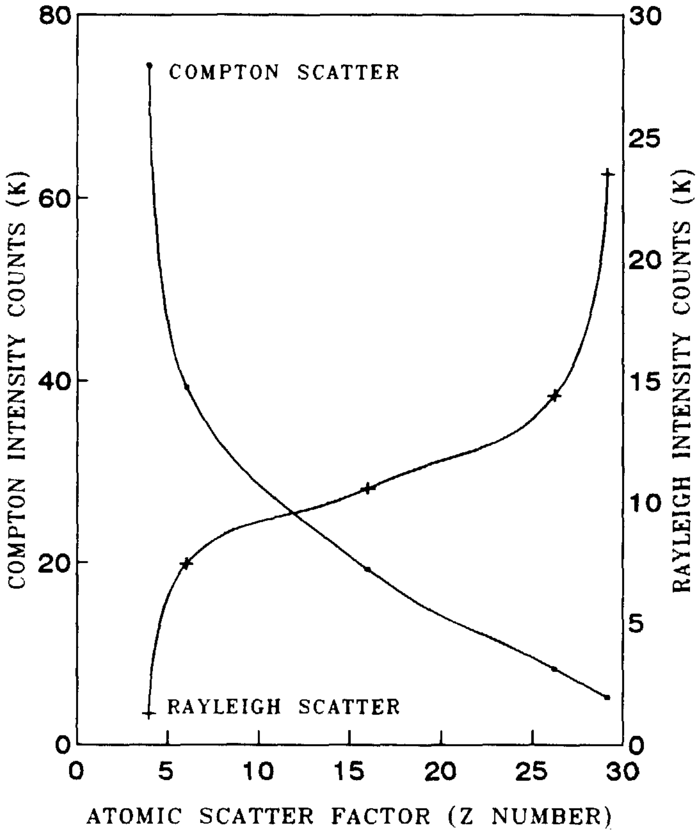


Figure 4: Partitioning of Scatter Intensities - The difference between the light element contributions (shaded area) and heavy element contributions (unshaded areas).



X-RAY SCATTER CALIBRATION CURVE FOR IRON CONTAINING COMPOUNDS

Figure 5: The calibration curve of the x-ray scatter algorithm, relating the Compton/Rayleigh scatter ratio to the ASF. The XRS measures undetected light elements in the sample matrix.



COMPTON RAYLEIGH SCATTER INTENSITIES

Figure 6: Molybdenum excitation at (50kV) as a function of sample atomic number. The Compton/Rayleigh signals are used to estimate the light element contribution to the total sample matrix.

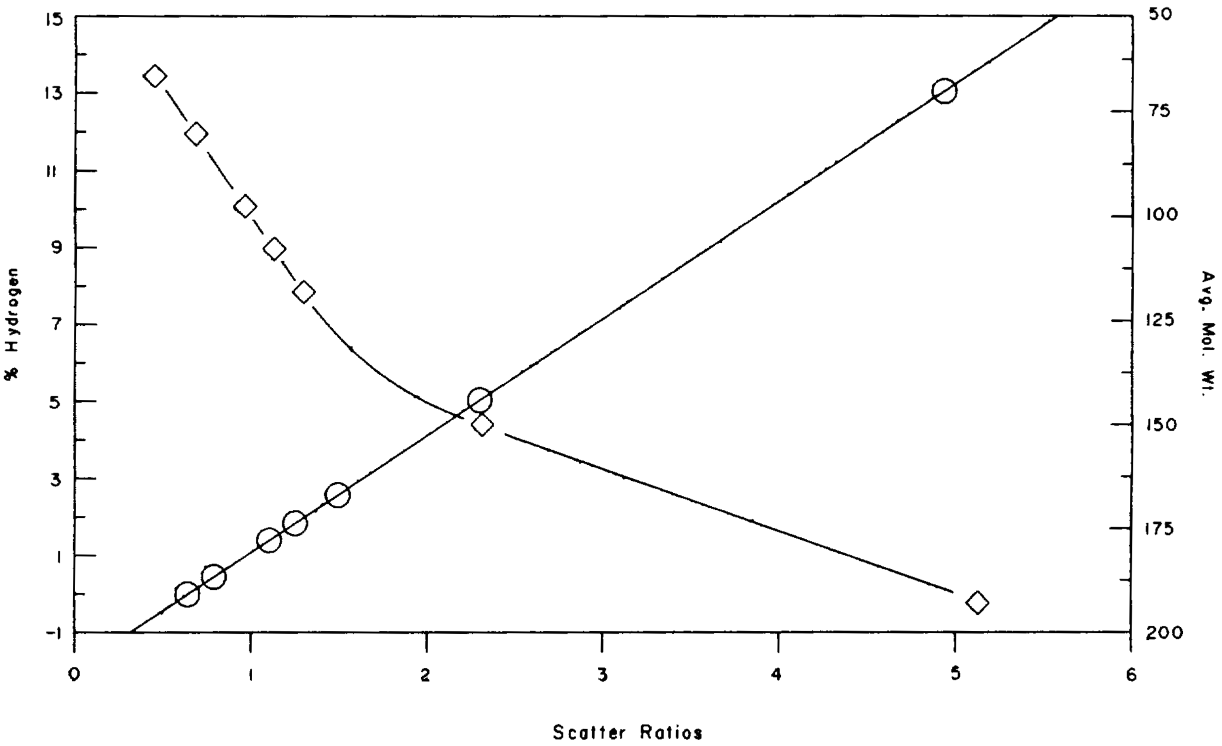


Figure 7: Scatter Ratios vs IPA/CFC 113 Mixtures.

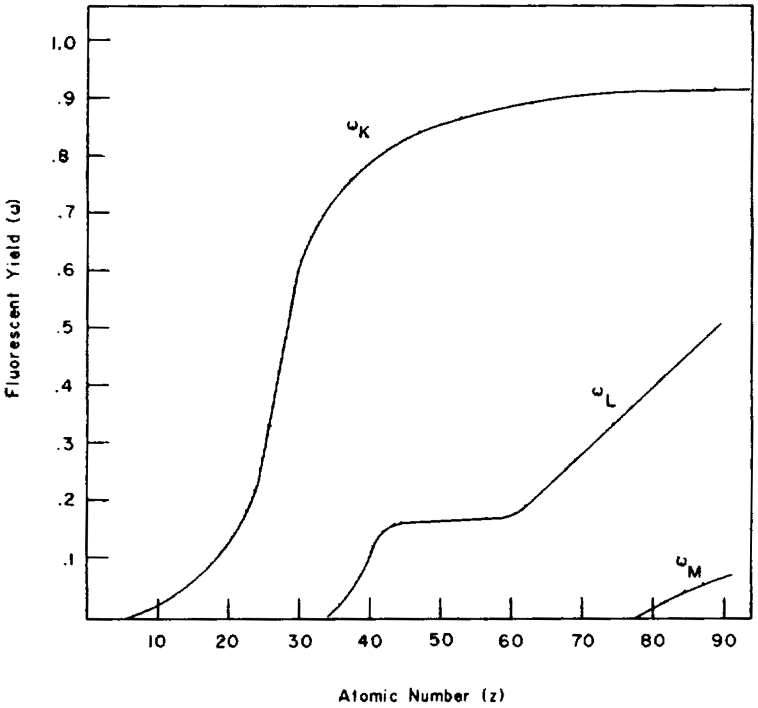


Figure 8: Variation in Fluorescent Yield with Atomic Number.

omit

ARTIFICIAL INTELLIGENCE
PART 3

AN ARTIFICIAL INTELLIGENCE-BASED STRUCTURAL HEALTH MONITORING SYSTEM FOR AGING AIRCRAFT

Joseph E. Grady
NASA Lewis Research Center
Cleveland, OH 44135

Stanley S. Tang and K.L. Chen
Structural Integrity Associates, Inc.
San Jose, CA 95118

ABSTRACT

To reduce operating expenses, airlines are now using the existing fleets of commercial aircraft well beyond their originally anticipated service lives. The repair and maintenance of these "aging aircraft" has therefore become a critical safety issue, both to the airlines and the Federal Aviation Administration.
This paper presents the results of an innovative research program to develop a structural monitoring system that will be used to evaluate the integrity of in-service aerospace structural components. Currently in the final phase of its development, this monitoring system will indicate when repair or maintenance of a damaged structural component is necessary.

INTRODUCTION

Cyclic mechanical loading causes the progressive development of damage in aircraft structures that can eventually lead to structural failure. If the initiation and development of this damage could be tracked nondestructively, the structure could be repaired or replaced prior to failure. Toward this end, a variety of non-destructive evaluation (NDE) techniques have been developed to detect damage in advanced aerospace composite materials [1-4].

An efficient alternative to these traditional NDE techniques that can be applied to in-service structural components was recently proposed [5,6]. The approach is to measure changes in *global* structural dynamics that result from damage-induced changes in the material properties [7-12]. In reference [12], a sensitive technique was developed to detect small changes in material properties of composite laminates, such as those caused by damage due to mechanical loading. Vibration of laboratory test specimens was monitored, and changes in measured vibration frequencies and damping properties were shown to result from the damage-induced microstructural changes in the composite material.

As part of a Small Business Innovation Research contract with the NASA Lewis Research Center, engineers at Structural Integrity Associates, Inc., recently demonstrated that a personal computer-based pattern recognition algorithm could be "trained," using laboratory test data, to recognize such characteristic changes in structural vibrations and to infer from those changes the type and amount of damage in a structural component. A potential application of this approach to an in-flight airframe monitoring system is shown schematically in Figure 1.

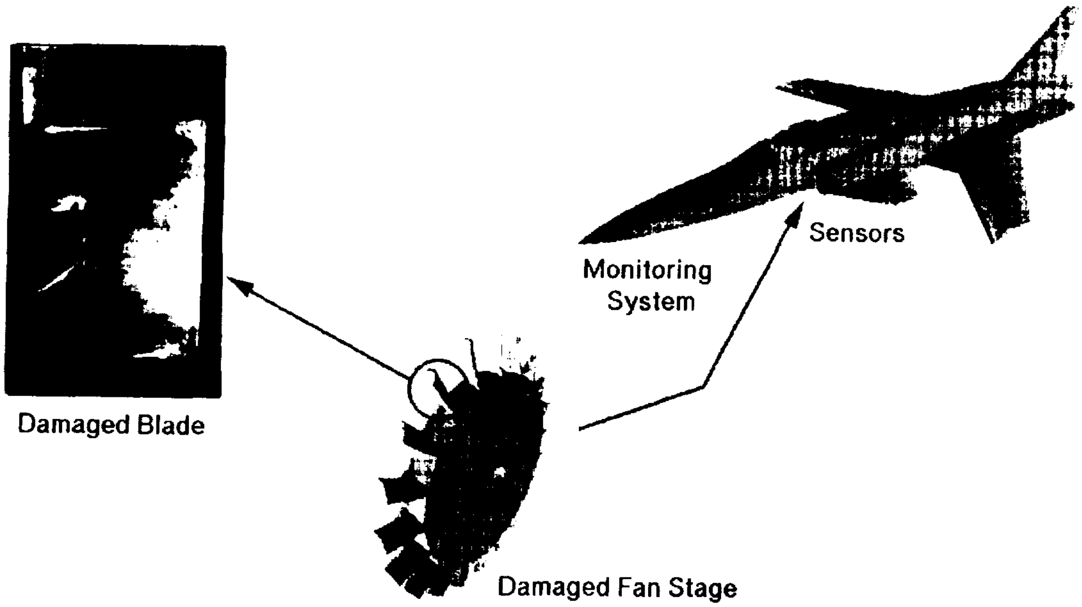


Figure 1: Conceptual Application of Structural Health Monitoring Technology to an In-Flight Airframe Monitoring System

The technology described in this paper will be used to monitor the damage development and resulting structural degradation of aging airframes that naturally occur as a result of the repeated takeoff/landing and pressurization/de-pressurization cycles that aircraft are routinely subjected to in the course of their duty cycles.

APPROACH

Vibration Testing

To evaluate the effects of ply debonding, or "delamination" on vibration measurements, a series of vibration tests of delaminated T300/934 graphite fiber/epoxy matrix composite beams was conducted. The test specimens were of dimension 5 x 0.5 x 0.04 inches, as shown Figure 2. Each specimen was 8 plies thick, and was laid up in a $[0°/90°]_{2S}$ cross-ply configuration. Ply disbonds

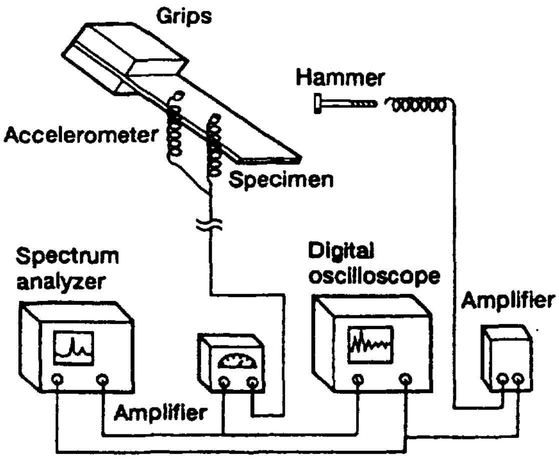


Figure 2: Experimental Apparatus for Vibration Testing

(delaminations) from one to four inches long were introduced into the material by inserting thin, non-adhesive teflon strips between selected piles of the laminates prior to curing.

The test data contained strain measurements during the first 2.5 seconds of free vibration of beams with delaminations of length 0, 1, 2, 3 and 4 inches along the midplane (neutral axis) [13]. Strain measurements were obtained from a single strain gage oriented longitudinally along the beam and located 0.5 inch from the clamped end. The strain data were digitally sampled at a rate of 800 hz (t = 1.25 msec).

Pattern Recognition

Application of pattern recognition to failure analysis and diagnostic evaluation has increased significantly during the last decade, [14]. Pattern recognition can be considered as one of the many forms in the artificial intelligence (AI) field. The mathematical approaches to pattern recognition may be divided into two general categories [14-16], namely, the syntactic (or linguistic) approach and the decision-theoretic (or statistical) approach.

The majority of the developments in the application of pattern recognition methods to failure detection and diagnostics has used the decision-theoretic approach. This is a process that is generally used to digest a vast amount of data, reduce it into a meaningful representation, and make decision on the outcome of the observation data using a classifier. The types of test data that are used by the pattern recognition algorithm to classify structural damage is shown in Figures 3 and 4.

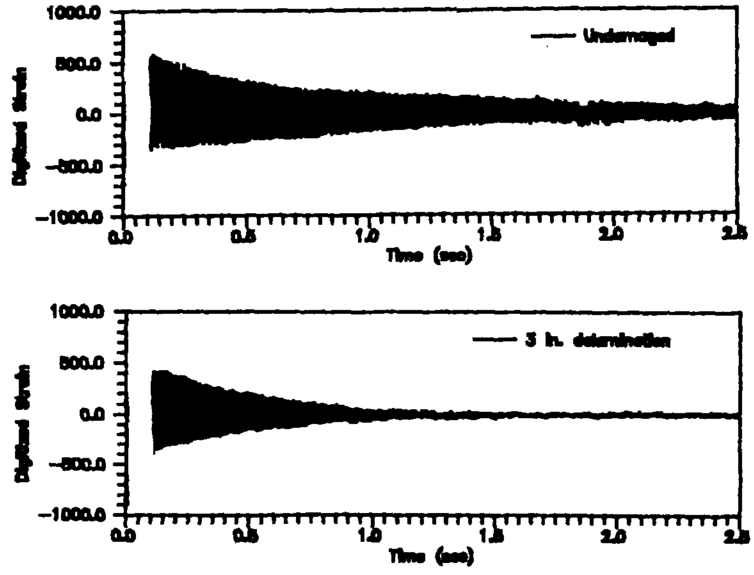


Figure 3: Strain History Measurements from a Vibrating Composite Beam Structure

Figure 3 shows two time domain measurements of the vibration response as measured by a strain gage mounted at some point on the vibrating structure. Comparison of the two signals shows how interply delamination affects the transient response. The vibration response of the damaged beam decays much more quickly due to the energy dissipation caused by friction between the delaminated surfaces. The rate at which the signal decays is dependent upon the extent of the delamination damage in the structure.

Figure 4 shows the results of similar measurements expressed in the frequency domain [17]. As damage develops, a loss in structural stiffness causes a corresponding decrease in the resonant frequencies of the structure, causing this data to shift along the x (frequency) axis. These shifts in frequencies are related to damage characteristics during the training phase. With sufficient training input, the pattern recognition algorithm can relate typical waveform characteristics (such as vibration decay times and shifts in resonant frequencies) to structural damage levels.

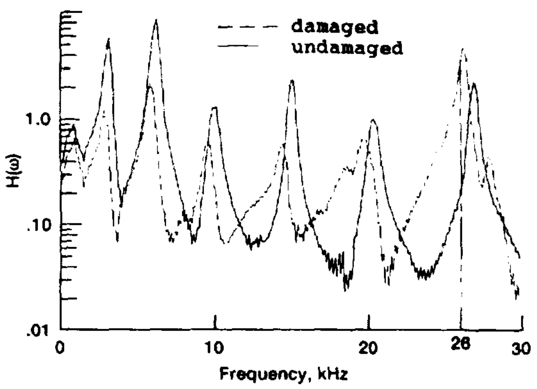


Figure 4: Structural Vibration Response in the Frequency Domain

Four fundamental steps are required to "train" the pattern recognition algorithm:

- Pattern Measurements
- Feature Extraction
- Learning
- Classification

After a set of features (e.g.; frequencies, damping properties) are calculated that characterize the pattern measurements (vibration signals), the classifier partitions the feature space into a number of regions, and associates each region with one of the known outcomes (e.g.; damage levels). Decision making ability is established through a learning process which compiles and retrieves information based on experiences where a priori knowledge of an outcome has been established.

Figure 5 presents a framework of the monitoring methodology for the material degradation of composites using pattern recognition. One key requirement of the methodology is the availability of appropriate dynamic response data of different damage levels. These measurements serve as a database to be used in the feature extraction and learning.

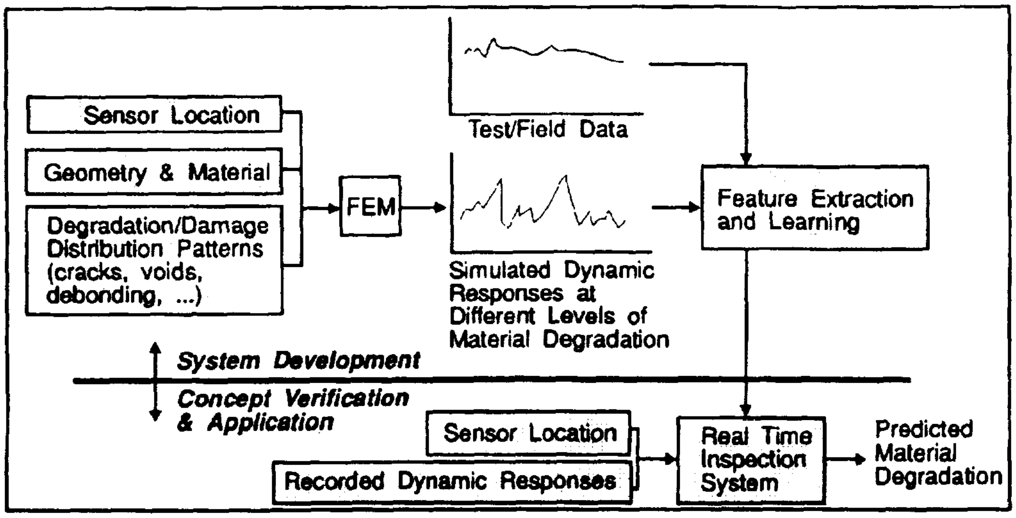

Sensor Location
Geometry & Material
Degradation/Damage Distribution Patterns (cracks, voids, debonding, ...)
FEM
Test/Field Data
Simulated Dynamic Responses at Different Levels of Material Degradation
Feature Extraction and Learning
System Development
Concept Verification & Application
Sensor Location
Recorded Dynamic Responses
Real Time Inspection System
Predicted Material Degradation

Figure 5: Application of Pattern Recognition Approach to Structural Health Monitoring

Training data can be obtained from actual operation environments of the system to be monitored, or it can be simulated from the dynamic analysis of the components or the structures. The training of the pattern recognition algorithm can be upgraded regularly as additional data with known failure status are added to the data base.

Computational Analysis

An extensive experimental database exists that shows the effect of delamination damage on vibration characteristics of composite laminates [5,6,13]. No such database exists that shows the effects of matrix cracking. Therefore, computational structural analysis was used to augment the existing experimental database to include the effects of matrix cracking on vibration behavior.
Free vibration analysis of the cantilevered composite beams was conducted using the modal superposition method available in the general purpose finite element code ANSYS [18]. Localized matrix cracking in the material was simulated by decreasing the flexural modulus over a specific region in the structure. These calculations were performed using a three-dimensional finite element model of the test specimen, with isoparametric solid elements that have orthotropic material properties. The finite element model had eight elements through the thickness and 10 elements along the longitudinal axis of the beam.

RESULTS

The initial step in applying the pattern recognition method is to conduct the system training (learning). During this phase, a priori knowledge of the correct output classification of the data for a given set of input is needed. In this case, the results of the vibration tests and finite element analyses were used as the training data. The knowledge gained during this learning process can then be used by the decision processor (classifier) to evaluate future input when the output status is unknown.

To develop this training base, the strain histories recorded from the vibration tests were characterized, both in the time domain and the frequency domain, and then correlated with the known levels of damage in the test specimens using the 71 different waveform classification features available in the TestPro monitoring system [19]. Time domain classifiers included mean, standard deviation, maximum amplitude and rise and fall time characteristics. Frequency domain classifiers included direct power spectrum features and cumulative distribution functions.

To assess the applicability of this approach to damage monitoring in composite structures, several different classification schemes were evaluated. These were the classification of

- Damage Modes
- Damage Severity
- Damage Location

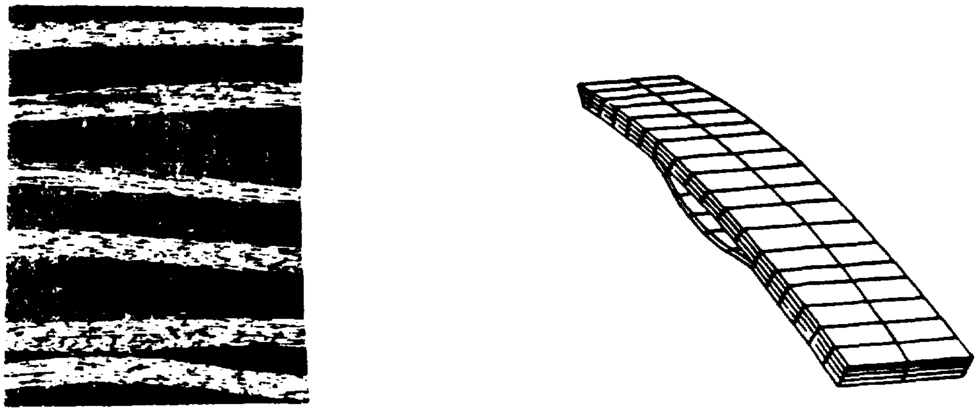
The objectives are, therefore, to train the system such that it can classify the different types of damage (damage modes) in the structure, quantify the severity of the damage, and determine the location of the damage. This section summarizes the effectiveness of the monitoring system in each of these areas.

Damage Modes

The data used to train the pattern recognition algorithm came from composite beam structures with three different categories of damage:

- Undamaged
- Localized matrix cracking
- Delamination

Each of these damage modes are depicted schematically in Figure 6.



matrix cracking

delamination

Figure 6: Typical Damage Modes in a Polymer Composite

The pattern recognition algorithm was used to identify, based on the vibration signal, the damage that exists in the structure. For the purpose of this investigation, each test specimen was assumed to be characterized by one of the three damage states listed above.

The data base was divided into two groups: Training and Analysis, Table III. The data in the training group, Table III, were put through the learning step to determine the optimum feature(s) to be used in the classifiers for damage status classification. The optimum waveform feature was determined in this manner to be "Mean Value of the Normalized Enveloped Function," a time domain feature. The enveloped function is represented graphically by a curve connecting the positive amplitude peaks of the waveform. It is therefore always positive and represents a low pass filter or integration process. The actual damage status of the structure was compared with that obtained using the pattern recognition algorithm. The results shown in Table I indicate that 98 percent of the total damage classifications were correct, using the nearest neighbor classifier.

Table 1: Monitoring System Indicates Damage Mode
with 98 Percent Accuracy

Training	Analysis	Total
27/28 *	22/22	49/50

* correct classifications / total cases analyzed

Damage Severity

After the damage mode has been identified, as described in the previous section, an evaluation of the extent of that damage can be made. To quantify the extent of localized matrix damage in the structure, the problem was again posed as a three-class problem:

- Undamaged
- Minor Damage ($E/E_0 > 0.9$)
- Major Damage ($E/E_0 < 0.9$)

Physically, a uniform degradation of the elastic moduli would represent distributed damage such as matrix cracking.

The data in the training group were assigned to the appropriate classes for the training exercise. After training, the "Mean Value of the Normalized Enveloped Function", was again determined to be the optimal discriminator for classification. Table 2 summaries the evaluation results for classification of the degree of modulus degradation. Using the nearest neighbor criteria, the pattern recognition algorithm correctly classified the level of modulus degradation in 41 of the 46 cases examined, an 89 percent average.

Table 2: Monitoring System Indicates Damage Severity with 89 Percent Accuracy

Training	Analysis	Total
22/24 *	19/22	41/46

* correct classifications / total cases analyzed

Damage Location

To conduct the system training, a two-class problem was defined, which classified the damage location as within either O inches to 3 inches or 3 inches to 5 inches of the clamped end of the cantilevered composite beam, as shown in Figure 7.

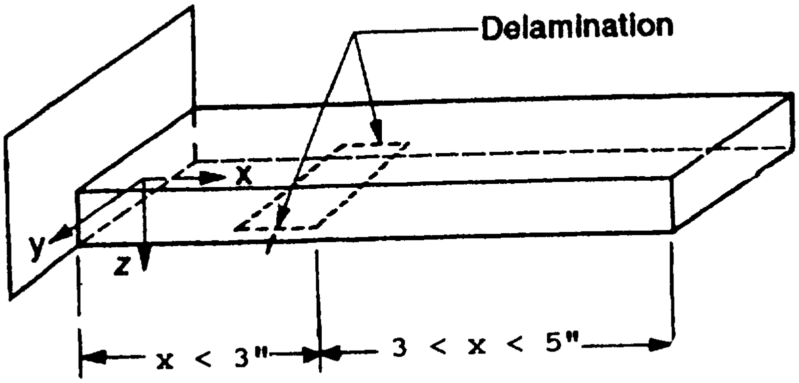


Figure 7: Classification of Damage Locations

The length of the damaged zone was not considered. The data in the analysis group has damaged zones overlapping the two defined regions.

The optimum feature was determined to be the "Difference between 50% Level and 25% Level" of the Waveform Cumulative Distribution. The results, summarized in Table 3, indicate that 80 percent of the damage locations were classified correctly by the pattern recognition algorithm using the nearest neighbor criteria classifier [17].

Table 3: Monitoring System Indicates Damage Location
with 80 Percent Accuracy

Training	Analysis	Total
22/24 *	19/22	41/46

* correct classifications / total cases analyzed

Since the primary objective of this project was to demonstrate the feasibility of using the pattern recognition approach as a means of damage detection, only a limited amount of system training was conducted. The percentage of correct classifications should improve significantly if a more extensive set of test data were used to train the system.

CONCLUSIONS

It was demonstrated that a pattern recognition algorithm can be trained to interpret structural vibration measurements in terms of damage characteristics in a composite structure. This approach can therefore be used together with a measurement system to monitor damage development in aerospace structural components. Potential applications include in-service structural monitoring, or routine material inspections for quality control applications during manufacturing. In either application, the results would provide information needed to schedule maintenance and to make decisions for repair or replacement.

Due to the success of this work, the project has recently received substantially increased funding from NASA to continue work on a Phase II program, which was awarded to Structural Integrity Associates, Inc. in August. During this two-year development program, a pattern recognition algorithm for a prototype "Structural Health Monitoring System" will be developed and demonstrated on a specific aerospace structural component.

ACKNOWLEDGEMENT

The artwork in Figure 1 was done by Mr. Matt Melis of the Structures Division at the Lewis Research Center.

REFERENCES

[1] J. Krautkramer and H. Krautkramer, Ultrasonic Testing of Materials, Springer-Verlag, New York, 1969

[2] J.H. Williams, Jr. and B. Doll, "Ultrasonic Attenuation as an Indicator of Fatigue Life of Graphite Fiber Epoxy Composite," Materials Evaluation 38, 1980, pp. 33-36

[3] S.S. Lee and J.H. Williams, Jr., "Stress Wave Attenuation in Thin Structures by Ultrasonic Through-Transmission," J. Nondestructive Evaluation 1, 1980, pp. 277-286

[4] E.G. Henneke, II, "Ultrasonic Nondestructive Evaluation of Advanced Composites," in Non-Destructive Testing of Fiber-Reinforced Plastics and Composites, Elsevier, 1990, pp. 55-160

[5] S.S. Tang and L.J. O'Brien, "A Novel Method for Fatigue Life Monitoring of Non-Airframe Components," Paper AIAA-91-1088-CP, 32nd Structures, Structural Dynamics and Materials Conference, AIAA, Baltimore, MD., April, 1991.

[6] S.S. Tang, K.L. Chen and J.E. Grady, "On the Monitoring of Degradation of Composite Materials using the Pattern Recognition Method," presented at the American Society of Composites 7th Technical Conference on Composite Materials, October, 1992

[7] Y. Weitsman, "On Wave Propagation and Energy Scattering in Materials Reinforced by Inextensible Fibers," Int. J. Solids and Structures 8, 1972, pp. 627-650

[8] A. Vary, "Concepts and Techniques for Ultrasonic Evaluation of Material Mechanical Properties," in Mechanics of Non-Destructive Testing, W.W. Stinchcomb, ed., Plenum Press, New York, 1980, pp. 123-141

[9] J.C. Duke, Jr., ed., Acousto-Ultrasonics: Theory and Application, Plenum Press, New York, 1988

[10] A. Vary, "Acousto-Ultrasonics," in Non-Destructive Testing of Fiber-Reinforced Plastics and Composites, Elsevier, 1990, pp. 55-160

[11] A. Wanner and K. Kromp, "Young's and Shear Moduli of Laminated Carbon/Carbon Composites by a Resonant Beam Method," Brittle Matrix Composites 2, Elsevier, 1989, pp. 280-289

[12] J.E. Grady and B.A. Lerch, "Evaluation of Thermomechanical Damage in Silicon Carbide/Titanium Composites," AIAA J. 29:6, 1991, pp. 992-997

[13] M.H. Shen and J.E. Grady, "Free Vibrations of Delaminated Beams," NASA TM-105582, AIAA Journal 30:5, 1992, pp. 1361-1370

[14] Fu, K S., "Application of Pattern Recognition," CRC Press, Inc., 1982.

[15] Fukunaga, K, "Introduction to Statistical Pattern Recognition," 2nd Edition, Academic Press, Inc., 1990.

[16] Bow, S. T., "Pattern Recognition, Application to Large Data Set Problems," Marcel Dekker, Inc., 1984. 12. Therrien, C. W., "Decision, Estimation and Classification, An Introduction to Pattern Recognition and Related Topics," John Wiley & Sons.

[17] G.M. Jenkins and D.G. Watts, Spectral Analysis and its Applications, Holden Day Publishing Company, 1968

[18] User Manual, ANSYS, Engineering Analysis System, Swanson Analysis System, Inc., Houston, PA., May, 1989.

[19] User Manual, TestPro, Version 2.21, Infometrics, Silver Spring, MD, May, 1991.

THE GROUND PROCESSING SCHEDULING SYSTEM

This paper was not submitted for inclusion in these proceedings. You may wish to purchase the audiocassette of this presentation by contacting:

The Technology Utilization Foundation
41 East 42nd Street, Suite 921
New York, NY 10017
Ph.: (212) 490-3999

If you would like further information on this presentation, please contact:

Michael J. Deale
Senior Engineer
Lockheed Space Operations Company
1100 Lockheed Way
LSO-459
Titusville, FL 32780
Ph.: (407) 861-5837

REACTIVE CONTROL AND REASONING ASSISTANCE FOR SCIENTIFIC LABORATORY INSTRUMENTS

David E. Thompson
Richard Levinson
Peter Robinson

NASA Ames Research Center
Mail Stop 269-2
Moffett Field, CA 94035

ABSTRACT

Scientific laboratory instruments that are involved in chemical or physical sample identification frequently require substantial human preparation, attention, and interactive control during their operation. Successful real-time analysis of incoming data that supports such interactive control requires (1) a clear recognition of variance of the data from expected results and (2) rapid diagnosis of possible alternative hypotheses which might explain the variance. Such analysis then aids in decisions about modifying the experiment protocol, as well as being a goal itself. This paper reports on a collaborative project at the NASA Ames Research Center between artificial intelligence researchers and planetary microbial ecologists. Our team is currently engaged in developing software that autonomously controls science laboratory instruments and that provides data analysis of the real-time data in support of dynamic refinement of the experiment control. The first two instruments to which this technology has been applied are a differential thermal analyzer (DTA) and a gas chromatograph (GC). Coupled together, they form a new geochemistry and microbial analysis tool that is capable of rapid identification of the organic and mineralogical constituents in soils. The thermal decomposition of the minerals and organics, and the attendant release of evolved gases, provides data about the structural and molecular chemistry of the soil samples.

INTRODUCTION

Over the past two years, researchers at NASA Ames have developed a new scientific laboratory instrument and have implemented intelligent control and analysis software to support the operations and data analysis of this new instrument. In particular, the authors, as researchers in artificial intelligence, have worked in close collaboration with two Ames microbial ecologists, Rocco Mancinelli and Lisa White, to affect this development. This paper focusses on the intelligent software technology part of that project and its potential generalization to other scientific laboratory instruments. Scientific laboratory instruments that are involved in chemical or physical sample identification frequently require substantial human preparation, attention, and interactive control during their operation. Our software is intended to alleviate the user of much of this attention and interactive control. Successful real-time analysis of incoming data that supports such interactive control requires (1) a clear recognition of variance of the data from expected results and (2) rapid diagnosis of possible alternative hypotheses which might explain the variance. Data analysis is a goal in its own right; real-time analysis, however, can support decisions about modifying the experiment protocol. Thus, the software both reactively controls science laboratory instruments and provides data analysis in support of dynamic refinement of the experiment control. The first two instruments to which this technology has been applied are a differential thermal analyzer (DTA) and a gas chromatograph (GC). Coupled together, they form a new geochemistry and microbial analysis tool that is capable of rapid, robust identification of the organic and mineralogical constituents in soils. The thermal decomposition of the minerals and organics, and the attendant release of evolved gases, provides data about the structural and molecular chemistry of the soil samples. The details of this new tool are provided in the following section.

The coupling of these analysis systems results in a more detailed characterization of the minerals and organics in the soil samples than has previously been available; their combined use has also required the development of new reasoning expertise, detailing how the data or results of the two types of analyses interrelate. This expertise has been gained through the construction of an integrated DTA-GC instrument itself, through development of the control and reasoning software in synchrony with this construction, and finally through the use of the system on soils and mixtures whose chemical decompositions provide clear examples of the interplay between thermophysical and chemical processes.

The DTA-GC software has been implemented in terms of three development levels. Level 1 represents functionality of the system as a reactive (that is, non-planning) controller. It requires the operation of the sensory perception, analysis, and control components, and the system reacts to evolved gas events by recognizing the event as an increase in oven pressure, and then exercising the GC sampling protocol. Both DTA and GC data is analyzed. At level 2, a predictive control loop is added by introducing an experiment planner, but all the components still operate sequentially. This means that in this first phase of operation (levels 1 and 2), the system is capable of controlling a single experiment run and then reasoning about the data after the run has completed. The received data is matched against encoded representations of data in mineral library records. The matches form explanations of the observed features in the data, and represent a best-guess identification of mineral and organic content. This explanation and identification can then be used to suggest follow-up experiment protocol needed to resolve any ambiguities in the identification. At level 3, all of the components can operate in parallel. This phase of implementation is a transition to operations in an interrupt mode, exploiting parallel reasoning, planning, and execution, whereby the system carries out partial matches of data with the library records while the data is "coming in" from a run. It thus allows re-programming of an experiment profile during that run, based on expectations of identification, if there are deadline limits. The status of our system with respect to each of these development levels is discussed in a later section of this paper.

The team is engaged in establishing performance criteria and evaluation standards for all the software, yielding performance metrics which can guide our extensions and help empirically determine the meaning of 'improvements' to the system. We are particularly interested in exploring the necessary trade-offs between speed of analysis and fidelity of analysis: accuracy in reporting identification versus speed and economy of the representation, noting discrimination errors. These metrics are currently being established only for the DTA-GC instrument, without consideration to generalization of the system to other instruments. The more improvements we make to affect robust control and reasoning, the more we will understand the possibilities for generalization of this software to other science instruments.

As a second prototype, it is our intention to apply the software system to a multistage bioreactor being developed at Ames during this next year. Our particular bioreactor will be used in part to evaluate the microbial paleo-environmental conditions and constraints that are implicitly represented in inhabited soil and mineral samples studied through DTA-GC. It will also be used separately to study the nutrient and environmental characteristics of natural carbon and nitrogen cycling in the Earth system. This work therefore supports NASA's interest in the role of nutrient cycles in Global Change studies, in the effects of planetary physico-chemical environments on early evolution of life, and in controlled ecological life support systems. The multistage bioreactor requires far more extensive reactive control and reasoning assistance during its operation than has been found necessary under DTA-GC, so this extension will help guide further software enhancements.

The long range commercial potential for the DTA-GC instrument itself is primarily for use as an analysis tool in laboratories (or in the field) that require rapid identification of solid samples without the need for refined wet-chemistry or scanning calorimetry. Additionally, the intelligent software developed for DTA-GC provides further commercial potential as a generic predictive/reactive control and reasoning architecture that can assist scientists in critical control and analysis decisions, and can allow for instrument operations and electronic-linked analysis under remote or hostile conditions.

BACKGROUND ON DTA AND GC, AND THE COUPLED SYSTEM

A differential thermal analyzer is an unpressurized programmable oven -- it heats up mineral or other solid samples at a controlled rate, from ambient temperature and pressure to 1200 degrees C. The heating causes the minerals to undergo chemical and structural changes. These changes include phase transformations in the mineral structure, melting, oxidation, nucleation and crystal reorganization, or simple breaking of chemical bonds and release of gases that are either physically adsorbed interstitially or are chemically bonded in the lattice structure of the particular minerals. Any organics that are contained in the sample of course undergo similar decompositions with attendant release of gas from the residue. Any substance put into the DTA oven will produce particular changes upon heating depending on its chemistry and crystal structure, thus allowing for partial identification of the substance. The temperature changes in the sample are measured against the temperature of an inert reference. The resulting difference in temperature, at ambient pressure, is proportional to the energy utilized or released in the sample during these thermophysical events. Hence, the changes in the sample are recorded by the DTA as "difference features" in the data stream. Any event which utilizes energy is *endothermic*. The sample then appears as "cooler" than the reference, and thus produces "valleys" in the data stream. Events that release energy are *exothermic* and show up as "peaks" in the data stream. The character of such a thermal event, such as its duration, intensity, onset temperature, and whether it is endo- or exothermic, is indicative of the mineral structure, proportion, and content in the sample, but it is not unambiguously diagnostic for identification. It is important to realize that DTA by itself does not provide any chemical information except from inference. Furthermore, because DTA measures only *temperature* differences, there is no direct measure of the actual heat involved in a given reaction, so no information is revealed concerning heat capacities of the minerals. Nor can one definitively measure relative proportions of different minerals contained in the sample cup, rather only their presence or absence. There is also no guarantee that presence of trace minerals will be detected unless significant amounts of these are contained in the sample to yield a signal, but of course then one does not know whether their presence is significant or only at trace levels in the parent sample. Certain variations in silicate structure, notably in clays, will not show up as differences in DTA signatures, yet this structure can be quite important since it controls the availability of lattice sites for certain ionic replacements or even preferential locations for organic compounds. DTA also provides no information on the grain size distribution of the sample or of the parent rock. All these points notwithstanding, DTA is an extremely robust thermal analysis instrument which faithfully identifies the presence or absence of diagnostic thermal events from which detailed mineral structure can be inferred, and from which much headway can be made concerning inference about processes and reaction pathways. This makes it an ideal system to be coupled with other analysis techniques, and it also may be made capable of operating in field conditions outside of a well-supplied laboratory, or even on planetary surfaces.

A gas chromatograph essentially consists of a column of material through which gas mixtures flow for purposes of constituent identification, plus a detector that quantifies the gases as they flow out of the column. When a gas mixture flows through the column, the individual gas compounds diffusively separate due to their differing affinities for the material packed in the column, and thus the compounds can be identified chemically according to their relative flow rates. This identification is at the molecular level, not elemental or ionic level; GC provides chemical information, not molecular weight information as might a mass spectrometer. The GC gives total proportional volume of gas compounds eluted through the column during one diffusion event or gas injection. By sending the sample through both polar and nonpolar columns and detectors in parallel, that are separately calibrated for particular gas compounds, the normally varying retention times can be compressed so that all the data becomes available at roughly the same time (on the order of minutes) without the various gases interferring with or masking each other. This is especially important when trace gases are sought, because the high resolution on the column necessary to detect the trace gas signatures would be swamped by even a small amount of water being eluted through the column.

The coupled DTA-GC instrument is itself a new research tool. By coupling a DTA to a GC, the scientist can determine both structural and evolved gas chemistry of a single sample. When both sources of information are combined, a more complete and less ambiguous characterization results. Typically, GCs have a pyrolitic "front-end", and rapidly heat up an entire sample. By using a programmable oven, the samples are heated slowly so that when gases are released during a thermophysical event, that release temperature is recorded and the gases are temperature- (time-) stamped according to when in the experiment run they evolved off the sample. Hence, if one observes carbon dioxide gas coming off at around 350 degrees C, then one knows it is from decomposition of organics and not from decomposition of a calcium carbonate like limestone since the limestone decomposition and its release of CO_2 occurs at near 600 degrees C. Thus, decisions can be made as to the amount and type of minerals that are present in the sample, and one can discriminate between the gases from minerals and from organics. If the sample is an unknown, then its DTA and GC "signatures" are compared by our software to characterizations in the database, along with geochemical domain knowledge and with expectations generated by the system. The system generates a set of hypotheses about what the sample contains, and it suggests and controls variations in the experiment run that will help to eliminate alternative hypotheses. Such a system can perform analysis either for target minerals and organics or for toxic compounds, and it can both verify expectations and suggest presence or absence of unanticipated species. The additional information available from the coupled DTA-GC system enhances independent DTA structural information or molecular chemistry from the GC. It also contributes to elucidation of reaction pathways and provides gas volume proportions, but not unambiguous mineral proportions. Were mineral proportions available, then one could map the DTA-GC information back to parent rock or even to geologic environment information. However, the problem remains that from knowing only presence or absence of minerals, only disjunctive possibilities of parent rocks or environments can be known. Rocks are identified not only by their chemistry but by the exact proportions and occurrence of those chemical constituents so that several vastly different kinds of rocks may still have identical chemistry. Furthermore, there is a critical sampling issue concerning whether the distribution of minerals or substance in the DTA sample cup is representative of the proportionate distribution in the parent rock. This indicates that in order to make the DTA-GC system into a functioning geologic analysis assistant, a different class of information is needed, specifically that concerning grain size of species and proportionate occurrence.

REQUIREMENTS FOR ANALYSIS AND CONTROL

The kinds of science instruments we are concerned with may be broadly classified as ones in which control decisions are made reactively, in real-time, based on incoming data. For example, we run the DTA-GC under mild vacuum so that release of gases from the sample during decomposition may be recognized by pressure sensors, thereby immediately triggering or changing the GC sampling strategy for that run. This reactive control notwithstanding, the DTA-GC currently operates best in a mode where decisions on sample identification are delayed until all relevant data has been acquired so that as much uncertainty as possible is eliminated before analysis. In this section, we discuss the capabilities required to support intelligent analysis and control.

The DTA-GC application requires *sensory perception* capabilities. We define these simply as the ability to acquire information about the external world via sensors. The system must interpret real-time DTA, GC, and pressure signals from the hardware sensors. These sensors provide results in the form of voltage streams that are typically plotted graphically and then visually interpreted by humans. Because our system operates semi-autonomously, it needs some signal processing capabilities for recognizing peak and valley features in the voltage streams. Even though it does not require graphical representation for its decisions, we provide graphical display of the data for use by the attending scientist. Furthermore, the system must address a form of limited perception since it is never certain which events will be encountered during the heating process. This uncertainty is compounded by signal/noise or figure/background discrimination tasks. For example, it may be difficult to discriminate between, or assign semantic meaning to, a single "valley" signal versus two "peaks". Thus, some heuristics are necessary to bias such decisions.

The application requires *data analysis* capabilities, which we define as any processing or reasoning over data that was acquired through sensory perception. The results of DTA-GC data analysis is a set of hypotheses that postulate mineral combinations that could be contained in the sample. When a single observed event can be explained by two different minerals because they both have events in the same temperature range, multiple hypotheses are produced. The result is a set of competing hypotheses that represent an ambiguous model of the unknown soil. Because this is the first combined "DTA-GC" system, the only experts on the analysis of this combined data are our microbial ecologists, who themselves are learning about the new system. However, experts in DTA and in GC separately often employ a variety of heuristic knowledge when they choose between alternative hypotheses or explanations. We need to model the expert's reasoning process using a high level language so that our results will make sense to these scientists. Ideally, the scientists should also be able to develop and maintain the knowledge base themselves. This need for a high-level knowledge-based representation combined with heuristic search are the typical motivations for expert system techniques. Since a given observation may not perfectly match the generalized characterization in our mineral library, the use of probabilistic techniques for assignment of matches is also needed. Further, belief revision techniques are motivated due to the system's limited perception and incremental data acquisition in an uncertain world.

This application requires *planning* capabilities, which we define as the ability to select actions by performing "look ahead" or "predictive" search. Because the constituents of the soil sample and its in situ environment are unknown, an appropriate set of experiments to validate or clarify identification cannot be designed in advance. Therefore, the system must perform on-line planning in order to design experiments based on knowledge gained. Also, since competing hypotheses will often exist, the system must take actions aimed at clarifying ambiguities. For example, consider a case in which the first sample run indicates only that gas evolved somewhere between the temperatures of 200 and 700 degrees. The data analysis results from that run might then induce two competing hypotheses: one assuming the gas was produced at 300 degrees and another assuming it happened from a different event at 600 degrees. A simple follow-up experiment on a second sample might collect gas only between 200 and 400 degrees. If the gas were again detected in that smaller interval, then the second hypothesis could be eliminated.

The use of planning techniques is further motivated by the need to contend with limited resources. In a remote planetary setting, the system might not always have enough time or soil sample for a complete second run. Therefore, the planner must reason about resources in order to choose its best experiment design strategy. For example, in the lab, a complete experiment involves heating the reference and sample up to 1200 degrees C at a rate of 10 degrees/minute, thus taking about two hours. If the system were to have only one hour in which to clarify ambiguities that occur at 1000 degrees, there would not be enough time for a complete second run. The planner could choose a strategy that uses a much faster heating rate to "skip" data collection in the first 900 degrees, stop and come to thermal equilibrium, then proceed at the desired 10 degrees/minute for data collection in the critical section. When there is not even enough time or soil for a partial second run, the planner might instead choose between strategies that seek to clarify the results by simply analyzing the data differently without requiring the hardware. In particular, it could rerun the analysis in order to (1) look for masking effects between two decomposition events that occur in similar temperature ranges, (2) clarify matches under different prior probability assignments of mineral groupings in the Bayes net, or (3) look for possible alternative assignments of endotherm/exotherm features in the data due to "single valley versus two peak" ambiguities or other figure/background assignments. These actions also represent an experiment, even though no science hardware is involved The knowledge representation used to model these strategies needs to be a high-level language so that scientists can develop such critical strategies themselves. Additionally, the language must support heuristic search techniques, and it must be procedurally expressive enough to represent the conditional and iterative control required for encoding arbitrarily complex strategies. One additional point: recall that the planner designs experiments based on the results of data analysis, which often contain competing hypotheses. However, those hypotheses may change at any time as unexpected exothermic, endothermic, or gas-release events are observed. Thus, the planner must operate in an uncertain and changing environment. In order to plan appropriate experiments in a chan͡ging world, the planner must be able to incorporate asynchronous sensor reports into its search process.

Finally, this application requires *real-time control* capabilities, which we define as the ability to take actions in bounded time. Our system must perform real-time control in order to react to unexpected thermal events, and to capture gas produced while heating the sample. Although the system cannot be certain in advance whether these events will occur or when, it must respond within seconds of their detection. If the planner cannot produce a plan within the available time, the controller must still operate with some intelligence. Thus, it must be able to generate experiments reactively by instantiating a design strategy according to heuristics that do not involve time consuming look-ahead search.

In summary, DTA-GC needs to combine a mineralogical expert system with integrated sensory perception, probabilistic data analysis, planning, and control. The next section describes our architecture and its components in terms of the software engineering and artificial intelligence techniques we have applied to these requirements.

THE DTA-GC SOFTWARE ARCHITECTURE

A simplified view of our software architecture is illustrated in Figure 1. It consists of three elements: a hardware relay, an analysis component, and a control component. The 'hardware relay' is responsible for sending effector commands to the hardware, and for receiving sensor reports from the hardware. The 'analysis' components provide the sensory perception capabilities that acquire information via hardware sensors, and the data analysis capabilities which reason about the sensory data. The 'control' components provide both the experiment planning and the real-time control capabilities. The software is written in LISP and C, and operates on a Sparc 2 Sun Workstation. The system accepts scientific goals and a time limit as input, includes both reactive and predictive control loops, and produces analytical results. The reactive control loop, indicated by the solid arrows in the figure, selects actions in bounded time by matching sensor readings against condition-action "reflex" rules. The predictive control loop, indicated by the dashed arrows, involves sending the analysis results to the experiment planner. The planner searches through a space of experiment design procedures either for a useful follow-up experiment, for modifications to the current experiment, or for analyzing the data differently. A successful search produces a new experiment in the form of condition-action rules to be passed to the experiment controller. We now briefly discuss each of the five software components in Figure 1, and the techniques we have used to address the requirements described in the previous section.

The job of the *hardware relay* is to receive sensor readings and transmit effector commands to the hardware. The DTA-GC hardware includes a programmable DTA oven, two GC columns and detectors, two pressure sensors, and four valves which control the gas flow between the DTA and the GC. The hardware relay currently receives nine real-time data streams from the hardware sensors, and it can transmit over 100 distinct effector commands to the hardware. All of the these instruments communicate with our Sparc 2 through a General Purpose Instrument Bus (GPIB), the IEEE-488 standard for byte serial, bit parallel interface. To facilitate this communication, we have developed a general LISP/GPIB interface written in C.

The job of the *sensory perception* component is to identify the qualitative features in the DT, pressure, and GC signals. We use a "Scale Space Filtering" technique originally developed by Witkin [5,6] for use in image processing domains. This technique detects peaks and valleys in a curve by convolving Gaussian filters of varying standard deviation with the input signal. As the size of the filter increases, the convolved signal becomes increasingly smoothed. Hence, the points of inflection that remain after applying the largest filters correspond to the most prominent variations in the input signal. Points of inflection at varying filter scales are then grouped into scale-space

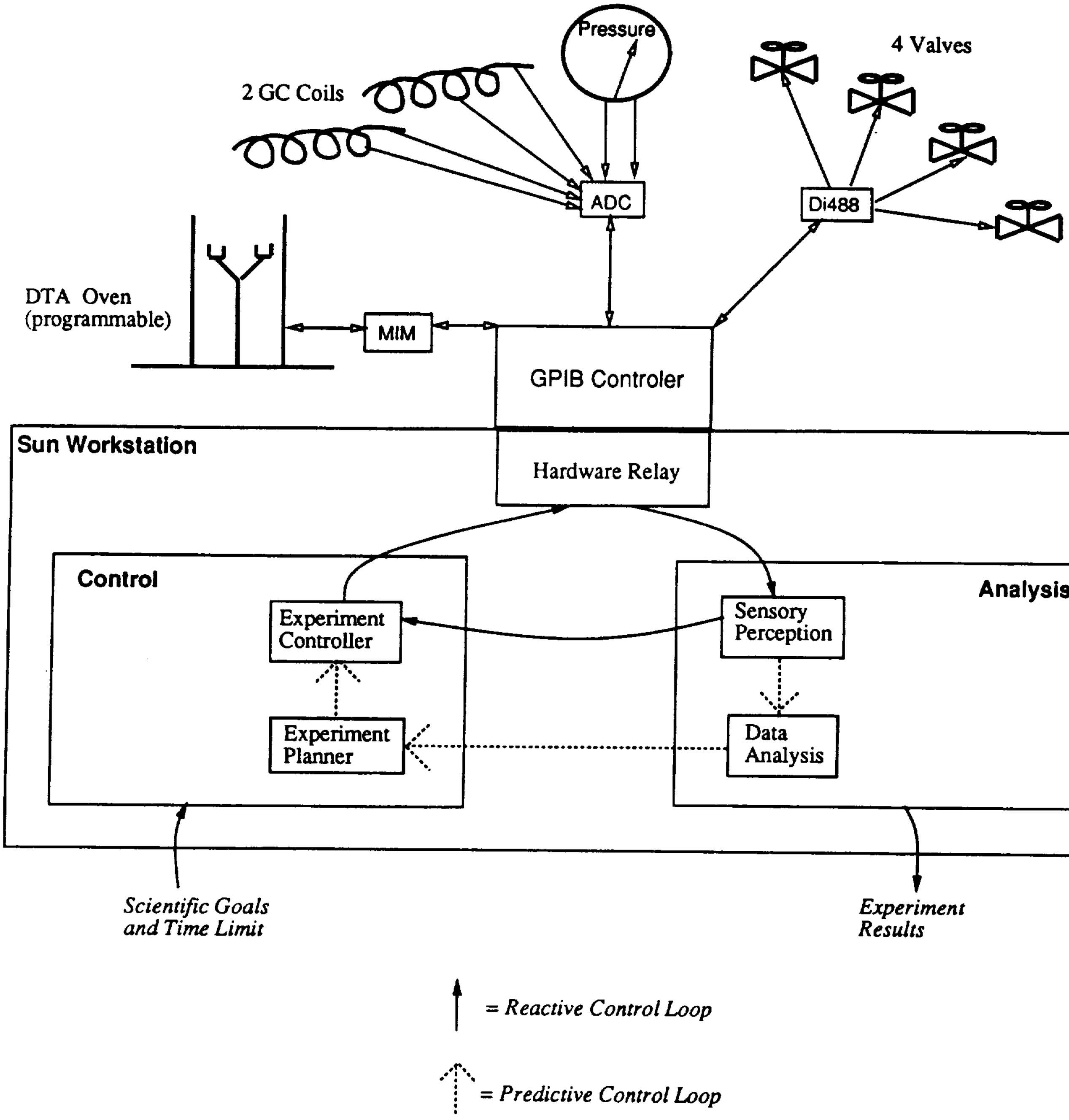


FIGURE 1. The DTA-GC Architecture

contours. The first derivative of the signal and its trend is used to determine whether a given contour group is a peak or a valley; it also aids in determining a degree of belief associated with the contour according to the probability that a feature observed really is a thermophysical event. This belief attribute helps to address the inherent perceptual uncertainty in our domain generated by signal/noise or figure/background discrimination issues, as well as the use of a sparse set of Gaussian filters. See [3] for a more complete description of our sensory perception component.

In the DTA-GC system, *data analysis* corresponds to generating hypotheses that postulate mineral combinations contained in the soil sample. We generate hypotheses through a two step method: Bayesian classification and heuristic search.

The classifier uses a Bayes tree to probabilistically match observations against events associated with known minerals in its library. The library contains knowledge of thermal and gas evolution events for over 30 classes of minerals including clays, carbonates, and salts. The classifier defines a Bayes tree for each mineral. Each child of a root mineral node defines a process node such as 'phase transition' or 'chemical reaction' which is produced by heating the minerals. Each of these process nodes has a terminal child node which corresponds to specific mineral decomposition events. These mineral event nodes test observations for membership in a class of endotherm, exotherm, or gas events that occur within a given temperature range. The classifier uses the probabilities generated during sensory perception to assign probabilities to the terminal nodes in the Bayes trees. Using the conditional probability links from mineral nodes to process nodes and from process nodes to mineral-event nodes, a standard Bayes tree propagation algorithm [4] is used to deduce the probabilities at all non-terminal nodes. The minerals are then ranked according to their associated degrees of belief. Here the belief attribute helps to address domain uncertainty by indicating the probability that the observation really is an instance of mineral decomposition event. Two issues arise with the output of the classifier. First, since the mineral events in our library may overlap in temperature range, the classifier may match a single observation to multiple mineral events, thus increasing the belief in both minerals based on the same piece of evidence. For example, both types of clays, montmorillonite and kaolinite, will match a single observed exotherm at 1000 °C. Second, each mineral model may only account for a subset of the total observations. Thus, another procedure is required to provide global explanations for the entire set of observations. In order to address these two issues, the classifier output is passed to an explainer that has the job of constructing systematic explanations for the set of observations as a whole.

The explainer is a general purpose inference engine that uses the local matches provided by the classifier to construct explanations or hypotheses for the set of observations as a whole. Each explanation contains a set of distinct mappings from each observation to a unique mineral decomposition event. This is done by reasoning about the matches provided by the classifier. The classifier can match a single observation to two different mineral events, or it can match a single mineral event to two different observations. Each of these cases produces disjunctive explanations. Thus, in our above example, one explanation will match the exotherm to the kaolinite decomposition event while another explanation matches it to the montmorillonite decomposition event. More disjunction is introduced to model cases where an observation is left unexplained. The explainer searches through this space of alternative explanations with the aid of a heuristic control function that combines multiple scoring dimensions. This heuristic is a form of Occam's Razor which prefers explanations that minimize the number of minerals used, the number of unmatched observations, and the number of unobserved events, while maximizing the combined probabilistic beliefs of the observations and the mineral events. The explainer currently uses two very simple hypothesis generation rules. The first rule defines a search space that matches each set of observations to a distinct set of classifications. The second rule completes the search space by allowing observations to remain unexplained. Even for simple examples, these rules can produce many distinct explanations. This ability to automatically and systematically construct and evaluate so many alternative, yet viable, explanations can provide a benefit to the human expert who may not be so rigorous in exploring alternatives. Our system includes closed loop control, which enables the system to design and perform its own experiments. The primary output of *data analysis* is a set of explanations, termed the 'result'.

The integration of *planning* and *control* components in this architecture is based on Drummond's Entropy Reduction Engine (ERE) [1,2]. We chose the ERE approach because it has the benefit that the controller operates independently from the planner so that real-time control is not dependent on the more expensive search behavior of the planner. Our system differs from ERE primarily in the style of search used by the planner component. Our planner generates a task decomposition space, whereas their planner generates a state-space search.

The experiment *controller* is a rule-based system that matches sensory enablement conditions to GPIB effector commands. Its job is to control the laboratory equipment in real-time according to a set of Experiment Control Rules (ECRs) that are either provided by the scientist or synthesized by the experiment planner. Our controller is based on the "Reactor" and "Situated Control Rule (SCR)" elements of the ERE architecture. Under this approach, the controller operates in a perpetual sense-act cycle, executing rules that function as quick reflexes to provide the reactive control capabilities of the system. In the DTA-GC system, the controller must be able to react to unexpected thermal and gas events within seconds of their detection in order to properly analyze them.

Although many types of low-level commands can be sent to the DTA-GC instrument, we have defined three abstract operations that characterize our required experiment control behavior. These commands are 'record', 'skip', and 'sniff'. 'Record' causes the oven to heat up at the regular rate of 10 degrees/minute, during which time data is collected. 'Skip' causes the oven to heat up quickly, during which time data is not collected. 'Sniff' causes gas to be passed to the GCs for analysis, and then reconfigure the valve system for acquiring the next event.

The job of the experiment *planner* is to produce an experiment that clarifies the ambiguous results of a current or a previous run. A 'clear result' contains only one explanation that explains all observations; this rarely occurs. More often, the result contains multiple explanations that use different minerals to explain the same observation. Additionally, the result often contains observations that cannot be explained, and events that were expected but not observed. These cases represent three distinct forms of ambiguity. The planner searches through a task decomposition space to generate a set of Experiment Control Rules (ECRs) that might clarify the given ambiguities. First, the experiment planner selects which ambiguities to clarify using heuristics that consider ambiguity type and resource availability. The planner then chooses among hypotheses that postulate experimental, chemical, sensory, or modelling causes for each ambiguity. Next the planner selects a strategy for proving the hypotheses. General strategies include designing a second run that skips uninteresting temperature intervals, modifying the current run, or modifying the data analysis procedure alone. Lower-level strategies produce specific ECRs by selecting specific temperature intervals for 'skipping', 'recording', or 'sniffing'. Experiment plans that do not violate resource constraints are passed to the controller.

The planner is implemented in *Propel*, a general-purpose language that we have designed to be procedurally expressive enough to represent real-world procedures, while maintaining the benefits of heuristic search. *Propel* procedures allow subgoals and other choice points to be embedded within the conditional and iterative control constructs of a LISP-like language. These procedures are used to represent our experiment design strategies. The *Propel* interpreter generates disjunctive experiment plans by heuristically searching through the task-decomposition space that is defined by these strategies. Even though *Propel* was primarily designed for search, our system performs closed loop control by actually executing the experiments it designs, and analyzing the results.

To address our deadline management requirements, the planner must ensure that results are returned within the given time limit. The planner first estimates the available computation time by subtracting an initial estimate of required execution time from the given time limit. During simultaneous planning and execution, this estimate of computation time is adjusted according to the projected durations of developing plans. If a plan is found within the available computation time, then it is passed to the controller for execution. Otherwise, the controller could begin execution of a default experiment, or it could reactively instantiate an experiment design strategy. This is facilitated by the *Propel* strategy representation which can be instantiated in bounded time using predetermined heuristics. This type of action representation, which can be used by both the planner and the controller, allows for a tighter integration between planning and execution.

Since the planner must operate in a changing environment, we developed a mechanism called 'dynamic dependencies' that integrates asynchronous perception and analysis into the planner's search process. With our mechanism, the planner performs dependency analysis on the projection paths to identify external conditions on which its plans rely. The analysis component is informed about these plan assumptions so that it can notify the planner as soon as their status changes. The planner can then adjust its search control to favor plans that are based on new beliefs instead of continuing to develop plans that are based on obsolete assumptions. This technique allows DTA-GC to break the typical planning system assumption that the world does not change during the planning process. This "static world assumption" does not hold when the system is planning changes to the current experiment. Performing dependency analysis on our procedurally expressive experiment strategies is a difficult task.

STATUS AND TRANSFER PROSPECTS

Much time has been spent building the coupled DTA-GC instrument hardware itself and our LISP/GPIB interface to it. We have also focussed extensively on building up the mineral library used by the Bayes classifier by running the DTA-GC on known samples; further work has concentrated on the sensory perception and explainer components, and on the development of *Propel*, the experiment method language, and its reactive dependency mechanism. The status of our system can be presented in terms of the three development levels described at the outset; work has progressed at all levels.

In April 1992, the first level of functionality for the reactive control loop of DTA-GC was successfully demonstrated and turned over to the scientists. Since then, the mineral library and classifier has been enhanced to include characterizations of over 30 classes of minerals, and the experiment control language has been greatly expanded. Currently, the system can execute default Experiment Control Rules which heat a sample slowly while monitoring the incoming DTA, GC, and pressure data. If the pressure in the oven reaches an assigned threshold, our system automatically reacts by evacuating the gas into the GC for analysis, and then it prepares for the next gas event. The operation of each component at this level, sensory perception, data analysis, and experiment controller, functions well. The sensory perception component is implemented, but we are exploring alternative methods, especially to identify the *onset* of DTA events rather than the *peak* of DTA events. Even though "peaks" are easier to identify, because DTA peak amplitudes may shift due to the amount of material present, we must focus on onset temperatures of events. This of course allows us to map our events directly to the traditional melting point or phase change literature on minerals. Our LISP/GPIB interface and hardware relay currently forwards all sensor data to signal processing, but it will soon perform data filtering so that only "significant" sensor data is relayed for evaluation. The data analysis component has been implemented and produces explanations, but the rules and heuristics it uses need to be tuned through additional knowledge engineering efforts. Capturing this knowledge is necessarily slow since no one has previously performed computer analysis of DTA data, let alone fusion of that data with asynchronous GC data. We intend to continue addressing issues of identifying, representing, and modelling thermophysical interactions between decomposing mineral combinations that tend to obscure data and hence confuse the classifier. Finally, the experiment controller has been implemented. We have demonstrated the ability to react to detected gas events to within one second. Since the controller is a rule-system, it has been straightforward to implement. However, the current default Experiment Control Rules have turned out to be rather brittle and as yet provide little coverage for unexpected events. Thus, we will be developing a more robust set of default ECRs through knowledge engineering efforts as we learn more about the system through our two collaborators' usage.

The second level primarily has involved the introduction of the experiment planner component and the development of better modelling and heuristic control techniques for data analysis. This level consists of *serial* predictive control. At this level, the planner can suggest follow-up runs that could produce better explanations. The experiment planner has been prototyped but needs further development. In particular, the *Propel* language for representing and searching through experiment strategies is implemented, but the knowledge engineering of these strategies has just begun. Since the DTA-GC is a new instrument, there are no existing strategies, and our experts will first have to develop them. At this level, we also introduce deadline limits into the problem. The deadline management mechanism has been partially designed but has not been completely implemented.

At the third level, termed *parallel* predictive control, all components operate in parallel, and this is the phase in which the dynamic dependency mechanism is required. The dynamic dependency mechanism is not fully implemented, but development has begun. We are currently converting the original ERE state-space search approach to work for *Propel*'s task-decomposition space.

We feel that our work on the DTA-GC system will yield several self-contained and general technological components that could transfer easily to other applications. Our general architecture, characterized by dual reactive and predictive control loops, can be applied to any scientific instrument that requires real-time control in conjunction with autonomous design of experiments that clarify previous results. Our LISP/GPIB interface is also a general tool that can be used by any LISP-based system to communicate with any of the more than 4000 instruments that use the GPIB protocol. On the data analysis side, the scale-space filtering, Bayesian classification, and development of disjunctive explanations are general techniques that could easily be instantiated for other applications. Our contribution has been primarily in the linking of these capabilities and producing code to affect unified data analysis. We have implemented these techniques as linked, general tools that could be instantiated for other applications. Such an architecture could also allow for more model-based analysis. The explainer itself is a standard production rule system that uses domain specific rules. The *Propel* language, which we use to represent and interpret the experiment design strategies, is specifically designed to be a general tool that can be transferred to many applications. *Propel* can be used by any application that requires control procedures to be represented in a heuristic search framework. *Propel* procedures can be reasoned with by a planner, and also executed directly by the controller. This allows the controller to execute procedures rather than simple if-then rules. This is a feature that can be used by a variety of real-time applications where execution of a control procedure may have to begin before it has been completely instantiated by the planner.

CONCLUSION

We have described an architecture designed to autonomously control a new geochemistry instrument. The system functions as an instance of a general class of autonomous scientific instruments, that integrate sensory perception, data analysis, experiment planning, and experiment control. We have described how these components function and how they interact to provide autonomous control of the DTA-GC instrument. The architecture itself is now being used as we extend the system to other instruments. The system we have described represents a synergy between AI applications and AI techniques. The DTA-GC application has stimulated the development of techniques for the integration of perception, planning, and control, which in turn allow us to tackle new real-world applications that are even more ambitious.

ACKNOWLEDGEMENTS

We gratefully acknowledge the extensive collaboration of our microbial ecologists Rocco Mancinelli and Lisa White. We also appreciate the efforts of Deepak Kulkarni who developed the structure of our classifier, and Kiriakos Kutulakos who aided Deepak in the development of the original sensory perception component. We also thank TA Instruments from whom we received (through proprietary agreement) the DuPont DTA internal command protocol that we used as a base for building our controller code.

REFERENCES

[1] Bresina, J., and M. Drummond (1990) Integrating Planning and Reaction: A Preliminary Report. Proc. AAAI Spring Symposium Series. Stanford University, CA.

[2] Drummond, M., and J. Bresina (1990) Anytime Synthetic Projection: Maximizing the Probability of Goal Satisfaction. Proc. 8th AAAI, Boston, MA.

[3] Kulkarni, D., K. Kutulakos, and P. Robinson (1992) Data Analysis using Scale-Space Filtering and Bayesian Probabilistic Reasoning. Computers and Chemistry, vol.16, no.1.

[4] Pearl, J. (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publ.

[5] Witkin, A. (1983) Scale-Space Filtering. Proc. 8th Int. Joint Conf. Artificial Intelligence, Karlsruhe, Germany.

[6] Witkin, A. (1987) Scale-Space Methods; in Encyclopedia of Artificial Intelligence. Edited by S. Shapiro. Wiley.

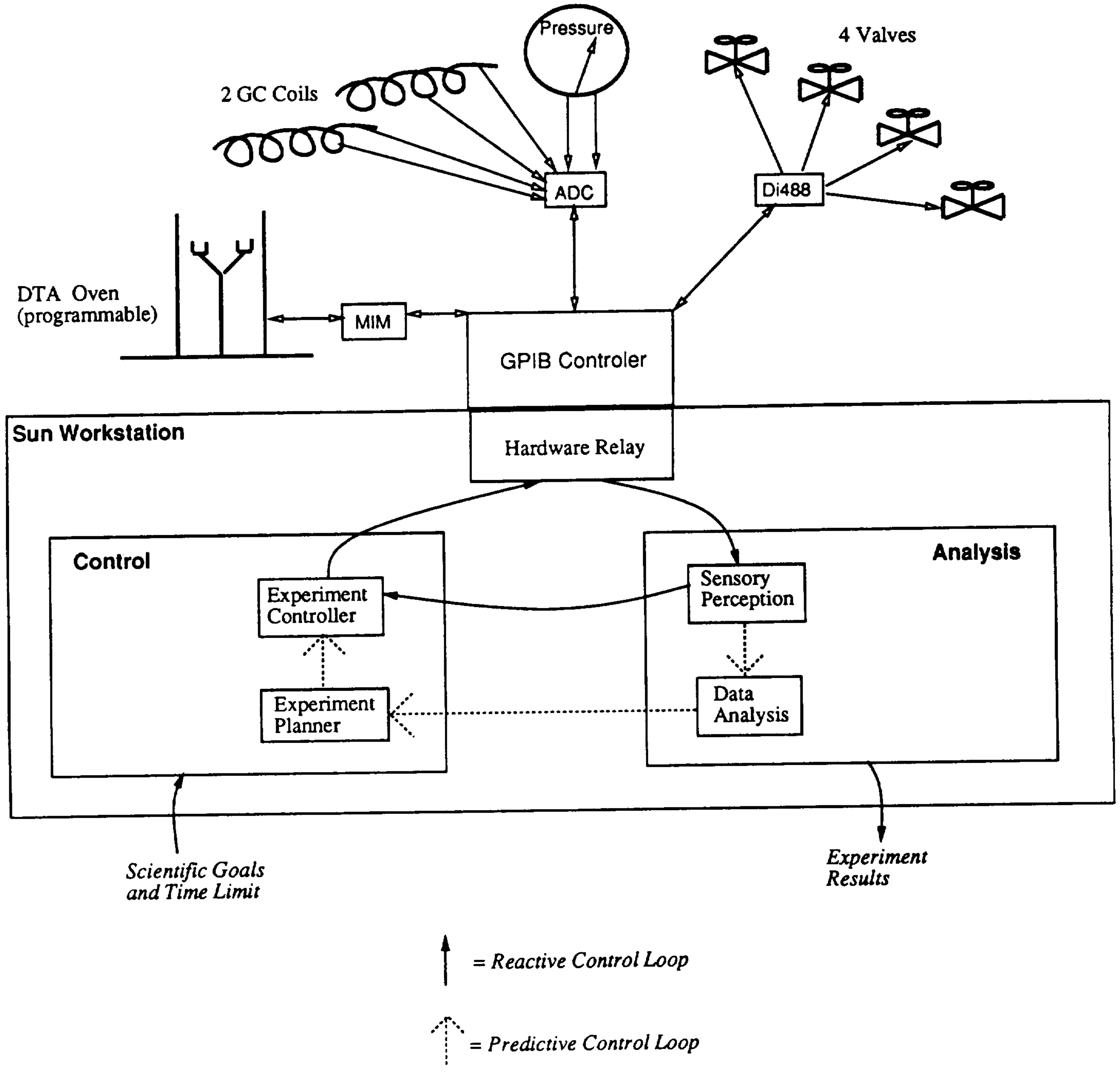


FIGURE 1. The DTA-GC Architecture

KNOWLEDGE FROM PICTURES (KFP)

Walt Truszkowski
Code 522.3
NASA Goddard Space Flight Center
Greenbelt, MD 20771

Frank Paterra, Sidney Bailin
CTA Incorporated
6116 Executive Boulevard, Suite 800
Rockville, MD 20852

ABSTRACT

The old maxim goes: "A picture is worth a thousand words". The objective of the research reported in this paper is to demonstrate this idea as it relates to the knowledge acquisition process and the automated development of an expert system's rule base. A prototype tool, the Knowledge From Pictures (KFP) tool, has been developed which configures an expert system's rule base by an automated analysis of and reasoning about a "picture", i.e., a graphical representation of some target system to be supported by the diagnostic capabilities of the expert system under development. This rule base, when refined, could then be used by the expert system for target system monitoring and fault analysis in an operational setting.

Most people, when faced with the problem of understanding the behavior of a complicated system, resort to the use of some picture or graphical representation of the system as an aid in thinking about it. This depiction provides a means of helping the individual to visualize the behavior and dynamics of the system under study. An analysis of the picture, augmented with the individual's background information, allows the problem solver to codify knowledge about the system. This knowledge can, in turn, be used to develop computer programs to automatically monitor the system's performance. The approach taken in this research was to mimic this knowledge acquisition paradigm. A prototype tool was developed which provides the user: 1. a mechanism for graphically representing sample system-configurations appropriate for the domain, and 2. a linguistic device for annotating the graphical representation with the behaviors and mutual influences of the components depicted in the graphic. The KFP tool, reasoning from the graphical depiction along with user-supplied annotations of component behaviors and inter-component influences, generates a rule base that could be used in automating the fault detection, isolation, and repair of the system.

INTRODUCTION

This paper details the results of the Knowledge From Pictures (KFP) work. The continuing objective of this work is to develop a system that can build a knowledge base to perform Fault Detection, Isolation, and Recovery (FDIR) from an annotated graphical description. Specifically, the KFP tool should take a user defined graphical image of a system's components and interconnections, and drawing from domain specific libraries for component behavior, develop an expert system to perform FDIR processing of the system defined. The user defined graphical image is also intended to be used as a user interface or front end to the generated knowledge base.

As stated above, this work is motivated by the observation that pictures are often drawn to describe the operation or behavior of many systems and problems that humans address. For example, describing the three basic parts of an atom and how they interact is most easily done with a picture showing protons and neutrons tied together in the center and electrons orbiting around them. Other examples are discussed by Musen et al in (6) and Montalvo in (5).

This work draws on results by Navinchandra et al in (7) and Barker-Plummer and Bailin in (1). Navinchandra et al have exploited the idea that components of a system often cooperate by reacting to each others' actions. The action-reaction function can be thought of as a collection of influence paths where the action of one component changes the state of another. For example, if the system being defined consisted of a power supply connected to a light bulb, the power supply generating electricity would change the state of the light bulb. In the KFP tool we use the idea of influence paths and their flows to determine when a failure has occurred and to isolate failed components.

C-5

Barker-Plummer and Bailin describe a system designed to perform theorem proving from graphic descriptions of proofs. Their system, called GROVER, analyzes an image to generate a set of assertions and develop a proof strategy for solving the theorem described in the image. The proof strategy is represented as a set of lemmas that GROVER builds from the assertions in the image. These lemmas are then fed to a "conventional" theorem prover to provide the complete proof.

In KFP we use the concept of assertions, both derived from components in the image and from domain specific knowledge captured in component libraries, to determine when a component is in a state other than those in its definition. When this situation has been detected, a fault has occurred. This observation aids the KFP tool in isolating a fault and beginning the fault recovery process.

In the remainder of this paper we describe the current system and its state. At present a working prototype exists that can be used to generate FDIR knowledge bases. The generated knowledge bases have a text based user interface. The KFP tool demonstrates approximately a 10 to 1 expansion factor for lines of code generated vs. components and influence paths entered by the user.

SYSTEM DESCRIPTION

The original goal of the KFP tool was to be able to generate FDIR rules from graphical images and then use those same images as the user interface to the FDIR system. It was noted early in the system's design, however, that the image alone may not provide enough information about the system to support fault isolation or recovery activities. Given this, the goal was changed slightly to allow non-image information, i.e., system and component states, and component influence relationships, to be used in the FDIR system generation and operation. This information can be entered by the user when the system is being defined, or extracted from libraries that describe the behavior of known components.

The tool was designed to solve the FDIR problem as three subproblems, i.e., detection, isolation, and recovery. Each of the solutions generate knowledge base components that, when taken together, perform FDIR. This is the approach often taken by a human programmer developing an FDIR system, so it seems reasonable to use the same approach in an automated system.

As discussed in the Introduction, influences between components of a system are used to isolate a failed component. The fault is detected when an alarm condition occurs. An example of such a condition would be a temperature sensitive object operating outside of its design temperature range. Figure 1 shows a system in which such a fault may occur. In this figure there are five components that make up a subsystem. The lens component is temperature sensitive and will register an alarm when its sensor reads above or below defined thresholds. An alarm is specified as a collection of component states. In this example the only component involved in the alarm condition is the lens itself; however, in a more complex system one might also need to check other components, such as the quality of communication signals being received, before it is known that an alarm condition exists. In Figure 1, the temperature driver controls the temperature of the lens by turning on and off the heater and cooler as needed.

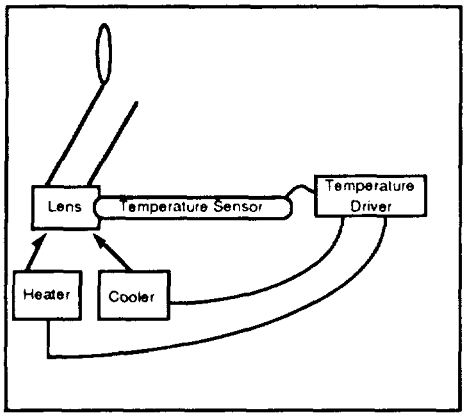


Figure 1: Example System

The cause of an alarm could be one of many failed components, and KFP uses the influence flows among components as well as their known behavior states to identify the component that has suffered a fault. The behavior states of a component describe how it will act when it experiences a particular set of influences. For example, in the system shown in Figure 1, the heater could have the behavior of producing heat as an output influence when it is experiencing the input influences of *power* and *heater on*. Behaviors are represented as component operation states. This implies that the same input influence combination may produce different output influences depending on the current state of the component. For example, when the heater component is off and it receives a *heater on* influence from the temperature driver, it begins to produce heat. If the heater is on already, however, and it receives a *heater on* influence, it does nothing in response. This allows an intelligent component to receive an input and remember the state that it has entered as a result, even after the influence has been removed.

In addition to states, objects can also have local variables. These variables can be manipulated with simple arithmetic operations and can be considered along with the current incoming influence flows to determine when a state change should occur.

Each alarm condition is represented by a CLIPS (C Language Integrated Production System - a NASA-developed inference engine) rule that uses facts about the state of the components contributing to an alarm to determine whether the condition exists. When an alarm is detected, a search begins for the faulted object causing the alarm. This search is performed by tracing back through the paths of influence that are input to the alarming object. The influence paths form a collection of chains of objects that either directly or indirectly influence the components contributing to the alarm. The tracing is performed via a collection of rules that examine the objects in each path. When these rules fire they use information about the current state of the object being examined, and the states of the objects that influence it, to determine whether it is behaving correctly. If the object being examined is not in the correct state then the fault has been isolated. If the object is in the correct state, the objects that influence it are examined next.

After a fault has been detected and isolated, the recovery phase begins. At present the recovery phase consists of notifying the operator and allowing him or her to take corrective action.

COMPONENTS OF THE KFP SYSTEM

There are three main components to the KFP system as shown in Figure 2, object libraries, the system analyst, and the KFP software. Object libraries contain behavior descriptions and associated influence information for known objects. The system analyst extracts objects from the libraries or defines new to describe the complete system to be monitored. The system analyst is responsible for identifying the influence paths among objects and indicating what influence types flow across those paths. Additionally, the system analyst describes the conditions that should cause alarms. The KFP software analyzes the information provided by the system analyst and generates an FDIR expert system. The remainder of this section describes the user interface and operation of the KFP software.

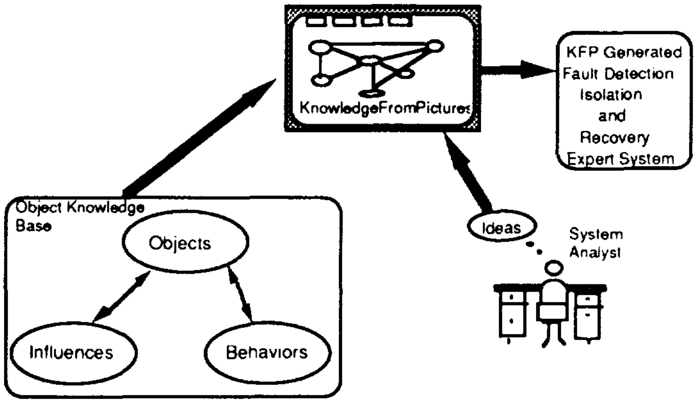


Figure 2: Components of the KFP System

Figure 3 shows the KFP main screen with an example system defined. The components of the system, shown as boxes, are *Temp Sensor*, *Lens*, *Temp Driver*, *Heater*, and *Cooler*. These components are connected via influence paths, shown as lines in the figure. The top part pf the screen contains a collection of pull down menus. The **Files** menu is used to load and save system descriptions, clear the current work space, generate FDIR systems, and quit the KFP system.

The **Objects** menu is used to add and remove objects from the system being designed. The analyst can select a known object from a library, define a new object, or remove an object from the work space. When defining a new object, the analyst only needs to provide a name initially; the object's behavior states and ports can be defined later with their respective menus.

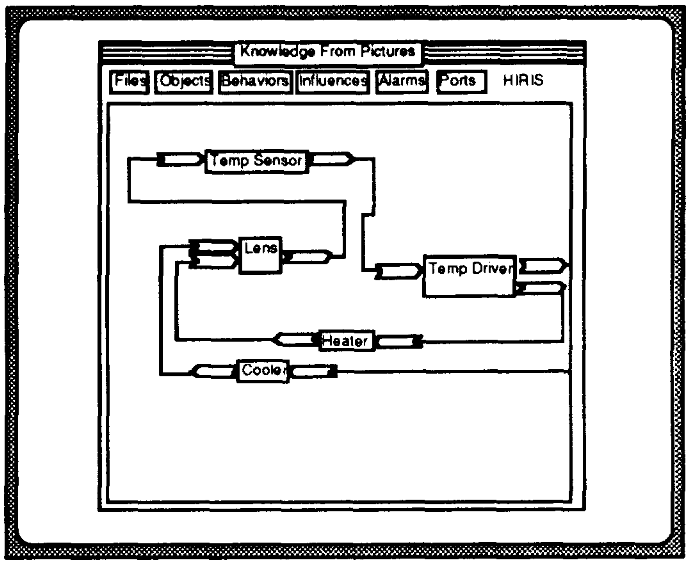


Figure 3: KFP Main Display

The **Behavior** menu is used to add, modify, or remove behavior states of objects in the system. When the analyst selects the *add behaviors* option, the menu shown in Figure 4 is presented The four menus in the figure are used to define the state transitions for an object. The *Present State* is the state that the object is in before the transition occurs. The *Cause* is the variable assignment or input influence that causes a state transition to occur.
The *New State* is the state to which the object transitions, and the *Result* is the variable assignment or influence type that the object will exhibit after the transition.

Under each of these menus there is a button labeled New that is used to define a new state, influence type. Under the Cause and Result menus there is a second button used to define any variable assignment used as part of the transition. When the Exit button is pressed the behavior definition is added to the object's description, and is available to KFP when generating the FDIR expert system.

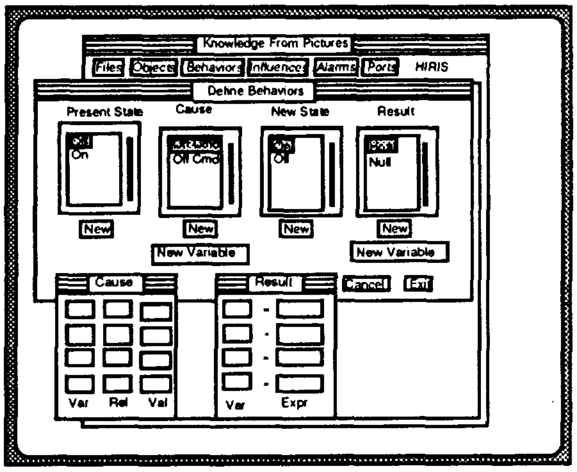


Figure 4: Behavior Definition

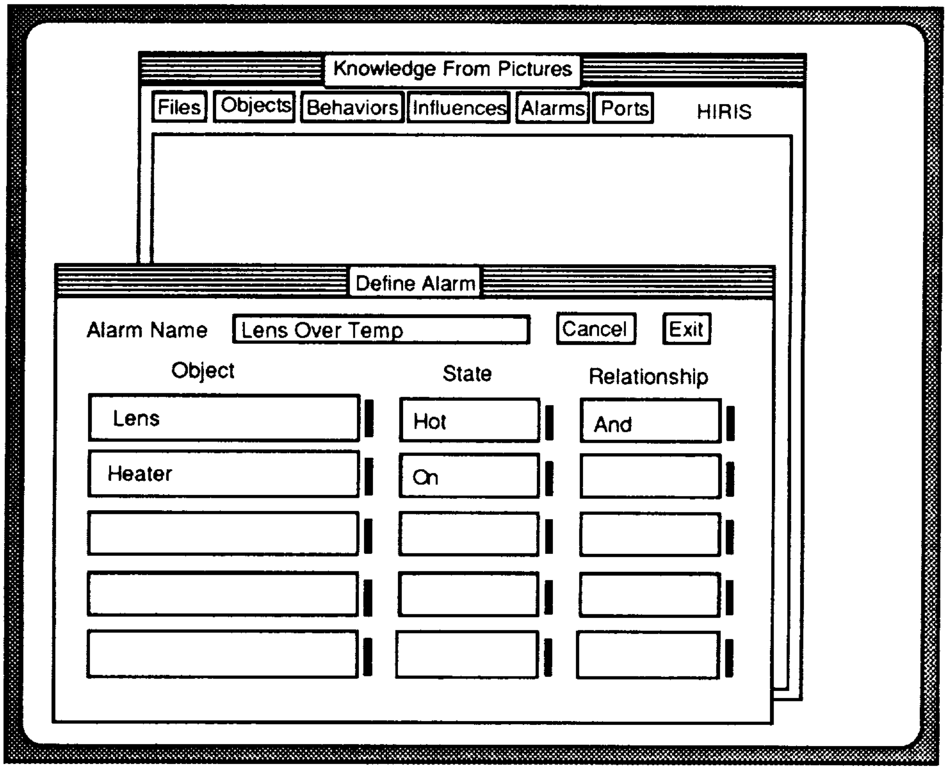
If the edit option is selected, the analyst is presented with a list of existing behaviors which can be selectively deleted. At present there is no way to modify an existing behavior other than to remove it and add a new behavior with the modifications.

The two small panel displays will be shown when the analyst selects the New Variable button for either the Cause or Result menus.

The **Influences** menu is used to define and remove influence paths among component ports. When an analyst selects *add influences* from this menu, the originator and receiver object ports for the influence flow can be selected by mouse clicks. The data type of the flow is determined by the connected ports. If the types of the ports differ, the user is notified and the flow is removed.

The influence path and type are then available to be used when generating the FDIR expert system. When an analyst selects *remove* from the **Influences** menu, a list of existing influences paths is presented allowing selective deletion by the analyst.

The next menu available on the main display is the **Alarms** menu. This menu allows the analyst to define alarm conditions for objects in the system. The analyst selects *add* from the **Alarms** menu and selects the object to which the alarm is attached. The display shown in Figure 5 is then presented, and the analyst can select the objects, states, and relationships that contribute to the alarm condition.



Alarm Name: Lens Over Temp

Object	State	Relationship
Lens	Hot	And
Heater	On	

Figure 5: Alarm Definition

The analyst also specifies the name of the alarm at this time. All of this information is used when performing the FDIR task. As with the previous menus, there is an option to remove existing alarm definitions. When this option is selected, the analyst is presented with a list of the names of existing alarm conditions and can select the ones to be deleted.

The last menu available is the ports menu. This menu is used to add incoming or out going ports to an object. To add a port, the analyst first selects the port direction from the menu and then clicks on the object that is to have the new port. After the object has been selected a menu like that shown in Figure 6 is display, and the analyst can select or define the data type of the port.

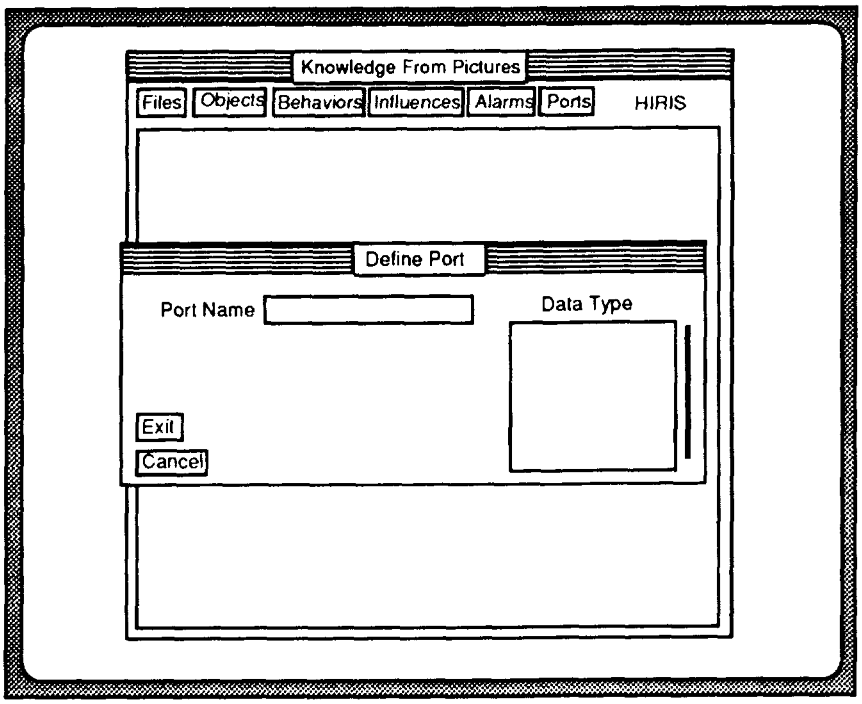
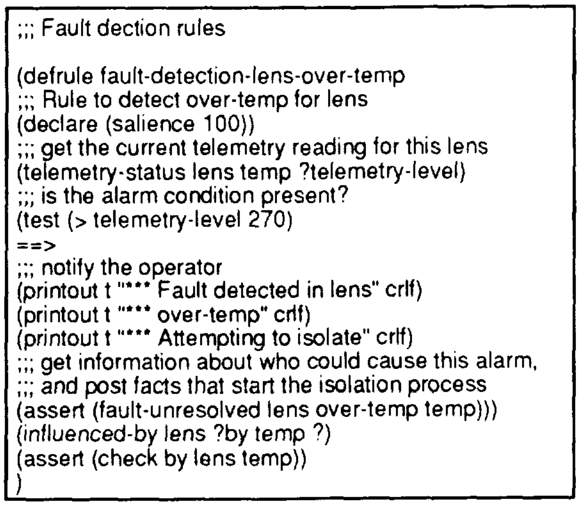


Figure 6: Port Definition

After adding all the components, behaviors, influences, alarms, and ports that are needed to define a system, the analyst selects the *generate* option from the **Files** menu to generate an FDIR expert system.

SAMPLE OUTPUT RULES

This section shows a sample of the rules and facts generated by the KFP tool. Separate rules are generated for each of the tasks: detection, isolation, and recovery. Facts are generated to describe the influence paths and behavior states of the system. Additionally, rules are generated to update local variable values when state changes occur. Each of these except the recovery and variable update rules is shown in this section. The recovery rules simply print out advice extracted from a library once the fault has been isolated.



```
;;; Fault dection rules

(defrule fault-detection-lens-over-temp
;;; Rule to detect over-temp for lens
(declare (salience 100))
;;; get the current telemetry reading for this lens
(telemetry-status lens temp ?telemetry-level)
;;; is the alarm condition present?
(test (> telemetry-level 270)
==>
;;; notify the operator
(printout t "*** Fault detected in lens" crlf)
(printout t "*** over-temp" crlf)
(printout t "*** Attempting to isolate" crlf)
;;; get information about who could cause this alarm,
;;; and post facts that start the isolation process
(assert (fault-unresolved lens over-temp temp)))
(influenced-by lens ?by temp ?)
(assert (check by lens temp))
)
```

Figure 7: Sample Fault Detection Rules

Figure 7 shows a sample fault detection rule, in which the present telemetry for the temperature of a lens is obtained and checked against a known threshold. If the threshold is exceeded, facts are asserted to begin the search for a failed component.

```
;;; Fault isolation rules

(defrule fault-isolation-heater-1
(declare (salience 90))
;;; is an alarm present
?notification = (fault-unresolved ?fault-object ?alarm-name ?inf-type)
;;; If so, should I be checked for causing it?
?check-me=(check heater fault-object inf-type)
;;; get the current state to see if heater is influencing mp now
(state heater ?cur-state)
;;; get the current telemetry so that behavior can be determined
(telemetry-flows heater p_temp ?telemetry-level)
;;; Get the behavior information for the heater
(behavior heater curstate ?cur-behavior)
(test (~= telemetry-level cur-behavior)
==>
;;; notify the operator
(printout t "*** Fault isolated in heater" crlf)
(printout 1 "*** Developing a recovery plan" crlf)
;;; removed the notice and the 'check-me' fact
(retract notification)
(retract check-me)
;;; post the fix-it fact
(assert(fix-it heater p_temp cur-state))
```

Figure 8: Sample Fault Isolation Rules

Figure 8 shows one of two rules generated to isolate a failed object. In this rule an object is check to determine if it is in the correct state for its input influences. If it is not, then the fault has been isolated to that object. If it is in the correct state, a second rule would fire that would assert the name of the next object to be examined.

Figure 9 shows a sample of the influence facts that are used to determine which objects are in an influence path. These facts are used by both the fault detection and the isolation rules.

```
;;; (influenced-by <from> <from's-state> <to> <influence-type>
(influenced-by sensor lens hot hot)
(influenced-by sensor lens cold cold)
(influenced-by driver sensor hot hot)
(influenced-by driver sensor cold cold)
(influenced-by heater driver on cold)
(influenced-by heater driver off hot)
(influenced-by cooler driver off cold)
(influenced-by cooler driver on hot)
(influenced-by lens heater hot on)
(influenced-by lens cooler cold on)
```

Figure 9: Sample Influence Facts

GRAPHICAL MODELLING AS A BASIS FOR SOFTWARE DEVELOPMENT

Simulation and diagnostics play a key role in a satellite control center. They support the two principal activities of the control center—commanding and monitoring the spacecraft. Commanding employs simulation to verify the acceptability of commands, prior to their being uplinked to the spacecraft. Monitoring involves fault detection, isolation, and recovery when telemetry values received from the spacecraft fall outside of defined limits. In our work implementing a testbed for an advanced control center, which we call the Intelligent Ground System (IGS), we found that simulation and diagnosis activities tend to derive from the same set of knowledge, namely models of the spacecraft components. An integrated approach, in which diagnosis and simulation are both driven by the same run-time models, seems feasible to us; at this point, however, we are aiming at a less ambitious goal, which is the generation of distinct programs to support the respective functions from the same graphical model. We can view this as "design time integration" rather than "run time integration."

Our main hypothesis is that modeling of a spacecraft and its subsystems, and reasoning about such models, can—and should—form the key activities of ground system software development; and that by using such models as inputs, the generation of code to perform various functions (such as simulation and diagnostics of spacecraft components) can be automated. Moreover, we contend that automation can provide significant support for reasoning about the software system at the diagram level.

The software models the states, behaviors, and interactions of elements in its environment. Given this role for the software, it seems appropriate to look for a language in which such information can be made explicit. Graphical modeling of objects, their behaviors, and their interactions is an obvious choice for such a language; there is nothing new in our advocacy of diagrams to express such information. Our contention, which may be more questionable, is that the real complexity of the software lies in the interactions expressed by the graphical models, not in the implementation details of the eventual code.

We contend that the structure of the implemented code, for at least certain functions of the ground system—specifically, simulation and diagnosis—is sufficiently well understood to permit us to generate it automatically, and therefore to allow us to redefine the development process as one of developing and reasoning about the graphical models. The previous sections described the progress we have made to date in demonstrating this idea. Similar ideas have been put forward in a recent article by Harel (2).

REASONING ABOUT DIAGRAMS

We have been working for several years on an automated reasoning system that takes diagrams as input. Recently we have begun to apply these ideas to the problem of reasoning about software. The graphical models that we discussed in the previous sections are interpreted by the Formal Interconnection Analysis Tool (FIAT) as plans for proving assertions about the software design.

The particular type of assertions processed by this tool grew out of an actual experience in debugging part of our ground system testbed. In testing a particular simulator program it was found that the behavior of the system was not as expected, but no errors could be found in any of the simulator components. The problem turned out to be one of missing connections between objects in the simulator. Since the simulator architecture keeps each object autonomous—completely ignorant of the objects to which it is connected in a given application—the absence of these connections did not result in any anomalous behavior on the part of any object, but the system itself was not behaving as expected.

Thus we decided to apply the planning concept to verifying statements of the form, "If event *x* occurs at object *A* then event *y* will occur at object *B*." The planner takes event *y* at *B* as a goal, and tries to construct a plan that starts from event *x* at *A* as an initial condition (typically, various other context conditions are specified as well). A goal is reduced to subgoals by traversing the connections specified in the diagram: if a goal state in an object *D* follows, according to *D*'s behavior description and the connections specified in the diagram, from a certain state in object *C*, then this state in object *C* becomes a subgoal of the goal state. A failed plan, when presented to the developer, serves to identify missing connections that may have been overlooked in defining the system.

We have noticed a similarity in the logic of this planner and that of the KFP tool, which similarly traces back through the influence paths in the diagram in generating fault isolation rules. We have not studied this similarity in enough detail to decide whether the two tools can make use of a single "influence traverser" mechanism, but there seems to be some promise of this.

NEXT STEPS AND FUTURE IMPROVEMENTS

The tool is currently a working prototype that achieves most of the task goals. There are some issues that still need to be considered, and extensions that would add to the system's power. At present the system provides only a text based interface for the generated FDIR system. Providing a graphical user interface, by reusing the analyst defined picture, is the next function that we will address in this prototype. After this is completed, all of our original goals will have been met. The remainder of this section lists the enhancements that will be added to the existing system to make it more widely usable.

Currently there are very few objects in the object libraries. Populating this library would make the generation of new system description images simpler because predefined objects could be reused.

Once a system has been defined, it might be advantageous to use the complete system as a component of a larger system. For example, a power system FDIR expert system may be usable for more than one spacecraft. Such complete systems could be represented as *smart icons* on the work space, and could be used just as if they were simple components.

There may be some refinements to the heuristics used to analyze the graph and generate an expert system, as well as those employed in the generated expert system itself. We intend to explore this issue in order to increase the quality and performance of the generated code.

The user interface for the current prototype was developed in TAE and is easy to use. There are some refinements, however, that would present more information to the user and would result in faster operation of the tool. For example, currently to display all the influence paths between two objects, the tool must generate an additional window and menu. It would be faster to display this information in the work space when the user clicks on an influence path line. The enhancements that we currently envision can be implemented in TAE.

Finally, we are concerned about KFP's ability to handle very complex system pictures. The current expansion factor of objects, behaviors, and influence paths to generated code is approximately 10 to 1. We believe that the expansion is linear, but it still may be too large for very complex systems. We intend like to address this issue by working with additional, more complex examples.

KFP AS A SOFTWARE ENGINEERING PARADIGM BASIS

We have made a start at what we hope will become an integrated graphical modeling and development system, in which software development becomes synonymous with defining and reasoning about graphical models. The prospects for such an integrated environment are based on a few empirically perceived similarities:

- Similarity between the information used to simulate a system and that used to diagnose faults
- Similarity between the logic used to reason about system behavior during development, and that used to diagnose faults during operation (backward chaining over influence paths)
- Similarity in the program structure of specific simulators and specific diagnostic systems, which has allowed us to define generic architectures for each of these applications

Within the scope of the current framework, there are perhaps two major open issues: 1) the impact of scale-up on the performance of the generated code, and 2) the feasibility of automated reasoning about additional aspects of the models.

The efficiency of the generated fault detection, isolation, and recovery rules for a large, complex system is an open issue. The examples we have worked with to date in KFP have been obtained from actual systems (either existing or being developed), but they are very small subsets of these systems. There is a solid basis of real-time scheduling theory (e.g., rate-monotonic scheduling) with which we can address scale-up performance issues for the generated simulator code, but we lack such a firm basis for a rule-based diagnostic system. The solution to this problem may be to evolve to a more thoroughly model-based approach to diagnosis, in which there is no production rule interpreter at all. This would, in addition, permit a greater degree of integration between the diagnostic and the simulator code.

An open issue concerning reasoning about the models is whether automation can support reasoning about issues other than the pre-condition/post-condition behaviors currently addressed. One major area that we would like to investigate is support for reducing the state space of a set of interacting components. This problem arises in "reachability analysis," in which one tries to prove (or at least to convince oneself) that no unexpected states are entered. In the area of communications protocols, this has proven to be a difficult but necessary process that can be supported by a variety of heuristic techniques, some of which are automated (3,4)

POTENTIAL COMMERCIAL APPLICATIONS

Certainly the development of new software and knowledge engineering paradigms based on graphical reasoning would have great commercial value. But there are many other potential applications of the technology discussed in this paper. Consider the application of the KFP in the field of Computer Aided Design (CAD), a field strongly related to the KFP work. Intelligent CAD (ICAD) systems exist which are indispensable in the development of many products in various industries. However, at the Eurographics Workshop on Intelligent CAD Systems (held in April, 1988 at Koningshof Congress Centre, Veldhoven, The Netherlands) several important implementation issues were raised and discussed. These included: the definition of design objects, the overall object design process, linguistic issues associated with representing design objects and their inter-object interactions, the relationship between object-oriented and logical paradigms, and the need for increasing the intelligence of ICAD systems. The KFP project is addressing these issues. Though the current focus of the KFP work is on expert system rule-base development the relevance of the emerging KFP concepts and approaches to advanced ICAD systems are obvious.

Another longer-term application of the KFP concepts could be in the area of highly intelligent robots. The graphical reasoning techniques being investigated in the KFP system could prove to be extremely appropriate for advances in the evolution of robotic vision and reasoning capabilities.

REFERENCES

1. Barker-Plummer, D. and Bailin S. "Graphical Theorem Proving" To be submitted to the Journal of Automated Reasoning.

2. Harel, D., 1992. Biting the silver bullet: Toward a brighter future for system development. IEEE Computer, January 1992.

3. Holzman, G., 1992. Protocol design: redefining the state of the art. IEEE Software, January 1992.

4. Lin, F. and Liu M., 1992. Protocol validation for large-scale applications. IEEE Software, January 1992.

5. Montalvo, F. "Diagram Understanding: Associating Symbolic Descriptions with Images." IEEE Computer Society Workshop on Visual Languages, held in Dallas, TX on June 25-27, 1986. IEEE Computer Society Press, pp. 4-11.

6. Musen, M., Fagan, L., Shortliffe, E. "Graphical Specification of Procedural Knowledge for an Expert System." IEEE Computer Society Workshop on Visual Languages, held in Dallas, TX on June 25-27, 1986. IEEE Computer Society Press, pp. 167-178.

7. Navinchandra, D., Sycara, K., and Narasimhan, S., "A Transformational Approach to Case Base Synthesis", Journal of AI in Engineering Design and Manufacturing, Vol. 5, #2, 1991.

Omit

BIOTECHNOLOGY AND LIFE SCIENCES PART 4: COMPUTERS IN MEDICINE

OPTIMAL DESIGN OF COMPOSITE HIP IMPLANTS USING NASA TECHNOLOGY

T.A. Blake*, D.A. Saravanos**, D.T. Davy*, S.A. Waters*, D.A. Hopkins**

*Orthopaedic Engineering Laboratory
Case Western Reserve University
Cleveland, OH 44106

**Structural Mechanics Branch
NASA Lewis Research Center
Cleveland, OH 44135

ABSTRACT

Using an adaptation of NASA software, we have investigated the use of numerical optimization techniques for the shape and material optimization of fiber composite hip implants. The original NASA in-house codes, were originally developed for the optimization of aerospace structures. The adapted code, which was called OPORIM, couples numerical optimization algorithms with finite element analysis and composite laminate theory to perform design optimization using both shape and material design variables.

The external and internal geometry of the implant and the surrounding bone is described with quintic spline curves. This geometric representation is then used to create an equivalent 2-D finite element model of the structure. Using laminate theory and the 3-D geometric information, equivalent stiffnesses are generated for each element of the 2-D finite element model, so that the 3-D stiffness of the structure can be approximated. The geometric information to construct the model of the femur was obtained from a CT scan. A variety of test cases were examined, incorporating several implant constructions and design variable sets.

Typically the code was able to produce optimized shape and/or material parameters which substantially reduced stress concentrations in the bone adjacent to the implant. The results indicate that this technology can provide meaningful insight into the design of fiber composite hip implants.

INTRODUCTION

In orthopaedics, one of the most successful and widely used procedures to treat joint disease is total joint replacement (TJR). Total joint replacement involves the use of prosthetic components to replace the biological joint surfaces. The objective of total joint replacement is to provide an artificial joint that is capable of reproducing "normal" joint kinematics, and that is able to withstand the stresses induced by everyday activities for as long as possible. Although there exists today a plethora of orthopaedic implants that are capable of dramatically improving the performance of pathological joints, their service life is limited. Joint replacement procedures are now being performed on young, active patients. It is evident that the service lives of implants must be extended in order to prevent implant failures and revision surgeries.

One of the most common modes of failure of orthopaedic implants, especially hip and knee prostheses, is loosening of the components. The mechanisms by which this loosening takes place are complex, and most likely involve the interface between the bone and the implant. Clinical studies have suggested that interfacial failure is one of the most common initiators of aseptic loosening in total hip replacements [9,17]. Studies such as these have prompted many investigators and designers of orthopaedic implants to concentrate on minimizing the stresses and stress concentrations in the interface, since these high stresses could ultimately lead to component loosening. Critical regions with regard to implant loosening are the bone adjacent to the implant and the bone/implant interface itself.

Another factor that has been proven to contribute to aseptic loosening of orthopaedic implants is the mismatch between the stiffness of cortical bone and the stiffnesses of traditional implant materials. In the load sharing that takes place between implant and bone, the stress carried by each material is proportional to its stiffness [7]. As a result, an overly stiff implant tends to carry nearly all of the stress in the bone/implant composite structure. This phenomenon is known as stress shielding. The response of bone to changes in stress, as predicted by Wolff's hypothesis is known as remodeling [21]. In many cases, when bone is insufficiently loaded, its adaptive remodeling leads to atrophy with subsequent thinning and increased porosity [13]. Bone degradation at the bone/implant interface usually results in aseptic loosening and implant failure.

Attempts have been made in recent years to reduce the effects of stress shielding by creating new, low stiffness materials for orthopaedic applications. Fiber composite materials seem to have great potential because of their high strength, their biocompatability, and because of the possibility of tailoring their material properties. In theory, fiber composite implants can be manufactured that meet the demanding strength requirements while closely matching the stiffness of the adjacent bone. Implants such as these would minimize stress shielding effects, which may help prevent implant failure due to aseptic loosening.

Designers of fiber composite implants are faced with a virtually infinite number of combinations of shapes and material properties that can be used for a given implant design. Therefore, the process of searching for optimum choices of shapes and material properties can be both complex and time-consuming. Work by several authors has demonstrated the utility of numerical optimization techniques for the solution of similar problems [8,11,12,15,23]. Some investigators have incorporated Finite Element algorithms into iterative numerical optimization schemes, and attempted to optimize the composite structure of total joint replacements for various parameters [8,11,12,23]. Other investigators have sought to optimize the material properties of orthopaedic implants [22]. The design variables included the Young's moduli of the implant and the cement layer. Good agreement was found to exist between the stresses predicted by the design sensitivity analysis and those obtained from the finite element model [22].

In previous work, our project has adapted computational procedures for shape and material tailoring of aerospace structures to the shape optimization of a total knee and/or hip replacement (TKR) or (THR) component [2,16]. The computational procedures were originally developed in NASA programs for composites analysis and structural optimization [3,4,5]. In the present work, we have extended the procedures to include both shape and material optimization. Here we describe the application of the procedure to a femoral component of a total hip replacement.

METHODS

The OPORIM program has evolved from an analysis code called STAT (Structural Tailoring of Advanced Turboprops), which was originally developed by Pratt & Whitney under a NASA contract [3,4,5]. OPORIM was adapted and enhanced to perform shape and material optimization of fiber composite hip endoprostheses.

The basic structure of the OPORIM program is shown in Fig. 1. Information supplied by the user includes geometric information, material properties, choice of design variables, and information related to the optimization scheme. First, a three-dimensional geometric model is constructed. Then, a dimensionless mesh of gridpoints is created and mapped onto the midplane of the 3-D model. The gridpoints are then used to construct a 2-D finite element mesh. Using laminate theory and the 3-D geometric information, equivalent stiffnesses are generated for each element of the 2-D finite element mesh. An objective function is formulated, usually based on some stress criterion. External loads and boundary conditions are applied to the mesh, and a finite element analysis is performed. The optimizer then changes the design variables in attempt to minimize the objective function. A remesh scheme is performed in response to the change in design variables, and another F.E.A. analysis is conducted. Design variable changes and F.E.A. analyses are performed until convergence on an optimal design has been achieved.

Geometry generation within OPORIM is accomplished by means of several design curves. Some of the design curves describe the external geometry of the model, and some describe the internal geometry of each material region within the model. The discrete geometric information contained in the input file is interpolated to produce piecewise quintic polynomial splines. The spline curves are used in the design optimization process, and are updated as design changes are made to the model [3].

The composite analysis section of OPORIM is based upon ICAN (Integrated Composite Analyzer), developed at NASA Lewis Research Center's Structural Mechanics Branch [14]. ICAN was designed for the analysis of multilayered fiber composites using micromechanics equations and laminate theory. Laminate theory provides OPORIM with the capability to represent complex composite 3-D structures with a relatively simple mesh of 2-D finite elements. In addition, it allows for the specification of composite properties as design variables in the optimization routine.

Each element of the finite element mesh can be thought of as a laminated plate, composed of material layers through the thickness of the plate. Each material layer is composed of a number of plies, whose thicknesses can be controlled by the input file of OPORIM. Using laminate theory, an equivalent stiffness can be calculated for the laminated plate to create a 2-D plane stress finite element [1,2].

OPORIM uses the relationships derived for composite plate stiffnesses to construct element stiffness matrices based on the three-dimensional geometry (thicknesses), and the material properties of each material region. Once a finite element solution has been obtained, the strains in the individual plies are back-calculated. From these, the laminate model is used to calculate the individual ply stresses [19].

The finite element analysis in OPORIM is based on the commercial code NASTRAN. OPORIM supports elements that are very similar to the NASTRAN TRIA3 element, a 3-node combined membrane-bending triangular plate element. Each node has six degrees of freedom.

The computational efficiency of the 2-D element makes it better suited to optimization algorithms than other more computationally costly elements, such as three-dimensional elements. To confirm the validity of the optimization results, it was necessary to compare the optimal 2-D designs to equivalent 3-D finite element models. This insured that design improvements were actually made.

The optimization module of OPORIM is ADS, a commercial optimization code written by Dr. G.N. Vanderplaats [18]. ADS separates the solution of the problem into three basic levels: the optimization strategy, the optimizer, and the one-dimensional search. Several choices of specific algorithms for the iterative optimization procedure are available in the code.

APPLICATION

The OPORIM code was adapted to perform shape and material optimization of a total hip prosthesis. Reference [1] covers in detail the changes that were made to the OPORIM source code, and the additional code that was developed for these studies.

A typical manufactured total hip prosthesis is illustrated in Fig. 2. It consists of an acetabular component (A), a femoral component (B) and an acetabular insert (C). The models that were considered in this study consisted of the femoral component of the prosthesis and the proximal eight inches of the femur. As a starting point for the model, a CT scan of the proximal twelve inches of a femur was obtained, with cross sections taken at every 5 mm. Measurements were taken from the CT scan of the width of the femur (in the coronal plane), the thickness of the femur (in the sagittal plane), and the thickness of the cortical shell.

The material properties for the bone of the proximal femur were taken from the literature [6,20,24]. All bone considered in this work was assumed to be isotropic. The proximal femur was broken down into three material regions: a cortical shell region, a high-density cancellous region, and a low-density cancellous region.

Two different implant materials were considered in these studies: a titanium alloy (Ti-6Al-4V), and carbon fiber-reinforced polyetheretherketone (PEEK) composite. The properties of titanium, carbon fiber and PEEK were obtained from available literature. As the shape of the implant is changed during the optimization, it becomes necessary to move the nodal lines to accurately represent the new shape of the implant. This is done by the remeshing scheme in OPORIM, which locates the design curves (splines) that define the edges of the implant, and which moves the nodal lines so that the shape of the implant and the core are defined. In addition, the remesh scheme redistributes the rest of the nodal lines to maintain favorable aspect ratios of the mesh.

The objective function for the design optimization was chosen to be a measure of the stresses in the region of the bone near the implant interface. This reflects the conviction that controlling interface stresses is one of the more important issues with regard to bone adaptation and passive loosening.

Several detailed objective functions were considered including the maximum Von Mises equivalent stresses in the bone elements adjacent to the implant. Others included the sum of the squares of the Von Mises stresses and the sum of the squares of the maximum shear stresses in these same elements.

The same loading conditions were applied to all of the models for all of the objective functions considered. The nodes at the distal (lower) end of the models were fixed.

One of the loading conditions was a force couple producing a bending moment of 1 lb-in applied to two nodes on the end of the "neck" portion of the implant. This loading scheme certainly does not accurately represent physiological loading. However, pure bending is attractive as a test case, since it does not depend on the orientation or the point of application of the load [10]. Results for this loading condition are described below.

RESULTS

Numerous results were generated in the study. To illustrate typical results, we cite the following 2 cases. For more information and additional results, the reader is referred to references [1] and [2].

Case 1

Case 1 involved shape optimization of an implant composed of one material region. The implant was constructed entirely of titanium.

The initial and optimum shapes of the implant in Case 1 are shown in Fig. 3. The optimization produced a proximal widening of the implant and a slight amount of distal narrowing. The initial and optimal distributions of von Mises stress along the normalized length of the implant are shown in Figures 4a and 4b. Note that the solid and dashed lines represent stresses along the lateral and medial edges of the implant, respectively. The stresses are presented in p.s.i., and appear to be extremely low, but the applied loading was a unit bending moment of 1 lb-in, and these stresses can be scaled up for higher applied loads. The peak cancellous von Mises stresses in the initial model were reduced by 77%, and the objective function was reduced by 65%.

Case 2

Case 2 involved shape optimization of an implant with two material regions. The outer region was composed of carbon fiber reinforced PEEK and the inner "core" region was composed of titanium. The design variables included width variables of both the inner and outer regions.

The optimization produced proximal widening in the carbon fiber/PEEK outer region. The optimization of the inner titanium region produced proximal widening, slightly distal to where the titanium region intersects the "neck" of the implant. A slight amount of distal narrowing was observed. The optimization reduced the peak cancellous von Mises stresses by 65% and the objective function by 64%.

DISCUSSION

All of the optimization cases produced significant reductions of the peak cancellous bone stresses in the vicinity of the bone/implant interface. The magnitudes of these reductions ranged between 45% and 84%. In addition, all of the cases resulted in a more uniform transfer of load at the interface, as illustrated by the plots of stress vs. length.

For all of the cases subjected to the same loading, certain similarities in the stress distribution were observed regardless of the implant construction. The peak stresses in the initial models always occurred in elements at the proximal end of the implants. These stresses seemed to dominate the optimizations, since they contributed the most to the objective functions. In every case, the high stresses at the proximal end of the implant were reduced so that the proximal stresses were closer in magnitude to the more distal stresses.

As expected, the choice of objective function had an impact on the outcome of the optimization cases. In some cases, dramatic differences in the optimal shape and material properties were observed when a different objective function was used for the same test case. In spite of the different outcomes, all of the objective functions produced comparable reductions in peak stresses. There were certain trends, however, that seemed to occur regardless of the type of objective function that was used. For instance, each one of the shape optimizations resulted in some type of proximal widening of the implant. The shapes that were obtained were similar to those of typical currently manufactured implants, which have fairly massive proximal ends.

Another trend seemed to occur when material optimization was performed on implants constructed of an outer and inner region of fiber composite. In all of these cases, the optimal fiber/volume ratio of the outer region was greater that of the inner region.

The most dramatic results were always obtained when shape optimization was combined with material optimization. If shape or material optimization was performed alone, the reductions in peak stresses were always less than in the combined optimizations. These results are logical, since the additional design variables in the combined optimization increased the size of the design space by adding new ways for the designs to change.

The optimization results demonstrate that the choice of objective function and loading scheme significantly affected the outcome of the optimizations. However, certain characteristics, such as stiff outer material regions and proximal to distal tapers, were produced for all of the cases, no matter which loading scheme or objective function was used. Although the results of optimization studies such as these are not comprehensive enough to be used as the basis for implant designs, they can provide some meaningful information that can help better define the sensitivities of implant designs to various parameters. As computational capabilities improve in years to come, more realistic anatomical models incorporating interface behavior and bone remodeling will be possible. In closing, at its present state the developed software may be a significant computer aided design tool for the improvement and possible customization of orthopaedic implants.

ACKNOWLEDGEMENT

This research was sponsored by the Technology Utilization Office at the NASA Lewis Research Center, Cleveland, OH under NASA Grant NAG 3-1027.

REFERENCES

1. Blake, T.A. (1992) Design Optimization of Composite Hip Endoprosthesis, Master of Science Thesis, Dept. of Mechanical and Aerospace Engineering, Case Western Reserve University.

2. Blake, T.A., Davy, D.T., Saravanos, D.A., Hopkins, D.A. (1992) Numerical Optimization of Composite Hip Endoprosthesis Under Different Loading Conditions, *Proc. 4th AIAA/UASF/NASA/OAI Symposium on Multidisciplinary Analysis and Optimization*, Cleveland, Ohio, 119-129.

3. Brown, K., Harvey, P., "*Structural Tailoring of Advanced Turboprops (STAT) Theoretical Manual*", Pratt & Whitney PWA-5967-42, March, 1987.

4. Brown, K., Harvey, P., "*Structural Tailoring of Advanced Turboprops (STAT) Interim Report*", NASA CR 180861, August, 1988.

5. Brown, K., Harvey, P., "*Structural Tailoring of Advanced Turboprops (STAT) Programmer's Manual*", NASA CR 182164, March, 1989.

6. Carter, D.R., Hayes, W.C. (1977) The Compressive Behavior of Bone as a Two-Phase Porous Structure, *J. Bone Jt. Surg.* **59A**, 954-962.

7. Christel, P., Meunier, A., Leclercq, S. (1987), Development of a Carbon-Carbon Hip Prosthesis, *J. Biomed. Mater. Res.* **21**, 191-218.

8. de Beus, A.M., Hoeltzel, D.A., Eftekhar, N.S. (1990) Design Optimization of a Prosthesis Stem Reinforcing Shell in a Total Hip Arthroplasty. *J. biomech. Engng.* **112**, 347-357.

9. Gruen, T.A., McNeice, G.M., and Amstutz, H.C., (1979) Model of failure of cemented stem-type femoral components - a radiographic analysis of loosening. *Clin. Orthop.* **141**, 17-27.

10. Huiskes, R. and Boeklagen, R. (1989) Mathematical Shape Optimization of Hip Prosthesis Design, *J. Biomechanics* **22**, 793-804.

11. Huiskes R., Boeklagen, R. (1988) The Application of Numerical Shape Optimization to Artificial Joint Design. "*Computational Methods in Bioengineering*", (Spilker, R.L.; Simon, B.R., eds.) *BED*, **9**, ASME, 185-197.

12. Huiskes, R., Kuiper, J.H. (1990) Numerical Shape Optimization of Prosthetic Implants, *Proc. of the First World Congress of Biomechanics*, La Jolla, CA.

13. Kusswetter, H., Gabriel, E., Stuhler, T., and Topfer, L. (1984), Remodeling of the femur in conventionally implanted hip prostheses, *The Cementless Fixation of Hip Endoprostheses*, Morsher, Springer Verlag, Berlin, pp. 17-20.

14. Murthy, P.L., Chamis, C. (1986) Integrated Composite Analyzer (ICAN). NASA TP 2515.

15. Saravanos, D.A., Chamis, C.C. (1990) Multi-Objective Shape and Material Optimization of Composite Structures Including Damping. *AIAA J.*, to appear. (Also NASA TM 102579).

16. Saravanos, D.A., Mraz, P.J., Davy, D.T. (1991), Shape Optimization of Tibial Prosthesis Components, NASA Contractor Report, in press.

17. Stauffer, R.N. (1982) Ten year follow-up study of total hip replacement - with particular reference to roentgenographic loosening of the components. *J. Bone Jt Surg.* **64-A**, 983-990.

18. Vanderplaats, G.N., Sugimoto, H., Sprague, C.M. (1983) ADS-1: A New General Purpose Optimization Program, *Proc. 24th AIAA/ASME/ASCE/AHS Structures, Structural Dynamics, and Materials Conference*, Lake Tahoe, NV.

19. Vinson, J.R., and Sierakowski, R.L., "*The Behavior of Structures Composed of Composite Materials*", Martinus Nijhoff Publishers, Dordrecht, The Netherlands, 1987.

20. Weaver, J.K., Chalmers, J. (1966) Cancellous Bone: Its Strength and Changes with Aging and Some Methods for Measuring Mineral Content, *J. Bone Jt. Surg.* **48A**, 289-298.

21. Wolff, J. (1870) Ueber die innere Architektuer der Knochen und ihre Bedeutung fuer die Frage vom Knochenwachstum. *Virchows Arch. path. Anat. Physiol.* **50**, 389.

22. Yang, R.J., Choi, K.K., Crowninshield, R.D., Brand, R.A. (1984), Design Sensitivity Analysis: A New Method for Implant Design and Comparason with Parametric Finite Element Analysis, *J. Biomechanics* **17**, 849-854.

23. Yoon, Y.S., Jang, G.H., and Kim, Y.Y. (1989) Shape optimal design of the stem of a cemented hip prosthesis to minimize stress concentration in the cement layer. *J. Biomechanics* **22**, 1279-1284.

24. Yoon, H.S., Katz, J.L. (1976), Ultrasonic Wave Propagation in Human Cortical Bone: Measurement of Elastic Properties and Microhardness, *J. Biomechanics* **9**, 459-464.

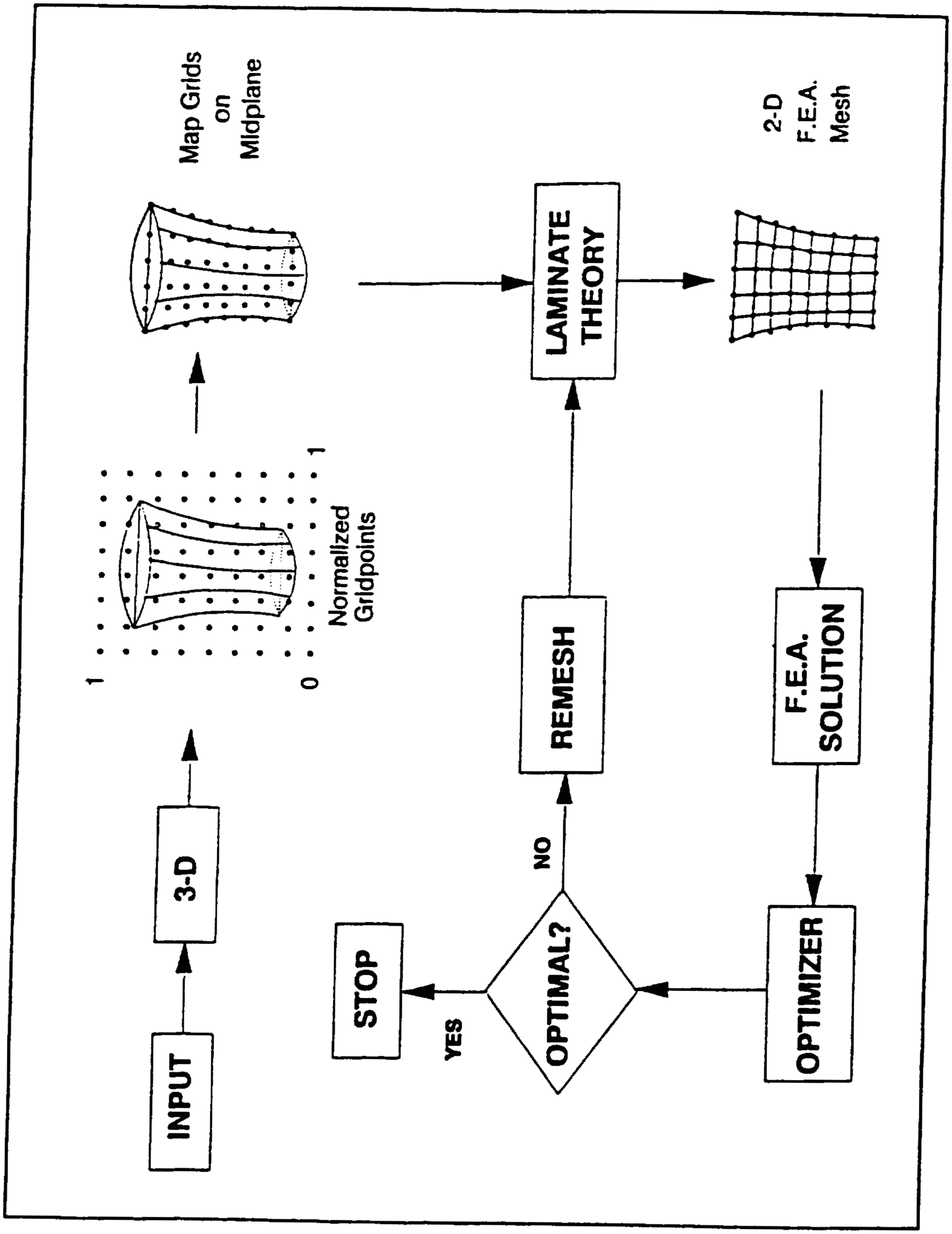


Figure 1: Flowchart of OPORIM

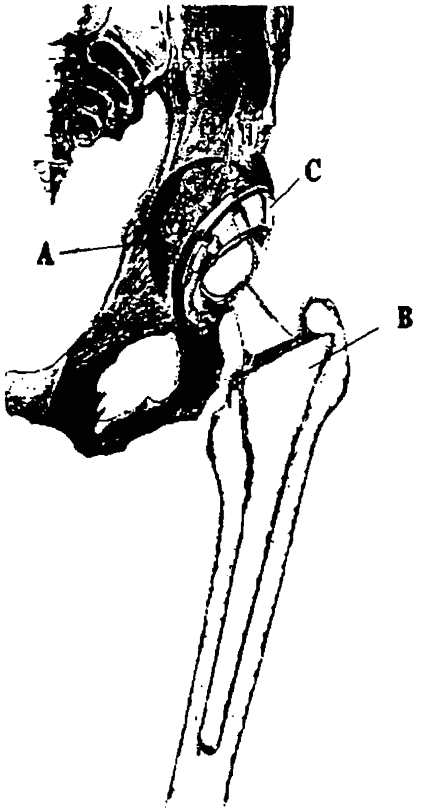


Figure 2: Typical Total Hip Replacement Joint

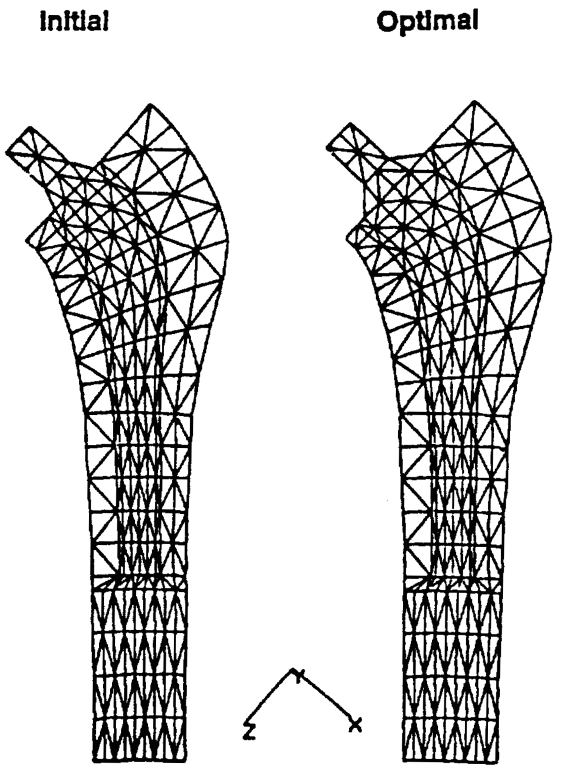
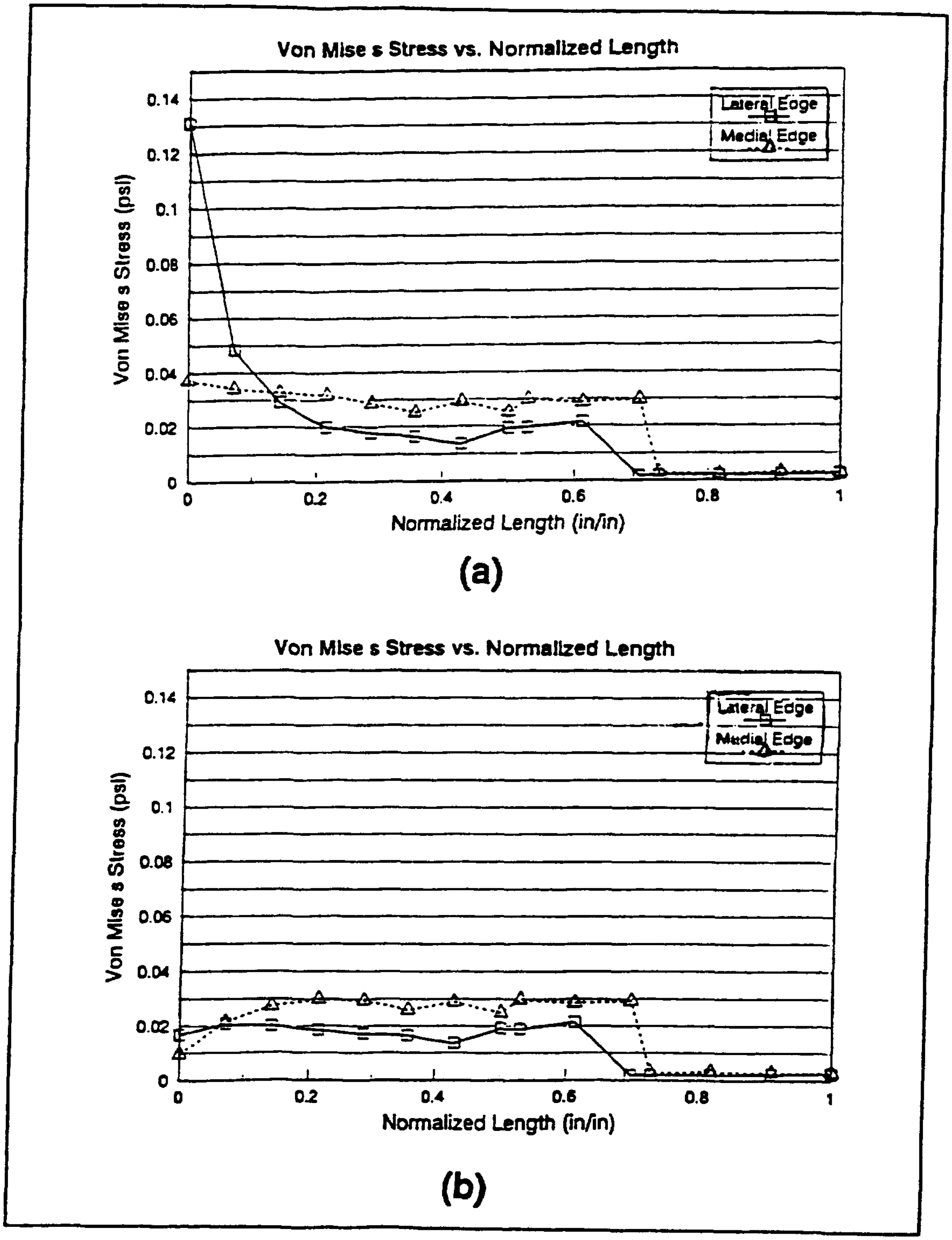


Figure 3: Initial and Optimal Implant Shapes



(a)

(b)

Figure 4: von Mises stresses, Case 1 (a) Initial Design (b) Optimal Design

FINITE ELEMENT ANALYSIS OF A COMPOSITE ARTIFICIAL ANKLE

Leigh Ann Perkins, Lawrence Johnston, and Charles Denniston
NASA Marshall Space Flight Center
Huntsville, AL 35812

Blaise E. Czekalski
Intergraph Corporation
Madison, AL 35894

ABSTRACT

Ultra-light carbon fiber composite materials are being utilized in artificial limbs with increasing frequency in recent years. Dr. Arthur Copes, an orthotist from Baton Rouge, Louisiana, has developed a graphite epoxy composite material artifical ankle (Copes/Bionic Ankle) that is intended to be used by amputees who require the most advanced above-and-below-the-knee prosthetic devices. The Copes/Bionic Ankle is designed to reproduce the function of the natural ankle joint by allowing the composite material to act as a spring mechanism without the use of metal mechanical parts. NASA Marshall Space Flight Center has agreed to participate in the design effort by providing the structural analysis of the artificial ankle design.

INTRODUCTION

This paper presents the structural analysis that was required to define the composite members of the Copes/Bionic Ankle. The finite element modeling expertise and extensive computer facility that resides at NASA Marshall Space Flight Center (NASA MSFC) were essential to ensure a good design of the Copes/Bionic Ankle. The utilization of this resident technology demonstrates the engineering potential that will soon become available through similar computer systems within many private companies to improve the quality of life for many people. Some of the potential uses of this technology are development of self-propelled vehicles for paraplegics, knee prosthetics, robotic structures for machine shops, and customized exoskeleton load carrying frames to be used by persons who must lift heavy loads. A thorough knowledge of structural and materials engineering is required to utilize the computer finite element codes for design engineering analyses. Material strength, ductility, stiffness, and fatigue life are primary areas of understanding to produce a suitable product of this nature.

STRUCTURAL ANALYSIS

The drawings of the design depict the foot adapter and ankle made of graphite epoxy composite material that is attached to an aluminum alloy vertical post. The interface of the ankle to the post is a spherical ball joint which provides for ankle rotations about all three mutually orthogonal axes.

The modeling of the Copes/Bionic Ankle was initiated by the interactive development of the ankle geometry by Blaise Czekalski. This interactive model was then converted to the ANSYS computer code. A 3-D finite element model (FEM) of the ankle was then developed from ANSYS layered shell elements for the composite material, isotropic shell elements for the aluminum support plate, and isoparametric 20-node solid elements for the aluminum post. This FEM is described in Figure 1. It is comprised of 2300 elements with 45,400 total degrees of freedom.

The composite material system for the design is T300/5208, which has Union Carbide Thornel 300 graphite filaments that are impregnated in Narmco 5208 epoxy resin. The orientation of the filament layup for the ankle is ±30 degrees. The structural properties for this material that were used in the analysis are shown in Figure 2. Fatigue life predictions were determined from the fatigue curves for the calculated maximum stresses in the composite material [1]. For fatigue life of approximately one million cycles, the maximum calculated bending stress in the composite material should not exceed 50,000 psi.

The Copes/Bionic Ankle assembly is installed into a foot structure. The foot adapter of the Ankle is restrained from flexing by the foot structure. The ankle flexes about the three mutually orthogonal axes like a human ankle. The ankle must be capable of rotating 20 degrees about the Y axis and 10 degrees about the X and Z axes, separately or simultaneously, with moments not to exceed 60 inch pounds. The number of the graphite layers and the orientation of the fibers in the composite elements must be adjusted to achieve the proper stiffness.

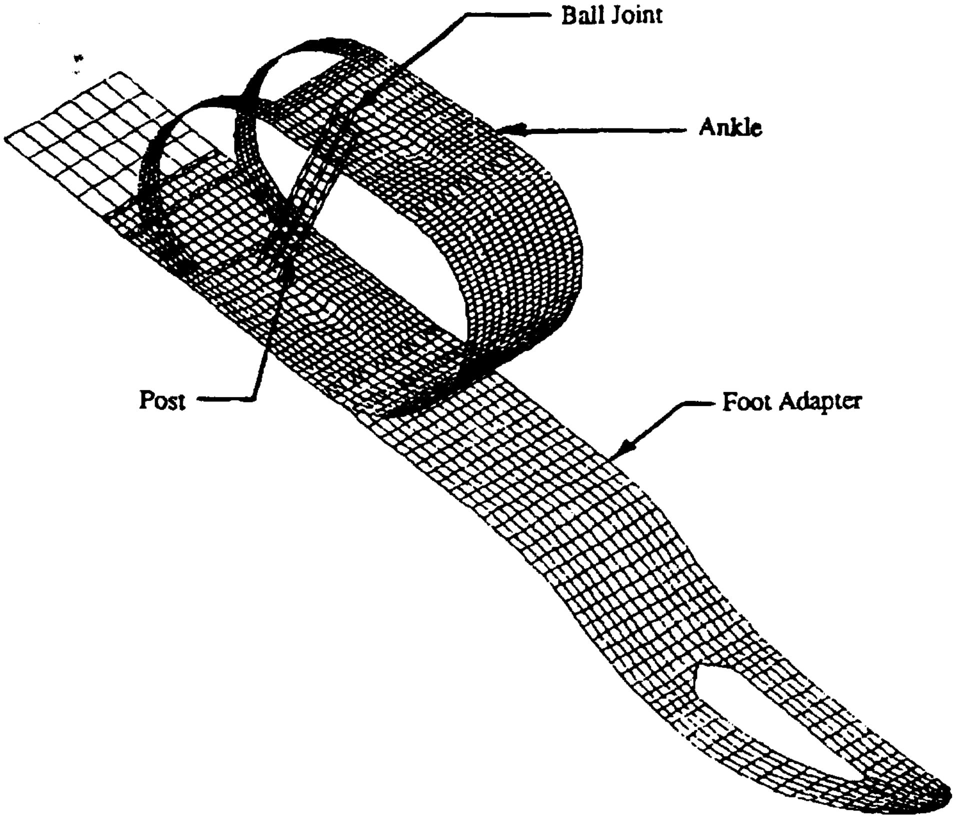


Figure 1. ANSYS Finite Element Model of Copes/Bionic Ankle

Elastic Moduli	EX = 33 $(10)^6$ PSI
	EY = EZ = 2.1 $(10)^6$ PSI
	GXY = 2.1 $(10)^6$ PSI
	GYZ = GXZ = 3.3 $(10)^6$ PSI
Poisson's Ratios	NUYZ = 2.7 $(10)^{-1}$
	NUXY = NUXZ = 2.7 $(10)^{-2}$

Thickness of Thornel 300 graphite ply = 0.005 inch

Figure 2. Structural Properties for the T300/5208 Composite Material [2]

ANALYTICAL RESULTS

The front surface of the initial design of the ankle acted like a shear beam for rotations about the vertical and longitudinal axes and was therefore much too stiff to allow rotations of 10 degrees about these axes.

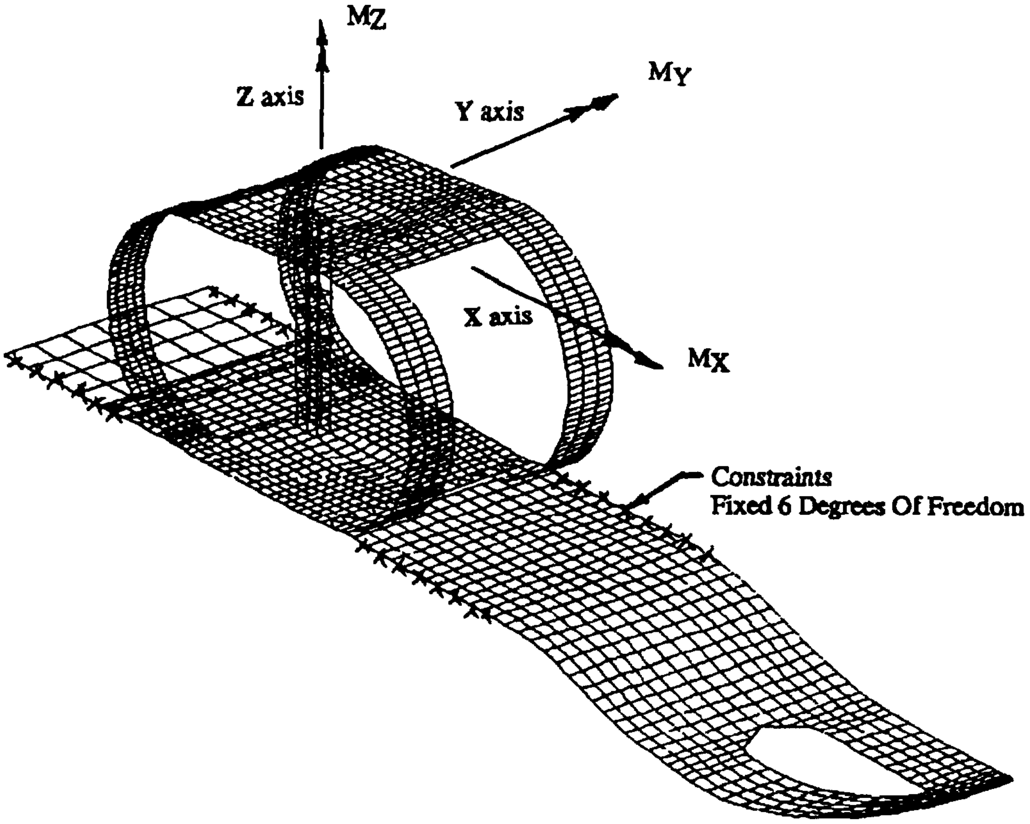


Figure 3. ANSYS FEM of the Modified Copes/Bionic Ankle, Showing Applied Loads and Constraints

The FEM of the Copes/Bionic Ankle was therefore modified to delete the center of the front curved part of the ankle so that it will resemble the aft curved part. The FEM of the proposed modification of the Copes/Bionic Ankle is shown in Figure 3. To employ an iterative approach, the 24 graphite filaments were reduced to 8 for all of the ankle to approximate the desired flexibilities. The purpose of this modification was to allow the required rotations of the ankle within the moment restraints.

The required 20 degree rotation about the Y axis was imposed on the FEM of the modified Copes/Bionic Ankle. The resultant bending moment about the Y axis was calculated to be about 670 inch pounds, which is an order of magnitude too high. The bending stresses were also about an order of magnitude too high. Additional design iterations will be explored to hopefully achieve a level of the proper stiffness and stresses.

CONCLUSIONS

The Computer Aided Design application of the ANSYS structural code at the NASA Marshall Space Flight Center is providing valuable and essential assistance in the development of the Copes/Bionic Ankle. The proof of the concept is greatly helped, and may result in the desired stiffness and adequate useful life.

References

1. Tsai, S.W., "Composites Design, 4th edition," Think Composites, 1988
2. Material Properties from Dr. Copes on 6-23-92

DESIGN OF A PORTABLE POWERED SEAT LIFT

Bruce Weddendorf
NASA Marshall Space Flight Center
MSFC, Alabama 35812

ABSTRACT

People suffering from degenerative hip or knee joints find sitting and rising from a seated position very difficult. These people can rely on large stationary chairs at home, but must ask others for assistance when rising from any other chair. An orthopedic surgeon identified to the MSFC Technology Utilization Office the need for development of a portable device that could perform a similar function to the stationary lift chairs. The MSFC Structural Development Branch answered the Technology Utilization Office's request for design of a portable powered seat lift. The device is a seat cushion that opens under power, lifting the user to near-standing positions. The largest challenge was developing a mechanism to provide a stable lift over the large range of motion needed, and fold flat enough to be comfortable to sit on. CAD 3-D modeling was used to generate complete drawings for the prototype, and a full-scale working model of the Seat lift was made based on the drawings. The working model is of low strength, but proves the function of the mechanism and the concept.

INTRODUCTION

This paper describes how the portable powered seat lift prototype was designed. It includes how requirements were derived and how they were met through the design of the prototype. Also included are the lessons learned from building and testing the working model and possible improvements to the design to make it lighter and less costly to manufacture.

BACKGROUND

People suffering from degenerative hip or knee joints find sitting and rising from a seated position very difficult. These people can rely on large stationary chairs at home, but must ask others for assistance when rising from any other chair. An orthopedic surgeon identified to the MSFC Technology Utilization Office the need for development of a portable device that could perform a similar function to the stationary lift chairs. The MSFC Structural Development Branch answered the Technology Utilization Office's request for design of a portable powered seat lift.

DESIGN REQUIREMENTS

Engineers in the MSFC Structural Development Branch began the design process by developing functional requirements for a portable seat lift. These requirements were generated with the help of the orthopedic surgeon requesting the technology. The first requirement established was strength: The portable powered seat lift must support a load of 300 pounds with a factor of safety of 2 on yield. This requirement was established to prevent collapse of the seat lift during use and applies throughout the lifting range. The seat lift mechanism also must lift a 300-pound weight from the fully closed position. Maximum lift time was decided to be 10 seconds. A maximum weight of ten pounds for the seat lift was established. This weight was not met with the prototype design, as meeting the requirement of low cost and ease of fabrication were considered more important. In addition to guidelines laid out by the orthopedic surgeon, this effort included research by MSFC engineers. The research included videotaping several people standing and sitting in front of a grid and entering the hip positions into the CAD system. This information was used for generating the range of motion requirement for the seat lift.

CONFIGURATION SELECTION

Several types of overall lifting schemes were considered, including crutch type lifts, a walker with a built in lift, and others. When the criteria of ease of use by most patients and portability were applied, the powered opening seat cushion configuration was selected. The seat lift configuration chosen has an unobtrusive, briefcase-like appearance when being transported, and operation should be familiar to anyone who has used a powered lift arm chair. This configuration requires a slim overall height, but this height is still significant and may pose problems in some cases when the user sits at a low table. This choice of configuration also requires the user always to sit on the lift when they need its assistance, making comfort of the seat over long periods a concern. Use of the device in a chair on casters may pose problems, as there may be a tendency for the chair to roll back away form the user when lifting.

PROTOTYPE DESIGN PHILOSOPHY

Design of this prototype was based upon meeting the requirements for lift angle and height, strength, and overall size as described above. In addition to these requirements, the design was based upon production of one or two units using commonly available machine tools. No molds, dies, or special tools were required to build the prototype design. Cost was considered to be a major design driver, and simplicity of fabrication took precedence over weight. Further influencing the design was the intended use of the prototype as a test article to prove if the portable powered seat lift concept is technically and medically feasible. Because the powered seat lift prototype is intended for use with patients in a controlled and supervised testing environment, no effort was made to generate or meet safety requirements beyond simple structural strength.

MECHANISM DESIGN

Design of the mechanism proved challenging because of the large range of motion, high forces, and thin packaging envelope. Many different overall mechanism types and layouts were considered before the final type was developed (see figure 3). This mechanism features front and rear facing arms, attached to shafts mounted in bearings on the stationary base, with the front arm crossing over the shaft for the rear arm. The front arm is pivoted behind the rear arm pivot, which allows both arms to be as long as possible. The two shafts on which the arms are pivoted are geared together with a ratio chosen to give the right amount of lift at the front and the rear of the seat. The front of the seat is pinned to the front arm, while the rear of the seat is supported by rollers pinned to the rear arm. The system of two arms and the seat has one degree of freedom and can therefore be controlled at any one location. The control location chosen for the mechanism is the rear arm shaft. Crank arms fixed to the rear arm shaft are pinned to connecting links that are pinned to a slider on a track fixed to the base frame. The links and crank arms convert the linear motion of the slider into rotation of the shaft. The position of the slider on the track is determined by an acme screw which runs through it. The screw is supported on thrust bearings and features a worm gear made onto its forward end. The worm gear is driven by a worm directly connected to the motor shaft. This system has a large gear reduction which allows the use of a small motor. The large gear reduction cannot be back driven, so it will remain stationary unless the motor turns.

VERIFICATION OF THE GEOMETRY

The geometry of the mechanism was tailored to match the lift curve generated by a 5 foot 10 inch subject from the data. To check the validity of the design, two stationary wood models of the lift at different heights were made and tested subjectively by people of varying size. The results were as expected, so no changes to the geometry were made. The tests demonstrated that the increasing tilt of the seat as the lift progresses allows shorter users to be supported closer to the front of the lift, and taller users closer to the rear. The wood models also allowed us to learn that handles were necessary on the rear corners of the seat lift, and that under one of these would be a good location for the control switch. The width of the seat lift was also determined by use of the wood models. Originally, the planned seat lift width was as close to the width of a chair as possible, with minimum clearance for the chair arms. This was found to be a poor assumption, as the user's hands must fit between the seat lift and the chair arms as they hold the handles at the rear of the seat. The width of the design was reduced to allow this clearance. See figure 1 for the overall dimensions and figure 2 to see the motion of the lift.

DESIGN METHODS AND DETAILS

In order to package the mechanism within the confines of the seat lift envelope without any interferences, a 3-D model of the entire assembly was made using Intergraph EMS computer aided design (CAD) software. Strength critical parts were first sketched and hand analyzed for strength, then sized and input into the CAD. Other parts which were not as highly loaded were sized in the CAD by clearance constraints. The solid modeling allowed clearances between all parts to be verified throughout the range of motion of the mechanism. As the design progressed, several design changes were incorporated as new problems surfaced. For example, a potential pinch hazard between the seat top and the base was eliminated by making the sides of the seat from soft closed cell foam (see figure 3). Pinching from the internal mechanism is also a consideration. The prototype design does not include protection for this type of pinching, as it is intended to be used as a development and test article only, and pinching could only occur with a deliberate and deep insertion of the hand into the opened seat lift. A production model may require a guard for certain parts of the mechanism.

DESIGN DRAWINGS AND DOCUMENTATION

Engineering drawings of all parts and assemblies of the prototype design were made. The drawings were made from projected views of the parts and assemblies of the CAD solid model. There are over 90 sheets of drawings documenting the prototype design, which meet the MSFC drawing standards.

WORKING MODEL

A working model of the prototype design has been built which is very close in appearance and operation to the prototype, except it is of low strength. It was built at the MSFC model shop using the prototype drawings as a guide. The overall dimensions of the model are made to the drawings, but the materials and tolerances are changed. The gears of the model are standard low-strength aluminum and brass, instead of the high-strength steel parts of the prototype. The model does use the same bearings, a similar aluminum frame, and identical electric parts as the prototype would. Similarity of the model to the prototype allows testing of the mechanism and electric system function for less cost than construction of the prototype itself. The model is a valuable demonstration tool for the portable powered seat lift as well as helping in the development of the prototype.

Construction and testing of the model taught several important lessons. Building the model verified the assembly of the mechanism and electric hardware was possible without interference. Testing the model uncovered a problem with the original location of the limit switches which interrupt the circuits of the control switch when the lift reaches fully closed or fully open. The closed position switch was damaged by the rear arm whose position it was suppose to sense. This was because the mechanism had enough momentum to move slightly even after the motor power was cut. This problem was solved by using a different type of limit switch and relocating both limit switches to operate by the position of the slider block, which has a much more controlled motion. The most important lesson of the model, however is that the motor originally selected is inadequate. This motor was selected based upon its advertised power output, light weight, small size and low cost alone. No curves of torque vs speed were available, as the motor came from a hobby shop. The motor makes very little torque at zero speed, where the powered seat lift requires maximum torque, and therefore cannot lift the necessary load. A different motor must be selected for the prototype which meets the torque requirement.

POWER REQUIREMENTS

The requirement of lifting a 300-pound load to the full extent of the lift in 10 seconds was used to calculate the required power of the portable powered seat lift. Power required to perform this task is 61 watts or about 1/12 horse power. Friction in the linkage can add significantly to the actual motor power required. No testing was available to quantify this friction, so an attempt was made to calculate it. Minimum motor power was determined to be about 150 watts, but the accuracy of the calculations is still untested at this writing. The gear ratio can be varied between the motor and the acme screw, depending upon the speed at which the motor will run and its torque output. The prototype design has a reduction of 20:1 from the motor shaft to the acme screw. The 20:1 reduction allows the motor to lift the seat in 720 shaft revolutions, with a speed of 4320 revolutions per minute achieving the desired 10-second lift. The torque requirement on the motor is 0.135 Nm (1.20 inch-pounds) minimum at startup to move the seat using the 20:1 ratio, and neglecting friction. With the calculated friction, the motor should produce no less than 0.33 Nm (2.93 inch- pounds). These numbers can be adjusted for different speed motors and their required gear ratios, given that they produce the sufficient power.

Large torsion springs were added to the prototype design to assist the motor, to balance the upward lift torque requirement and the lowering requirement. These springs are not required for the system to function if a motor is selected using the above criteria for motor power. Springs may prove beneficial in reducing wear in the mechanism by off loading about 70 pounds of the user's weight. The springs are then energized by the motor in the closing cycle of the seat lift making a more even split of work for the motor during opening and closing of the lift.

RECOMMENDED PROTOTYPE DESIGN CHANGES

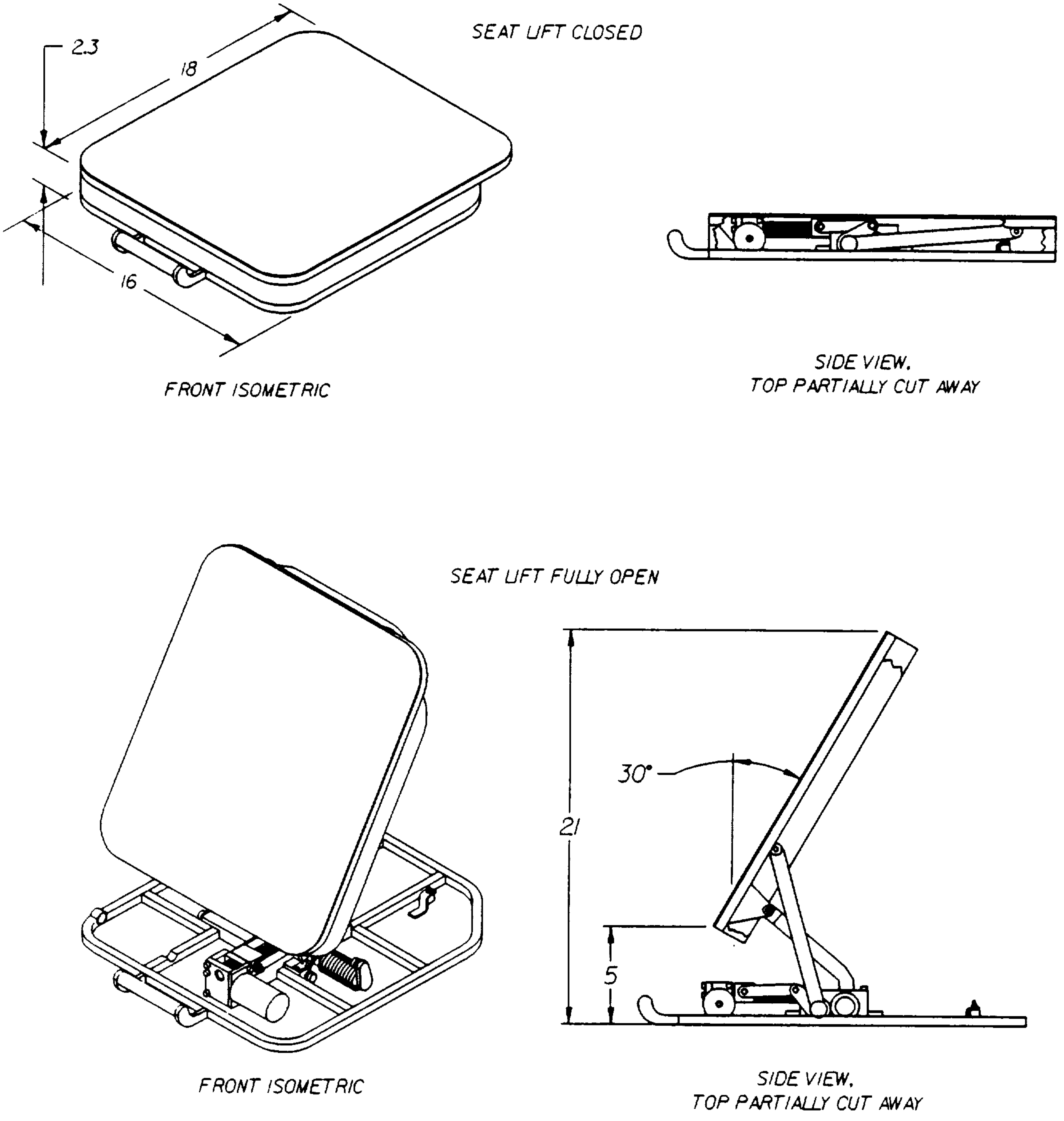
The portable powered seat lift prototype design presented here is a good starting point for any effort to produce a production portable seat lift. The prototype design drawings are complete, but changes to the design should be made before even a prototype unit is built from them. These changes are mostly the result of learning from the model. The most important change is the selection of a different motor with proven torque capability and the modification of the motor mount and gear ratio to install it. The limit switches should be installed as they are now in the working model. It may be better to eliminate the springs provided the motor chosen has enough torque to start the lift when loaded. An area of concern is the friction in the plain thrust bearings retaining the acme screw. This friction has not been measured at this time, but a ball thrust bearing of sufficient strength should be considered as a replacement in this area.

DESIGN IMPROVEMENTS FOR PRODUCTION

Many changes in the design should be made to make it more suitable for mass production. These changes should be made in conjunction with other changes made based on knowledge gained from construction and testing of the prototype. Changes made for production only should focus on the areas of weight and cost reduction first. To help meet these goals for the production design, molded reinforced plastic top and bottom halves for the seat lift should be considered. These parts are complex assemblies on the prototype and can be made easily in one piece after investing in the molds required to make them. The top may require the tracks on which the rear arm rollers rest and the bracket to which the front arm pins to be die cast aluminum parts co-molded into the plastic shell. A similar solution could work on the base plastic shell and the metal frame which supports the mechanism. The aluminum frame supporting the shaft bearings and motor could be cast in one piece instead of machined from billet and welded into an assembly as in the prototype. Each arm, with its attached shaft and gear could be made in one net shape piece using powdered metallurgy. The large shafts could be made hollow, and the shaft gears do not have to go all the way around, as only a partial rotation of each is made. Smaller shaft bearings and bearing supports may also be possible. A study should be made to determine if springs in conjunction with a smaller motor would have a weight and cost advantage to a larger motor without assist springs. Taking into account all the areas for possible weight savings, the target weight of 10 pounds can be met, and perhaps significantly undercut.

DESIGN LICENSING

The portable powered seat lift prototype design is the property of NASA and a patent application will be filed to protect the key design features. Prospective manufacturers are encouraged to contact the MSFC Chief Patent Council, CC01, MSFC Alabama 35812 for licensing information.



SEAT LIFT CLOSED

FRONT ISOMETRIC

SIDE VIEW,
TOP PARTIALLY CUT AWAY

SEAT LIFT FULLY OPEN

FRONT ISOMETRIC

SIDE VIEW,
TOP PARTIALLY CUT AWAY

FIGURE 1

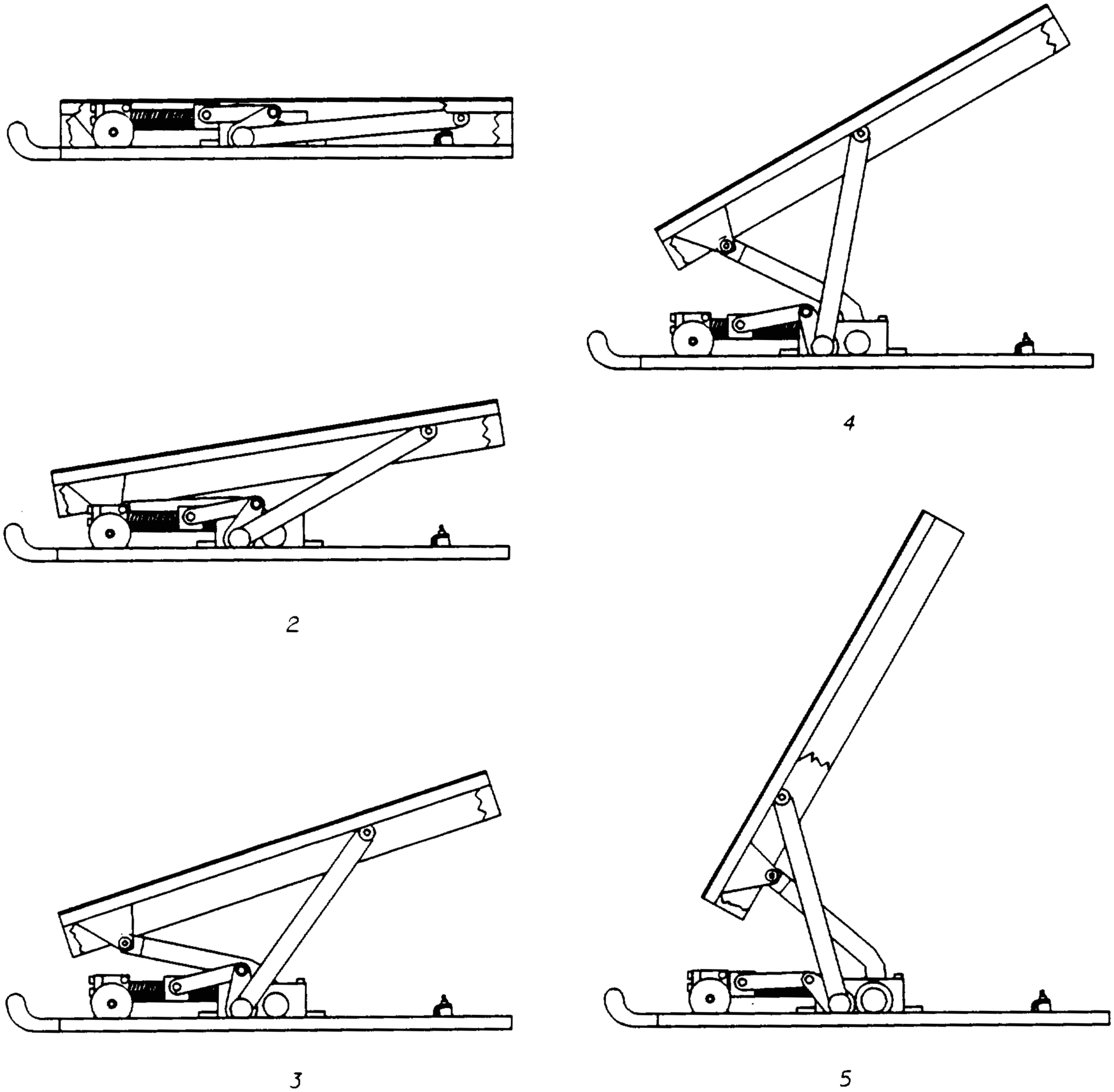


FIGURE 2

PORTABLE POWERED SEAT LIFT
SHOWN IN FIVE POSITIONS INCLUDING
CLOSED (1) AND FULLY OPEN (5).
ALL VIEWS FROM SIDE, WITH TOP
PARTIALLY CUT AWAY.

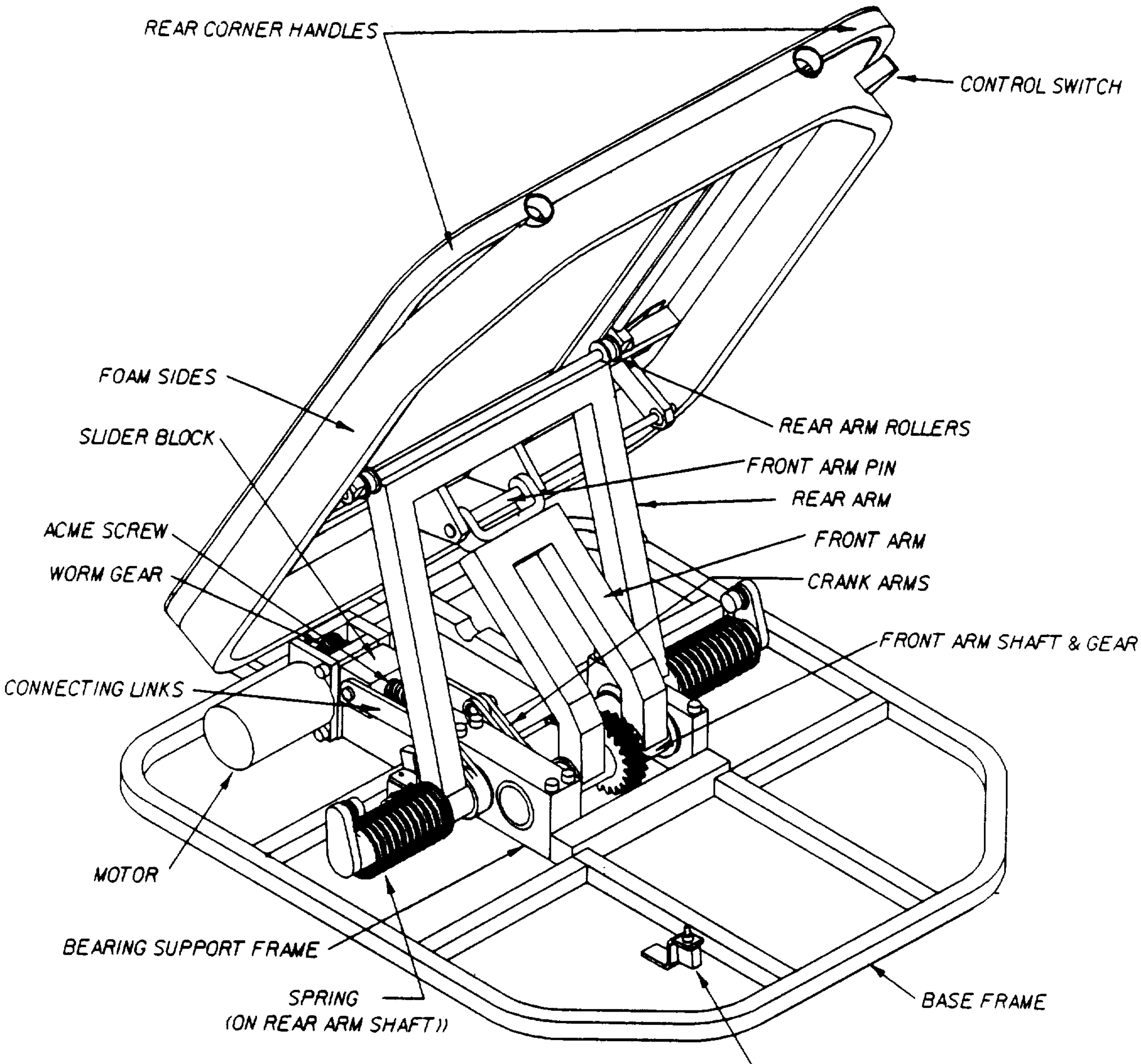


FIGURE 3

ISOMETRIC VIEW. SEAT LIFT FULLY OPEN

MICROCOMPUTER BASED SOFTWARE FOR BIODYNAMIC SIMULATION

N. Rangarajan and T. Shams
GESAC, Inc.
Route 2, Box 339A
Kearneysville, WV 25430

ABSTRACT

This paper presents a description of a microcomputer based software package, called DYNAMAN, which has been developed to allow an analyst to simulate the dynamics of a system consisting of a number of mass segments linked by joints. One primary application is in predicting the motion of a human occupant in a vehicle under the influence of a variety of external forces, specially those generated during a crash event. Extensive use of a graphical user interface has been made to aid the user in setting up the input data for the simulation and in viewing the results from the simulation. Among its many applications, it has been successfully used in the prototype design of a moving seat that aids in occupant protection during a crash, by aircraft designers in evaluating occupant injury in airplane crashes, and by users in accident reconstruction for reconstructing the motion of the occupant and correlating the impacts with observed injuries.

DESCRIPTION OF SIMULATION SOFTWARE

The software developed and used by us in occupant simulation is called DYNAMAN. This package consists of the following modules:

1. A preprocessor that enables the analyst to interactively set up an input data file or to modify an existing data file that is needed to carry out the simulation.

2. A simulation module which accepts the input file that was created using the preprocessor, and produces output files that contain various dynamic variables that describe the three-dimensional motion of the occupant, e.g. accelerations, displacements, contact forces, etc. (The simulation module is essentially the ATB Version 4.2)

3. A postprocessor that can be used to view the output of the simulation module in pictorial, graphical, and tabular forms.

4. A program to estimate dimensions of a human occupant based on sex, weight and height.

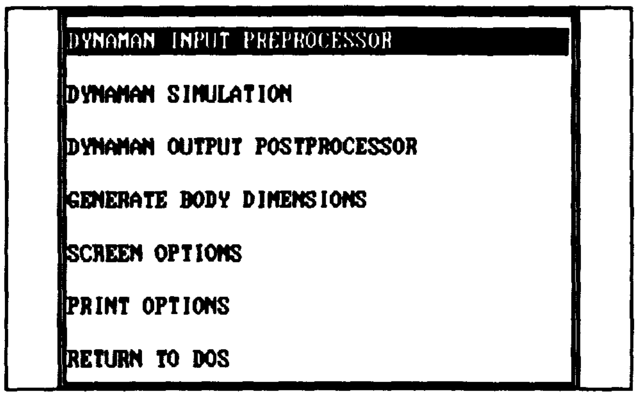
The software package will run on 80386- and 80486-based personal computers under DOS 3.xx and above. Both 16-bit and 32-bit versions are available. In addition, a workstation version has been developed, which will work on a Silicon Graphics Iris workstation.

Elements of Simulation Input

In order to simulate the motion of a vehicle occupant or pedestrian, the following broad categories of information are required:

1. Geometric and inertial properties of the occupant;
2. Motion of the vehicle;
3. Environment around the occupant;
4. Definition of functions for the interactions;
5. Definition of contacts between occupant and environment;
6. Initial position of the occupant;
7. Integration and output parameters to run the simulation module.

The DYNAMAN preprocessor is used to set up the various input data required to run a simulation. Typically these input data are read in from previously created files, and the input parameters are changed to produce a new input file. The preprocessor use a variety of menus, tables, dialog boxes, and graphics for both displaying the data and for accepting the user input. Figure 1. shows the menu from the primary screen of the DYNAMAN package which provides access to the different DYNAMAN programs.



DYNAMAN INPUT PREPROCESSOR

DYNAMAN SIMULATION

DYNAMAN OUTPUT POSTPROCESSOR

GENERATE BODY DIMENSIONS

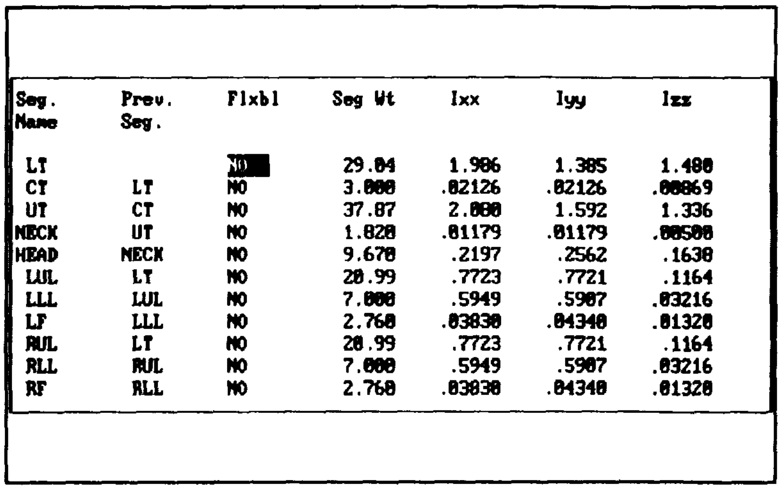
SCREEN OPTIONS

PRINT OPTIONS

RETURN TO DOS

Figure 1: Primary Menu of DYNAMAN

Figure 2. shows a screen displaying a data table of information describing the various segments defined in a particular simulation. It shows the values of such things as segment masses and moments of inertia. The user can move to any field (like in a spreadsheet) and change the value by entering a new number.



Seg. Name	Prev. Seg.	Flxbl	Seg Wt	Ixx	Iyy	Izz
LT		NO	29.04	1.906	1.305	1.400
CT	LT	NO	3.000	.02126	.02126	.00069
UT	CT	NO	37.87	2.000	1.592	1.336
NECK	UT	NO	1.020	.01179	.01179	.00500
HEAD	NECK	NO	9.670	.2197	.2562	.1630
LUL	LT	NO	20.99	.7723	.7721	.1164
LLL	LUL	NO	7.000	.5949	.5907	.03216
LF	LLL	NO	2.760	.03030	.04340	.01320
RUL	LT	NO	20.99	.7723	.7721	.1164
RLL	RUL	NO	7.000	.5949	.5907	.03216
RF	RLL	NO	2.760	.03030	.04340	.01320

Figure 2: Screen for Defining Segment Data

In the next sections, we will discuss each of the broad categories of data listed above.

Geometric and Inertial Properties of the Occupant

The vehicle occupant in DYNAMAN is modeled as a number of segments that are connected by joints. Each body segment and joint is identified by a number and a mnemonic assigned to it. The maximum number of segments you can use currently to model the crash victim is 60.

The principal source of validated data for occupants come from testing done on anthropometric test devices (ATD) or crash test dummies. Occupant input data may also be obtained from the BODGEN program which accesses a database of occupant size data. This database was created from a study of several thousand male and female volunteers of different age groups and from a large sample of children. The software estimates the dimensions of an occupant of specified sex, weight and height from regression equations set up for each of the body segments that are used in the simulation software. The BODGEN program is derived from the GEBOD program.

Each body segment can be described fully by defining its weight, moments of inertia, and the orientation of its principal axes. In addition each body segment also has one or more contact ellipsoids attached to it. These ellipsoids are described by their semiaxes and location of their centers. The ellipsoids are used in determining the contact forces generated when contacts between body segments and vehicle interior planes exist.

Joints are used to connect body segments to each other. There may a number of different linked systems, each system consisting of a set of segments connected together, but the systems themselves being distinct from each other. Different kinds ofjoints can be defined to constrain the relative motion between the connected segments. The joints can model a hinge, a ball and socket or a more complicated type of motion. The torques required to rotate the adjacent segments at a joint in various directions are input into the model.

Vehicle Motion

The motions of upto six different segments can be specified in DYNAMAN. These motions determine the crash event in which the occupant is placed. The motion can be specified as unidirectional or with full 6 degrees of freedom. The initial location and orientation of the segments undergoing the prescribed motion can also be specified.

Environment Around the Occupant

The environment to which the occupant is exposed may consist of one or many of the following:

1. Vehicle contact planes: Each contact plane that is defined can be allowed to contact any defined segment during the course of the simulation.

2. Belt restraint systems: A belt restraint system can be made up of several harnesses each consisting of a number of belts. Several belts may be joined together at tie-points. Each belt can be in contact with a number of segments at several points.

4. Airbag restraint systems: The airbag is modeled as a stretchless ellipsoidal bag and it interacts with contact ellipsoids attached to selected occupant segments and reaction panels on the vehicle.

5. Constraints: These are distance constraints imposed on the relative motion between a pair of segments. You can constrain a specified point (a) on a segment to move, in such a way that there is a constant distance (this distance can be zero) between it and a point (b) on another segment.

6. Spring-dampers: You can use a spring-damper combination to connect two segments. One situation where you might like to use a spring-damper combination is when you want to model the thorax as two segments (spine and sternum) connected by a spring-damper combination. You can then evaluate chest deflection.

7. External forces and torques: You can apply specified forces and torques on prescribed points of a segment.

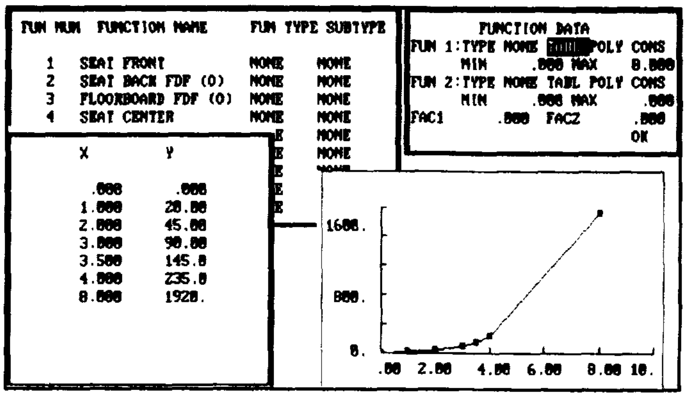
8. Additional contact ellipsoids: These may used to model the contact between certain segments with greater fidelity.

9. Joint restoring forces: Joint restoring torques are defined as functions of the joint flexure angle for specified joint azimuth angles and are used to model the joint torques with greater detail than given by a simple joint torque coefficient.

Definition of Functions

A number of functions are required which define the interaction between various body segments and the environment, and possibly between two different body segments.

The contact between a plane and a segment (or between two segments) is governed by the force-deflection, inertial spike, energy absorption, permanent deflection, and friction coefficient functions that must be defined by the user. Functions which describe the stretch characteristics of the belts can also be defined, as well as, the deflection characteristic of a segment with a belt. The screen setup for entering function data is shown in Figure 3.



```
FUN NUM  FUNCTION NAME        FUN TYPE SUBTYPE

   1     SEAT FRONT           NONE     NONE
   2     SEAT BACK FDF (O)    NONE     NONE
   3     FLOORBOARD FDF (O)   NONE     NONE
   4     SEAT CENTER          NONE     NONE
                                 E     NONE
                                 E     NONE
```

```
          FUNCTION DATA
FUN 1:TYPE NONE TABL POLY CONS
      MIN      .000 MAX    8.000
FUN 2:TYPE NONE TABL POLY CONS
      MIN      .000 MAX     .000
FAC1       .000  FAC2       .000
                            OK
```

```
      X          Y

    .000       .000
   1.000      20.00
   2.000      45.00
   3.000      90.00
   3.500      145.0
   4.000      235.0
   8.000      1920.
```

Figure 3: Windows Defining Function Values

Contact Definitions

The actual contacts that will be allowed between different planes and contact ellipsoids have to be defined by the user. Similarly the characteristics of each contact point of a belt have also to be defined. Contacts between segments and the airbag are also defined.

Initial Position and Belt Position

The preprocessor allows the user to interactively set up an initial configuration where the reaction forces from the initial contacts with vehicle planes are reduced to a minimum. The appearance of the screen for this procedure is shown in Figure 4.

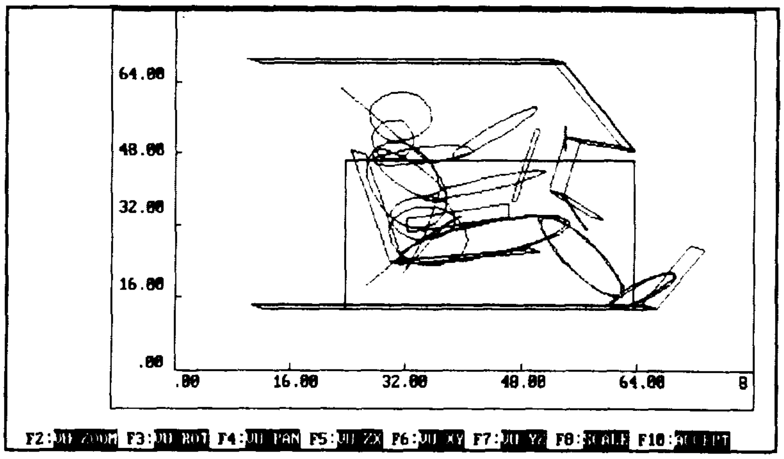


Figure 4: Initial Position Screen

A similar graphical procedure is used to create and position harness belts about occupant segments. The user can insert, delete and move points defining a belt using a fully graphical interface.

Integration and Output Parameters

In order to properly control the numerical features of the simulation model, the program requires the user to define several parameters. These are the initial integration step size, and the maximum and minimum integration step size and the length of the simulation.

Postprocessing

The DYNAMAN postprocessor allows the user to view the output in a variety of ways. One can get pictorial, graphical and tabular output information from the simulation module. The output can be tailored to one's requirements by defining the type of information needed, and the interval between two successive points when output is produced. Pictures from two simulations can be compared, e.g. to see the effect of varying a parameter. Plots from one or more simulations can be compared with experimental data directly through the postprocessor. Options are available to produce pictures and plots according a number of user defined formats. Hard copies of the pictures and plots can be made on laser printers and plotters.

APPLICATIONS

DYNAMAN has been used in a number of different areas by both government researchers and commercial clients. Some of the major areas of application are in accident reconstruction, as a tool for measuring injury potential during a crash event, and in the design of vehicle components such as seats and airbags, and

Reconstruction of Accidents

One major application of DYNAMAN is in the reconstruction of accidents. The flowchart given in Figure 5. describes the steps usually involved in setting up a simulation with the DYNAMAN occupant simulation program in order to model a real life accident.

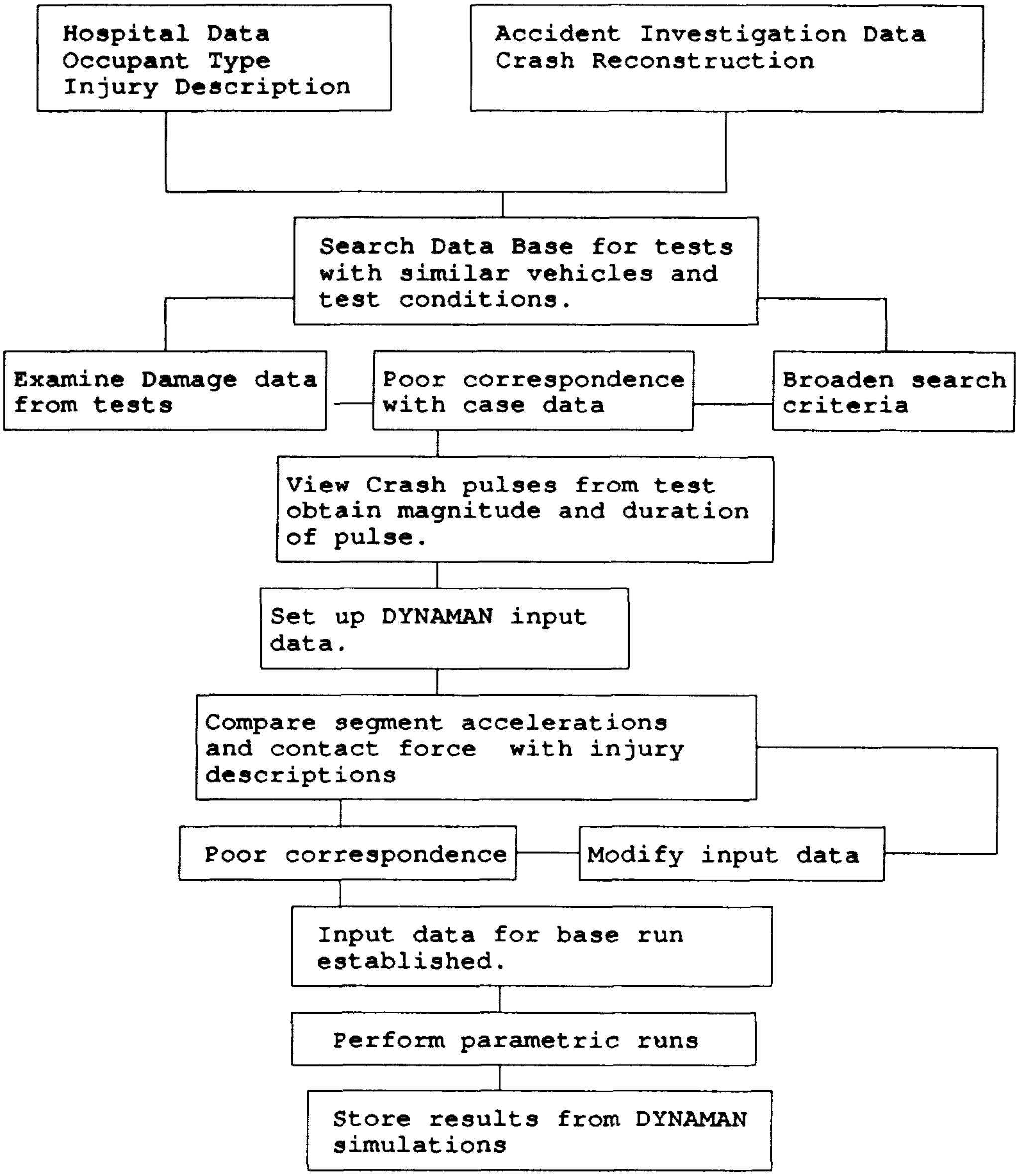

Hospital Data
Occupant Type
Injury Description
Accident Investigation Data
Crash Reconstruction
Search Data Base for tests with similar vehicles and test conditions.
Examine Damage data from tests
Poor correspondence with case data
Broaden search criteria
View Crash pulses from test obtain magnitude and duration of pulse.
Set up DYNAMAN input data.
Compare segment accelerations and contact force with injury descriptions
Poor correspondence
Modify input data
Input data for base run established.
Perform parametric runs
Store results from DYNAMAN simulations

Figure 5: Methodology for Accident Reconstruction

An example of an accident reconstruction simulation of a belted driver of a pickup truck which hit the side of another vehicle at about 40 mph. Figures 6 and 7 show the state of the occupant at a time just prior to the crash and 150 msec after the crash.

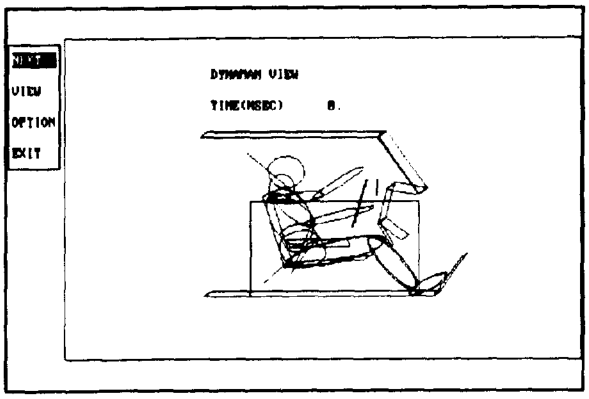


Figure 6: Side View at 0 msec

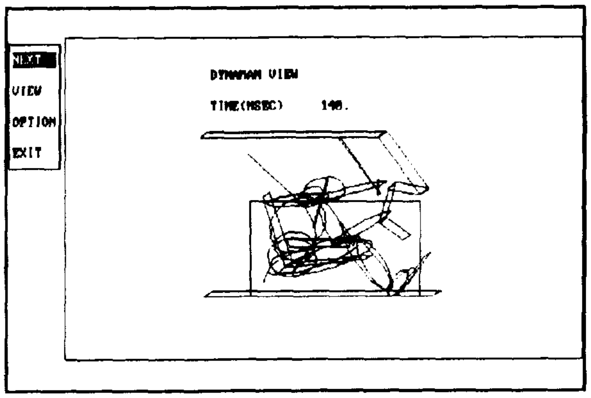


Figure 7: Side View at 150 msec

Injury Evaluation

DYNAMAN has been employed in evaluating the injury potential during a variety of crash events. Apart from vehicle crashes, it has been used in aircraft and helicopter crashes, as well as, pilot ejection. Figure 8 shows an example of the motion of helicopter pilot sitting in an energy absorbing seat during vertical crash. The position of the pilot and seat at two points in time are shown.

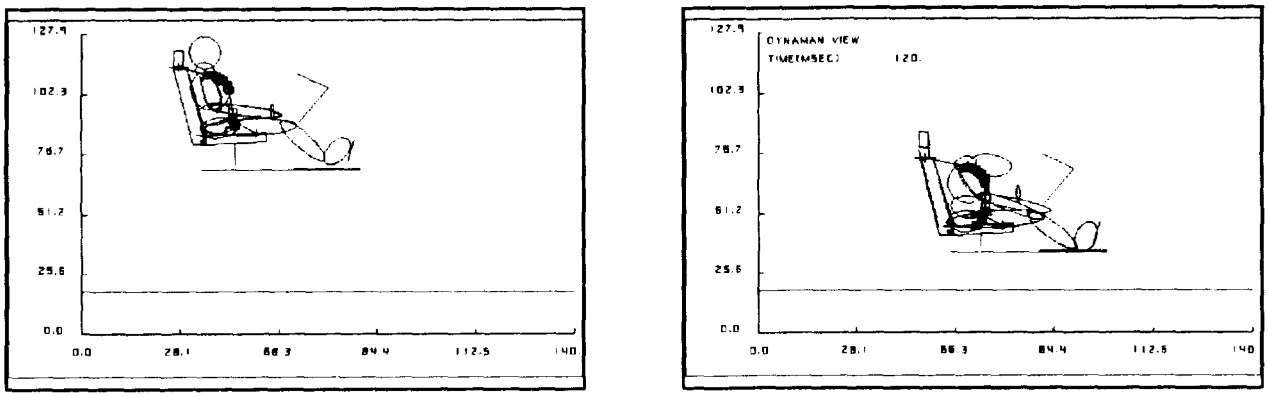


Figure 8: Simulation of Pilot during a Helicopter Crash

Figure 9 shows the motion of a pilot being ejected from an aircraft. Again, it shows the pilot and ejection seat at two time positions, but this time the two positions are superimposed on the same frame.

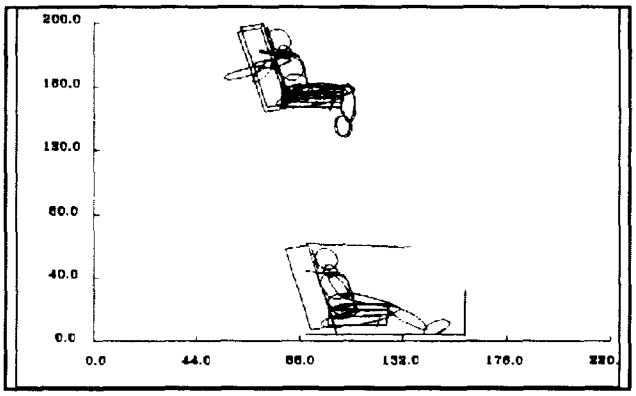


Figure 9: Simulation of Pilot Ejecting from Aircraft

DYNAMAN as Design Tool

DYNAMAN can be used as a tool for formulating the basic design of such devices as airbags, child seats, and belts. For such a purpose, the basic simulation is conducted with input based on known parameters for the system under consideration. A number of design parameters are then identified, and a series of simulations are done by varying the values of the specific design parameters. From the matrix of simulations, the set of design parameters that provide the best degree of safety with an optimum level of comfort are then selected for producing a prototype of the device.

As an example, in the design of a child seat, DYNAMAN can be used to determine appropriate ranges for child seat weight and geometry, its seat cushion characteristics and its restraint system. Figure 10 shows output from a simulation with the child seat at the point of maximum excursion.

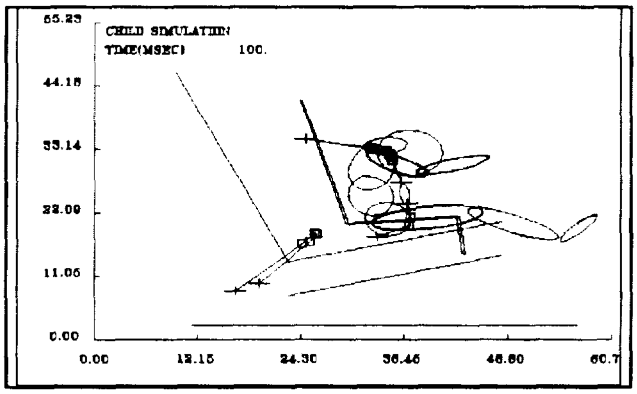


Figure 10: Simulation of Child in Child Seat

DYNAMAN has been successfully used in the design of a car seat which will undergo a motion during a crash event. The seat motion was designed to produce a lowering of the injury potential of the occupant, as compared to a non-moving seat. The seat motion can be optimized to work with other restraint systems such as two-point belts and airbags. Figure 11 shows a setup of an unbelted driver with a driver side airbag.

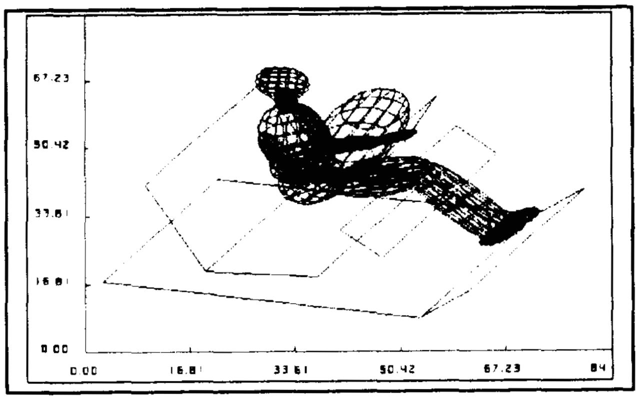


Figure 11: Simulation of Driver with Airbag

omit

DATA MANAGEMENT, STORAGE, AND PROCESSING PART 2

THE DATA EGG: A NEW SOLUTION TO TEXT ENTRY BARRIERS

Gary L. Friedman
Technical Group Leader
Advanced Engineering and Prototype Group
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109

ABSTRACT

A unit that allows text entry with only one hand has been developed, and holds the promise of allowing computers to be truly portable. It is unique in that it allows operation in any position, freeing the user from the traditional constraints of having to be seated near a desk. This handheld, chord-key-based unit can be used either autonomously for idea capturing, or tethered to a personal computer and used as an auxiliary keyboard. Astronauts, journalists, the bedridden, and anyone else normally barred from using a computer while on the job could also benefit from this new form of man-machine interface, which has been dubbed the "Data Egg".

INTRODUCTION

As computers continue to shrink in physical size and grow in raw CPU power and memory, a new problem has been thrust upon designers within the last five years: traditional keyboards dictate that the unit must retain the minimum physical dimensions of the keyboard, lest the keyboard becomes unusable. In short, the packaging becomes the machine's own I/O bottleneck. Examples of this limitation can be found in today's popular "pocket organizers", such as the Sharp Wizard, the Casio B.O.S.S. and Hewlett Packard's HP 95LX palmtop PC. In all these units, the inadequate keyboards severely curtail access to their otherwise powerful features.

Many attempts to attack the problem of general keyboard inefficiency have been made in the past, most notably by IBM who developed a new method of text entry for the stenographer called the Chordian keyboard. In addition, voice recognition, handwriting analysis, and several variations of the Dvorak keyboard (which re-arranges the letters in a standard QWERTY keyboard so that the most common letters are actuated by the strongest fingers) have all been studied.

Most of these keyboard alternatives do not address the newest technology-induced problem: how to insure that a computer's accessibility is not proportional to its size.

A UNIQUE SOLUTION: THE AgendA

One solution to the input problems of these smaller computers came from a firm in the UK called Microwriter Systems, plc. They've created a unique 'personal organizer' that could be operated with one hand. The device, called the 'AgendA', was similar in function to the popular Sharp Wizard and Casio B.O.S.S. organizers, and attempted a brilliant work-around to these products' biggest deficiencies: their tiny and unusable keyboards. (Figure 1.) The AgendA possessed seven large buttons (three for the thumb, one for each of the remaining fingers). Pressing different combinations of these buttons resulted in the generation of all characters, numbers, and commands necessary to enter and extract information. The alphabet was easy to learn, thanks to clever mnemonics and other memory jogs which associate finger position with the character's shape (see Figure 2).

The AgendA's biggest problem was that its design needlessly anchored the user to a desk and a chair, a fault shared by the rest of today's text-entry schemes. Despite its small size and portability, it cannot be used extemporaneously while walking, driving, or other times when thoughts pop into users' heads. A solution to this long-standing problem sprung up last year when I proposed a version of the AgendA that didn't require a flat surface. Basically, the seven-button scheme was kept intact and wrapped around a shape that was easy for the hand to hold, such as an egg.

The resulting solution, dubbed the 'Data Egg', turned out to solve many problems unaddressed by today's technologies. When used by itself, it can capture ideas that pop up while in transit - ideas that would normally evaporate by the time one got around to writing them down. When attached to a PC and another display device, it can provide a superior computer interface to those who are bedridden. The resulting 'Bedridden Workstation' has also been prototyped, and is described later in this article.

To date, two working versions of the Data Egg have been constructed. The software which drives it accommodates both the autonomous and tethered modes described above, and provides a general framework for application expansion.

THE EGG EVOLVES

The original Data Egg idea took a plastic Easter egg and glued seven buttons and a strap around it, a device custom-tailored to my hand (see Figure 3). After months of typing on it, the egg-shaped device soon was deemed too bulky and the strap, although helpful, was a nuisance. The newer version resembles a beeper, and is worn on the belt when not in use, always handy.

Autonomous Mode

Often while working on important projects, one of my biggest frustrations would be that, at the most unpredictable times, my already over-burdened mind would come up with the infamous "Oh, one more thing..." or "Whoops! I forgot to...". These thoughts usually occurred during inconvenient times, such as while driving or walking to and from the office. The big irony was that during those times I had a laptop computer close by in my briefcase, but I couldn't access it because I wasn't sitting down and immobile.

Using the Egg's Autonomous mode, it is possible to capture just about any idea regardless of the activity. The user's eyes never have to leave what they're doing. When a mindstorm occurs at three in the morning, the Data Egg allows the semi-conscious mind to record thoughts with a minimum of movement and effort, something not possible with pen and pad.

After a year of practice, my "typing" speed on the Data Egg has hit an average of 30 words per minute. While this can in no way compete speed-wise with conventional typing or talking into tape recorders, it does provide a silent and non-burdening alternative to these standard solutions.

I have personally used the Data Egg in the field for over 18 months to capture my random ideas as well as important factiods that come up in conversation. It also allows me to type complete memos and letters during my otherwise monotonous commute to and from home.

Bedridden Workstation

The other Data Egg mode, 'tethered', provides for the Bedridden Workstation, and allows those lying down to have complete access to a standard PC and all the software it runs. The idea was inspired about two years ago when a fellow programmer had back trouble and wasn't able to use a computer until he recovered.

Using a computer while lying down is a pain. The head must be propped up by a pillow to see the screen, and the keyboard has to rest on the user's stomach, which requires the hands to type at incredibly fatiguing angles. Realizing that it shouldn't be necessary to have a healthy back in order to use a computer, the environment illustrated in the top of Figure 4 was envisioned.

The Bedridden Workstation is formed by tethering the Data Egg to a larger computer, and incorporating an innovative display device (described below) for full-screen feedback. In use, text is typed in with the hand lying comfortably at the user's side, while a TSR (Terminate and Stay Resident) program on the PC takes the ASCII and function codes generated by the Data Egg and "presses" the appropriate character on the computer's keyboard. The TSR program, written in Turbo Pascal, is general enough to allow popular software like Wordperfect (which uses obscure ALT- and SHIFT-Fn key combinations), Lotus 1-2-3, Procomm, and Framework to be used by remote control.

The key component to the Bedridden Workstation is the Private Eye display device. Rather than placing a CRT in front of the user, the Private Eye instead places a small box an inch in front of the user's eye, which projects a virtual image of the PC's screen that "hovers" about five feet in front of the user (but remains invisible to outside observers). When combined with a PC and the Private Eye, the Data Egg allows the user to perform information editing in addition to the information capturing possible in Autonomous mode. This combination of peripherals finally allows the bedridden to have comfortable access to the PC and all of its software.

CONSTRUCTION AND EVOLUTION

As previously described, the first working Data Egg model consisted of a plastic Easter egg and an elastic strap to secure the egg to my hand while typing. A small cable took the pushbutton signals to a small, battery-powered single-board computer, which wasn't quite small enough to fit inside the egg.

Although this setup made for a convenient development environment, some of the egg's drawbacks soon surfaced. The elastic strap made it easy to type while lying down, but also made it difficult during everyday life to quickly put on and take off. This is good for certain applications, but to create an instantly accessible tool, it became clear that a new design without a strap would be necessary.

The Egg also had another problem: the shape was far from universal when it came to accommodating different hand sizes, and could be used only by those who were right-handed. These two drawbacks would be disastrous if such a one-handed text entry product were to become a viable commercial product.

Clay models

About a dozen clay prototypes of shapes were created to try to solve the problems of accommodating two hands without a strap, and yet be just as comfortable to use. Figure 5 shows some of the shapes which emerged from the brainstorming phase.

One of the experimental shapes didn't rely on fingertips at all, but rather was actuated by the first joints of the fingers (plus two buttons for the ball of the thumb). This meant that a smaller shape could accommodate a larger diversity of hand sizes, and make the device both easier to hold and less cumbersome to carry. The clay model demonstrating this concept eventually led to the development of the "Data Beeper", shown in Figure 6.

The Beeper Hardware

The Beeper consists of a hollowed-out Motorola pager, with finger buttons along its long edge and three buttons on top for the thumb, one of which is actuated by the thumb's first joint. All of the buttons have been physically customized to some extent, giving just the right "feel" and travel to ensure that the beeper could be securely held without accidentally typing a character.

Inside the beeper is an 8051-derivative CMOS microprocessor, 32 KB of battery-backed, non-volatile RAM, an analog-to-digital converter for checking supply voltage, a serial port, and two LEDs for local feedback. Located on the top plate of the unit, these LEDs indicate battery strength, confirmation of commands, and acceptance of text input. Because society instills a great deal of power in personal pagers, the beeper version of the Data Egg also includes a "beeper", an audio oscillator that emulates the personal pagers and empowers its owner to escape boring meetings.

The only thing missing from the beeper design is the inclusion of a liquid crystal display for local feedback; this was left out of the first prototype because of space constraints and the extra software complexity it would entail. It will definitely be included in the next version.

Normally, when the device is being carried, it is only used to generate ASCII text. The alphabet has been expanded, however, to include every character and key combination recognizable by the BIOS of an IBM PC for the times it is tethered to a desktop computer. (To bypass the 128-character limitations inherent in the seven-button scheme, some of these combinations are achieved via two-keystroke commands.) Six additional commands are responsible for switching modes, dumping text to the built-in serial port, clearing memory and running diagnostics. An on-board analog-to-digital converter also keeps track of the battery's voltage, and gives about a week's worth of warning before it goes "dead" (below 5.2 volts) by flashing either a red or green LED during power-up.

WHY NOT A TAPE RECORDER?

The Data Egg has a few advantages over using a pocket tape recorder, which is the current tool of choice for people who work in creative fields:

- No transcribing is needed to achieve paper output (although the text often has to be polished once downloaded to a PC),
- Can interface directly to a computer (when tethered to a PC, as in the Bedridden Workstation),
- Discreet operation. The user doesn't call attention to himself while at the symphony.

The problem of sorting the idea fragments once they get downloaded into a PC is greatly aided by a software package called Lotus Agenda (not related to the original Microwriter AgendA from which the Data Egg idea came.) It automatically searches the input stream and recognizes notes to make calls, meet with people, and even picks out familiar names. It then allows all this linked information to be sorted and viewed in many different ways. The two make an ideal team for taking scraps of thoughts and turning them into useful lists.

In the long run, I believe that voice input, coupled with computer conversational skills being developed by linguist researchers worldwide, will be the ultimate in "intuitive" and friendly user interfaces in the future, with no typing skills to master and no narrow command sets to memorize. The Data Egg is not meant to compete with voice recognition technology. Rather, I see it as a useful complement to facilitate the quiet capture of ideas while away from the workplace.

FUTURE PLANS

There are many other features that an ideal computing companion should possess, but which modest fabrication resources preclude. Some of the items on my wish list, in order of importance, include:

Incorporate a multi-line LCD screen for local feedback.

- Embed a speech synthesizer chip, which would provide speaking-impaired individuals with a text- or phoneme-based synthesizer that is not bulky and cumbersome as are today's offerings.

- An improved shape which is operated by the first joints of the fingers (and the ball of the thumb); this would make it easier to hold, smaller, and more likely to accommodate different hand sizes.

- Integrate a mouse function into the hand-held device as shown in Figure 7; it will then be possible to operate mouse-based applications with just one hand instead of three.

- Software improvements:

  - Password protection for secure files.

  - Routine for pocket modems. The user should be able to hook the Data Egg to a phone line and have it automatically log in and download your notes to your main computer.

    - Clock/calendar/alarm; which would provide standard alarm/programmable timer functions in addition to the current ability of providing a date and time stamp on all data dumps.

    - Far future: When memory becomes even denser than it is today, the Egg's information capturing abilities should be expanded to include digitized voice recording and, with the inclusion of a lens and charge-coupled device array, a point-and-shoot electronic imaging camera. The ability to capture text, sound, and images would make it the dream tool of a journalist or anyone wishing to easily document the times of their life.

CONCLUSION

Over the past year and a half, the Data Egg's Autonomous mode has become as important for me as the pocket tape recorder is for poets, producers, and those in other creative professions. About 90% of its value lies in its instant accessibility, the other 10% in its discreetness of operation.

Let me emphasize that I am not of the mindset that "every spare minute must be filled up with something productive or I'll explode". Mentalities like that only serve to raise blood pressure and reduce the quality of life outside the office. The Data Egg is driven by quite the reverse philosophy: If your mind is going to be racing with a billion ideas anyway, it would be a waste to allow them to evaporate. As long as my brain insists on coming up with important thoughts at inconvenient times, I will continue to want such an idea-capturing device at my service.

- 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. Special thanks to Jeff Shaw from the Art Center College of Design for his work on the more marketable shape.

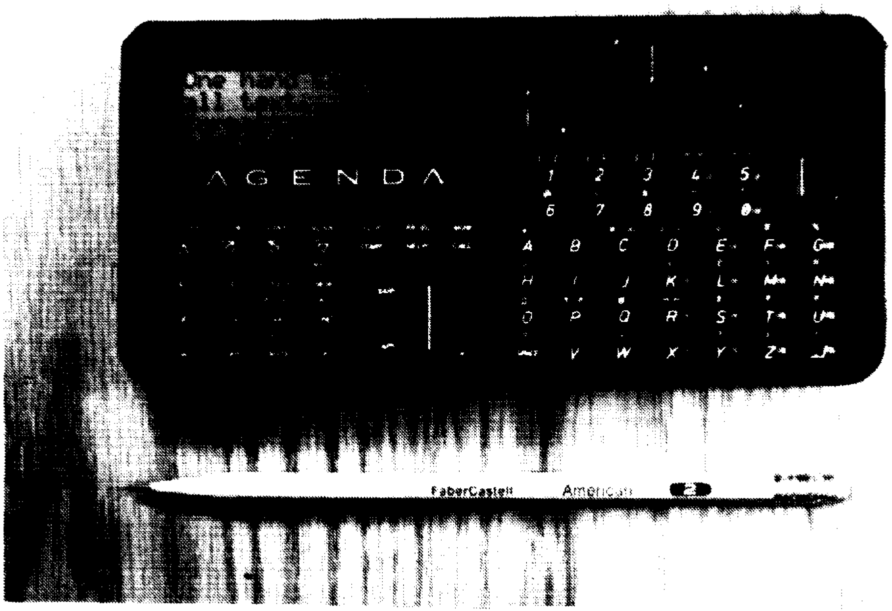
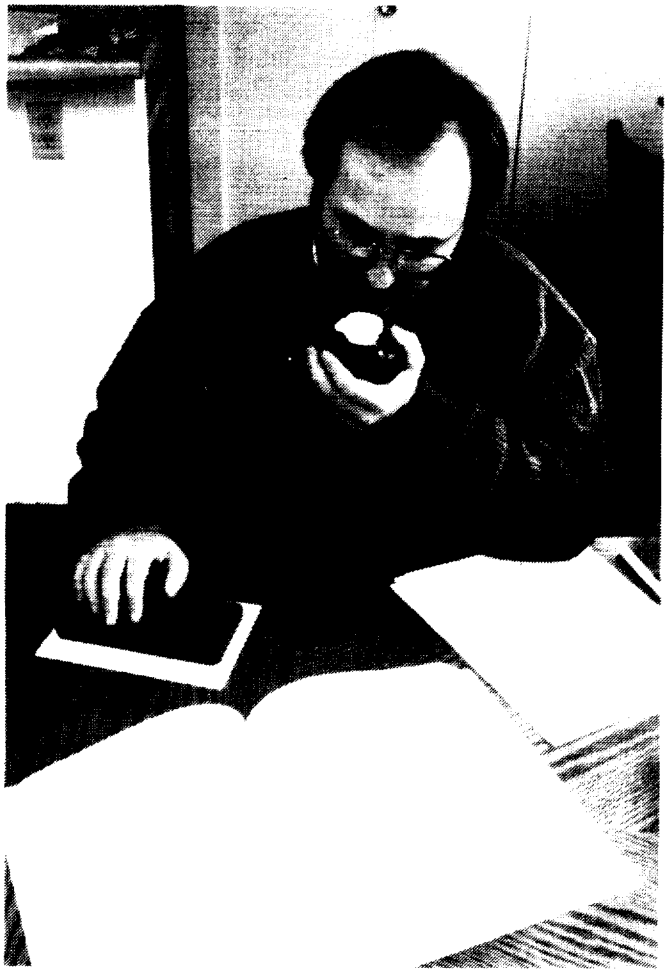
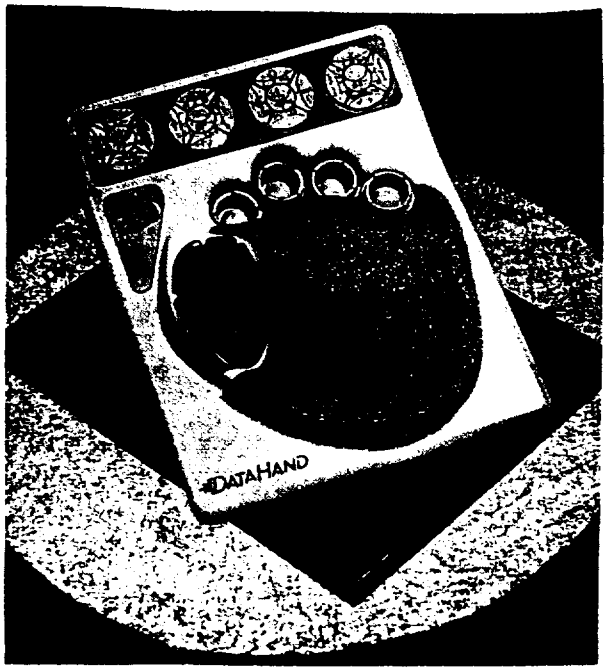


Figure 1
The AgendA's (above) method of typing allowed simultaneous support of scholarly and culinary activities. All single-handed solutions to date (including Industrial Innovations' experimental "Data Hand", below) share the same arbitrary limitation: they can only be used while sitting down.





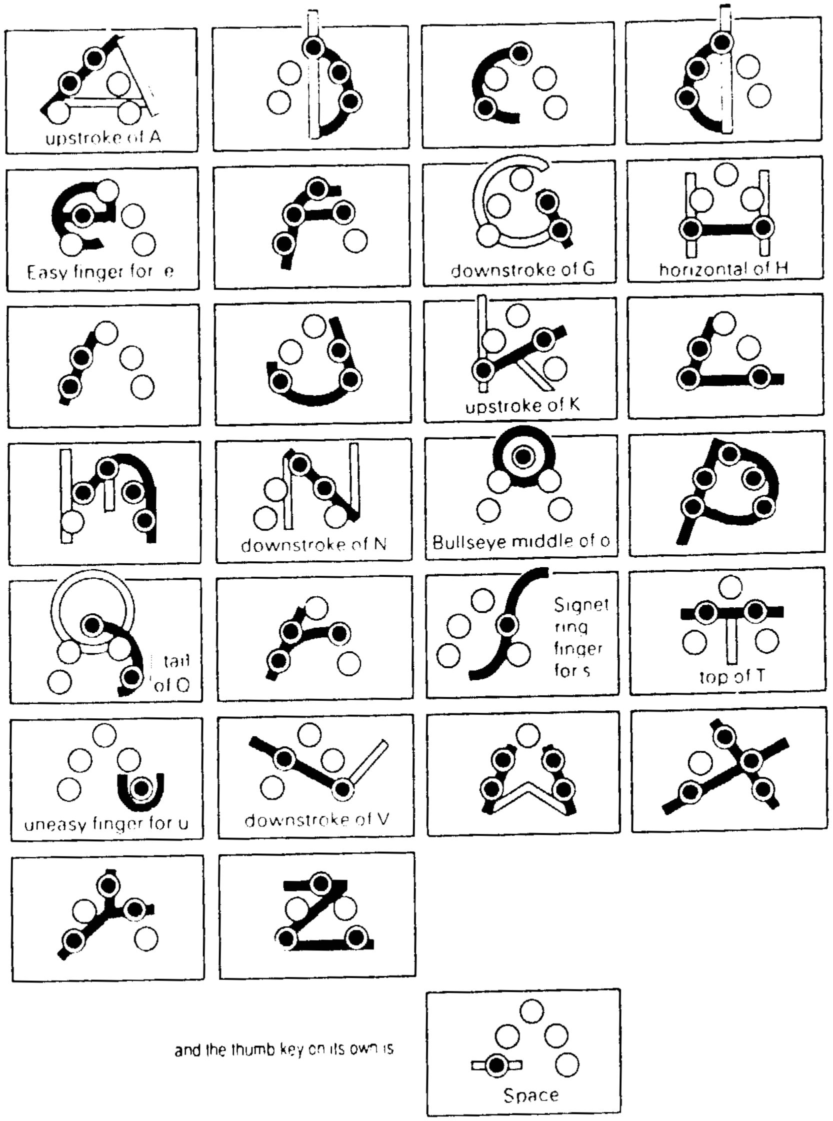


Figure 2
A sample of the Microwriter Alphabet

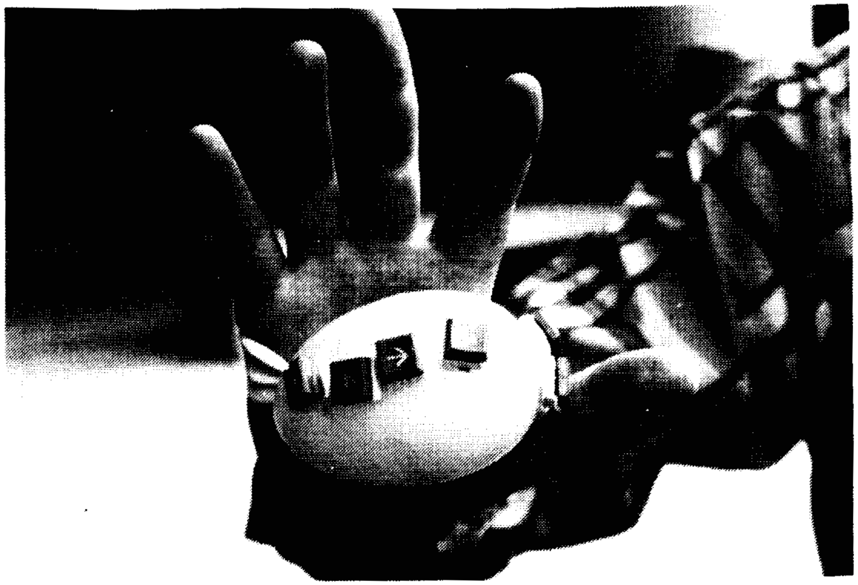


Figure 3
The Data Egg, a one-handed text entry device, consists of seven buttons wraped around a shape that's easy for the hand to hold. Pressing different combinations of these buttons (three for the thumb, one for each of the remaining fingers) allows typing in any position, free of the historic positional constraints of desk and chair.

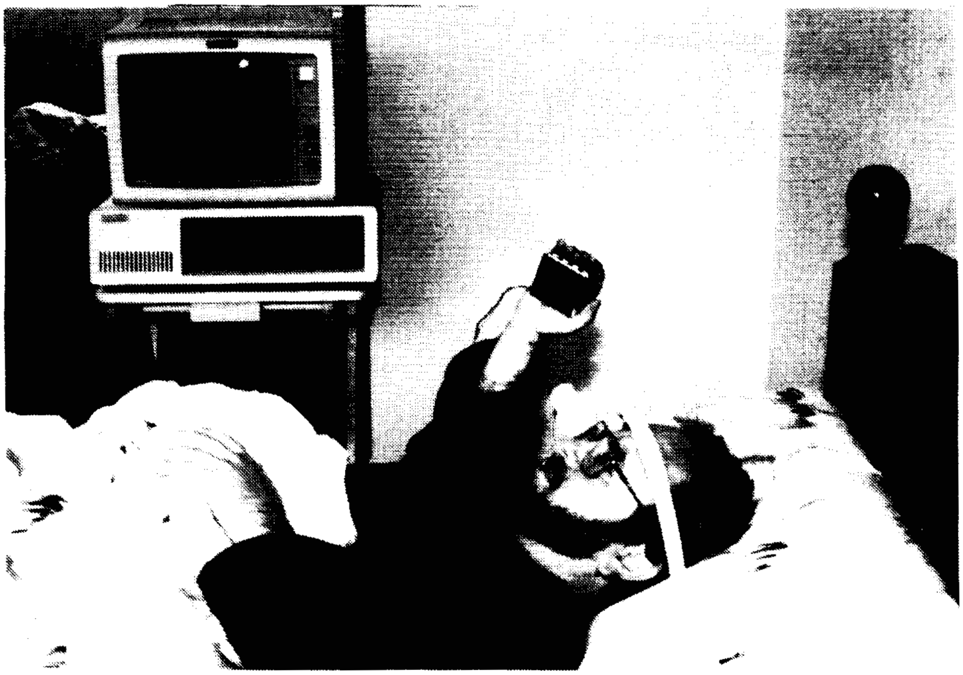
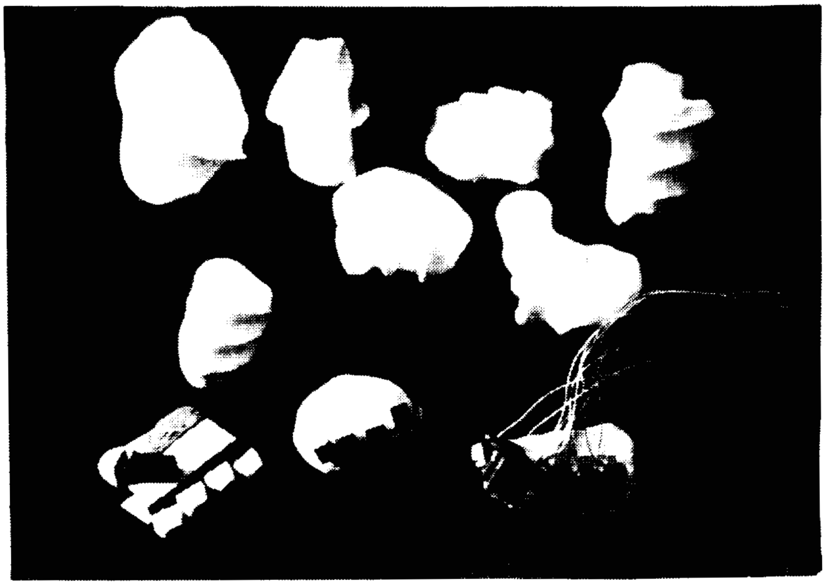
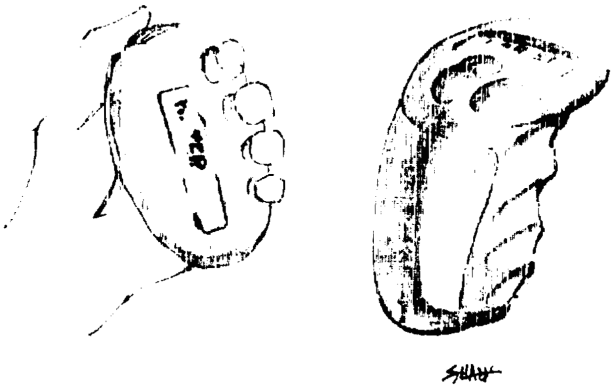


Figure 4
A computer workstation for the bedridden emerges when the Data Egg is combined with a virtual display device called the Private Eye. Text is typed in via one hand lying at the user's side, while a TSR (Terminate and Stay Resident) program on the PC "presses" the appropriate character on the computer's keyboard. The Private Eye projects a virtual image of the PC's screen which "floats" about five feet in front of the user. The resulting Bedridden Workstation allows those with back problems to have complete access to any commercial software for the PC.

Figure 5
Examples of shape ideas resulting from the brainstorming phase.

(Drawings and fiberglass model designed by Jeff Shaw, Art Center College of Design.)





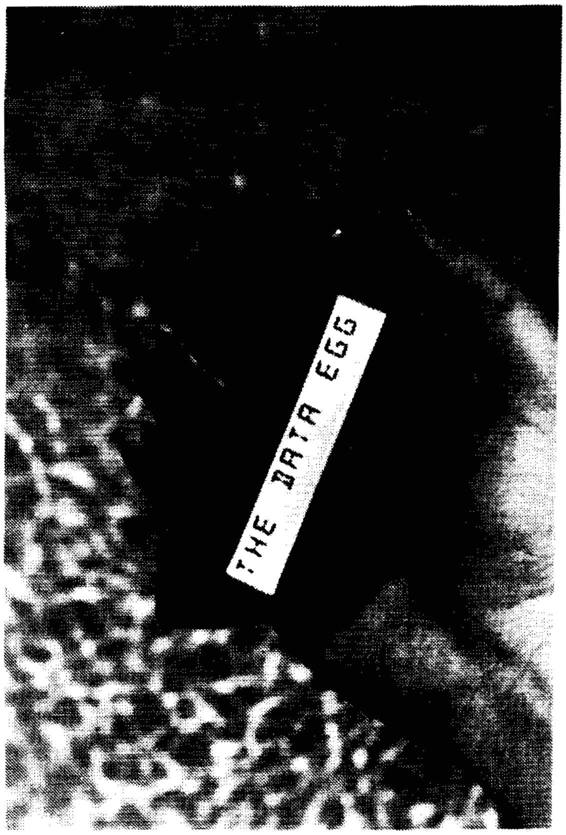
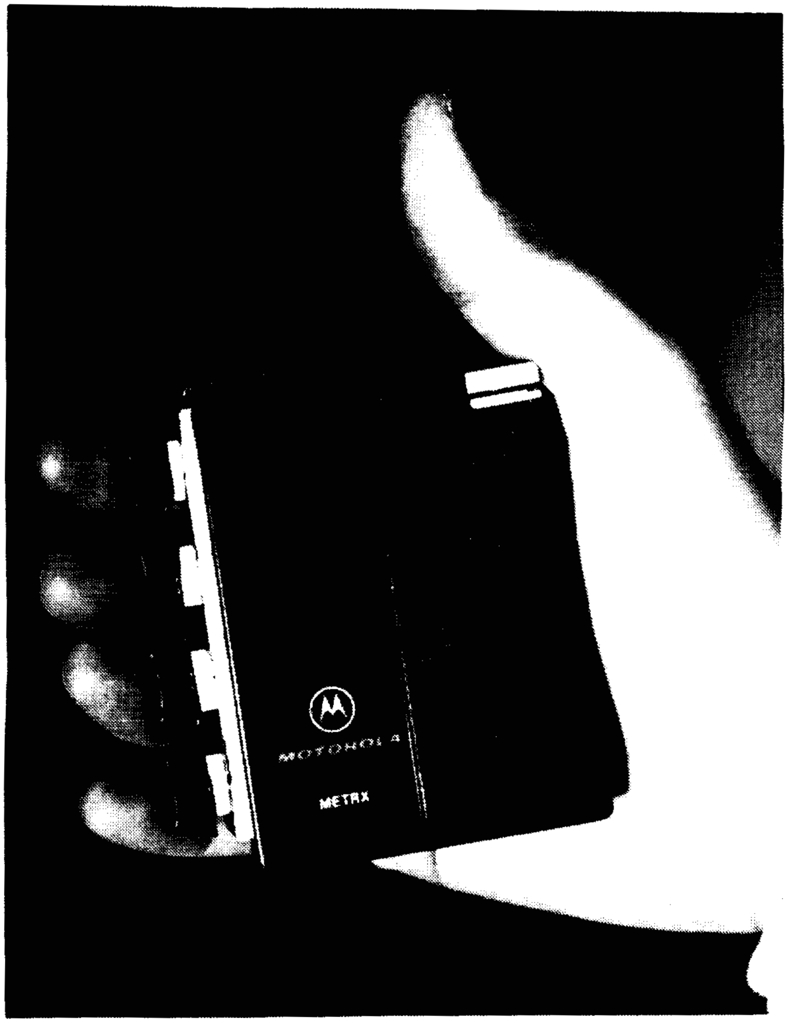


Figure 6
A functioning prototype of the Data Egg is disguised as a beeper, which is a socially acceptable device to carry. Inside the unit is an 8051 microprocessor and 32K of non-volatile RAM, which can capture text and download it to a computer via a bult-in serial port. Because of its enhanced portability, the Data Egg can capture ideas wherever the user might be; ideas that would normally evaporate while walking, driving, or resting.



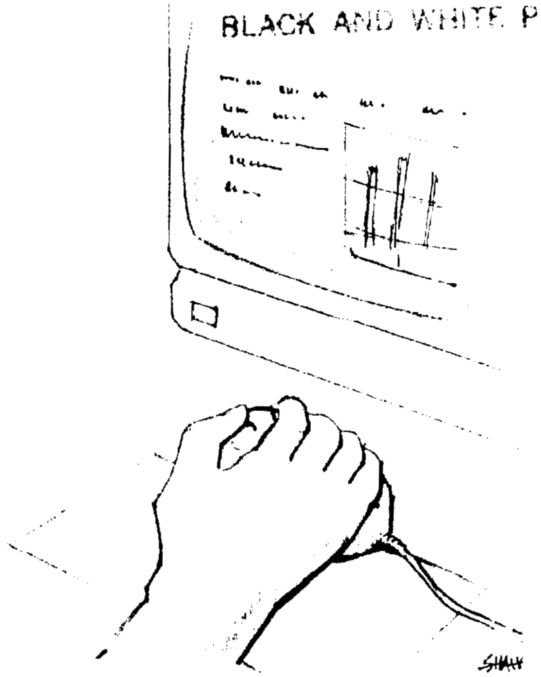


Figure 7
Data Egg/Mouse hybrid allows the use of a graphical user interface with one hand instead of three.

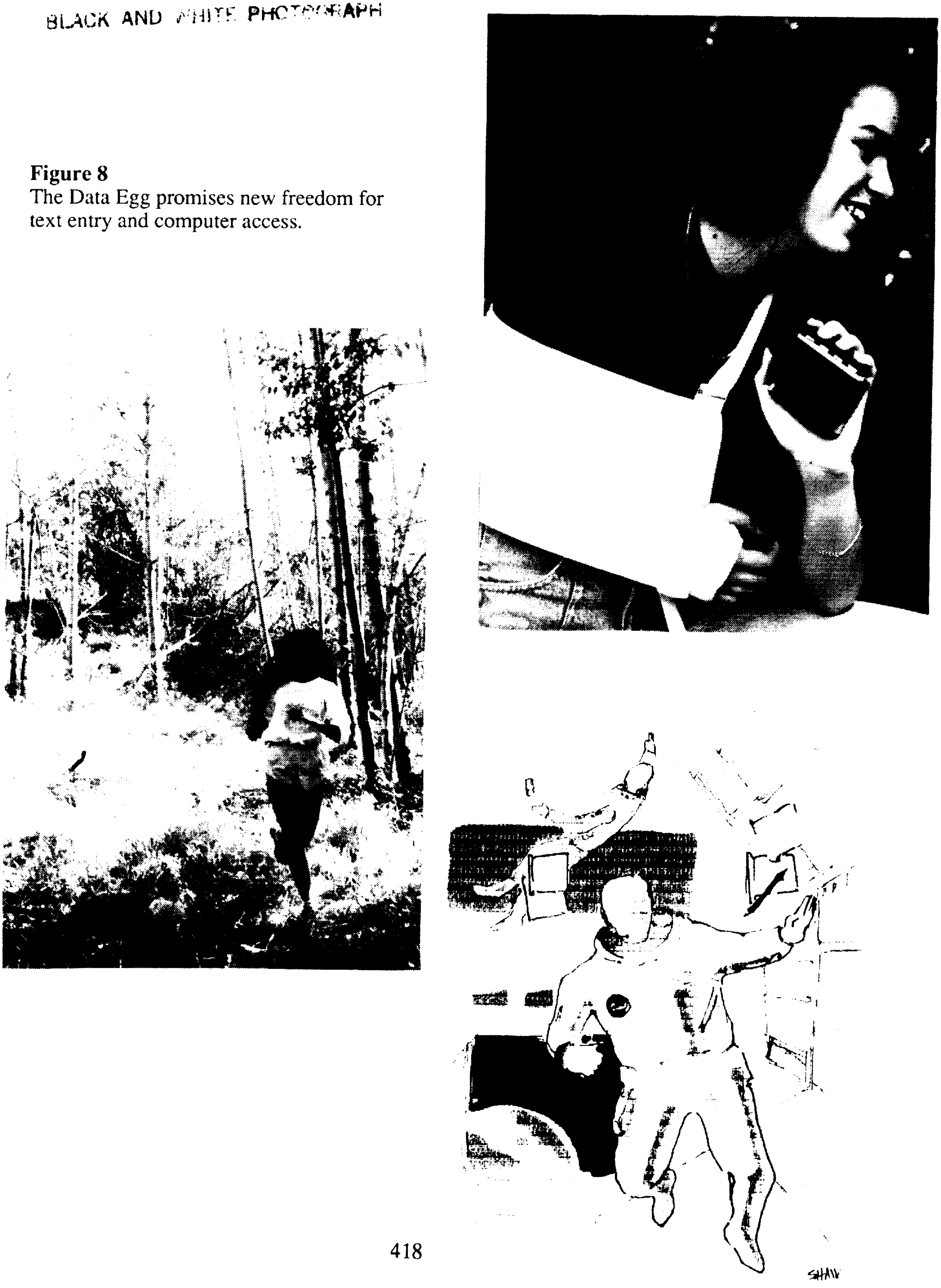


Figure 8
The Data Egg promises new freedom for text entry and computer access.

MIRAGE: THE DATA ACQUISITION, ANALYSIS & DISPLAY SYSTEM

Robert S. Rosser
NASA Johnson Space Center
GE Government Services
Houston, TX 77058

Hasan H. Rahman
NASA Johnson Space Center
GE Government Services
Houston, TX 77058

ABSTRACT

Developed for the NASA Johnson Space Center Space and Life Sciences Directorate by GE Government Services, the Microcomputer Integrated Real-time Acquisition Ground Equipment (MIRAGE) system is a portable ground support system for Spacelab life sciences experiments. The MIRAGE system can acquire digital or analog data. Digital data may be NRZ-formatted telemetry packets or packets from a network interface. Analog signals are digitized and stored in experiment packet format. Data packets from any acquisition source are archived to a disk as they are received. Meta-parameters are generated from the data packet parameters by applying mathematical and logical operators. Parameters are displayed in text and graphical form or output to analog devices. Experiment data packets may be retransmitted through the network interface. Data stream definition, experiment parameter format, parameter displays, and other variables are configured using spreadsheet databases. A database can be developed to support virtually any data packet format. The user interface provides menu- and icon-driven program control. The MIRAGE system can be integrated with other workstations to perform a variety of functions. The generic capabilities, adaptability, and ease of use make the MIRAGE a cost-effective solution to many experiment data processing requirements.

INTRODUCTION

This paper describes the overall design, major features, and possible applications of the Microcomputer Integrated Real-time Acquisition Ground Equipment (MIRAGE) system. The MIRAGE system provides a portable, self-contained unit capable of data acquisition, monitoring, analysis, archival, playback, and network transmission. The MIRAGE can acquire RS449 synchronous serial NRZ-formatted Spacelab downlink telemetry data transmitted from a High-Rate Demultiplexer Interface (HRDI) or as Consultative Committee for Space Data Standards (CCSDS) packetized data from a network interface. Analog input signals can be acquired and inserted into data packets. Meta-parameters are generated from the data packet parameters by applying mathematical and logical operators. Data parameters are displayed in text and graphical form in a Macintosh window environment. Selected parameters are output to strip chart recorders or other analog devices through a digital-to-analog interface. Experiment data packets may be transmitted through the network interface. The MIRAGE also accepts IRIG-A formatted time input through a Macintosh serial port. Data stream definition, experiment parameter formats, and other variables are read into the program from spreadsheet databases. The Macintosh user interface provides menu- and icon-driven program control. Experiment data acquisition and processing are supported during baseline data collection, experiment hardware bench testing, and real-time support of flight experiments. Archived data are played back and analyzed postflight. The MIRAGE can be integrated with other data acquisition and analysis systems to perform a variety of experiment data processing functions.

The MIRAGE is being used initially to support the Baroreflex experiment on the German D-2 Spacelab mission (STS-55) scheduled to fly in February, 1993. The objective of the Baroreflex experiment is to measure the sensitivity of the carotid sinus baroreceptor reflex during spaceflight to determine the effect of weightlessness on normal cardiovascular reflex control mechanisms. Principal Investigators at the German Space Operations Center will use a MIRAGE system for real-time data acquisition, display and analysis during the mission [1]. The MIRAGE will also be used in bench testing of experiment hardware, and during preflight and postflight baseline data collection. It will also be used to play back, analyze, and transform archived data.

OVERALL DESIGN

Design Considerations

The original design specifications for the MIRAGE system were created to support the Baroreflex experiment. During the early design phase, it became clear that by selecting the right hardware and using a modular, object-oriented approach to software development, the MIRAGE could become a flexible, powerful system capable of operating in a wide range of data acquisition environments. Some of the desired features of the MIRAGE system that were combined to meet this goal are listed below.

- Easy to use
- Easy to maintain and modify
- Possess real-time, multifunction capabilities
- Acquire data from several sources
- Support multiple experiment data streams
- Generic data displays to handle a wide variety of data
- Use preexisting software and off-the-shelf hardware where possible
- Provide real-time and post-time data analysis
- Provide data playback capability

Input Data Format

The original design specifications for the MIRAGE required that it support the acquisition and processing of a synchronous serial NRZ-formatted data stream generated by an experiment payload microcomputer at a minimum bandwidth of 32 kilobits per second. The data stream is formatted into High-Rate Multiplexer (HRM) frames [2]. In this format the data bits are formatted into 12 or 16 bit words. The words are grouped into minor frames. A minimum of four minor frames are grouped into major frames. Each minor frame begins with a standard 6-byte header. The first 32 bits of the header are a 24 bit sync word and 8 bit minor frame number used for frame synchronization. Data parameters are stored in the remainder of the minor frames. Data parameters may or may not be major-frame repetitive; ones that are not repetitive are indicated by bits set to indicate the presence or absence of particular parameters in the major frame. Upon acquisition by a ground system, major frames are grouped into packets of one or more major frame per packet. The MIRAGE system can be configured to acquire packetized digital data in other formats.

Up to sixteen analog channels can be input into the MIRAGE system. Samples of the signals are digitized and stored in digital data packets.

Hardware

The MIRAGE system was initially developed on a Macintosh IIfx platform. The MIRAGE software will run on any Macintosh with a Motorola 83020 or higher processor and at least 2 MB of internal RAM memory; however, at least five NuBus slots are necessary to install the boards required for a full-function MIRAGE system. Macintosh system software 6.0.5 or higher is required. A HRDI box is necessary to acquire Spacelab downlink telemetry data.

Digital and analog data acquisition and analog data output are supported with four NuBus boards manufactured by National Instruments [3]· An NB-DIO-32F is used to receive data directly from a HRDI. A slight modification to this board is made to provide handshaking capability with the HRDI. An NB-MIO-16L provides up to eight differential or sixteen single-ended analog input channels. An NB-AO-6 analog output board provides up to six channels of analog output. An NB-DMA-8-G provides Direct Memory Access (DMA) transfers to speed up the digital and analog acquisition processes and the analog output process. National Instruments provides low-level drivers for the boards.

An Ethernet controller card is used to support data acquisition and transmittal over Ethernet using DECnet protocol. An 8 bit, 256-color graphics video board and color monitor (16 inch or larger preferred) provide high-resolution graphics display. An internal hard disk is used as a boot disk, and also holds the MIRAGE software. An external, removable-cartridge Small Computer System Interface (SCSI) disk is used for data archival. A Graphtec WR7700 eight-channel analog strip chart recorder is used to provide hardcopy strip charts. Any printer connected to the Macintosh locally or on a Local Area Network (LAN) can be used for printed output.

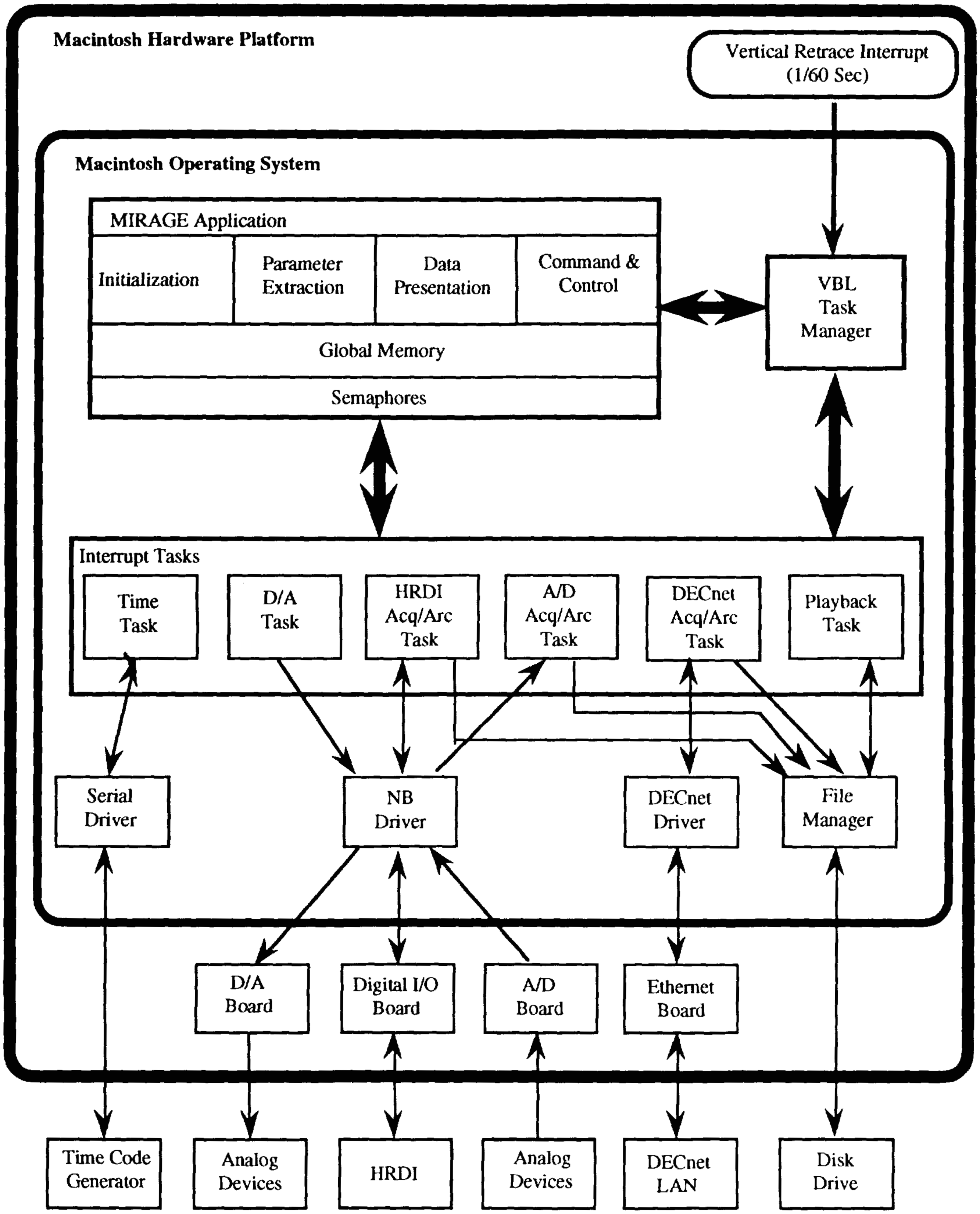


Figure 1: The MIRAGE System Architecture

Software

The MIRAGE system software was developed on the Macintosh platform using Apple's Macintosh Programmers' Workshop C. A modular, object-oriented approach to software design was used to assure ease of modifiability and maintainability. Necessary use of the Macintosh toolbox, however, means the MIRAGE software is not directly transportable to other systems such as DOS-based Personal Computers.

The MIRAGE software makes use of the vertical retrace interrupt service provided by the Macintosh toolbox to support real-time functionality. The MIRAGE software also uses National Instrument's library of function calls to control the NB boards. DECnet for Macintosh software is used to provide DECnet protocol interface to the Ethernet controller hardware. A third-party library of charting functions integrated with the MIRAGE software to provides X-Y plots.

See Figure 1 for an illustration of the MIRAGE software and hardware architecture.

Configuration Database

The MIRAGE configuration database is created using Microsoft Excel. The database consists of several tab-delimited text spreadsheets, arranged into folders on the disk the MIRAGE software resides on. The MIRAGE database is used to define the MIRAGE system defaults (fonts, colors, etc.), experiment-specific hardware configuration, data stream format (stream data rate, packet frequency, etc.), acquisition defaults, experiment parameter format (packet location, extract masking information, etc.), display format, and analog input and output characteristics.

Data Flow

Figure 2 gives a representation of the data flow through the MIRAGE system. There are three external sources: the Macintosh user interface, the experiment database, and the experiment data source. The user enters experiment stream, parameter, and display data into the experiment database. Through the Macintosh interface to the MIRAGE application, the user controls at runtime the experiment data source and other application functions. The user can choose the HRDI, DECnet, or analog acquisition functions. Data can also be played back from an archived disk file.

The data acquisition portion of the program reads data from the specified external data source and, based upon the contents of the stream database, extracts and stores the major frames in the primary buffers. If the user has the archival function turned on, the primary buffers are read by the archival process, a header is generated, and the data packet is written to the archive file.

The parameter extraction function then extracts the data values for each parameter specified in the display databases from the primary buffers. The parameter database is used to locate and process the data values. The extracted and processed data values are stored in the parameter buffers.

If analog output is enabled, the data values for the parameters to be output are extracted from the parameter buffers and stored in the analog output buffer. The buffer is then passed to the analog output process.

For each display and graph window defined in the display databases, the data values to be displayed are extracted from the parameter buffers and output in the specified window in textual or graphical form based upon input from the display databases.

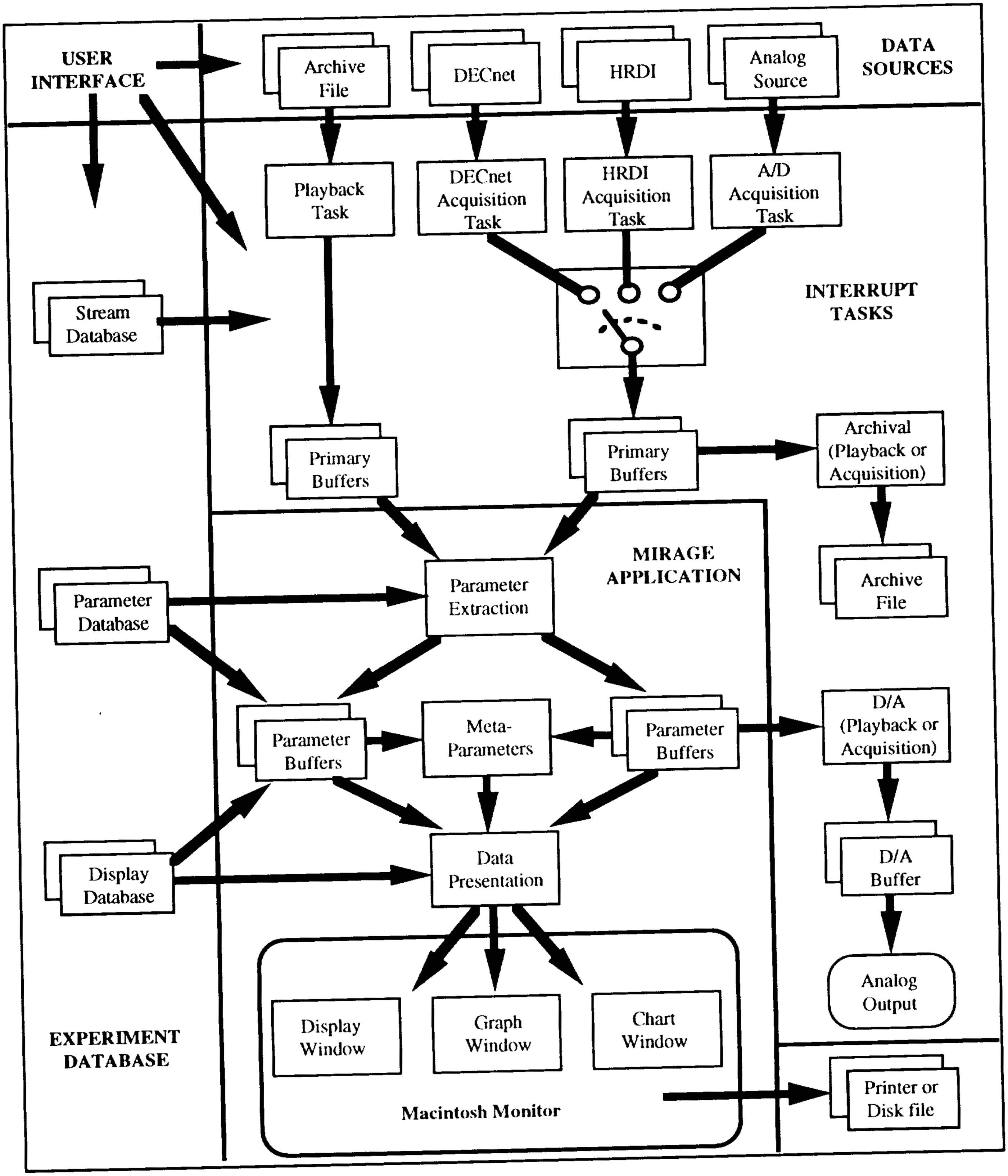


Figure 2: Data Flow Through the MIRAGE System

SYSTEM FEATURES

Acquisition

The MIRAGE system can acquire data from three interfaces: HRDI, DECnet, and analog.

HRDI Acquisition

The MIRAGE system uses the HRDI interface to acquire data during bench testing of Spacelab experiment payload microcomputers and during flight of a life sciences Spacelab mission. During flight, several data streams generated by experiment payload microcomputers are multiplexed to form the HRM Spacelab downlink telemetry data stream. On the ground, hardware external to the MIRAGE system demultiplexes this data stream into the separate experiment data streams, or channels. The HRDI interface can handle one of these channels at a time. During bench testing, the data stream generated by the experiment payload microcomputer can be connected directly to the HRDI.

The HRDI acquisition is achieved using an external HRDI box connected to a NB-DIO-32F input/output card via a 50-pin ribbon cable.

The HRDI is an interface board that resides in a box with its own bus and power supply developed with funding from the General Services Administration by GE Government Services at JSC [4]. The HRDI card receives balanced current serial signals and clock from the experiment payload microcomputer on two twinaxial cables. The line receivers compatible with this input are employed to convert current to standard transistor-to-transistor logic voltage levels. The input data is clocked into an Input Shift register whose length is 32 bits. This length is determined from the requirement to align itself to a 32 bit field (24 bit sync pattern + 8 bit frame count) for establishing and maintaining frame synchronization. The synchronization word for the experiment stream is set using thumbwheel switches. Some operational characteristics of the HRDI are preprogrammed by the host computer through a latched 16 bit Digital Output port. The HRDI board has an effective bandwidth of 512 kilobits per second.

All data transmitted by the HRDI flow through a 64 word first in, first out (FIFO) buffer (a 2K word FIFO buffer is currently available). The output from the HRDI is a sequence of 16 Bit words along with a transfer REQUEST signal. This REQUEST signal generates a DMA request for the host computer. The host acknowledges this REQUEST with a separate ACKNOWLEDGE signal, which also resets the old REQUEST signal.

The HRDI board requires two identical 40 pin ribbon cables compatible with the DEC DR11-W interface module. A Mac-to HRDI Interface board interfaces between the 50-pin NB-DIO-32F connector and the HRDI board.

The National Instrument NB-DIO-32F card is modified to interface with the HRDI board. The NB-DIO-32F is interfaced to the National Instrument Real-Time System Integration (RTSI) bus so that DMA transfers to Macintosh memory can occur using the NB-DMA-8-G. The 32 lines of digital I/O of this board are divided into four bytes, each of which can be programmed to function as input or output. The maximum transfer rate is 360K 32 bit words per second, more than adequate to support life sciences experiment data bandwidth. This card uses one DMA channel from the NB-DMA-8-G board.

DECnet Acquisition

During DECnet acquisition, a ground acquisition computer receives the HRM data stream, formats it into CCSDS packets, and transmits the packets to the MIRAGE node. The data transfer between the ground acquisition computer and the MIRAGE is accomplished over an Ethernet LAN using the DECnet transparent task-to-task communication services.

The MIRAGE system can acquire network packets transmitted over Ethernet using DECnet protocol. An Ethernet controller card in the Macintosh allows it to connect to thinwire or thickwire Ethernet media. Digital's DECnet for Macintosh supplies software support for the DECnet protocol.

The link between the MIRAGE system and the ground acquisition computer can be initiated in two ways. When the MIRAGE network acquisition is in MASTER mode, the MIRAGE network software searches for a designated object on the LAN and, upon finding it, initiates the link. In SLAV᠌E mode, the MIRAGE system registers itself as a network object and waits for the ground acquisition computer to link to it.

Analog Acquisition

The MIRAGE can acquire up to eight differential or sixteen single-ended analog signals, extract samples from the signals at different frequencies, and pack the samples into experiment major frames. Analog acquisition is used in preflight and postflight baseline data collection for Spacelab life sciences experiments.

A National Instruments NB-MIO-16L board is used for analog-to-digital conversions. The board handles up to sixteen single-ended or eight differential 12 bit analog channels at a maximum sampling rate of 100 kHz. The NB-MIO-16L is interfaced to the RTSI bus so that DMA transfers to Macintosh memory can occur using the NB-DMA-8-G.

After the acquisition of each buffer of analog input signals, the buffer containing the samples is demultiplexed and packed into an experiment major frame. Each channel will represent one analog parameter for the experiment. The MIRAGE parameter database maps the samples of the different parameters into the major frame.

Other parameters necessary for the processing of the experiment major frame are inserted into the major frame after the analog data has been acquired. Locations and values of these parameters are derived from the MIRAGE experiment parameter database.

Data Displays

The Macintosh user interface is used to display digital and analog experiment parameters and the MIRAGE status parameters in a window environment. The MIRAGE displays are database-configurable. The display control interactive user interface allows the real-time modification of the data display windows. Some operations allowed are zooming in and out on graphs and charts, changing display colors, and changing fonts and font sizes. See Figure 3 for an example of a MIRAGE display window.

Each MIRAGE display window will have a number of items, or objects, used to display data in one of three formats: text, graph, or chart.

Text objects display discrete data in alphanumeric format.

Graph objects display analog experiment parameters in scrolling strip chart format. Each graph object will have one or more channel objects. One or more analog parameter trace is drawn to each channel. X- and Y-axis labeling is provided.

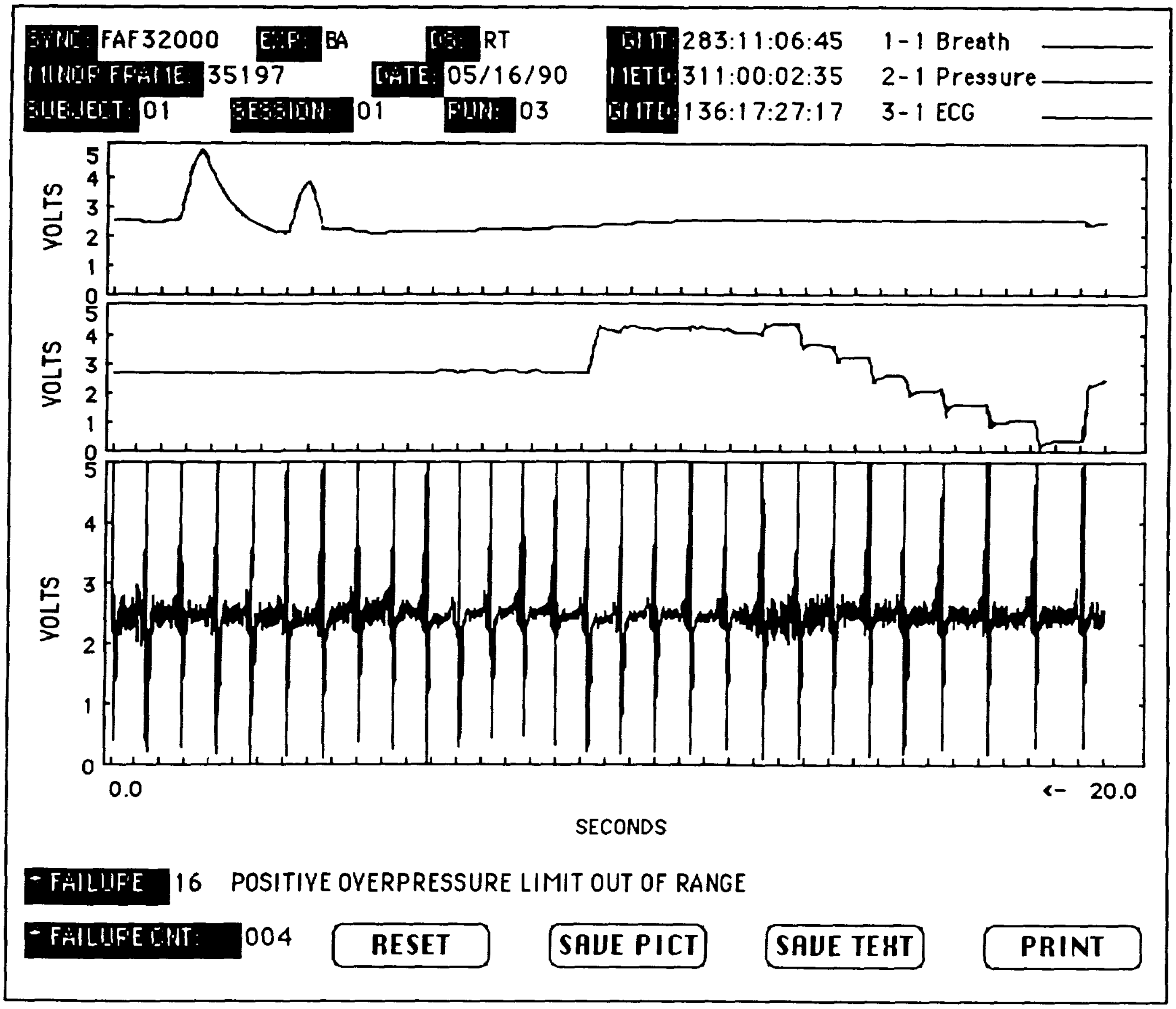
Chart objects display X-Y plots of selected parameters in real-time. Plots can be point-only or line plots. Recurring plots can be overlaid or cleared before replotting.

Data Analysis

The MIRAGE system provides several tools to aid the experiment scientist in data analysis. Some data analysis is performed in real-time.

The experiment meta-parameter database defines parameters that are derived from experiment parameters in the major frame. Meta-parameters can also be constants to use in the derivation of other meta-parameters. A meta-parameter definition consists of a type declaration, an operator, and a set of one or more operands. Meta-parameters can be 8, 16, or 32 bit integers or 32 bit real numbers. Several mathematical and logical operators are available including add, subtract, divide, multiply, logical AND, and logical OR. Meta-parameter operands may include experiment parameters or other meta-parameters. Meta-parameters are displayed in text, graph, or chart display objects.

The chart function, as noted above, allows for the plotting of X-Y plots during real-time.



SYNC: FAF32000 EXP: BA OS: RT GMT: 283:11:06:45
MINOR FRAME: 35197 DATE: 05/16/90 METD: 311:00:02:35
SUBJECT: 01 SESSION: 01 RUN: 03 GMTD: 136:17:27:17

FAILURE 16 POSITIVE OVERPRESSURE LIMIT OUT OF RANGE

FAILURE CNT: 004

Figure 3: Typical Display Window with Graph

The information on displays can be saved in either Pict or text spreadsheet files for post-session analysis.

The MIRAGE can output selected parameters in digital packets over the network or as analog signals using the NB-AO-6. Other workstations or analog devices can receive and process this data.

Analog Output

Analog output of selected parameters is available through the NB-AO-6 analog output card. Up to six single-ended output channels are available. The NB-AO-6 is interfaced to the RTSI bus so that DMA transfers to Macintosh memory can occur using the NB-DMA-8-G.

During life sciences experiment support, the output signals are routed to a strip chart recorder. The strip chart recorder presently being used is a Graphtec WR7700. A Macintosh Serial port is connected to the recorder's RS232 port for periodic data time annotation. The serial port connection is also used to program the operational characteristics of the recorder such as speed, channel setup, etc. The output signals can be received by other workstations or analog devices to perform waveform analysis in real-time.

Archival

Experiment data streams acquired from any source can be archived to any disk drive connected to the MIRAGE system. Packets are archived exactly as they are received except for the addition of a 90-byte header. Archival can be suspended and resumed. Markers can be interactively inserted in the archive file to mark significant events during the run of the experiment.

As noted above, displays can be saved in Pict or spreadsheet file formats when updated.

Playback

The MIRAGE can play back previously archived data streams. Playback can be controlled interactively. Some of the interactive control capabilities of the playback system are speed, reverse, and jump to a frame or mark.

User Interface

The Macintosh system software gives the MIRAGE a menu-and-icon-driven graphical user interface. The user can control many aspects of a MIRAGE session by selecting menu items, clicking the mouse button on icons or windows, and typing text into window text items.

Event Logging and Status Displays

Significant events that occur during a MIRAGE session are displayed in an event window and written to a MIRAGE session log file on disk. Status windows display the status of acquisition, archival, playback, and analog output in real-time.

Adding New Experiment Data Stream Support

The MIRAGE system can be customized to support most experiment data stream formats and displays. To add support for a data stream, the experiment databases and data display windows are created. The databases are created as Excel text spreadsheets. A copy of an existing experiment database can be used as a template. Display windows are created using ResEdit, a graphical resource creating and editing program provided by Apple. A small library of experiment-specific functions written in C is created, compiled, and linked with the MIRAGE software. These functions can often be copied from existing experiment function libraries and modifed for the new experiment. Since the MIRAGE system supports a wide range of generic data display and analysis functions, custom functions are necessary only for unique data display and analysis requirements.

APPLICATIONS

The MIRAGE system is designed to support the acquisition, archival, and processing of Spacelab life sciences experiment data streams in HRM format. The considerations that went into the design of the MIRAGE system make it adaptable to a wide range of applications and data formats. For instance, the NB-DIO-32F can be used to acquire almost any 8, 16, or 32 bit parallel digital data stream. Frame synchronization can be ignored during acquisition to eliminate the sync word requirements. The MIRAGE system can be customized to acquire network data packets in a variety of formats.

Another useful feature of the MIRAGE system is its ability to transform data received from any source and retransmit the data in either packetized digital form over the network interface or as up to six analog output signals. This ability of the MIRAGE system to act as a standalone data acquisition and analysis system or to work in conjunction with a variety of other systems give it a wide range of applications. Some of the possible applications are listed below.

- Experiment support
- Analog data acquisition workstation

- Data analysis and transformation
- Network playback
- Analog output system
- Archival system

CONCLUSION

The MIRAGE system has met or exceeded all of its original design requirements. The system has been used in the NASA Space and Life Sciences directorate in a number of experiment ground support roles and has performed beyond expectations.

In March of 1991, NASA presented the MIRAGE development team with its Public Service Group Achievement Award "in recognition of their outstanding contribution to the design, integration, test, and fabrication of the technologically advanced portable MIRAGE system."

The MIRAGE system concept continues to grow. Future enhancements of the MIRAGE system may include GPIB and RS232 data acquisition, expanded data analysis capabilities, and support of other network protocols (TCP/IP, FDDI). DOS, UNIX, and VAX workstation versions of the MIRAGE system are also possibilities.

ACKNOWLEDGEMENT

The MIRAGE system described in this paper was developed by Robert S. Rosser of the General Electric Government Services, at the NASA Johnson Space Center in Houston, Texas, under a contract with the Life Sciences Project Division.

BIBLIOGRAPHY

[1] Spacelab D-2 GSOC - PLE BA EGSE Interface Specification
NASA Doc. No.: D2-OP-SP-440-WT, Issue: Draft, Revision: 0, Date: 28 Sept. 1989

[2] Spacelab High Rate Multiplexer (HRM) Format Standard for Spacelab Payloads
Doc. No.: MSFC-STD-630A, Date: 1 Sept. 1988
Order No.: LS-90027, NASA Johnson Space Center

[3] IEEE 488 and VXIbus Control, Data Acquisition, and Analysis
1993 Edition, National Instruments Corporation

[4] Maintenance and Operation Manual, High Rate Demux Interface (HRDI) and Simulated High Rate Demux (SHRDM)
Order No.: LS-40003-6, NASA Johnson Space Center

TUNNELING MAGNETIC FORCE MICROSCOPY

Dr. Edward R. Burke
Laboratory For Physical Sciences
8040 Greenmead Dr.
College Park, MD 20740

Dr. Romel D. Gomez, Dr. Amr A. Adly, Dr. Isaak D. Mayergoyz
Department of Electrical Engineering, Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742

ABSTRACT

We have developed a powerful new tool for studying the magnetic patterns on magnetic recording media. This was accomplished by modifying a conventional scanning tunneling microscope. The fine-wire probe that is used to image surface topography was replaced with a flexible magnetic probe. Images obtained with these probes reveal both the surface topography and the magnetic structure. We have made a thorough theoretical analysis of the interaction between the probe and the magnetic fields emanating from a typical recorded surface. Quantitative data about the constituent magnetic fields can then be obtained. We have employed these techniques in studies of two of the most important issues of magnetic recording: data overwrite and maximizing data-density. These studies have shown: (i) Overwritten data can be retrieved under certain conditions, and (ii) Improvements in data-density will require new magnetic materials. In the course of these studies we have developed new techniques to analyze the magnetic fields of recorded media. These studies are both theoretical and experimental and combined with the use of our magnetic force scanning tunneling microscope should lead to further breakthroughs in the field of magnetic recording.

INTRODUCTION

This paper updates and summarizes the work that we have performed [refs. 1-7] in developing a magnetic force scanning tunneling microscope. We have developed this device as a tool to study important questions in the rapidly evolving field of magnetic recording. Two of the questions that we have addressed are: (1) What happens when data on magnetic media is overwritten with new data? and; (2) What are the limits of data-density (the amount of data that can be recorded in a given area) in magnetic recording? Ideally, one might hope that the answer to the first question is: When old data is overwritten with new data, the old data is completely erased and only the new data is present. If the real world is less than the ideal, then new questions are posed: Under what conditions can the old data be retrieved? How much can be retrieved? Do different types of data respond differently? What can one do to insure complete erasure? Is the new data corrupted in any way by the overwrite process? In our studies we have tried to address all of these questions. The most important discovery of our work is that under certain conditions *all* of the overwritten data can be retrieved.

For the question of data density, we have demonstrated that there are limits to the density that can be obtained with present day materials. This then leads to the question: What are the important parameters that can be changed to increase the data density? This is the most important question in magnetic recording technology and it has been discussed extensively in the literature. What we have demonstrated with our studies is how the magnetic force scanning tunneling microscope can be used to study the processes that limit the data density and how the new theoretical studies that we have developed will lead to an improved understanding of these processes.

TECHNIQUE

Rice and Moreland [8] have shown that magnetic data on a hard disk can be imaged with a tunneling microscope by using a flexible triangular probe cut from a thin film of magnetic material. We have assembled a similar device [1]. This technique is a straightforward and useful extension of scanning tunneling microscopy (STM) [9]. In this new technique, a flexible magnetic probe is used in place of the fine metallic tip employed in STM for imaging of surface topography. The magnetic probe is deflected as it interacts with the local magnetic fields. The deflections change the tunneling gap (the probe-sample separation), which correspondingly change the tunneling current. A feedback system continuously adjusts the vertical displacement of the probe to keep the current constant as it is rastered across the surface. The changes in the vertical displacement are measured and recorded 400 times on a single scan. The image is constructed from 400 rastered scans. Thus, as schematically shown in Fig. 1(a), the

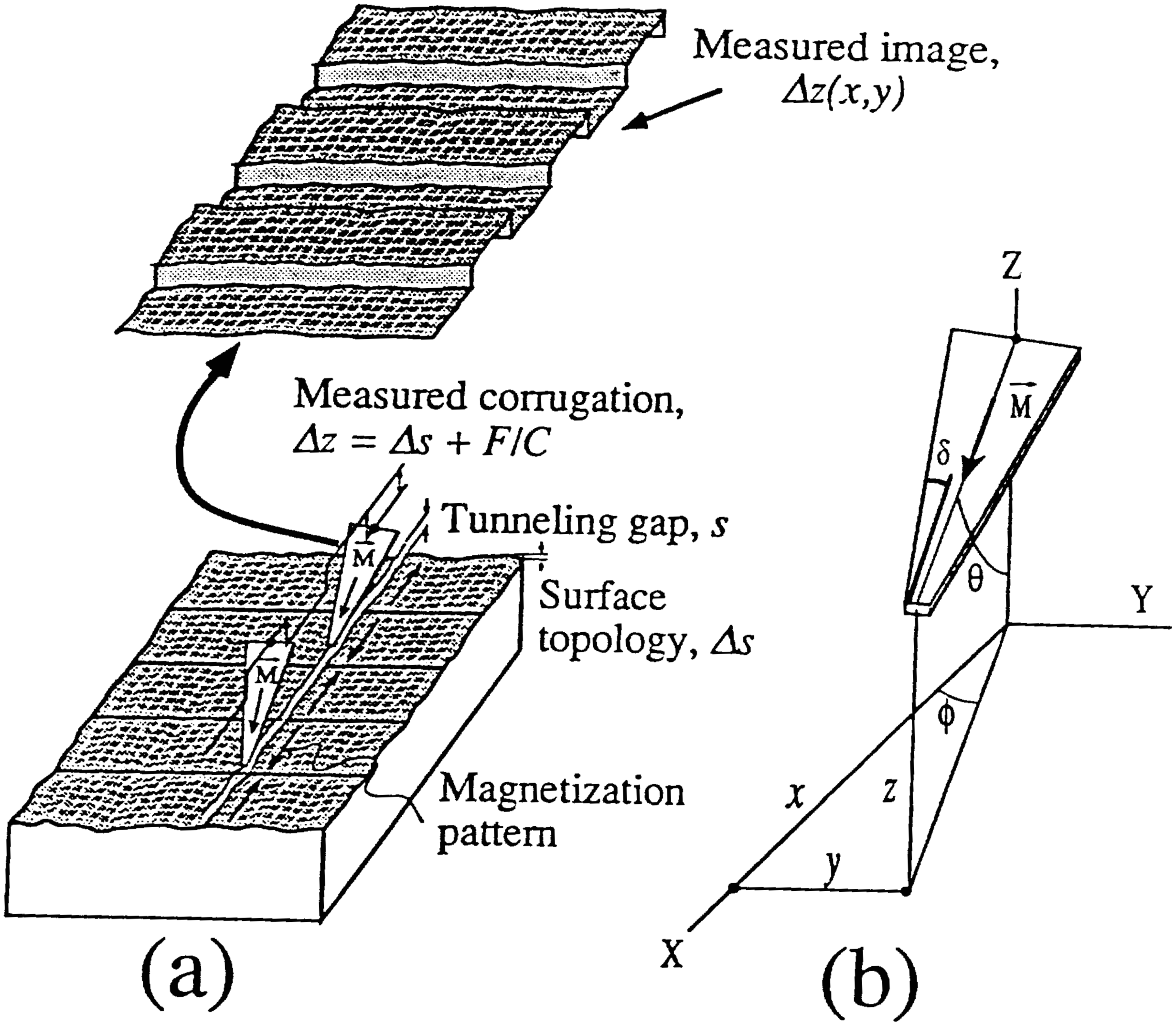


FIG. 1. (a) Schematic diagram of magnetic force scanning tunneling microscopy. The deflection of the probe due to its interaction with the local surface magnetic fields is mapped as a function of its lateral position (see text) and (b) probe geometry used in the analysis.

resulting image of vertical displacements represents the local magnetic field variations combined with the surface topography variations. The vertical displacements, and the in-plane rastering position are all controlled and measured by accurate piezoelectric elements.

A typical image is shown in Fig. 2. The image shows magnetically recorded data on a commercial hard disk. Three distinct data tracts are visible here, each having a width of approximately 45 μ. The tracks are separated by about 12 μ of non magnetized region. Magnetized regions appear as broad (~ 6 μ) depressions bounded by relatively narrow (~ 2 μ) bright protrusions. The "height" of these magnetized regions are of the order of 150-200 nm, which is roughly a factor of 10 larger than the surface roughness. The surface roughness shows up as the fine lines running perpendicular to the magnetic data. These fine lines are caused by the final machining of the aluminum disks. The disks appear mirror-smooth to the naked eye and the fine lines, which are readily apparent here, can only be observed with the most sophisticated optical microscopes. In Fig. 2(b), we show a high-resolution magnification roughly corresponding to the upper right-hand corner of the image. Two different types of tracks are clearly visible, distinguished by the change in relative sizes of depressions and protrusions; which demonstrates the ability of our device to distinguish between different directions of surface magnetization. The 3D image shown in Fig. 2(b) was constructed through system software. The variations in the amplitude of the displacements can be seen along the lower edge of the image in Fig. 2(b). To obtain quantitative information about the magnetic fields emanating from the surface, one would have to know how the displacement amplitude is related to the magnetic fields. In order to obtain this information, we have made a complete theoretical analysis [2] of the interaction between a flexible triangular probe and a typical magnetic pattern on a recorded surface. The use of this analysis allows the measurement of the magnetic fields of the recorded patterns imaged by a magnetic force scanning tunneling microscope.

THEORY

Magnetic Fields

Fig. 1(b) shows the geometry for our calculations. We assume that the recorded signal is a repetitive, symmetric pattern of wavelength λ in the x direction, with infinite extent in the y direction. The magnetic field **H** from the pattern can be expressed as the gradient of a scalar potential Φ,

$$H = -\nabla\Phi \tag{1}$$

The scalar potential will be the solution of Laplace's equation and can be written as,

$$\Phi(x,z) = \sum_{n=1}^{\infty} \Phi_n e^{-kz} \cos kx, \tag{2}$$

where $k = 2\pi n/\lambda$, and the coefficients Φ_n match the series solution to the particular field pattern. The field pattern will of course depend on the magnetization distribution within the recorded media. We have found it convenient to express the magnetization in Fourier series. If we assume that the recording media is so thin that the magnetization is uniform through the thickness of the film, then there are two different magnetization patterns that will lead to the scalar potential given by (2). The first pattern is a magnetization in the plane of the film given by,

$$M_x = M_s \sum_{n=1}^{\infty} m_{x,n} \sin kx, \tag{3}$$

where M_s is the saturation magnetization and the $m_{x,n}$ are the normalized Fourier coefficients. The other magnetization pattern that would lead to the scalar potential (2) is a magnetization perpendicular to the plane of the film given by,

$$M_z = -M_s \sum_{n=1}^{\infty} m_{z,n} \cos kx, \tag{4}$$

where the minus sign is a mathematical convenience. The fields can be constructed from linear combinations of (3) and (4), and if we use Maxwell's equations and make the transverse component of **H** and the normal component of **B**

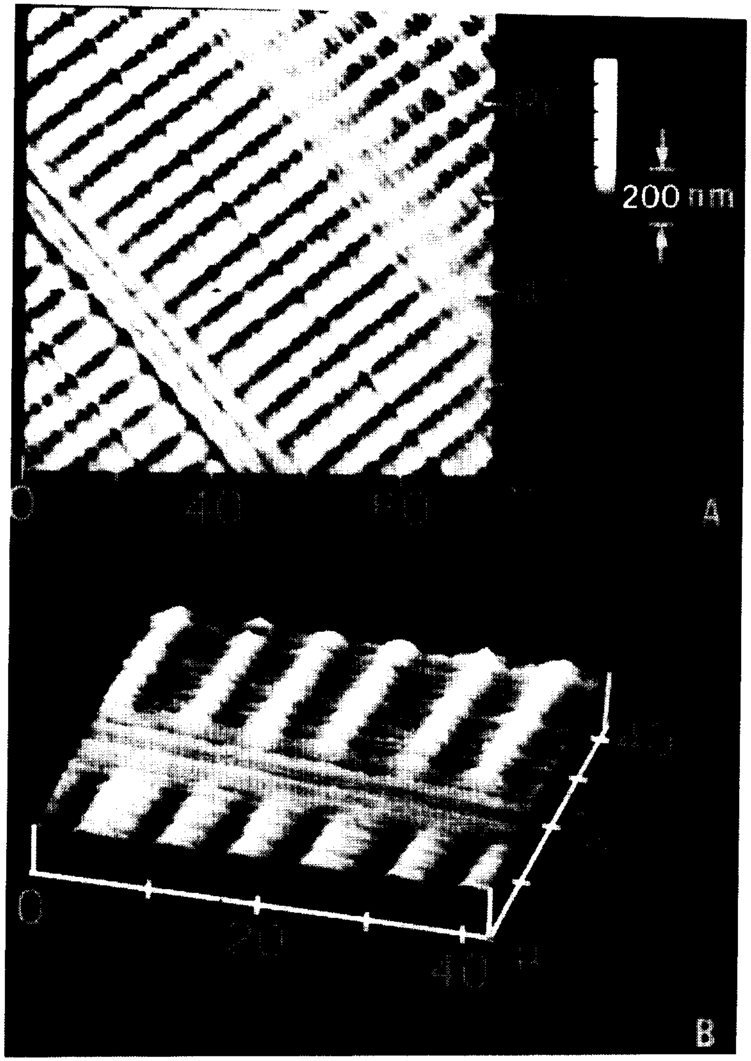


FIG. 2. (a) A large field-of-view MFSTM image showing the recording tracks of a computer hard disk drive magnetic media, and (b) perspective surface plot of a magnified view of the upper right-hand corner showing two different track types, distinguished by the lengths of the protrusions and depressions.

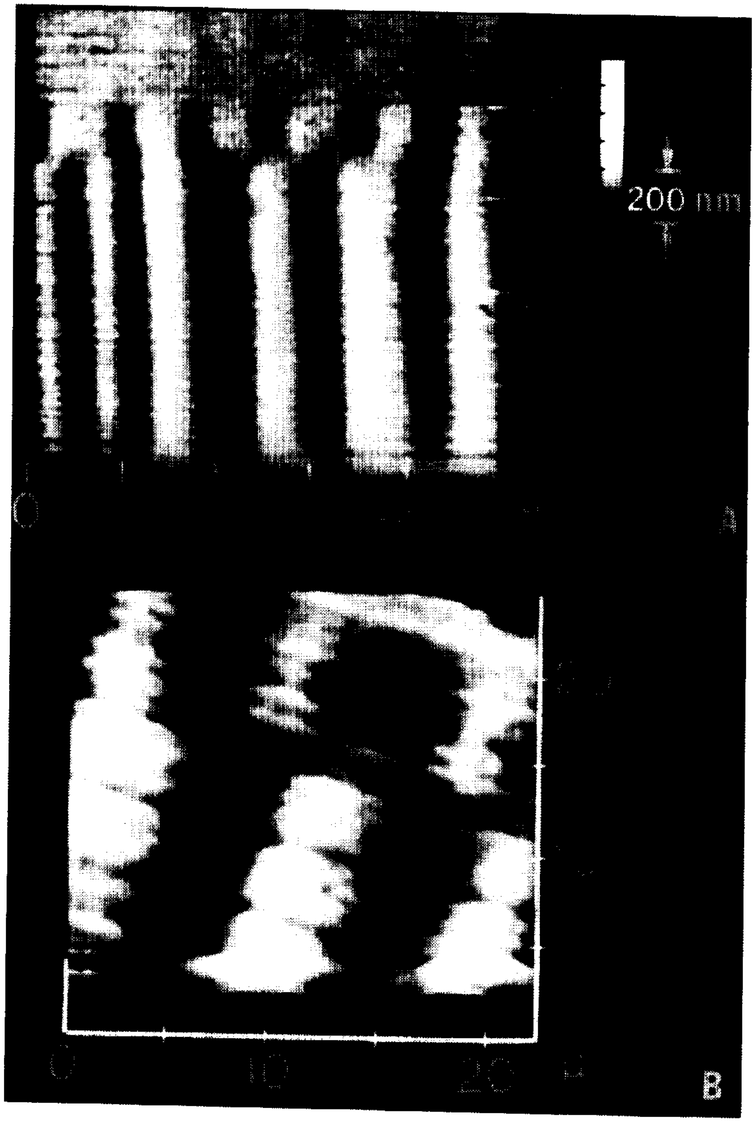


Fig. 3. Overwritten data on conventional commercial rigid disk showing: (a) remnant of previous data, and (b) erase band.

(B=H+4πM) continuous at the media surface, then we can solve for the coefficients Φ_n. We will leave the details of these calculations to a later paper. For now we will simply give the result which is,

$$\Phi_n = -2\pi M_s(1-e^{-kd})(m_{x,n}+m_{z,n})/k, \tag{5}$$

where d is the thickness of the recording media. We have used (5) to construct numerous field distributions, including all the ones we could find in the literature [10], [11], [12]. The point is that we can use these techniques to find the magnetic fields from virtually any distribution of magnetization. We now return to the major problem at hand: the interaction between these fields and the probe tip.

Energy of Interaction

The Energy of interaction between the field from the pattern and the last domain on the probe tip can be expressed as [10]

$$E = -\int H \bullet M \ dV, \tag{6}$$

where M is the magnetization of the last domain on the probe tip, and V is the volume of the domain. To perform the integral of (6) we make the following assumptions: (i) the domain is magnetized along the probe axis by shape anisotropy, (ii) the domain is much longer than λ so that the limit of integration in the z direction can be extended to infinity, and (iii) the thickness of the probe, t, is much less than the wavelength λ. Rugar et al. [12] have shown that the last domain on their probe tip was about 20 μ in length, and since most patterns on modern recording surfaces have a wavelength smaller than this, assumption (ii) is not unreasonable. The thickness of the probe is much less than a micron which is about the smallest wavelength currently available. Using these assumptions, the integral (6) was evaluated in [2] with the result,

$$E = Mtw\sum_{n=1}^{\infty}\Phi_n e^{-kz}\left\{\cos kx\frac{\sin((kw\sin\phi)/2)}{(kw\sin\phi)/2}+\frac{\tan\delta}{wk}\left[A_+\cos k(x-x_+)+A_-\cos k(x-x_-)\right]\right\}, \tag{7}$$

where,

$$A_\pm = 1/\sqrt{\cos^2\theta+(\sin\theta\cos\phi\pm\tan\delta\sin\phi)^2}, \tag{8}$$

and,

$$x_\pm = \pm\frac{w}{2}\sin\phi+\frac{1}{k}\tan^{-1}\left(\frac{\sin\theta\cos\phi\pm\tan\delta\sin\phi}{\cos\theta}\right). \tag{9}$$

The integrals were performed so that the point (x,z) is the coordinate of the probe tip. The first term in (7) is due to a magnetic charge Mtw at the tip of the probe. The magnetic potential is weighted by a sampling factor caused by the variation in the field across the width w of the probe tip. The next two terms can be thought of as the contributions from the magnetic charges on the sides of the probe, separated from the tip by the distances $x_\pm$. The equations (7-9) give a complete expression for the energy of interaction between the probe and the fields from the recorded media.

Probe-Tip Displacement

The quantity that is measured by the tunneling microscope is the displacement Δz of the probe tip. The displacement is caused by both the surface topography and the magnetic interaction between the probe and the magnetic field from the surface pattern. If the probe tip is properly designed, the interaction will predominate and the surface roughness will appear as a background noise.

If the probe is constrained to rotate in the θ direction, the displacement will be given by $l\sin\theta\Delta\theta$, where l is the length of the probe's moment-arm. A force F_N normal to the probe's tip will cause a rotation in the θ direction such that $lF_N = -K\Delta\theta$ where K is the tip torque constant. The displacement Δz is then given by

$$\Delta z = -\frac{l^2 F_N \sin\theta}{K} \; . \tag{10}$$

The force acting on the tip is the gradient of the energy $F = -\nabla E$ so that (10) becomes

$$\Delta z = \frac{l^2}{K}\left(\cos\theta\cos\phi\frac{\partial E}{\partial x} + \sin\theta\frac{\partial E}{\partial z}\right)\sin\theta . \tag{11}$$

Using (7), (11) becomes, after some manipulation,

$$\Delta z = -\frac{l^2 Mtw\sin\theta}{K}\sqrt{\cos^2\theta\cos^2\phi + \sin^2\theta}\sum_{n=1}^{\infty}\Phi_n kCe^{-kz}\sin\left(kx - \beta + \tan^{-1}\frac{\sin\theta}{\cos\theta\cos\phi}\right), \tag{12}$$

where,

$$C = \sqrt{\left[\frac{\sin((kw\sin\phi)/2)}{(kw\sin\phi)/2} + \frac{\tan\delta}{kw}(A_+\cos kx_+ + A_-\cos kx_-)\right]^2 + \left[\frac{\tan\delta}{kw}(A_+\sin kx_+ + A_-\sin kx_-)\right]^2} , \tag{13}$$

and,

$$\beta = \tan^{-1}\frac{\tan\delta(A_+\sin kx_+ + A_-\sin kx_-)}{\dfrac{\sin((kw\sin\phi)/2)}{(\sin\phi)/2} + \tan\delta(A_+\cos kx_+ + A_-\cos kx_-)} . \tag{14}$$

Equations (12)-(14) give a complete description of the interaction between the probe and the recorded pattern. In general, the equations are quite complicated and their usefulness is not readily apparent. In the case when the probe lines up with the pattern ($\phi = 0$) the equations reduce to a simple form,

$$\Delta z = -\frac{l^2 Mtw\sin\theta}{K}\left[H_x\cos\theta + H_z\sin\theta + 2\frac{\tan\delta}{w}\int_0^x H_z dx'\right] . \tag{15}$$

The first two terms give the interaction between the magnetic field and the magnetic charge at the tip. The next term gives the effect of the charges on the sides of the probe. This last term was written in the integral form so that it could be expressed in terms of the magnetic field H_Z. It could have been written in terms of H_X in which case it would have been identical to the expression for the flux picked up by a conventional recording head. Equation (15) is an important result because it shows that, if the third term can be made small, then the images can be related to the magnetic fields at a point. We call this "the point charge model" and we have expanded on it in [6] and [7].

Equation (15) can be used to obtain relative values of the magnetic field components H_X and H_Z. To obtain absolute values, the probe would have to be calibrated in a known field to obtain the factor l^2Mtw/K. One way to obtain the fields from (15) is to obtain three images at three different values of the angle θ. The fields H_X and H_Z can then be obtained at every point from a linear combination of the three images. For example, if three images, $\Delta z(\theta)$, were obtained at the angles of 30, 45, and 60 degrees, then H_X and H_Z could be solved for, to obtain,

$$H_x = \frac{K}{l^2 Mtw}\left[18.02\Delta z(30^\circ) - 29.35\Delta z(45^\circ) + 13.56\Delta z(60^\circ)\right], \tag{16}$$

$$H_z = \frac{K}{l^2 Mtw}\left[23.48\Delta z(30^\circ) - 29.35\Delta z(45^\circ) + 10.40\Delta z(60^\circ)\right]. \tag{17}$$

Since all the data used to construct the images is available in digital form, it is a simple matter to combine the images using (16)-(17) to obtain the magnetic fields. The main experimental difficulty with this procedure would be in obtaining the images at exactly the same location every time. One way to alleviate this problem would be to use the topological features of the surface as guide-points. We have thus made a complete theoretical analysis of how the magnetic force scanning tunneling microscope can be used to obtain the magnetic fields from recorded media as a function of all of the relevant parameters.

DATA OVERWRITE

Overwrite performance is a major concern in magnetic recording since data detection can be corrupted by previously recorded patterns when sufficient overwrite is not achieved. Even with direct overwrite, portions of previously recorded data can persist and be detectable. Tracking misregistration, or slight deviations in positioning of the recording head from the original track, could leave even more significant portions of previous data along the track edge [3]. Fig. 3(a) shows a $50\mu\times50\mu$ image of a commercial rigid disk with overwritten data. The new data appears as the long alternating bright protrusions and dark depressions representing oppositely magnetized regions along the track. Remnants of the previously recorded data appear as localized regions extending by a few microns from the upper track edge. It should be emphasized that this pattern was not deliberately constructed but was found on a previously used disk. The overwritten data can be completely recovered by simply reading it off on a bit-by-bit basis.

The high resolution image of a different overwritten region in fig. 3(b) suggests some interesting characteristics of the erase band. the regions where the old and new data coincide create continuous magnetization between the old and new tracks, as exhibited by the extreme left transition. This is not the case for the two succeeding transitions, however, where the new set is out-of-phase with the old set. Here, the old data are truncated prior to the emergence of the new data, leaving about a micron wide gap with no definite magnetization. This behavior is consistent with current notions of the erase band [13,14]. The write field within this narrow band was above the coercivity of the media to reduce the magnetization at those areas (which truncated the bright stripes of the old data) but the magnitude was not high enough to create new well defined magnetizations.

We have continued these investigations by obtaining images of deliberately overwritten data on thin film disk media. We have examined the relationship between the persistence of overwritten data and the radial offset of the recording head, as well as the effects of the recording density [5]. The effect of a previously recorded pattern can be detected even when distinct remnant transitions can not be identified. At recording frequencies in the range of 10 MHz, the previously recorded pattern affects the newly recorded track even at small ($>2\mu$) offsets, by introducing apparent lengthening or shortening of the track-width depending upon the relative phases of old and new patterns. Presumably, one could still extract the overwritten data in this case but it would require much more analysis. Distinct portions of overwritten data remain on the surface for offsets in excess of 2 μ. At low recording frequencies in the range of 1 MHz, larger offsets ($>4\mu$) are needed to detect previously recorded data. In this case, the non-uniform magnetization introduces recorded cells of "trapezoidal" cross sectional area. This effect is less severe at high frequencies, and plays a crucial role in extending the required minimum offset.

DATA DENSITY

Despite recent advances in media processing technology and the demonstration of storage densities beyond 1 Gigabit/square inch [14], our understanding of processes that lead to reduction in signal strength as the wavelength is decreased is still under development. The process involves a complex interplay between media and recording head properties. While a great deal of theoretical and experimental work has been performed, one of the difficulties in determining the roles played by the different mechanisms is the lack of systematic experimental data to characterize magnetization patterns with sufficient spacial resolution. What we have done is to show how the magnetic force scanning tunneling microscope (MFSTM) can be used as a powerful tool to study this problem.

In this work, we are concerned with direct real space imaging of recorded patterns. We performed a series of MFSTM measurements of magnetization patterns which were written with progressively increasing densities. We then analyzed the pattern behavior as the wavelength was reduced. This work extends previously reported investigations of high density recording on longitudinal recording media by using magnetic force microscopy [13]. In contrast with those MFM measurements, the current MFSTM based technique allows a more straightforward interpretation of the images and thus facilitates quantitative analysis. Specifically, we make quantitative estimates of the transition length by comparing image profiles with calculated lineshapes based upon an arctangent model for the transitions. We then discuss possible mechanisms that play major roles in causing self-erasure at high recording densities.

Experiment

Measurements were made on a commercially available rigid disk with patterns recorded on a precision spin stand system. Fig. 4 shows a series of recorded transitions in the range of 100 to 2000 FR/mm, which correspond to recorded wavelengths λ in the range from 20 μ to 1 μ. All images were obtained using a single imaging probe so that comparisons between different tracks are independent of probe specific properties.

In addition to trackwidth reduction [16], we find significant variations in the behavior of these patterns with increasing density. In the range from 100 to 600 FR/mm ($20\mu \le \lambda \le 3.1\mu$), the transitions are well defined and exhibit very little zigzag across the track. This is consistent with previously reported magnetic force microscope measurements on longitudinal recorded patterns [13]. Similarly, the amplitudes of the magnetic features are more or less constant, indicating that the relative strength of the fields do not vary significantly from the longest wavelength down to about 3.1 μ. At higher densities, the amplitudes decrease sharply and the patterns gradually lose their detail. The transitions start to become fuzzy, and in certain areas, they appear to merge together. The bits coalesce to form localized patches, becoming more noticeable for $\lambda<2.2\mu$. At 1 μ wavelength, the track edges become indistinct and the size of the coalesced regions has increased considerably, leaving individual transitions barely discernible. This effect could be associated with either poor high density performance of the media or the frequency response of the recording head. It is quite possible that recording fields perturb neighboring previously recorded bits, which reduces their magnetizations.

Discussion

We begin our analysis by deriving an estimate of the transition length. Close inspection of the images in Fig. 4 show that the lineshapes vary with the wavelength. This is illustrated clearly in Fig. 5 where a series of average line profiles from representative images in Fig. 4 are presented as solid curves. The peak occurs very near the left transition edge at the lowest recording frequency, and gradually moves to the center as the wavelength is decreased. This can be explained using a model that allows the transition length parameter to become a substantial fraction of the bit length.

As shown previously, the deflection Δz is given by (15). For the probe that we used, δ=15 degrees and w=2.5μ. To use (15), expressions must be found for the magnetic field caused by the recorded magnetization. For this study we will assume that the magnetization is a symmetric series of alternating polarities with arctangent transitions. The Fourier series that approximates the magnetization pattern is given by

$$M_x = 8M_s \sum_{n=odd}^{\infty} \frac{e^{-ka}}{k\lambda} \sin kx, \tag{18}$$

where a is the transition length. Strictly speaking, this series represents the arctangent transitions only in the limit where $\lambda >> 4a$. We have found, however, that (18) gives a useful (if not exact) fit to the data even when this limit is not satisfied. This series underestimates the peak magnetization as the transitions are brought close together but has the property of eliminating the sharp junctions at $x = n\lambda / 4$, for n odd, which appear by directly matching arctangent transitions. We computed the fields by using (1), (2), (5), and (18), and then substituting them into (15) to obtain lineshape profiles. The lineshapes were then fit to the experimental data using a as an adjustable parameter. The resulting curves are shown as dashed lines in Fig. (5). We find that the best fit to the data corresponds to a transition of a=0.7μ. The best fit was obtained at the longest wavelength, where this theory would be most applicable.

It should be emphasized that we attempted to fit the experimental data to numerous magnetization distributions, but the arctan distribution gave the best results. This is not surprising since the arctan distribution is the most widely accepted distribution seen in the literature. What actually happens, however, when the transitions are brought close together, might, according to our calculations, be somewhat surprising. It is commonly accepted that data density is limited by the increasing demagnetizing fields caused by the transitions being brought close together. Our calculations show, however, that the demagnetizing fields *decrease* as the transitions are brought close together. This is caused by the long tails of the arctan transitions which have the effect of drastically reducing the magnetization, and hence the demagnetizing fields, as the transitions are brought close together. It is this breakdown in the magnetization patterns that leads to the limits in data density. The usual arguments about increasing the coercivity to increase data density are still valid, because these arguments are made on the basis of isolated transitions. To increase data density you have to decrease the transition length. The transition length can only be decreased by increasing the coercivity, or the squareness of the hysterisis loop. Decreasing the magnetization, or the

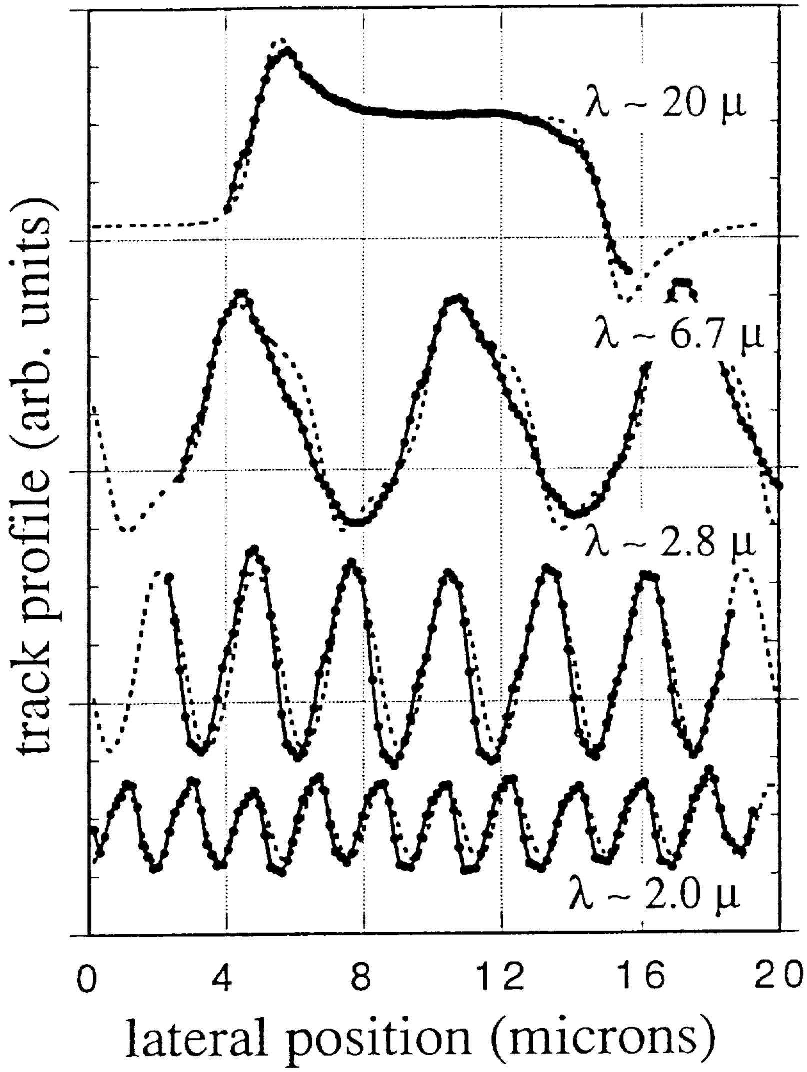


Figure 5. Solid curves: Average line profiles of representative images in Fig. 4; Dashed curves: calculated lineshapes for a constant transition length, a= 0.7 microns.

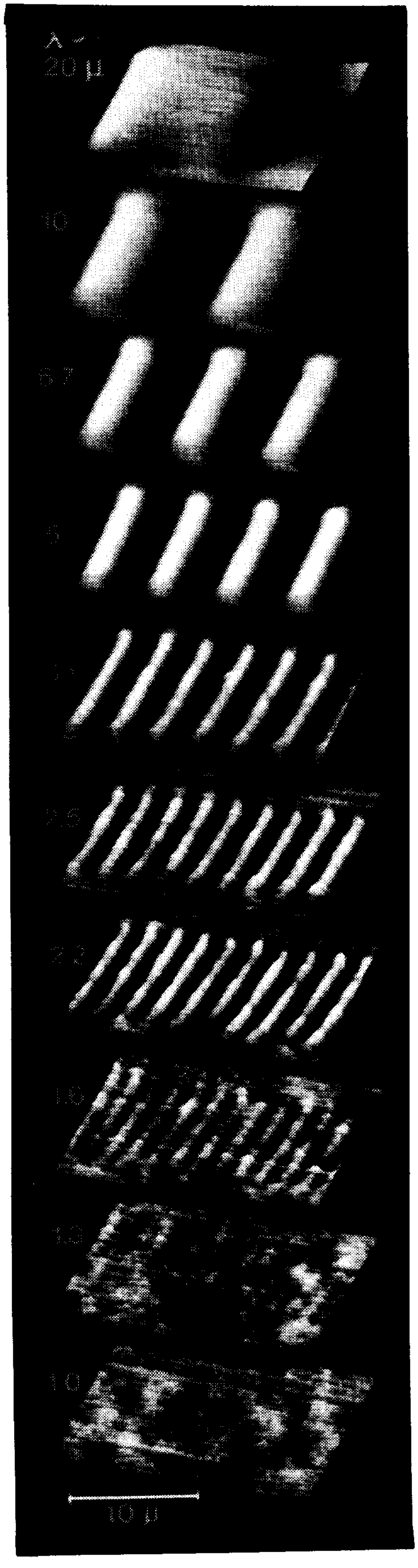


Figure 4. A series of MFSTM images of recorded patterns on thin-film media with progressively decreasing wavelengths.

media thickness, will also decrease the transition length but this will also decrease the readback signal. These conclusions are not new, but have been known for a long time [17], [18]. What is new is the actual observation of what is happening as the data density is increased. We have demonstrated that the magnetic force scanning tunneling microscope can be used as a powerful tool in studying these problems.

CONCLUSIONS

The MFSTM is shown to be a powerful technique for generating images of magnetization patterns. In particular, it has been used to yield images of persisting remanent data with minute details, and to investigate subtle features of overwritten data. We have demonstrated that the MFSTM is a powerful tool, useful in obtaining qualitative images and in deriving quantitative results. We used the technique in a study of data density and showed that the measured profiles could be adequately described by a model of arctan transitions. The best fit to the data gave a transition length of a=0.7μ. Our theoretical analysis shows how the constituent magnetic fields from recorded magnetic patterns can be obtained from the images. We also show how the sensitivity of the microscope varies with the orientation of the probe, and how this relates to experimental data.

ACKNOWLEDGMENT

We are grateful to Mr. John Gorczyca and Dr. Mark Kryder at Carnegie Mellon University, Data Storage Systems Center for their cooperation and for providing recorded disks used in these experiments. Mr. Gorczyca and Dr. Kryder were also co-authors of two of our publications [4],[5].

REFERENCES

[1] R. D. Gomez, E. R. Burke, A. A. Adly and I. D. Mayergoyz, "Magnetic Field Imaging by Using Magnetic Force Scanning Tunneling Microscopy," *Appl. Phys. Lett.* 60 (7), 17 February 1992 pp 906-908.

[2] E. R. Burke, R. D. Gomez, A. A. Adly, and I. D. Mayergoyz, "Analysis of Tunneling Magnetic Force Microscopy Using a Flexible Triangular Probe," *IEEE Transactions on Magnetics* Vol. 28(5), September 1992 pp 3135-3137.

[3] R. D. Gomez, E. R. Burke, A. A. Adly, and I. D. Mayergoyz, "Magnetic Force Scanning Tunneling Microscope Imaging of Overwritten Data," *IEEE Transactions on Magnetics* Vol. 28(5), September 1992 pp 3141-3143.

[4] R. D. Gomez, E. R. Burke, A. A. Adly, I. D. Mayergoyz, J. A. Gorczyca and M. H. Kryder, "Magnetic Force Scanning Tunneling Microscopy of High Density Recording," Accepted for publication in the *Journal of Applied Physics.*

[5] R. D. Gomez, E. R. Burke, A. A. Adly, I. D. Mayergoyz, J. A. Gorczyca and M. H. Kryder, "Microscopic Investigations of Overwritten Data," Accepted for publication in the *Journal of Applied Physics.*

[6] I. D. Mayergoyz, A. A. Adly, R. D. Gomez, and E. R. Burke, "Experimental Testing of Point Charge Model of Magnetic Force Scanning Tunneling Microscopy," Accepted for publication in the *Journal of Applied Physics.*

[7] I. D. Mayergoyz, A. A. Adly, R. D. Gomez, and E. R. Burke, "Magnetization Image Reconstruction from Magnetic Force Scanning Tunneling Microscopy Images," Accepted for publication in the *Journal of Applied Physics.*

[8] P. Rice and J. Moreland, "Tunneling-stabilized Magnetic Force Microscopy of Bit Tracks on a Hard Disk," *IEEE Trans. Magn.*, vol. Mag-27, pp 3452-3454 (1991).

[9] G. Binnig, H. Roher, Ch. Gerber, and E. Weibel, *Phys. Rev. Lett.* 49, 57 (1982).

[10] A. Wadas, and P. Grutter, "Theoretical approach to magnetic force microscopy," *Phys. Rev. B*, vol. 39, no. 16, pp 12013-12017, June 1989.

[11] A. Wadas, P. Grutter, and H.-J. Guntherolt, "Analysis of in-plane bit structure by magnetic force microscopy" *J. Appl. Phys.* **67** (7), pp 3462-3467 (1990).

[12] R. L. Wallace, "The Reproduction of Magnetically Recorded Signals," *Bell Syst. Tech. J.*, vol. 30,p.1145, 1951.

[13] D. Rugar, H. J. Mamin, P. Guethner, S. E. Lambert, J. E. Stern, I. McFadyen, and T. Yogi, "Magnetic force microscopy: General principles and application to longitudinal recording media" *J. Appl. Phys.* 68 (3), pp 1169-1183 (1990).

[14] H. Edelman and M. Covault, "Design of Magnetic Recording Heads for High Track Densities," *IEEE Trans, Magn.*, vol 21, p 2583, 1985.

[15] T. Yogi, C. Tsang, T. A. Nguyen, K. Ju, G. L. Gorman and G. Castillo, *IEEE Trans. Magn.* 26, 2271 (1990)

[16] T. Lin, J. A. Christner, T. B. Mitchell, J-S. Gau, and P. K. Gorge, *IEEE Trans. Magn.* Vol. 25, 710 (1989).

[17] R. I. Potter, "Analysis of saturation magnetic recording based on arctangent magnetization transitions, "*J. Appl. Phys.*, vol. 41, no. 4, pp. 1647-1651, Mar. 1970.

[18] M. L. Williams and R. L. Comstock, "An analytical model of the write process in digital magnetic recording." in *AIP Conf. Proc.* no. 5. pp. 738-742, 1971.

Operator Performance Support System (OPSS)

Marlen Z. Conklin (Systems Engineer)
Naval Command, Control, and Ocean Surveillance Center
Research, Development, Test and Evaluation Division
Code 793, 271 Catalina Blvd. San Diego, CA 92152-5000
(enrolled in the Doctor of Science in Information Systems
at NOVA University)

ABSTRACT

In the complex and fast reaction world of military operations, present technologies, combined with tactical situations, have flooded the operator with assorted information that he is expected to process instantly. As technologies progress, this flow of data and information have both guided and overwhelmed the operator. However, the technologies that have confounded many operators today can be used to assist him -- thus the Operator Performance Support System. In this paper we propose an operator support station that incorporates the elements of Video and Image Databases, Productivity Software, Interactive Computer Based Training, Hypertext/Hypermedia Databases, Expert Programs, and Human Factors Engineering. The Operator Performance Support System will provide the operator with an integrating on-line information/knowledge system that will guide expert or novice to correct systems operations. Although the OPSS is being developed for the Navy, the performance of the workforce in today's competitive industry is of major concern. The concepts presented in this paper which address ASW systems software design issues are also directly applicable to industry. They will make a dramatic impact on the way we work in the future, both the military and the industry. The OPSS will propose practical applications in how to more closely align the relationships between technical knowledge and equipment operator performance.

INTRODUCTION

As the basis of our society moves from the Agricultural, Industrial, and the Information Ages into the Knowledge Age it becomes apparent that the work environment must follow these progressions. The Navy's Combat Systems that were developed some years ago attempted to integrate information and systems, and provided the operator with the opportunity to successfully perform expected tasks. However, with extra capabilities, options, and "nice to have" features added to these basic systems, the flow of data and information to the operator was vastly increased. The resultant more complex systems and equipments have flooded the operator with assorted information that he is now expected to process instantly. Combat systems do not allow for performance that is less than perfect.

The purpose of this paper is to describe an on-line or independent support station that focuses on the information/knowledge available to the Anti Submarine Warfare (ASW) systems operators onboard U.S. Navy ships. The OPSS will systematically incorporate modern informational display techniques and tools so the operator has assess to the proper information he needs at the time he needs it. By providing the operator with this support he will respond more predictably and properly to the requirements of the situation as it is presented to him at his combat systems position. Today's systems must be directed towards user-oriented environments. The design of the OPSS must allow the operator to freely interact with all the CBT, hyperdocumentation and other materials available within the system. The user must have access to appropriate information to accurately integrate a broad range of processing functions.

The OPSS station hardware is identified as Device S10H7. This computer based hardware is configured with Intel 486/33 CPU, 1.2MB 5.25" floppy disk drive, 1.44MB 3.5" floppy disk drive, CD-ROM disk drive, 520 MB Internal Hard Disk Drive, 32 MB RAM, 17" color monitor, Keyboard, Track Ball, Stereo Ear Phones, Video Graphics Accelerator card, Sound Blaster Pro board, DVI playback board and modem. The system is the latest in technology as applied to a multimedia support and training delivery systems for U.S. Navy ships. Device S10H7 is designed to withstand the rigors of use onboard ships and meets all ELF/VLF low emission standards, EMI/RDI low radiation standards, and is certified safe for use. The software is MS-DOS-based, supports Microsoft Windows 3.1 and will be delivered on CD-ROM discs. These standardized commercial programs provide flexibility to support existing and future courseware and hardware enhancements.

In the development process of the OPSS software, one or more members of the end user community were included. Users' members have contribute valuable prospective during system design, assisted in the Fleet introduction of OPSS, and provide feedback to the OPSS design engineers during early production. They are also responsible for the ongoing system "sanity checks" as well as maintaining focus on the operator.

This paper will address the following steps of the systematic implementation of an operator centered workstation system:

- Knowledge acquisition
- Selection of information presentation
- Development of models
- Database system
- Development of OPSS

KNOWLEDGE ACQUISITION

The objective of knowledge acquisition is to identify all the information and knowledge that the operator must possess or have available to perform satisfactorily. For selecting data, one should ask "if I am the operator, what information is needed to get the job done?".

For centuries written material has been "the" conduit of information. We all know what a book or a manual looks like and how to use it to obtain information. One method of knowledge acquisition which OPSS utilizes is to use existing manuals and documentation and, with the help of the end user community, develop an outline of operation and maintenance of ASW equipment. This outline includes a Table of Contents, an Index, and a Glossary. The Table of Contents provides the information and knowledge domain in a hierarchical, linear format. However, the advantage of the OPSS system is that it is computer based, and therefore, information does not need to be linear! Items from the Table of Contents are "linked" and replicate "book browsing" behavior. One way to link the different contents of information from the Table of Contents is to develop a concept map in the format of a tree map. Tree map graphically depicts the information architecture. This is a useful technique to graphically represent meaningful linking relationships between two or more concepts. It also represents the relationship among concepts and the relationships across levels of the hierarchy. Tree mapping possesses a structural plasticity and externalizes concepts and propositions as they are organized in the mind. Changes to the system do not present problems as the open architecture style of development can be easily modified. Concept mapping is also a good technique for negotiating a problem domain with another individual. Tree-maps are designed for visualization of the hierarchical structure, and linking of knowledge structures and their relationships. An example of the mapping, showing simple hierarchical and linking formats, is shown in Figure 1.

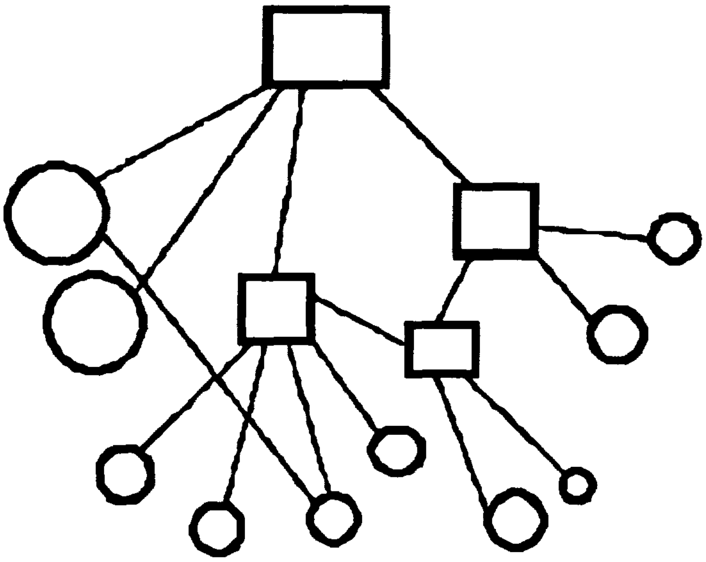


Figure 1. Concept tree mapping

Tree mapping is also very useful to convert the information into a model suitable for a system display such as the menus display format. Menus relate to user needs or problems, different user circumstances, different models of information, and allows the operator the flexibility to exit from one point and to link a different section of the system. For example the OPSS allows the operator to perform maintenance with the support of the on-line electronic maintenance documentation. If at any time he has problems performing a specific task, he may call up an interactive computer based training (CBT) which will explain how to perform the task.

The pragmatic user oriented approach is used to develop this system. Only what is useful for the operator is extracted and retained. Nice to have flexibility is scrutinized. If it does not meet operator needs, it is deleted. Needless information and knowledge are therefore discarded. What the system retains is limited to essential functions, and is justified by the end user. We keep it simple and efficient. The boundaries of information are not absolute because there is a modem network capability that establishes, with some operational restrictions, a network interface between U.S. Navy ships, school houses, developers and sponcers. The concept is to start with an essential kernel of information/knowledge and only add to the system if the requirement is established by the field operators.

The knowledge represented by concept mapping may be classified as formal rule-based processing of information, or it may be represented by non rule-based processing. Rule-based knowledge is the foundation of expert systems. The OPSS therefore has a rule-based front end that is designed to quickly navigate the operator through the hierarchy of menus. Non rule-based knowledge is utilized in the hypertext and hypermedia systems of the OPSS. Hypertext and hypermedia are used a vast diversity of conceptual frameworks and can be referenced while using metaphors, similes, analogies, diagrams, images, animations, sound and video all of which an operator can use but the computer cannot. Consequently the graphic user interface (GUI) representation of the ASW OPSS is an initial menu selection that eventually links the operator with hypermedia documentation or interactive CBT.

SELECTION OF INFORMATION PRESENTATION

Selected information will be presented in various media. The ASW OPSS programs incorporates elements of text, graphics, animation, picture, sound, video and software modules, while OPSS hardware is the latest in available technology applied to a multimedia workstation for U.S. Navy ships. Members of the end user community have reviewed the media selection process and have helped determine how the information should be expressed to support the operator's needs. The paradigm of how to represent knowledge is defined by the domain experts. We have attempt to minimize the incongruity between the developers and the system operators, with the objective of aligning relationships between technical knowledge and equipment operator.

Information and knowledge of the OPSS may be presented in several formats, such as electronic documentation, information retrieval system, hyperdocuments, content sensitive help, on-line advisory, interactive computer based training, simulation and scenario playback capability. Those formats may be references, advisor, or tutor as the operator works to solve his problem.

Operator needs will vary depending what function is being performed. For example the operator may be assigned to a surface combatant Maintenance Division, or may be involved in ship-wide training mode. Information to support those activities is time dependent and requires a different treatment then if an operator needed to perform a specific task immediately. Again we try to be pragmatic and apply the user oriented approach for the development of the system. A conscious effort has been used to keep the system tight and simple. Whatever becomes incorporated into the system must help the operator's performance. The system is therefore performance-oriented, rather than information or transaction-centered.

DEVELOPMENT OF MODELS

Models of the OPSS interface are built with simple application software and are empty shell representations of what the final system display will be. The models are based on the knowledge acquisition process and the proposed selection of information presentation. They are representations of the proposed GUI and how the human/computer interaction will be executed. There are two purposes for the development of models. One is for the end user community to verify the useability of the designed item and to ensure that the user's perspective has been taken into consideration before system development. The second purpose is for system requirement definitions to be used by the developers.

Design characteristics involve typical forms of computer interactions such as menu selection, command manipulation, forms fill and direct manipulation. Elements are accessed directly by offering hot spots in the displays such as a word, a group of words, a marked area in a picture and jump ahead command. Tools for information search are provided so the operator may search for information using words, combination of words and multiple selection. Nontrivial feedback dialogue are provided only when needed. The design is operator centered. With design features that allow the operator's logical intuitive interaction with the system. Motivational factors such as attention, relevance, confidence and satisfaction have also been considered. The operator will become involved and will be confident that progress towards his goals are being made.

DATABASE SYSTEM

The database system definition is performed by a systematic analysis and definition of the specification of the database management system (DBMS), the database (stored data) requirements, and the complete set of application programs (tools) used in the OPSS.

Database management systems (DBMS) have proven to be cost-effective tools for organizing and maintaining large volumes of data in the OPSS. Its primary function is to store data and provide operations on the databases. The operations required for OPSS are: create, delete, update and search (ad hoc query) of data. OPSS data processing requires databases that store large quantities of information having complex structures. A database schema or class hierarchy is developed describing the logical structure of the database supporting the overall system design, the relationship between individual components, and the operations that must be performed. We had to analyze the relative merits of relational and object-oriented database management systems (RDBMS and ODBMS) and the available commercial DBMS. A systematic approach to this analysis entailed a careful evaluation of needs, and how well those needs were met by available products. An ODBMS was eventually chosen for this project.

OPSS ODBMS is able to handle inheritance linking, polymorphism, run time binding, dynamic binding, ad hoc query, security and semantic integrity. It also has a seamless integration to C++ programming language interface. The application programming interface (API) adheres to the industry standards for this language. The ODBMS has a database browser, debugger and a graphical schema design utility that allow for quick design and modification of the database schema.

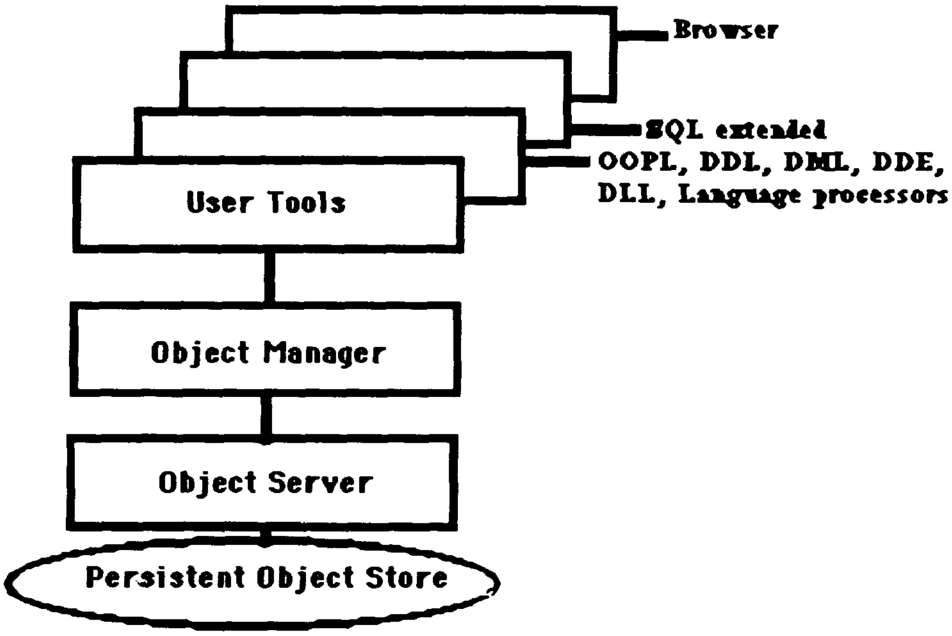


Figure 2. Overall Architecture of the OPSS ODBMS

terminology: Object -oriented Programming Languages (OOPL)
Dynamic Link Library (DLL)
Data Manipulation Language (DML)
Dynamic Data Exchange (DDE)
Dynamic Link Library (DLL)

DEVELOPMENT OF OPSS

The system is developed in modules. Most of those modules are developed with commercial available application software or tools. Many of the OPSS application software are Macintosh based, however the OPSS user program is Windows 3.1 on an MS-DOS operating system. This has not presented any problems because software is available to allow Macintosh files to be ported to the Windows MS-DOS environment. Those modules are then encapsulated with their internal state hidden, to be called up by the ODBMS. Intuitively one may visualize the encapsulation of an object as making it into a "black box" and the DBMS as a pointer that calls on it's functionalaties .

The ongoing OPSS development effort is performed by a group of specialized experts. The developer has the details of the requirement definition established by the concept tree mapping and from the models described earlier. As the modules are developed, the end user participates in the operational test and evaluation process. The implementation of OPSS is being performed in batches.

CONCLUSION

New technologies have become available to us today that provide us with tools to develop a system with complex databases. Those complex databases are composed of information and knowledge in an array of media. Today's database management system allows information and knowledge to be retrieved and manipulated quickly. By incorporating members from the end user community in the project, and by developing the system in an "end-user environment", OPSS promises to provide the operator with the electronic support required to allow him improved performance in the complex and fast reaction world of the military.

BIBLIOGRAPHY

Chin, John P. and Norman, Kent L.. "The menu metaphor: food for thought". Behavior and Information Technology, 1989, Vol. 8, No. 2, p. 125-134.

Colley, Grant "Retain the object model and retain the benefits". Object Magazine May/June 1992, p.27-28.

Dadrowski, Christopher E., Fong, Elizabeth N. and Yang, Deyuan. Object Database Management Systems: Concepts and Features. National Institute of Standards and Technology Special Publication 500-179, April 1990.

Deitel, H. M. and Kogan, M. S.. The Design of OS/2. Addison-Wesley: Reading, MA, 1992.

Gancarz, Robert and Colley, Grant "Consideration for evaluating object database management systems". Object Magazine March/April 1992, p. 57-61.

Gery, Gloria J.. Electronic Performance Support System. Weingarten Publication: Boston, 1991.

Interrante, L. and Biegel, J. (1989). "Automatic Knowledge Acquisition for an Intelligent Simulation Training System". Proceedings 1989 Annual Conference of the International Association of Knowledge Engineers.

Keller, John M.. "Development and Use of the ARCS Model of Instructional Design". Journal of Instructional Development, 1987, Vol. 10, No. 3, p. 2-10.

Laverson, Alan, Norman, Kent and Schneiderman, Ben. "An evaluation of jump-ahead technology in menu selection". Behavior and Information Technology, 1987, Vol. 6, No. 2, p. 97-108.

Martin, James. Hyperdocuments and How to Create Them. Prentice Hall: Englewood Cliffs, New Jersey, 1990.

Mandelkern, Dave. "Visual programming". Object Magazine September/October 1992, p. 39-43.

Mills, Carol B. and Weldon, Linda J.. "Reading Text from Computer Screens". ACM Computing Surveys, 1987, Vol. 19, No. 4, p. 329-358.

McNeese, Michael D. and Zaff, Brian S.. "Design Acquisition: Translating User Knowledge into Design Solutions". Proceedings of the Human Factors Society - Interface 1991, (p. 42-49).

McNeese, Michael D., Zaff, Brian S., Peio, Karen J., Snyder, Daniel E., Duncan, John C., McFarren, Michael R. An Advance Knowledge and Design Acquisition Methodology: Application for the Pilot's Associate. Defense Technical Information Center AD-AA233 700. Armstrong Aerospace Medical Research Laboratory Technical Report # 90-060.

Moshiri, E. (1989). "An Integrated Knowledge Acquisition Environment". Proceedings 1989 Annual Conference of the International Association of Knowledge Engineers.

Norman, Kent L., Weldon, Linda J., and Shneiderman. "Cognitive layouts of windows and screens for user interfaces". International Journal Man-Machine Studies, 1986, Vol.25, p. 229-248.

Shneiderman, Ben. "Designing Menu Selection Systems". Journal of the American Society for Information Science, 1986, Vol. 37, No. 2, p. 57-70.

Taylor, David "The coming convergence of object and relational databases". Object Magazine September/October 1992, p. 16-18.

Vasan, Robin "Integrating object and relational databases". Object Magazine July/August 1992, p. 59-61.

OML

MANUFACTURING TECHNOLOGY
PART 5

[redacted]

A NEW TECHNOLOGY FOR MANUFACTURING SCHEDULING DERIVED FROM SPACE SYSTEM OPERATIONS

R.S. Hornstein
Office of Space Communications
NASA Headquarters
Washington, DC, USA

J.K. Willoughby
President
Avyx, Inc.
Englewood, CO, USA

ABSTRACT

A new technology for producing finite capacity schedules has been developed in response to complex requirements for operating space systems such as the Space Shuttle, the Space Station, and the Deep Space Network for telecommunications. This technology has proven its effectiveness in manufacturing environments where popular scheduling techniques associated with Materials Resources Planning (MRP II) and with factory simulation are not adequate for shop-floor work planning and control.

The technology has three components. The first is a set of data structures that accommodate an extremely general description of a factory's resources, its manufacturing activities, and the constraints imposed by the environment. The second component is a language and set of software utilities that enable a rapid synthesis of functional capabilities. The third component is an algorithmic architecture called the Five Ruleset Model which accommodates the unique needs of each factory.

Using the new technology, systems can model activities that generate, consume, and/or obligate resources. This allows work-in-process (WIP) to be generated and used; it permits constraints to be imposed on intermediate as well as finished goods inventories. It is also possible to match as closely as possible both the current factory state and future conditions such as promise dates. Schedule revisions can be accommodated without impacting the entire production schedule.

Applications have been successful in both discrete and process manufacturing environments. The availability of a high-quality finite capacity production planning capability enhances the data management capabilities of MRP II systems. These schedules can be integrated with shop-floor data collection systems and accounting systems. Using the new technology, semi-custom systems can be developed at costs that are comparable to products that do not have equivalent functional capabilities and/or extensibility.

BACKGROUND

The operations of a space system such as the Space Shuttle, the Space Station, or a telecommunications satellite network have surprising similarities to running a manufacturing facility. Both space systems and manufacturing plants use scarce and expensive resources to satisfy objectives as efficiently as possible. Both must revise their expected activities when equipment malfunctions. Both must respond to opportunities that present themselves unexpectedly. In the realm of space systems operations, the resources may be satellites, antennae, or astronauts; whereas in a factory the resources are production lines, machines, and skilled laborers. A target of opportunity such as a solar flare, the creation of a distant black hole, or an Atlantic Ocean hurricane are to space operations what special orders or unforecasted sales demands are to manufacturing. In both domains, the goal is to get as much from the limited resources as possible, and to do so in a manner that responds to a changing environment.

Recently, the needs for "Finite Capacity Planning" and "Finite Capacity Scheduling" have been recognized by the manufacturing community [1]. Although scheduling has long been a part of manufacturing support software such as an MRP II system, the logic used in those systems does not adequately model the limitations that exist hour-by-hour on the shop floor. These limitations represent the finite capacity that must be modeled accurately and updated frequently in order to plan and replan the production activities.

The finite capacity of space systems has been the driver for developing a new technology for scheduling and rescheduling operational activities. This technology has been applied successfully to several complex manufacturing environments. Very few requirements from manufacturing environments have stretched the generality and completeness of the technology that has emerged from the space operations domain. Therefore, manufacturers can benefit by inheriting the capabilities embodied in the new technology with minimum costs for customization.

There are several components to the scheduling technology that has emerged from space operations. The single greatest design driver for all of these components has been the need to adapt easily from one application to another. In the 1960s and 1970s, each new space system required a start-from-scratch design of the software systems needed to support operations. It was generally conceded that the approaches used for the Apollo (Moon Landing) Program would not work well for operating the new Space Station Freedom, or that planning the communications with satellites using orbiting relay satellites would require different software than that used to plan the communications with ground-based antennae. The costs for each new application were very high. As a result, NASA sought ways to abstract the planing and scheduling problem, i.e., to find a generic way to describe and solve these problems that could be applied to any new space program.

The analogy to manufacturing is again apparent. Analysts have regarded the differences among production environments to be sufficient to justify custom development of finite capacity shop floor planning and scheduling systems. For example, the details of an aircraft brake manufacturing plant were not seen as similar to the processes for producing soups and canned vegetables. The search for generalizations and descriptive abstractions was not seen as a feasible task.

DESCRIPTIVE ELEMENTS

In Table 1, elements of a descriptive vocabulary are shown which are domain independent. The Table is not a complete presentation of all possible descriptive elements, but gives some examples from both manufacturing and the space operations world. The authors have been involved in the development and the application of planning and scheduling techniques to several space and manufacturing systems. Although the use of one single system for all of these environments is not (yet) possible, the degree of reusability of concepts, data structures, system architectures, algorithmic components, and software modules is remarkably high, and still increasing rapidly. Knowledge gained from one application suggests an approach that can be generalized; the result is that each successive application benefits from a rapid accumulation of reusable software features and modules.

GENERIC DESCRIPTIVE ELEMENT	EXAMPLE FROM SPACE OPERATIONS	EXAMPLE FROM MANUFACTURING
Resource (Pooled)	Propellant Bandwidth Electrical Power	Labor Skills Chemicals Work-in-process (WIP) Base Brand
Resource (Individual)	Crew Person Antenna Tape Recorder	Milling Machine Furnace Crane
Activity	Playback Recorder Crews Sleep Send Command Sequence	Make Subassembly X Ship Product X Perform Preventive Maintenance
Condition	Daylight Only Alternate Orbits Every 3 Earth Days	First Shift Only Not on Weekends
Temporal Relations	Record Before Playback Exercise Before Eating	Routings PM at Least Every 3 Days

Table 1: Analogies Between Space Operations and Manufacturing

The accumulation of generic insights has produced a set of descriptive data structures that are inherently hierarchic and asymmetric. For example, the generalized description of an activity to be scheduled, whether it be in space operations or manufacturing, can be captured in the data structure shown in Figure 1. Experience has shown that different application domains will require more or less information at any level in this tree-like structure, but will not require new levels in the data structure. Some applications will have activities that result in broad bushy activity trees; others will use narrow or sparce structures. Note that the number of branches and levels in some parts of the data structure are not the same as those in other parts. Hence, the observation that the data structures are asymmetric. This characteristic makes scheduling data difficult to manipulate in traditional tables or matrices which are the fundamental data structures of modern relational data base systems. We have found repeatedly that although relational data base systems are very appropriate for storing, retrieving, and reporting the inputs and outputs from a scheduling process, they are inappropriate for supporting the computational process of generating or revising a schedule.

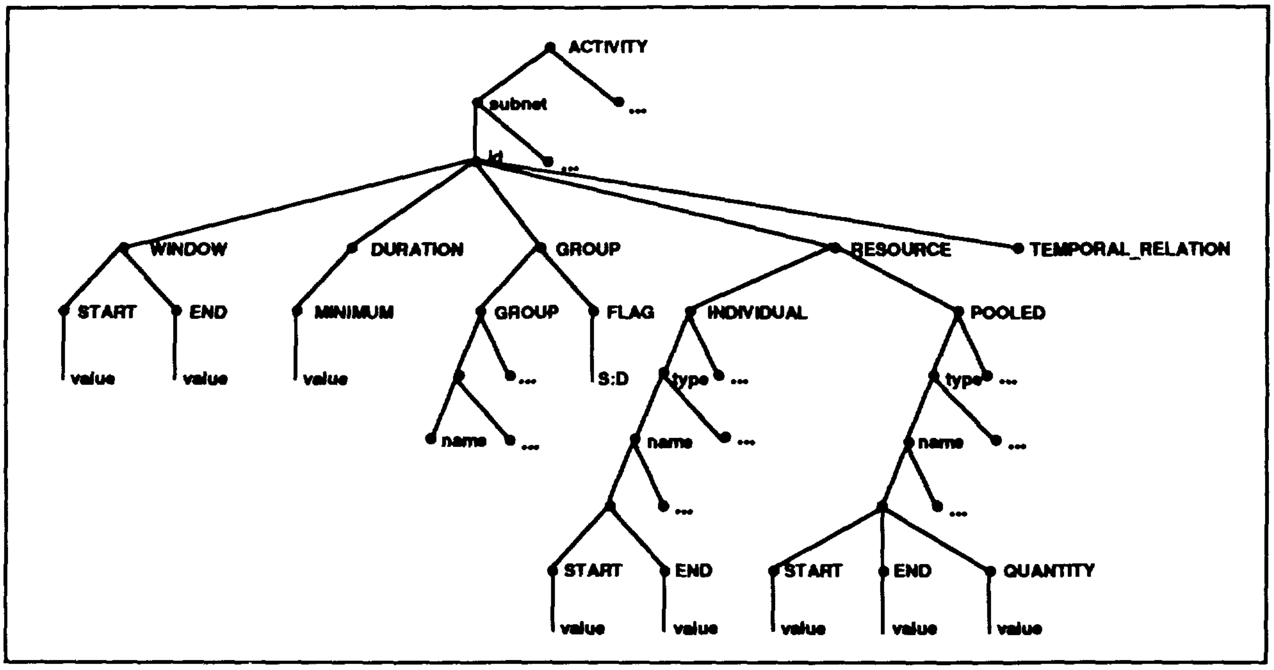


Figure 1: Hierarchic Asymmetric Structure of Activities to be Scheduled

Information about resource limitations (finite capacity descriptions) and timing OR sequencing constraints are also easily represented by asymmetric hierarchic data structures. Some of these structures are shown generically in Figures 2a and 2b. As a result, the descriptive mechanisms that have proven to be the most transportable, i.e., the easiest to apply in a very broad range of application domains are these hierarchic tree-like structures. These structures fit well with the concepts of object-oriented software development.

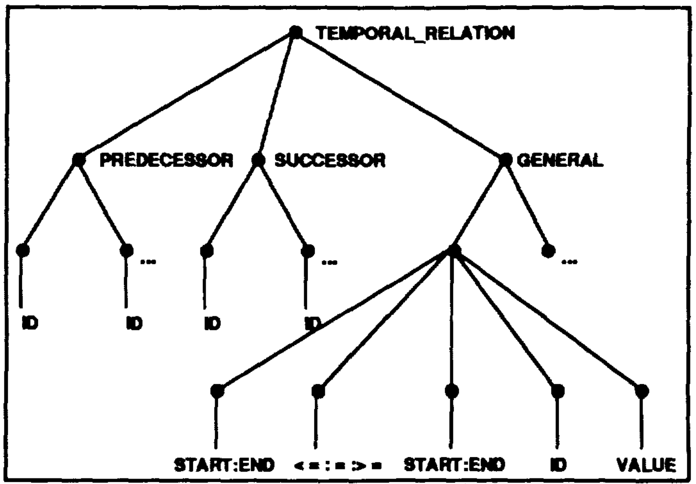


Figure 2a: Hierarchic Asymmetric Structure of Temporal Constraint Data

The generic descriptive framework provided by these data structures provides several advantages specifically for manufacturing environments. Among them is the generality of the resource modeling. Most manufacturing support software makes distinctions among different types of resources that the generic data structures do not. There need not be modeling differences between inventories, machines, power, raw materials, supplies, work-in-process (WIP), or labor using the generic resource data model. Since any resource can be obligated, generated, consumed, or have its attributes transformed, the modeling flexibility is enormous. For example, one step in a manufacturing process can create a new resource which a subsequent step can consume. This allows modeling of flexible routings without complex sequence relationships. In applications of generic resource data models, the authors have found significant efficiencies that could not have been possible if fixed routings had been imposed by the limitations of a scheduling system.

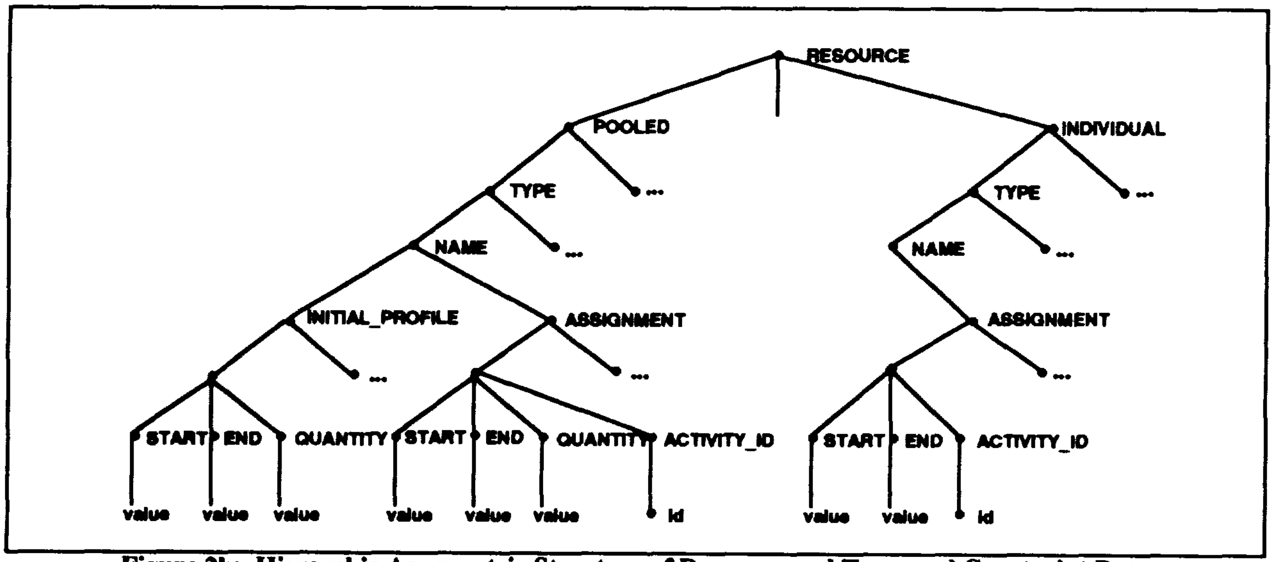
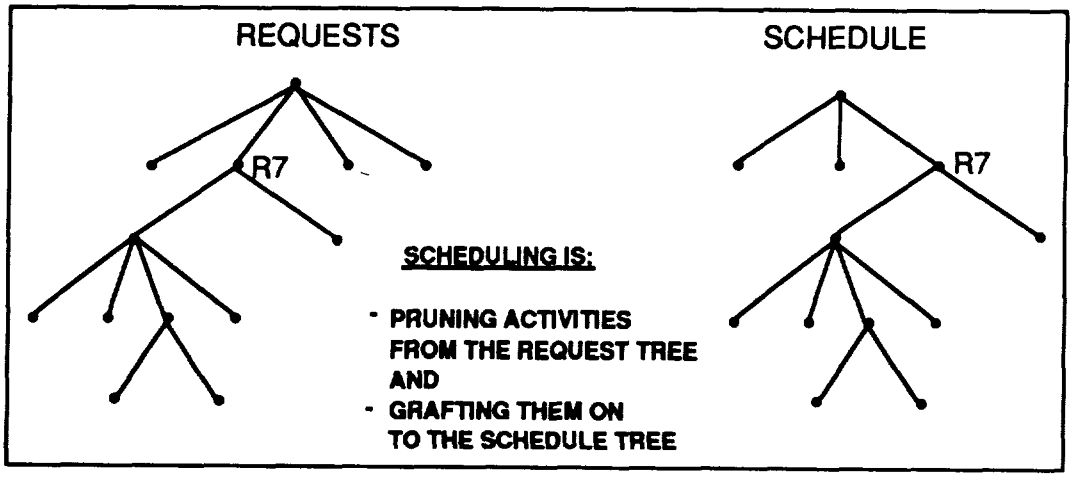


Figure 2b: Hierarchic Asymmetric Structure of Resource and Temporal Constraint Data

Another example of the resource data model flexibility is the use of generic descriptors with any resource. These descriptors can cause a part of any order to be tracked from one "location" to another or from one state of "completion" to another. The location and/or completion descriptors are simply attached to the resource. Any number of such descriptors can be used on any resource.

A LANGUAGE AND UTILITY LIBRARY

The solution to planning and scheduling problems in either space operations or manufacturing requires the manipulation of these tree-like data structures. Goal-directed research in the 1970s and 1980s along with field testing and revision led to a simple realization. If the output of a scheduling process, i.e., a schedule, was a hierarchic asymmetric data structure that looked a lot like the input data structures, then the scheduling process should be describable as a systematic manipulation of input data structures into output data structures. The concept is illustrated in Figure 3.



SCHEDULING IS:

- PRUNING ACTIVITIES FROM THE REQUEST TREE AND
- GRAFTING THEM ON TO THE SCHEDULE TREE

Figure 3: Scheduling as Tree Manipulation

Our premise has become the following:

> Scheduling can be described as the systematic manipulation of tree-like data structures in such a way that objectives are met and constraints are satisfied.

A software programming language that was idealized for manipulating these tree-like structures was devised in order to test this premise. Over twenty scheduling applications have been developed using this language [2] with the result that the average size and development time for applications have been reduced by a factor of approximately twenty when compared to custom-built systems from the 1970s and 1980s. This language has now evolved to be a set of data-structure manipulators written in C++ [3]. Scheduling systems are currently under development using these tree-manipulation capabilities in C++ for both space operations and manufacturing applications.

ALGORITHMIC ARCHITECTURE

To complement the data manipulation capabilities, a set of scheduling utilities has also been developed. These utilities are software modules that find frequent use in all scheduling applications. These reusable modules perform constraint checking, interval and set algebra, and data management operations that are independent of any application domain. The names of these modules are shown in Table 2. The module names suggest their functionality.

Scheduling Routines

- FindEarliestTime
- FindLatestTime
- ScheduleActivity
- UnscheduleActivity

Resource Management Routines

- AddResource Assignment
- AssignResources
- DetermineConsumableAvailability
- DetermineReusableAvailability
- FreeResource
- UpdateResourceAvailability
- UpdateResourceStates

Interval Algebra Routines

- CollapseIntervals
- ComplementInterval
- ComputeIntervalDuration
- IntersectIntervals
- UnionIntervals

Miscellaneous Routines

- GetForestVersion

Activity Management Routines

- InvertRelations
- OrderByRelations

Constraint Enforcement Routines

- EnforceRelations
- EnforceConditions
- EnforceResources

Profile Management Routines

- AddProfile
- CollapseProfile
- ComplementProfile
- ComputeProfileEnd
- ComputeSegmentArea
- ComputeSlope
- GetSegmentTime
- GetSegmentQuantity
- IntersectProfileSegments
- IntersectProfiles
- SetProfile
- SubtractProfile
- UnionProfiles

Table 2: Contents of a Scheduling Utility Library

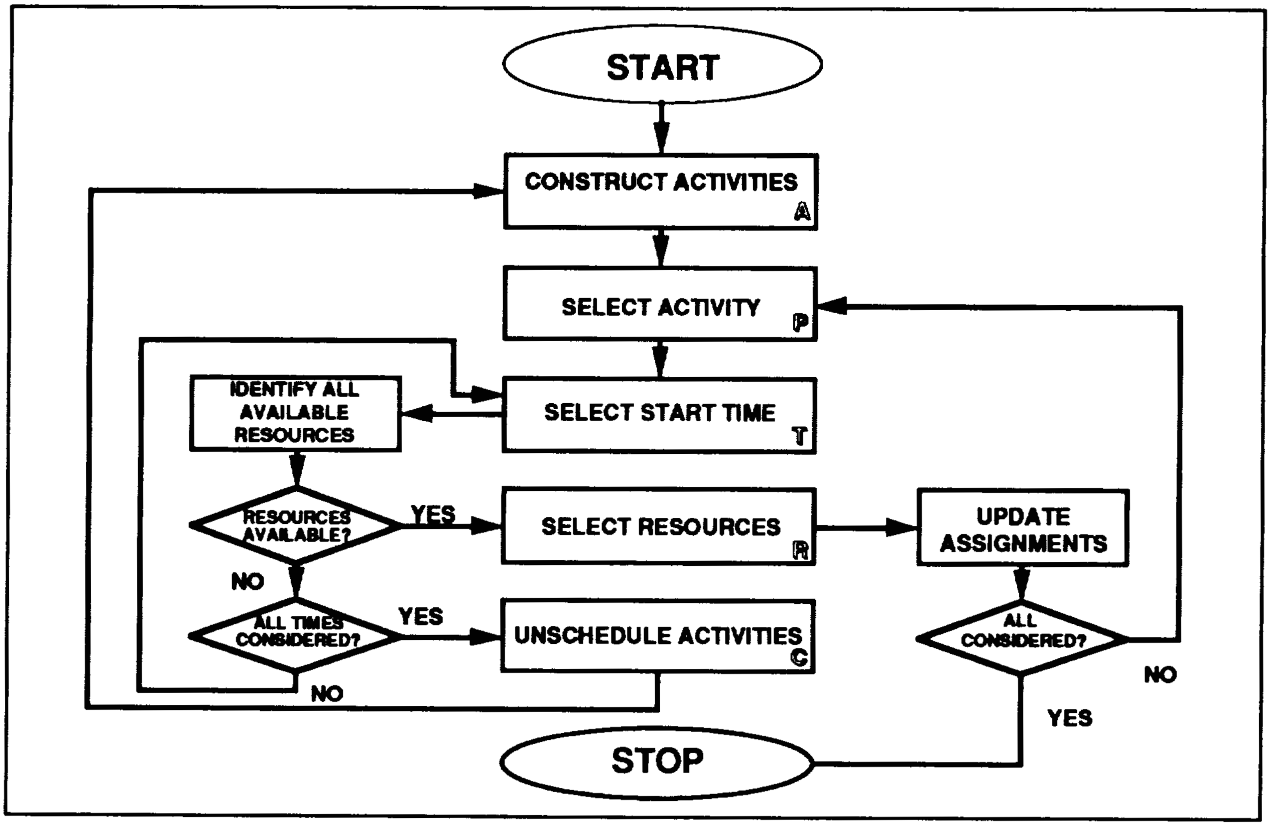


Figure 4: The Five Ruleset Model Architecture

In addition to the descriptive data structures and the language for application building already described, a third component of the technology is an algorithmic architecture called the Five Ruleset Model. This Model is a framework for describing the decision-making processes used in a broad range of scheduling algorithms. As shown in Figure 4, the Five Ruleset Model decomposes the solution logic of an algorithm into five nearly-decoupled sets of decisions called rulesets. The concept is that once each of these rulesets is specified, a unique algorithm is completely specified. All of the non-decision-making software can be pre-built and available as reusable code. The majority of most scheduling applications has proven to be in constraint checking and bookkeeping of the obligations of the resources. These portions of the scheduling algorithm can be pre-built using the data structures, the language, and the utility library approach previously described. The remaining tasks in implementing a system for a new environment are the specification and implementation of the unique decision-making rulesets for the Five Ruleset Model. Once determined, these rulesets can be inserted into the architecture as shown in Figure 5.

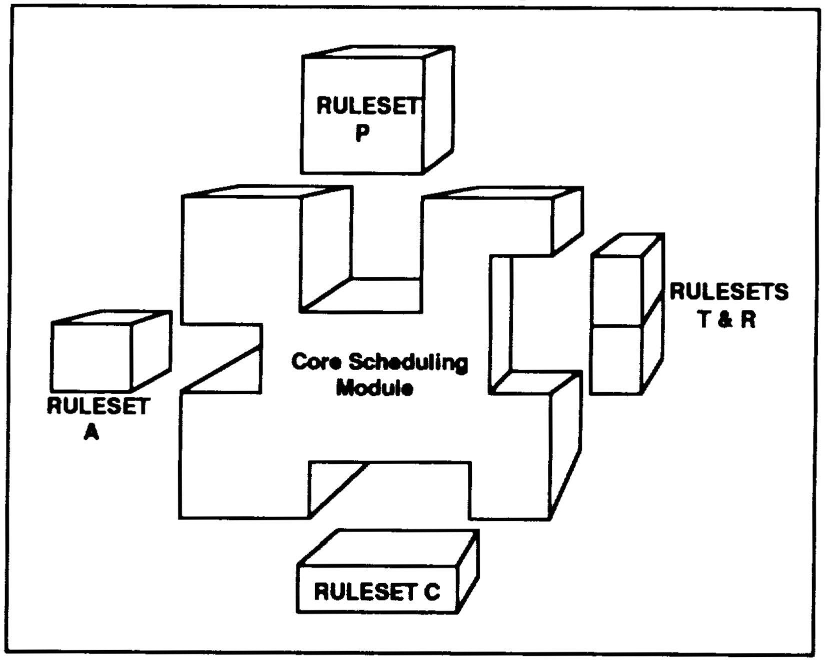


Figure 5: Reuse of the Five Ruleset Model for Different Applications

An important feature of the Five Ruleset Model architecture is its ability to support time-transcendent scheduling. This means that activities can be put on a timeline in any order, not strictly earliest to latest (as in simulation or queuing models) or latest to earliest (as in MRP II scheduling). With time-transcendent algorithms (sometimes called serial), activities can be inserted between other activities. This capability has two important ramifications:

1. Boundary conditions at both ends of a scheduling horizon can be met. For example, in manufacturing, the current state of the factory and the promised delivery dates can be taken as constraints.

2. Rescheduling does not require the revision of the entire timeline; simulations require the unraveling of a timeline in order to make a change. When this is done, the effect of the change will ripple through the timeline causing undesirable disruption to the operational process.

The Five Ruleset Model provides benefits for rapid application development as well as enhanced functionality over techniques common in manufacturing support software today.

SUMMARY

The three components of the technology have been applied successfully in both discrete and process manufacturing environments. The finite capacity scheduling applications can be interfaced with the information management capabilities of modern MRP II systems, with process control systems, and/or with Data Base Management Systems. The emphasis on generality, reusability, and extensibility has led to systems that are rapidly deployable and are easily modifiable when business rules change or when the manufacturing enterprise grows. The concepts transferred from the space operations world appear to be robust enough to contribute immediately to manufacturing environments. The costs are much lower than those associated with build-from-scratch solutions and the results include functionality that is not available in currently-used manufacturing support systems.

REFERENCES

1. *Just in Time: Implementing the New Strategy*, Maskell, Brian H., Hitchcock Publishing Company, Carol Stream, IL, 1989.

2. *Advanced Scheduling Environment (ASE™) Reference Manual*, Avyx, Inc., Englewood, CO, 1992.

3. *Seamless Transitions from Early Prototypes to Mature Operational Software: A Technology that Enables the Process for Planning and Scheduling Applications*, Hornstein, R.S., J.K. Willoughby and D.A. Wunderlich, Information Sciences, Inc., Englewood, CO, 1992.

THREE-DIMENSIONAL LASER WINDOW FORMATION

FOR INDUSTRIAL APPLICATION

Vincent G. Verhoff and David Kowalski
National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135

ABSTRACT

The NASA Lewis Research Center has developed and implemented a unique process for forming flawless three-dimensional, compound-curvature laser windows to extreme accuracies. These windows represent an integral component of specialized nonintrusive laser data acquisition systems that are used in a variety of compressor and turbine research testing facilities. These windows are molded to the flow surface profile of turbine and compressor casings and are required to withstand extremely high pressures and temperatures. This method of glass formation could also be used to form compound-curvature mirrors that would require little polishing and for a variety of industrial applications, including research view ports for testing devices and view ports for factory machines with compound-curvature casings. Currently, sodium-alumino-silicate glass is recommended for three-dimensional laser windows because of its high strength due to chemical strengthening and its optical clarity. This paper discusses the main aspects of three-dimensional laser window formation. It focuses on the unique methodology and the peculiarities that are associated with the formation of these windows.

INTRODUCTION

An increased interest in fundamental research in turbines and compressors has created a need for nonintrusive optical flow measurement systems. The state-of-the-art systems used in obtaining detailed velocity data are called nonintrusive laser data acquisition systems. These systems seed the airflow with small particles that flow through a fringe pattern which is created by intersecting laser beams. Data are obtained by measuring the pulsating light that is reflected as the seed particles pass through the fringe pattern.

Optically clear laser windows are used so that the laser beams and reflected light can pass through. Normally the laser window glass that is used in wind tunnels is approximately 1.0-in.-thick flat quartz. For turbine and compressor testing facilities, however, the windows are approximately 0.100 in. thick and have three dimensions. The difference between the two windows is shown in figure 1.

Two types of errors can be introduced to the system by the laser window: spatial error of the measurement volume and reduced signal amplitude. Spatial error of the measurement volume is caused when the laser beams pass through the window at incident angles. The actual focal point is then skewed from the desired focal point, resulting in an error [1], as depicted in figure 2. Errors associated with reduced signal amplitude are caused by reflection of the laser beam as it passes through the window. Window size, curvature, thickness, surface quality, and contour tolerance are the major factors that control the magnitude of error.

Because laser windows are molded to the flow surface profile of the turbine and compressor casings, their size and curvature cannot be altered. Window thickness, on the other hand, can vary. Thinner windows are desirable because they minimize error, but the glass must maintain high strength with the reduced thickness. In addition, these thinner curved windows must be able to withstand high pressure and temperature differentials while preserving surface quality. This paper narratively addresses three-dimensional laser window formation and the process that maximizes the surface quality and the contour accuracy. A more detailed explanation of three-dimensional laser window formation is given in [2]. In this paper three-dimensional laser windows will be referred to as "windows."

BACKGROUND

Laser systems are commonly used at NASA Lewis. Most of these systems are located in the engine component test facilities that test compressor and turbine rotors and components. The windows are usually located over rotating hardware where typical instrumentation cannot be used.

Overall safety is of primary concern. Therefore, the windows are hydrostatically pressure tested to 1.5 times the maximum operating pressures of the facilities. Operating pressures for various facilities range between 1.3 and 72 psia. Windows are also thermally qualified at facility temperatures and pressures if thermal differentials are significant. The windows are safety tested on both concave and convex surfaces.

GLASS SELECTION

Several types of glass have been used for window formation at NASA Lewis. Sodium-lime, borosilicate, and sodium-alumino-silicate glasses are among these types. On the basis of the Corning glass code 0317 [2], sodium-alumino-silicate glass has been determined to be the preferred glass for windows. This glass is recommended because of its ultrahigh strength through chemical strengthening, which is unavailable with other glass types. This increased strength allows the use of thinner windows, which reduce spatial error and produce higher quality laser data.

In addition, failure of a chemically strengthened sodium-alumino-silicate laser window will cause the window to shatter in tiny particles, only millimeters in cross-sectional area. This characteristic is advantageous because the smaller particles contain less kinetic energy and are less likely to damage blades or rotors. The particles from other failed glass windows are relatively large in cross-sectional area. A failed chemically strengthened sodium-alumino-silicate window is represented in figure 3.

MOLD DESIGN

The mold material that offers the best results for forming windows is machinable ceramic. The molds for window formation consist of a male and female mold. The male mold is machined to match the internal flow path surfaces, whereas the female mold is machined to these coordinates plus the glass thickness. The overall dimensions of the molds are 1 in. greater than the actual final size of the window glass. Both molds have threaded holes on the perimeter to fit alignment bars.

In order to maintain surface quality, accurate machining and polishing of the molds that are used for window formation are essential. The molds are machined and polished to a tolerance of 0.005 in. of the desired contour with a 16 Ra surface.

All slumping components are machined out of the same material to ensure similar coefficients of thermal expansion. Orientation of the slumping component configuration is shown in figure 4.

GLASS MOLDING

Inert gas furnaces are preferred for slumping because of the heating characteristics of the molds. Other types of furnaces usually introduce contaminants into the slumping environment that degrade glass surface quality. Cleanliness of the molds, glass, and furnace is critical to glass quality. These components should be thoroughly cleaned at the beginning of every slumping operation.

The forming temperature of windows is found by an iterative process. The mean annealing temperature of about 1100 °F is the theoretical starting temperature for new window formation. Adequate visual inspection of the window after slumping will indicate whether the temperature should be increased or decreased. The ideal slumping temperature for three-dimensional window formation is usually within 2 percent of the annealing temperature.

The procedure for window formation is as follows:

(1) Cut the glass to overall mold dimensions (i.e., 1 in. greater than the final window design dimensions).
(2) Thoroughly clean the glass with soap and water. Oil from fingers will develop into surface imperfections during the slumping operation.
(3) Thoroughly clean the male and female molds with an alcohol-based cleaner.
(4) Thoroughly clean the inert gas furnace.
(5) Bolt the alignment bars onto the molds.
(6) Position the male mold into the inert gas furnace.
(7) Insert the glass on top of the male mold.
(8) Position the female mold onto the male mold with the glass positioned in between them.
(9) Heat the furnace to slumping temperature.
(10) Soak the glass at slumping temperature for 4 to 6 hr.
(11) Cool the glass down to ambient temperature.
(12) Examine the slumped glass for proper curvature and quality.
(13) After desired curvature and quality are obtained, anneal the glass to relieve residual stresses.

Visual inspection of the slumped window will provide adequate information for altering the slumping temperature. If the slumping temperature is excessive, surface imperfections will be apparent. These surface imperfections will appear along the plane of severe three-dimensional contour or at the inflection point. This problem can be solved by decreasing the slumping temperature by 1 to 2 percent of the annealing temperature.

The combination of insufficient slumping temperature and excessive mold pressure may result in low-quality windows. These windows will appear wavy in the area where the glass was stretched. The solution to this problem is to increase both the slumping temperature and duration.

The windows can be molded in several steps by alternating slumping temperatures, modifying slump soak time, varying molding force, inverting mold positions, or any combination of these methods. When molding glass in several steps the female mold is always used first and by itself. After the glass is partially formed, the male mold is added, the entire assembly is inverted, and the slumping process is then repeated. Windows with extreme curvature or compound curvature may require repeated slumping operations. These complex windows often require the addition of weight to the molds to adequately form the glass.

Window development can be a lengthy process, but once the parameters are found for a particular window, reproduction is routine. The last process in the formation of windows is to anneal the glass. Annealing relieves residual stresses that build up in the glass during the development process.

GLASS EDGING

Previous methods of edging the windows to size involved scoring the glass to the desired size, then breaking the glass along this scored edge. During this process, small fragments of glass are broken away along the line of scoring, creating voids in the glass. Minute cracks remain along the edge of the glass after it is broken. These cracks weaken the glass and cause it to fail under load.

Minimizing cracks is important during the edging process. A perfected process for edging glass by using a numerically controlled water-jet cutting machine has been developed and is recommended for cutting windows. The machine abrades glass away along the desired cut line, reducing glass fragmentation. The water-jet cut edge of the glass is less densely populated with cracks, and crack penetration is less severe than with scored glass.

The desired window dimensions are programmed into the numerically controlled water-jet cutting machine. In order to alleviate window glass damage from the backsplash of the water jet, a 0.0625-in.-thick aluminum plate is placed across the ways of the water-jet cutting machine bed. In order to protect the glass from fracturing, duct seal is placed over the surface of the aluminum plate to absorb the high cutting frequencies of the water jet. The window glass is then secured into position by pressing it into the duct seal. This entire assembly is then submersed under water to further assist in absorbing the high cutting frequencies of the water jet. The maximum desired water-jet cutting rate for edging is 4.75 in./min, with a nozzle pressure of 30 000 psig.

After the glass is cut to size, the edges are rounded. Using aluminum oxide sanding belts helps protect the glass from damage because glass edges are most susceptible to impairment. The minimum radius of the rounded edges is equivalent to half the glass thickness, as shown in figure 5.

GLASS STRENGTHENING

Chemical strengthening of the glass through ion exchange is a readily available technology. The process will only be highlighted in this paper; detailed information is given in references 3 to 8.

Windows are chemically strengthened through ion exchange so that they can endure facility operating pressures and temperatures and to increase their relative impact strength. These windows are chemically strengthened to contain a 0.010-in. or greater compression layer. The specially designed chemical composition of Corning glass code 0317 sodium-alumino-silicate glass enhances this ion exchange to ensure maximized glass strength. This is the principal reason why this glass is used for window development.

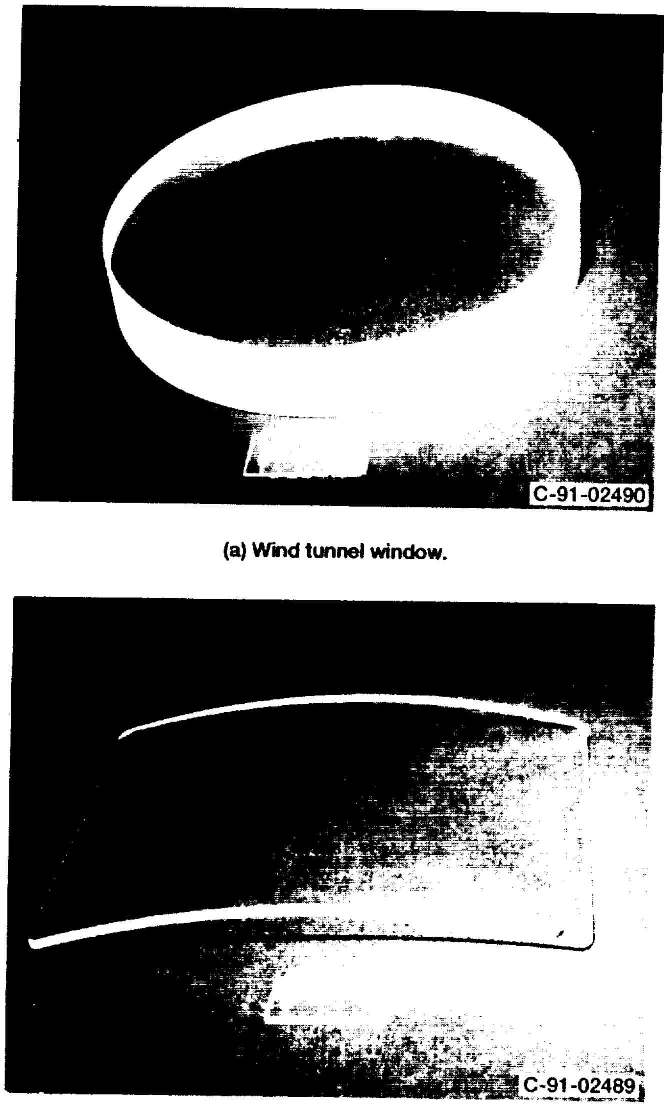
CONCLUDING REMARKS

High-quality contoured windows are a key part of laser anemometry systems for compressor and turbine facilities. The window development process that was described herein ensures accurate tolerances and flawless quality that could only be previously obtained from grinding and polishing. Although they have been tested for flaws, windows should still be considered and treated as a fragile material. Routine visual inspection and hydrostatic evaluations are considered important to their integrity.

Future experiments will need larger windows to increase the fields of unobstructed view. Future windows must also endure higher pressures and temperature gradients. With the present development processes these windows should successfully withstand these future requirements.

REFERENCES

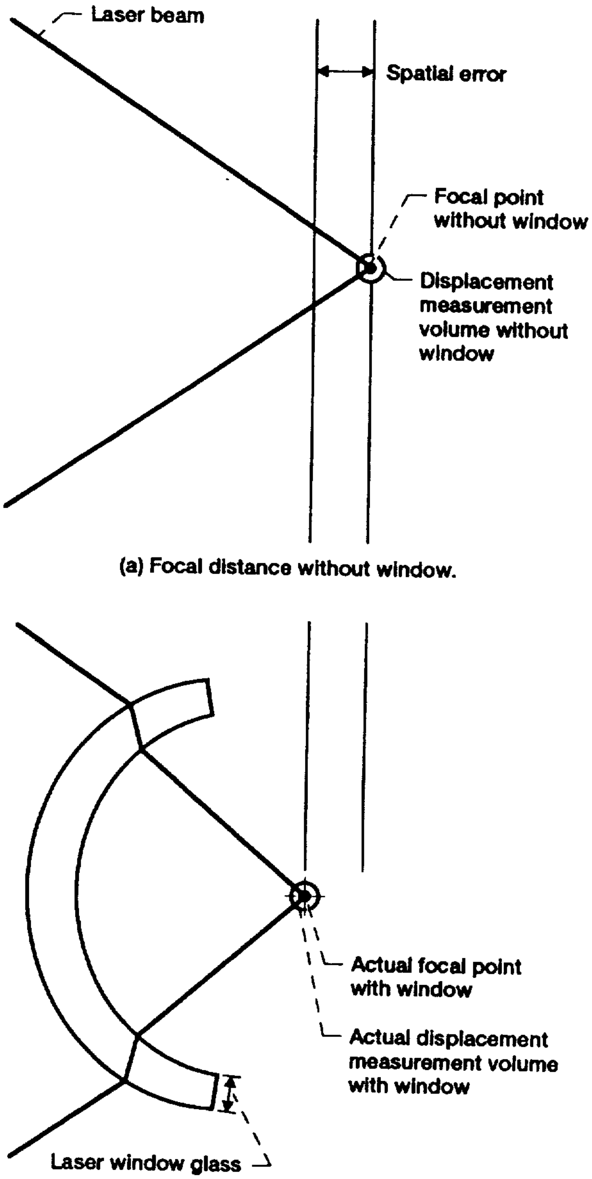
1. Bueche, F.J.: Introduction to Physics for Scientists and Engineers. Third ed. McGraw Hill, 1980, pp. 614-634.
2. Verhoff, V.G.: Three-Dimensional Laser Window Formation, NASA RP-1280, 1992.
3. Corning, Inc.: Corning Glass Code 0317 Specification Sheet. Corning, New York, 1990.
4. McLellan, I.; and Shgnd, G.W. II: Glass Engineering Handbook. Third ed. McGraw Hill, 1984.
5. Nordberg, M.E., et al.: Strengthening by Ion Exchange. Am. Ceram. Soc. J., vol. 47, no. 5, May 1964, pp. 215-219.
6. Hagy, H.E.: Design Strength of a Chemically Strengthened Glass. Central Glass Ceram. Res. Inst. Bull., vol. 13, no. 1, 1966, pp. 29-31.
7. Blizard, J.R.: Chemically Strengthened Glass. SAE Paper 680485, 1968.
8. Miska, H.A.: Understanding the Basics of Chemically Strengthened Glass. Mater. Eng., vol. 83, no. 6, 1976, pp. 38-40.
9. Olcott, J.S.: Strengthening of Glass. Science, vol. 140, no. 3572, June 14, 1963, pp. 1189-1193.



(a) Wind tunnel window.

(b) Turbine or compressor window.

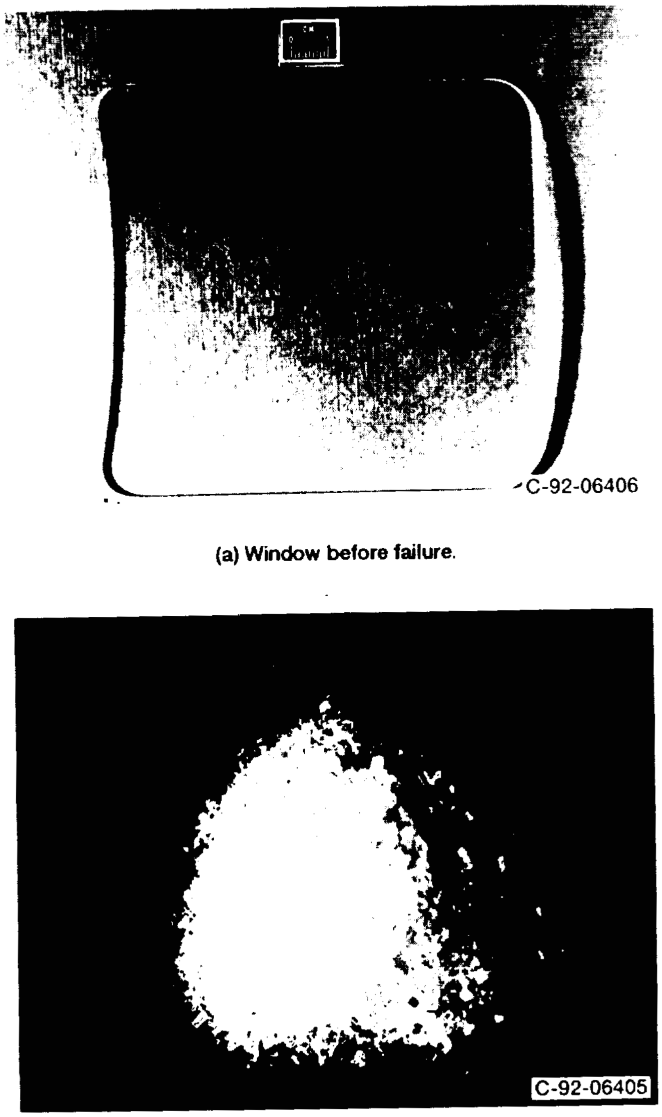
Figure 1.—Typical laser windows.


Laser beam
Spatial error
Focal point without window
Displacement measurement volume without window
Actual focal point with window
Actual displacement measurement volume with window
Laser window glass

(a) Focal distance without window.

(b) Focal distance with window.

Figure 2.—Spatial error associated with laser windows.



(a) Window before failure.

(b) Window after failure.

Figure 3.—Chemically strengthened sodium-alumino-silicate laser window.

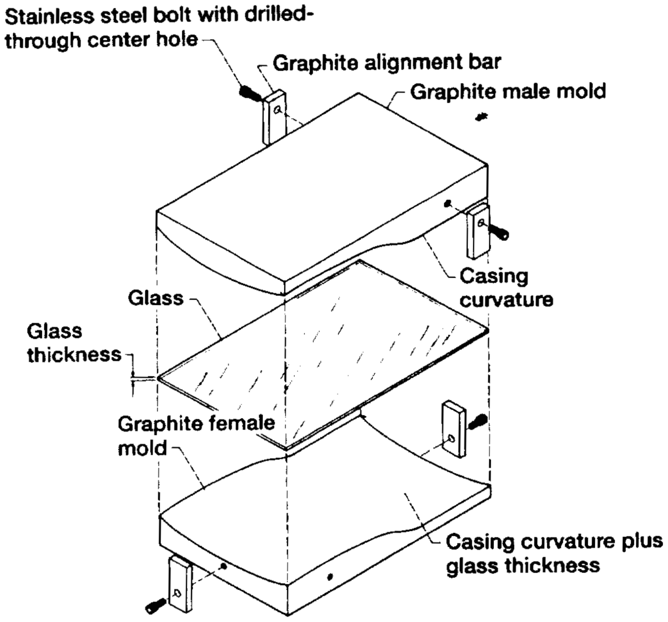


Figure 4.—Slumping components configuration.

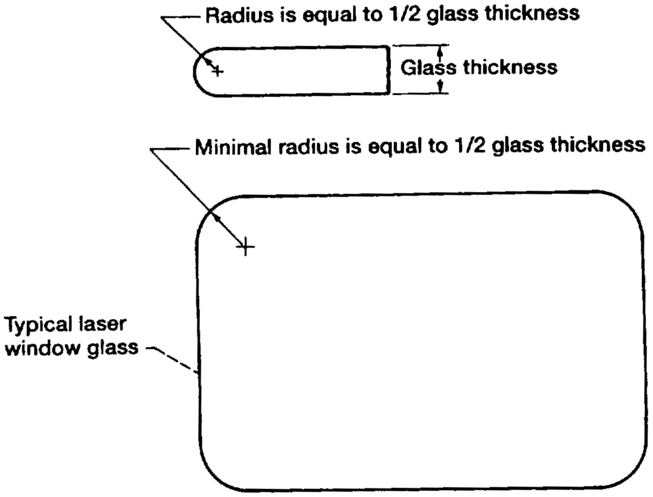

Radius is equal to 1/2 glass thickness
Glass thickness
Minimal radius is equal to 1/2 glass thickness
Typical laser window glass

Figure 5.—Rounded edges of typical laser window.

HIGH PERFORMANCE SAPPHIRE WINDOWS

Stephen C. Bates
Advanced Fuel Research
87 Church Street
East Hartford, CT 06108

Larry Liou
NASA Lewis Research Center
Cleveland, Ohio 44135-3191

ABSTRACT

High-quality, wide-aperture optical access is usually required for the advanced laser diagnostics that can now make a wide variety of non-intrusive measurements of combustion processes. Specially processed and mounted sapphire windows are proposed to provide this optical access to extreme environments. Through surface treatments and proper thermal stress design, single crystal sapphire can be a mechanically equivalent replacement for high strength steel. A prototype sapphire window and mounting system have been developed in a successful NASA SBIR Phase I project. A large and reliable increase in sapphire design strength (as much as 10x) has been achieved, and the initial specifications necessary for these gains have been defined. Failure testing of small windows has conclusively demonstrated the increased sapphire strength, indicating that a nearly flawless surface polish is the primary cause of strengthening, while an unusual mounting arrangement also significantly contributes to a larger effective strength. Phase II work will complete specification and demonstration of these windows, and will fabricate a set for use at NASA. The enhanced capabilities of these high performance sapphire windows will lead to many diagnostic capabilities not previously possible, as well as new applications for sapphire.

INTRODUCTION

The study of the physics and chemistry of combustion and other high temperature and pressure processes is increasingly being performed by advanced laser diagnostics that have extensive and steadily broadening capabilities. These diagnostics often require high-quality wide-aperture optical access that is not currently possible using conventional materials. The combined pressure and temperature constraints of the contained environment are beyond current window materials technology. Glass has too low a mechanical strength, sapphire is thought to be too sensitive to thermal stress and shock, zirconia cannot be manufactured in large pieces, and diamond is eroded too rapidly by an oxidizing atmosphere and is too expensive in large sizes.

SAPPHIRE

Of all of the commonly available materials sapphire is the hardest and has the highest melting point. The weakness of sapphire as a material and a major cause of its limited application has always been its poor thermal stress behavior. Because sapphire is such a strong crystalline solid it is difficult to shape and polish. Its strength, enhanced in single crystals, is far superior to that of glass. The potential of strong sapphire, however, is even greater if ways can be found to extend to large scales the strength that has been demonstrated for small single crystals. It has long been known that the relatively low practical strength of sapphire is a direct result of surface flaws that are generated during the fabrication of sapphire in a particular shape; the bulk crystal itself can be made essentially perfect by using modern crystal growing techniques.

Stoichiometrically there is only one oxide of aluminum: alumina, Al_2O_3. However, alumina can form various polymorphs, hydrated species, etc., depending on the conditions of its preparation. There are two forms of anhydrous Al_2O_3: α-Al_2O_3 and γ-Al_2O_3. In α-Al_2O_3 (sapphire) the oxide ions form a hexagonal close-packed array where the aluminum ions are distributed symmetrically among the octahedral interstices. The α-Al_2O_3 material is stable at high temperatures and is also indefinitely metastable at low temperatures. It occurs naturally as the mineral corundum. The Al_2O_3 that is formed on the surface of aluminum metal has still another structure: a defect rock-salt structure with an arrangement of Al and O ions in a rock-salt ordering with every third Al ion missing. Sapphire is a unique material in a number of ways. Its extreme hardness and chemical inertness is complemented by its very broad bandwidth for optical transmission (0.16 - 5.5 μm). Sapphire has special optical properties in that it has a large surface reflection and is optically active as well as birefringent. The forming and polishing of single crystal sapphire is difficult and time consuming. Common properties are given by manufacturers [1], while the primary reference for sapphire is a Russian book by Belyaev [2].

Mechanical Characteristics of Sapphire

The strength of brittle materials, which at normal temperatures includes single crystal sapphire, depends on a number of factors: rate of testing, temperature, quality of the surface of the specimen, ambient conditions, size of specimen, etc. Sapphire is a single crystal; its properties are orientation dependent and determined by the properties of the crystal itself. Its macroscopic behavior depends strongly on the perfection of the crystal, both internally and at its surface. Stress failure occurs as a result of the formation and propagation of cracks in the crystal. Near 900 °C sapphire becomes plastic, as the thermal energy in the crystal becomes large enough to enable the weakest slip planes in the crystal to move. Of all of the properties of sapphire, those related to its strength are the least well defined because the failure of sapphire is statistical; design values must be based on the minimum possible strength. The mechanical properties of sapphire are given as:

Tensile strength: at 25 °C: 410 MPa (60 kpsi) (design criterion)
at 500 °C: 280 MPa (design criterion), at 1000 °C: 350 MPa (design criterion)
Bulk modulus: 2.4 x 10^{11} Pa, Young's modulus (60° to c-axis): 3.45 x 10^{11}
Modulus of rigidity: 1.5 x 10^{11} Pa, Modulus of rupture: 450 - 700 MPa, Poisson's ratio: 0.25

There are four primary characteristics of a sapphire window that together determine the design strength of the sapphire. These are 1) Bulk crystal quality, 2) Crystal orientation relative to the window planes, 3) Surface preparation, and 4) Avoidance of mechanical design features that cause stress concentration. Defining the current standard practices with respect to these factors determines the baseline case for any strengthening comparison and leads to an understanding of how the fabrication and use of sapphire can be optimized.

Mechanically, sapphire is currently characterized by the optical quality of the bulk single crystal. There are no strict standards for describing the crystal, only approximate grades. The primary reason for this accepted imprecision is the lack of correlation of identifiable defects with the macroscopic behavior of a crystal - except for optical clarity. Optical grading is done both because large sapphire is usually used as an optical material, and because optical testing is the simplest and easiest technique that is used for identifying crystal defects. A microscopic property that does translate to macroscopic behavior is the crystal orientation, where bond strength translates to directional crystal strength.

For the purpose of determining the global mechanical strength of a high quality single crystal of sapphire the condition of the surface is the determining factor. The characterization of the surface finish must be supplemented by a knowledge of the subsurface damage layer that has been created by the mechanical deformation of the surface during the shaping or polishing of the piece. Often sapphire pieces are rapidly diamond polished to high quality without removing the associated damage layer in the crystal. This damage layer often cannot be easily detected, except by a separate chemical surface etching process.

Stress concentration is usually caused by holes or sharp edges. Window design excludes holes, but sharp edges are a common feature of windows simply because edges are rarely specified by the design engineer. Sharp edges produce chips easily, and these chips can scratch the window surface and significantly reduce strength. Sharp corners also occur when a window is made in one piece with two diameters; the large diameter is used for mounting and sealing purposes. The corner between the two diameters is not only a location of stress concentration but a location where there is likely to be surface damage from machining and no polish at all. This edge is a primary cause for the failure of windows using this design, and should be radiussed and polished.

Surface Strengthening

The process of strengthening sapphire by modifying its surface has long been known and practiced in the form of fire polishing. More recently research using glazing to strengthen glasses has been extended to the strengthening of sapphire by a similar glazing process [4,5]. The precise mechanism for the strengthening in either case is unknown, except that it is known that surface flaws are either eliminated (fire polishing) or their effects on overall mechanical strength are somehow diminished. One of the purposes of the research described here has been to identify the important mechanism involved in the strengthening of sapphire, so that a practical process can be developed to achieve a reliable increase in the strength of any appropriately processed piece. Fire polishing itself cannot be used, since the thermal stress inherent in this process breaks large pieces. Unfortunately for the purpose of identifying important strengthening mechanisms, practical experimental strengthening techniques almost always improve the surface through a number of mechanisms simultaneously.

Since the condition of the surface determines the overall mechanical strength of the piece, techniques have long been sought that can reliably improve the condition of this surface. The following mechanisms have been demonstrated to be associated with improvements in the strength of sapphire: 1) Polishing, 2) Healing of surface flaws (chemically or through high temperatures), 3) Protecting the surface 4) Sealing surface flaws (solid solution layers), 5) Compressive surface layers, and 6) Crack propagation prevention (dislocation pinning). The most used of these techniques has been compressive surface layers. Such layers improve the strength of brittle materials by preventing surface flaws from acting to cause failure; compressive stresses are created in the surface that cause much larger tensile stresses to be required for a crack to grow and propagate.

In the work of Kirchner [4], a series of methods were used to obtain compressive surface layers on a variety of ceramic materials. Emphasis was placed on strengthening by quenching, and by glazing and quenching, because these experimental methods yielded the highest strengths. Strengthening sapphire was emphasized as a result of the inherent strength and ready availability of sapphire. Substantial improvements in flexural strengths were observed. In some cases improved tensile strength, thermal shock resistance and delayed fracture properties also were demonstrated. Treatment of sapphire single crystals resulted in three-fold improvements in strength. A glazed and quenched alumina rod 3.2 mm in diameter with a glaze layer 0.04 mm thick may have a measured flexural strength of 820 MPa. At failure, the tensile stress in the outermost portion of the alumina under the glaze on the tension side of the rod was at least 807 MPa. Much of the improvement in strength was retained when the samples were abraded. The stresses in the glaze are much lower because of the initial compressive stress and the low elastic modulus. It is apparent that the bulk of the material is inherently strong, but that the untreated material fails at low stress levels because of processes originating at the surface. More recently, Dr. Bates used this glass glazing process to strengthen a large (100 mm diam., 5 mm thick, and 150 mm long) sapphire cylinder for use as the transparent cylinder shell in a research internal combustion engine [5,6].

An important issue associated with surface strengthening is surface protection. There are two important contributors of the environment to the strength of sapphire. One is environmental chemistry, and the other is handling or mounting damage. The effect of environment on sapphire surfaces is usually that of moisture in combination with stress. Other effects arise from corrosive atmospheres (they must be very corrosive to affect sapphire), very high temperatures, or a combination of these factors. Handling protection prevents hard particles from coming into contact with the surface and causing damage that would weaken the piece. Although sapphire is very hard and very scratch resistant, it is easy to microscopically scratch an unprotected surface. The reason for this is the omnipresence of hard particles in the form of alumina on "sand" paper, and chips from the sapphire piece itself. Simple aluminum cannot scratch sapphire; although there is an aluminum oxide film on the surface, this film is in a different chemical form than sapphire, and is much softer. Protective layers can consist of any coating ranging from a glass glaze applied to strengthen the piece, to an antireflective coating applied for optical reasons. Even plastic films can be used. The coating need not be very thick to protect the surface; coatings that have poor transmission in wavelength regions where the transparency of sapphire is required can still be used because the coating is too thin to have significant absorption.

For surface strengthening to be effective the condition of the surface must be the determining factor for the macroscopic strength of the single crystal, rather than any bulk flaws. The bulk material of a window by definition has only a few flaws, since it is good optically. Such a window must be of even better quality as an imaging or laser diagnostic window. Current commercial window sapphire is thus always of a bulk quality such that the surface finish probably determines overall material strength. The surface quality is most important at locations where there are large local tensile or shear stresses present. The surface condition at the exact position of maximum stress may not be relevant, however, because some other location on the surface may have half the stress but be unpolished (typically the side of the window). Stress distribution is highly three dimensional and determined by the combination of mechanical and thermal loading with residual stresses.

Almost all of the applications relevant to the present work are those where the surface which controls the strength of the piece is in a benign environment. These windows face a hostile environment - an environment that almost always includes high temperatures. Thus the side toward the hostile environment is hot, and the face away from it is cooler; the hot face is in compression and the cooler face is in tension as a result of thermal expansion differences. The face in tension is the face whose surface condition determines strength, and it is on the benign side. For a window into a combustion chamber, the benign side is the outside. For a window on a supersonic plane the benign side is the inside. In both cases the surface condition of the important side can be easily and passively maintained with the proper care.

SAPPHIRE WINDOW FAILURE TESTING AND ANALYSIS

The Phase I program [7] had a goal of providing an experimental demonstration of the strengthening of sapphire as an isolated material for the specific application of a window that provides optical access to a high pressure, high temperature environment. The mechanical strength of a sapphire window was to be significantly increased by appropriate processing of the surface that is in tension during use. In another sense the project goal was to greatly increase the design strength of sapphire with respect to current design practice and accepted engineering criteria. As will be shown, large additional gains in effective design strength can also be achieved by using improved mounting design techniques that are appropriate only to the material properties of sapphire.

Sapphire Surface Processing

The goal of the surface processing effort was twofold. The original goal of Phase I was to demonstrate sapphire strengthening through surface processing, with a secondary goal of identifying the mechanism of strengthening. A critical part of the work was the examination of the window surfaces by a scanning electron microscope (SEM), which led to a correlation of window failure strength with surface finish. Seven different types of surface processing were performed: 1) Standard 80/50 polishing, 2) "Epi" polishing, 3) Antireflection coating, 4) Molecular beam implantation, 5) Glass glazing, 6) Surface annealing, and 7) CO_2 laser melting. Different strengthening mechanisms were evaluated by using a surface processing technique that isolated a particular mechanism. Of the surface processing techniques epi polishing was found to be most effective for strengthening sapphire; the present discussion will center on polish strengthening experiments. The other techniques were found to be less effective, but were used to confirm the importance of the polishing as the most important effect among the many that could contribute to strengthening.

Small sapphire windows were used in an inexpensive hydraulic failure pressure testing facility. These windows were 2.5 cm diameter and 1 mm thick, sized such that the pressure required to break them is within the accepted rating for a commercial medium-pressure hydraulic assembly (up to 140 MPa). A set of 30 standard grade, random orientation windows with an 80/50 scratch/dig optical polish was obtained from Meller Optics (Providence, RI). This polish is standard when no specific finish is requested except that it be an optical finish. Final polishing is performed using a fine grit diamond compound. SEM images of this polish indicate that the polished surface is a mass of randomly oriented scratches on the order of 1 μm wide. Another set of 32 best quality, 0° orientation windows with an epi polish was obtained from Crystal Systems (Salem, MA). An epi finish is the best available polish for sapphire, where final polishing is performed by chemical removal of the sapphire surface using a colloidal silica solution at elevated temperature. SEM images of these windows reveal a variety of types and concentrations of defects. Characteristic of the colloidal silica polish is pits of varying size randomly distributed over the surface. The pits can occur in isolation or in clusters of varying size. A high quality epi polished surface has few defects widely spaced, while a low quality epi finish not only leaves remnants of scratches in place, but introduces quite a few streaks of deeper erosion. It should be noted that features visible in the SEM may not be actual surface features, but rather subsurface scattering sites, since they are actually identified only by changes in the deflection of electrons from the SEM beam. No polishing was required or performed on the edges of these windows, since the edges were not stressed at all, and chipping could not damage the window surface before testing.

Sapphire Window Failure Testing

A hydraulic facility as shown in Fig. 1 was constructed to failure test the processed sapphire windows. The configuration consisted of a hydraulic hand pump, medium pressure hydraulic tubing, a pressure gauge, and a window mounting fixture surrounded by a metal safety shield. The primary pressure containment was done by a pair of standard "Conflat" vacuum flanges that seals using copper gaskets. One flange was modified to simulate a clear-aperture fixture for optical access, and the other was drilled and tapped to accept a fitting from the hydraulic system. The sapphire windows were not clamped, but held against the O-ring seal by the internal pressure of the system. This prevented any possible complications from clamping, and turned out to be instrumental in the development of the high-pressure window mounting technique. The fixture included two safety shields; a case around the entire window mount fixture to contain pieces and a thin plate on the top of the aperture flange itself. Any significant air volume at high pressures would otherwise result in an extreme safety hazard when a windows failed.

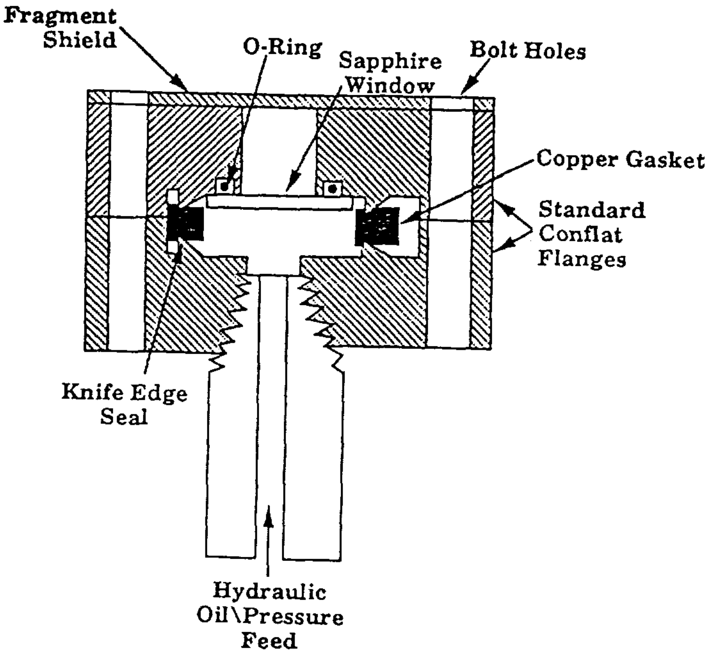


Figure 1. Sapphire window hydraulic failure testing facility; detailed drawing of window mount fixture.

Failure testing of 31 windows was done, sampled from a set of 62 windows that had undergone combinations of 6 different surface processing techniques (laser melting was an isolated test). The pressure behind each window was increased until the window broke, and the peak pressure was recorded. The results of failure testing that demonstrate the feasibility of strengthening of sapphire by surface processing are shown in Fig. 2, where the strength of the 80/50 polish windows is compared with that of epi polished windows. Also indicated is the predicted window failure strength based on the design failure strength of sapphire as specified by the manufacturer. It is crucial to note that epi polish windows were selected on the basis of minimal flaws under SEM inspection, and then failure tested. Every window chosen for minimal flaws had a large failure pressure.

After demonstrating a correlation between minimal surface flaws and strengthening, two epi polished windows were identified as having major flaws, and these were tested, expecting them to have significantly less strength. One of these two windows did have a strength comparable to the 80/50 polish windows, demonstrating that the polish was responsible for the increased strength. The other one was shown by the SEM not to have been polished with colloidal silica at all. Apparently the rough diamond machining of this window did not reduce its strength, an intriguing fact. The correlation between minimal flaws and increased strength was 100%, while the correlation between flaws and reduced strength was not perfect because there was not a large enough flaw density to guarantee that a failure causing flaw would occur in the highly stress center of a window. This contrasts with the 80/50 polish, which was continuously flawed.

The absolute magnitude of the failure stress for these 2.5 cm diameter windows is very large: 25-27.5 MPa (almost 4,000 psi), although the windows are only 1 mm thick. The high contained pressures before failure demonstrates the potential of sapphire windows. The clear viewing aperture was 1.9 cm. As will be shown, a significant part of the load capacity of this window arises from the mounting system.

Strengthening ratios vary from 4.7 (maximum strength of strong samples/minimum strength of weak samples) to 1.87 (minimum strength of strong samples/maximum strength of weak samples). The most appropriate factor is that of a design strength, which can be taken as the minimum strength of the stronger samples relative to the predicted failure strength based on manufacturer specifications. As expected, this predicted failure strength is

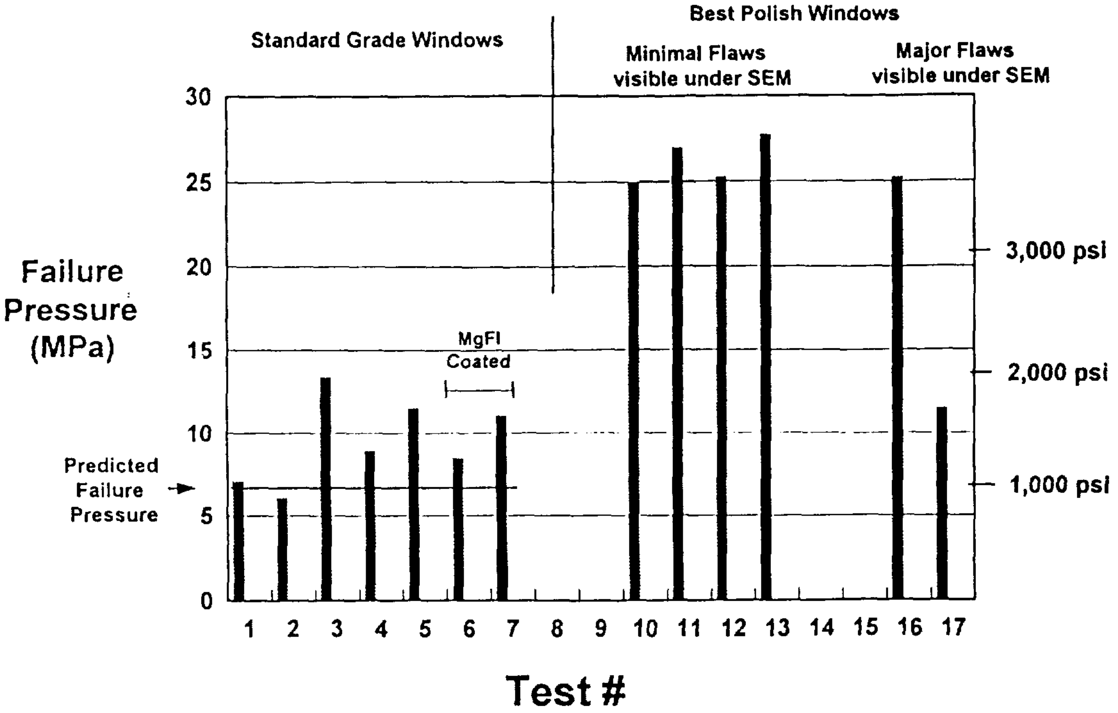


Figure 2. Sapphire window strengthening feasibility demonstration pressure testing results.

close to the minimum tested strength, because the design strength must be based on the worst case analysis. Thus, a design strengthening factor of 3.4 has been achieved through surface processing. This design factor increase implies that a new design strength of 1.4 GPa (203,000 psi) has been achieved for sapphire.

Other failure tests were performed to try to isolate the important strengthening processes. A MgFl antireflection coating isolated the surface from any environmental chemistry that might affect window strength, but did not cause strengthening. Both argon ion implantation and glass glazing resulted in strengthening of the 80/50 polished windows but to a lesser degree than epi polishing. Furthermore these processes did not increase the strength of epi polished windows. Testing of all of these techniques on both the 80/50 and epi polish windows separately also confirmed the role of the epi polish in providing the major strengthening effects measured.

There are two other possibilities for the difference in strength of the windows. The first is that the 80/50 polish windows are of lower bulk crystal quality. The second is the probable difference in crystal orientation of the two different types of polished windows. That these are not the cause of the strength difference can be shown from the data. Considering the quality of the bulk crystal first, if that were controlling the strength, the perfect crystals would always be stronger than the standard grade crystals. That is not the case is shown by case # 17 in Fig. 2. In this case the strength of the perfect and standard crystals are identical - only the surface processing has changed. That the orientation is almost certainly not the cause is demonstrated by the same case, where the degradation of the strength of the C-axis (0°) windows to that of the low strength 80/50 polish, and probably 90 degree orientation windows. Other data support this argument in a similar manner. The strengthening effect must be caused by surface finish.

CO_2 laser melting was attempted on a test window to determine if this could be a practical means of more reliably achieving a high quality surface. A 25W CW CO_2 laser emitting radiation absorbed by sapphire (10.6 μm wavelength) was focussed to a spot on a 80/50 polish window. Melt polishing was clearly visible in SEM images, smoothing out the scratches from the 80/50 polish, but fracture lines were always present. These fractures were caused by the thermal stress associated with the large local thermal gradients of the heating. This polishing process may be successful if it is performed while the sapphire piece is at a high enough temperature to allow the plasticity of the material to absorb the thermal deflections and prevent fracture.

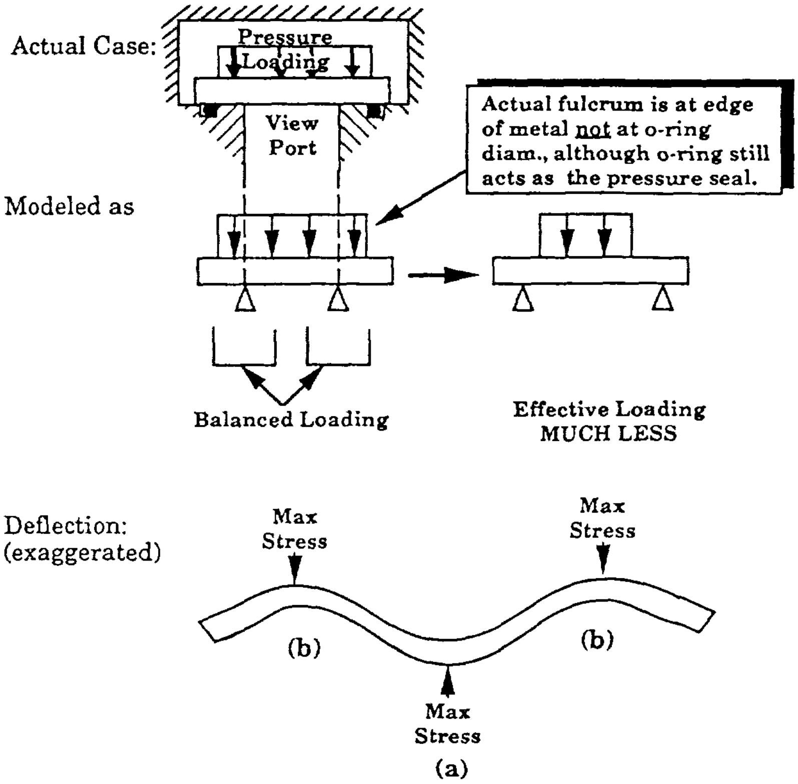
Window Mechanical Strength and Mounting Analysis

Although the standard technique for strength testing discs is the ASME four-point bending technique, it was decided to use an equivalent real window mounting system to perform the failure tests. The window can be modeled as a thin, uniform-thickness disk surrounded by a pressurized fluid up to the O-ring seal. Without pressure, the window rests flat on the metal support structure between the inner edge of the clear aperture and the outer diameter of the window. As the pressure is increased, the center of the window is pushed into the aperture and the outer diameter lifts off the metal. This is possible because the O-ring is pressurized and deforms to fill the space between the metal O-ring groove and the window, even if this space grows slightly. The load modeling is as shown in Fig. 3, together with the window deflection. For a uniformly loaded thin disc the maximum stress is related to pressure by:

$$S_m = k\,(wr^2/\delta^2) \qquad (1)$$

where S_m is the maximum stress, k is a constant equal to 1.27 for a thin circular plate, w is the pressure, r is the disc radius, and δ is the disc thickness. For a 410 MPa tensile strength of sapphire at room temperature, a 11.7 mm O-ring radius, and a 1 mm thickness, the predicted failure pressure is 2.55 MPa; a pressure far below even the worst test results.

However, as discussed above, the actual case of the loading on the window is a disc where the deflection is constrained by the inner edge of the clear aperture. As long as the maximum stress at (a) is significantly larger than at (b) the piece will fail at the center and surface processing will result in strengthening. This applies only to sapphire, since it is so hard that the inner metal edge does not cause the unusually high line stress and piece failure that occurs using glass. The loading outside the fulcrum diameter balances some of the load inside of the fulcrum diameter and causes a reduction in the effective load at the center. Using the simple area balance indicated in Fig. 3, a new effective loading area can be calculated, giving an effective radius for failure loading of 6.85 mm. This compares with the initially assumed radius of 11.7 mm. It also leads to an effective reduction in loaded area by a factor of 2.9 and an increase in predicted failure pressure by the same factor to 7.4 MPa. This prediction is in excellent agreement with the window testing data, as indicated in Fig. 2.



Actual fulcrum is at edge of metal not at o-ring diam., although o-ring still acts as the pressure seal.

Figure 3. Window failure stress analysis.

Furthermore, the new effective load area is almost a factor of 2 less than the clear aperture, illustrating the advantage of such a mounting scheme for increasing the design aperture of a given window.

Sapphire Thermal Analysis

The methods for increasing the strength of sapphire windows are predicted to be independent of temperature, but thermal stress effects are known to be an important contributor to the overall stress developed in a typical high pressure, high temperature window application. Thermal stress analysis tends to be complex, but it can be monitored experimentally. Basic thermal stress analyses have been performed, and these, together with previous research for the sapphire cylinder engine project [5], confirm the validity of the feasibility demonstration for harsh thermal environments but indicate the necessity of detailed work in Phase II. Experiments have been designed and prepared for thermal testing of strengthened sapphire.

Sapphire Window Strengthening Assessment

An assessment of the Phase I program to develop high performance sapphire windows must begin with the fact that current sapphire window design practice is very poor in general. This poor practice begins with a lack of adequate specification of the bulk material and the surface quality of sapphire windows, and a lack of design work to eliminate geometries that lead to stress concentration. Furthermore, the elimination of sharp edges is rarely specified. Poor practice continues with a total lack of the use of thermal isolation to minimize thermal stresses, and no consideration of modifying experimental operating procedure to lessen the severity of thermal transients. Mounting procedures that are used are those appropriate to glass but not to sapphire, and these procedures usually degrade the thermal as well as the mechanical performance of any type of window.

Another factor in the misuse of sapphire (and glass) windows is the lack of an understanding of statistical failure of brittle materials. Commercial design strengths must be based on the minimum strength of a material. For the case of statistical failure, such as is the case for sapphire, the majority of pieces are stronger than the quoted value. This means first that proof testing can easily be done to increase the safety factor of expendable pieces. It also means that the strength of a replacement will not be the same as the original - one window may survive for extended periods, whereas the next may fail immediately. Also ignored is the effect of environment on strength. It has been mentioned previously that it is relatively easy to scratch unprotected sapphire, and that scratches can lead to the failure of the entire piece.

The advances made in this program must be considered relative to these facts. Strengthened sapphire windows with improved mounting can only result in superior performance in the context of a thorough understanding of their proper use. The full benefits of the excellent properties of sapphire can only be obtained with proper manufacture, design, construction, and operating procedures.

Large Sapphire Window Design

A preliminary design of a large-aperture (15 cm) sapphire window on a combustion chamber was performed based on Phase I advances. Although such a window was previously thought not to be practical, the new design implies a window that is only moderately thick. A large aperture window to high temperature and pressure environments is a major goal for this program and for NASA; it would allow much more extensive use of current advanced laser diagnostics, and new information on difficult and important problems.

APPLICATIONS

Sapphire is widely used because of its attractive characteristics: excellent strength, hardness, chemical inertness, temperature resistance, and broadband transmissivity. It is also known for its high cost and its reputation for poor response to thermal stress and shock. Therefore, sapphire must be used with planning, both in its specification and design details, in order to utilize its attractive characteristics while avoiding the thermal stress and shock problems. A description is given below of the applications of sapphire as well as the potential of the improved sapphire. The listed applications provide the context for understanding the potential of improved sapphire.

General Applications

The following is a comprehensive list of the uses of sapphire as supplied by Meller Optics, Inc. Many applications are of specialty nature. Among the listed uses, note that sapphire fiber optics applications are expanding rapidly to provide diagnostic access to high temperature environments. Also, in addition to the single crystal sapphire on which this work is based, polycrystalline sapphire is used extensively in aerospace structures and for armor.

MECHANICAL

1. Fluid and gas nozzles
2. Hole and cap jewels for instruments
3. Wire and thread guides for electro erosion and printing machines
4. Blades for equalizing magnetic sound recorders
5. Various orifices for air, gas, and liquid meters
6. Capillaries for the semi-conductor industry
7. Chromatography pistons, pumps and valves for laboratory use
8. Probes for measuring instruments
9. Magnetic tape cleaners
10. Insulators
11. Pivots and machining stops
12. Balls-bearings, flow meters, check valves
13. Washers, valve seals (for particular flows)
14. Tubes
15. Narrow tolerance pieces
16. Point of sale windows for cash registers
17. Micrometer rotors
18. Microtome blades and knives
19. Tape guides

ELECTRONIC

1. Wafer carriers thin film deposition:
   a. Silicon on Sapphire
   b. Gallium Arsenide for field effect transistors
   c. Mercury Cadmium Telluride for detector arrays
2. Base substrates for thick film deposition of various metals for a wide range of circuitry
3. Acoustic delay lines

OPTICAL (Windows, lenses and prisms)

1. CO2 and O2 blood gas analysis
2. Environmental smokestack and auto emissions
3. Other toxic gas and fluid analysis
4. Industrial oven and furnace windows
5. Cryogenic analysis
6. Ultraviolet industrial lamps
7. Hermetic vacuum ports and seals
8. Thickness guides for paper machines
9. Fire and smoke detection instruments
10. Optical wavelength detectors and filters
11. Infrared missile seeker and camera covers
12. Centrifuge cell windows
13. Laser optics, etalons and reflectors; high power
14. Sight glasses
15. High pressure windows
16. Polarization optics
17. Engine turbine pyrometry
18. Refractometry
19. Sight Reticles
20. Telescope Optics
21. Fiber Optics
22. Lenses and prisms
23. Solar cell cover plates
24. Radiation damage environmental optics

CHEMICAL

1. Reactor components
2. Corrosion resistant cells, crucibles and tubes

MISCELLANEOUS

1. Transparent armor
2. Hollow wave guide for laser systems
3. Fiber optic tips for surgical lasers
4. Watch crystals and jewelry
5. Charges for vacuum coating

Windows Applications

Sapphire windows are used in a wide variety of research applications to provide optical diagnostic access to pressurized apparatus. They are also used in deep submersion vehicles. Commercial window applications include: 1) High pressure optical cells, 2) High temperature optical cells, 3) High pressure and high temperature optical cells, 4) Diagnostic cylinder wall for Internal Combustion engines.

NASA has many uses for sapphire windows, including a wide variety of diagnostic windows which provide the access to fluid and combustion studies. Future applications include structural windows in hypersonic vehicles.

One major application in defense is for missile and aircraft radomes. A radome consists of a hemispherical sapphire crystal mounted to permit tracking and guidance. The domes must be transparent to the proper wavelengths (infrared) and robust enough to survive the impacts and aerodynamic heating encountered in this application. Sapphire is the most competitive material for this application.

Potential

The potential is great for the strengthened sapphire; any improvement of a factor of 10 in strength will lead not only to improvements in present applications, but will also open the door to many new ones. For example, experiments in the more extreme environments such as those in high speed aerospace research can now be studied for longer duration, and new applications in structural windows, transparent armor, and visible high temperature furnaces, can now be considered.

CONCLUSIONS

Phase I research has successfully demonstrated the feasibility of greatly strengthened sapphire windows. Through surface processing, improved mounting designs, and minimization of thermal stresses, single crystal sapphire can be a mechanically equivalent replacement for high-strength steel. A factor of 10 increase in reliable design strength has been demonstrated for a strengthened, properly mounted sapphire window. The highest quality surface polishing resulted in experimental strengthening of sapphire windows by a factor of 3.4 relative to the strength of windows with a standard optical finish, which failed at the manufacturer's design strength. A new window mounting scheme provided a further increase of a factor of 2.9 in effective strength. The clear aperture that can safely be designed for a prespecified window thickness can be improved by this factor using the new mounting design. Guidelines have also been developed for specifying strengthened sapphire and minimizing thermal stress.

REFERENCES

1. Crystal Systems Technical Brochure, Crystal Systems Inc., Shetland Industrial Park, 35 Congress St., Salem, MA, 01970.
2. L.M. Belyaev (Editor), **Ruby and Sapphire**, Nauka Publishers, Moscow, Available from National Technical Info. Service, Springfield, VA, 22161, (1974).
3. H.P. Kirchner, R.M. Gruver, R.E. Walker, "Strengthening Sapphire by Compressive Surface Layers," J. of Appl. Phys., Vol. 40, No. 9, 3445-3452, (1969).
4. R.L. Gentilman, E.A. Maguire, H.S. Starrett, T.M. Harnett, and H.P. Kirchner, "Strength and Transmittance of Sapphire and Strengthened Sapphire," Comm. of the Am. Chem. Soc., Sept., C-116,117, (1981).
5. S.C. Bates, "A Transparent Engine for Flow and Combustion Visualization Studies," SAE Transactions, Vol. 97, Section 6, 892-908, (1988).
6. S.C. Bates, "Insights into Spark-Ignition Four-Stroke Combustion Using Direct Flame Imaging," Combustion and Flame, Vol. 85, Nos. 3 & 4, 331-352 (1991).
7. S.C. Bates, "High Performance Sapphire Windows," SBIR Phase I Final Report, NASA Contract # NAS3-26330, August, (1992).

A DUAL OUTPUT PRESSURE, HIGH RELIABILITY, LONG STORAGE LIFE GAS DELIVERY VESSEL ASSEMBLY

Isaac Maya, Joe McKee and Rajiv Rajpurkar
ARRAL Industries, Inc.
2101 Carrillo Privado
Ontario, CA 91761
909-947-6585

ABSTRACT

A Gas Vessel Assembly has been developed that delivers purified, very low moisture content gas at two different output pressures. High pressure gas is delivered at up to 6,700 psi, and low pressure gas regulated to 130 psi is also delivered via a second outlet over a wide range of flow rates. The device is extremely lightweight (less than 1 lb) and compact, affords maximum mechanical integrity, high reliability (0.9999 at 95% confidence level), and offers extremely long storage life. Specialized design and fabrication techniques are employed that guarantee gas purity and negligible leakage for more than 20 years, in widely varying conditions of storage temperature, humidity, altitude, and vibration environments. The technology offers unique advantages in fast, high pressure discharge applications. For example, when combined with a cryostat, cryogenic temperatures can be achieved such as those used in missile seeker technology. The technology has many additional applications such as: emergency power sources for safety devices such as those needed in nuclear power plants, refineries, collision cushioning devices, superconductor cooling devices, emergency egress systems, miniature mechanical devices that employ gas bearings, and other areas where long storage, extremely high reliability and/or high energy density power sources are required.

INTRODUCTION

ARRAL is a world-class manufacturer of precision electromechanical systems and assemblies, stored energy devices, and fluid systems for the international aerospace and defense communities. The company specializes in the manufacture, assembly, and testing of a broad range of complex mechanical, electromechanical and electronic devices. Particular expertise rests in special and unusual applications where our turnkey engineering and R&D capabilities can apply creative and innovative solutions to the design and manufacturing operations.

Our premier areas of specialization include on-board gas bottles and accumulators for sustained cooling, and stored energy devices that provide pneumatic pressure for fin stabilization and actuation control. A wide array of specialty and high production items have been manufactured, as shown in Figure 1. The latest product line is the Gas Vessel Assembly discussed below.

GAS VESSEL ASSEMBLY DESCRIPTION

The Gas Vessel Assembly (GVA) is shown in Figure 2. Its major components are: the manifold assembly, a partial toroid-shaped gas storage vessel, and a squib valve-flex circuit connection to an external firing signal. Key elements in the design of the gas storage vessel which were critical in achieving technical objectives were the development of a miniature pressure regulator in the manifold assembly, the use of a high strength specialty steel for the storage vessel, and a proprietary gas release system.

Manifold Assembly

The manifold assembly is shown in Figure 3. It is fabricated from high strength aluminum alloy, and provides mounting and interconnection for the squib valve and the regulator. Additionally, the manifold provides routing and discharge connections for the two different gas outputs. The dual output nozzles are shown in Figure 4.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

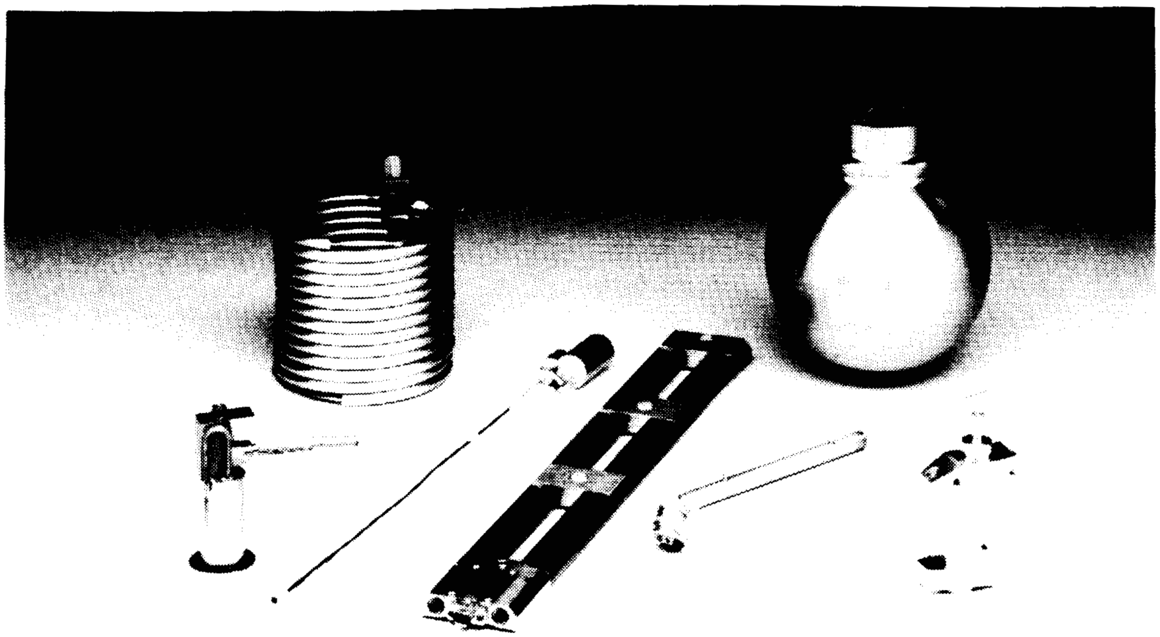


FIGURE 1. ARRAL-produced on-board gas bottles, accumulators, and stored energy devices.

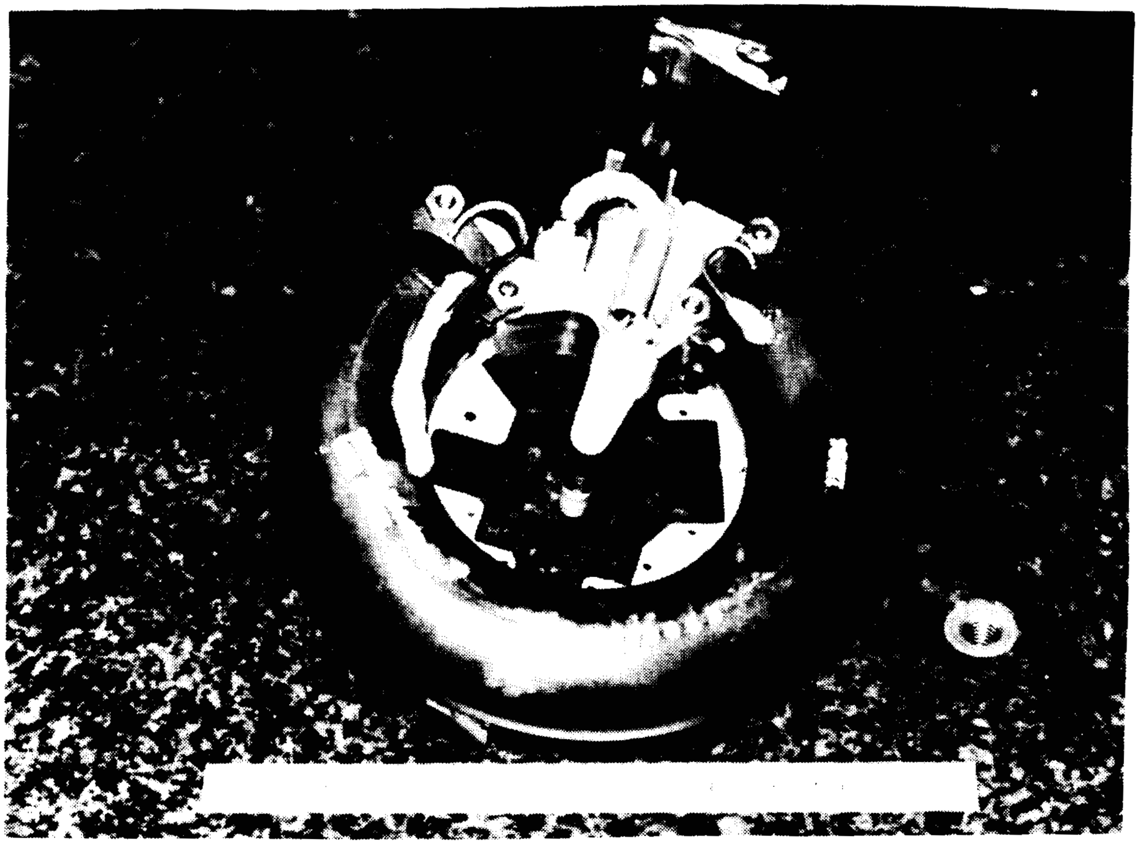


FIGURE 2. The Gas Vessel Assembly is extremely lightweight, compact, and reliable.

One of the keys to the success of the reservoir configuration is its minimum pressure boundary surface area design. The manifold is not a reservoir boundary, and sees gas pressure only after squib valve actuation or a test port is used. The manifold and the mounted components are in contact with the stored gas only during discharge. This insures stored gas purity and a minimum chance of potential leak paths. The manifold is separate from the gas storage vessel. Its fittings are adjusted for precise location without the risk of potential leaks, or the introduction of contaminants upstream of the filter.

The regulator is of a conventional piston-driven spool design, of very compact physical dimensions, and very lightweight. It is constructed of an aluminum alloy and stainless steel. The design incorporates a spring-actuated Teflon seal and a bellows seal. The regulator and the squib actuator are mounted on the manifold and sealed. Regulated gas is directed through a filter to the gas bearing fitting.

Gas Storage Vessel

The gas storage vessel shown in Figure 2 is constructed of specialty stainless steel toroid halves and spherical end caps, machined from solid material. The gas storage vessel is formed by welding two identical thin wall half torus rings together, removing a section of the resulting torus ring, and welding hemispherical end caps on each end of the resulting toroid. In addition, brackets are welded to the vessel for mounting the manifold, as well as to support the entire assembly. The gas storage vessel is heat treated, cleaned, charged with gas to 6,700 psi and checked for leakage prior to mating with the manifold. Under testing, the vessel has displayed a capability to withstand test pressures of almost 20,000 psi and a leakage rate roughly equivalent to a pressure loss of 3½ psi over 20 years.

The end caps have small diameter tubes for filling and discharging the pressure vessel. One end cap holds the discharge tube which is inserted into the manifold and becomes part of the squib valve. The storage vessel component currently in production holds 8.0 cubic inches of pressurized ultra-pure argon gas.

A key element in the design of the gas storage vessel was the use of a specialty steel for its higher strength. Heat treatment and surface treatment are performed after welding, returning the material to its near homogeneous, pre-welded condition. In thousands of applications, there has been no evidence of stress corrosion problems.

MANUFACTURING PROCESSES

A substantial inventory of manufacturing and test equipment was developed specifically for this application. The equipment and procedures provide a very accurate determination of the leak rates of the reservoir, and thus pressure as a function of storage time. These were developed specifically for the high volume, high reliability requirements imposed on these units. Highly specialized equipment designed by company engineers is required for compression, purification, and verification of the gas quality on a production basis.

To minimize "touch labor" cost to an absolute minimum and to minimize the required skill levels, specific design approaches were employed to facilitate automated manufacture and assembly of the tight tolerance, high reliability parts and subassemblies. For example, the regulator employs a variety of miniature screw machine parts that emerge from their primary fabrication step as completed parts, requiring no secondary operations. Furthermore, the parts are compatible with vibratory parts feeders and simplified assembly and fastening techniques. The completely self-contained regulator is then inventoried as a finished item that is subsequently mounted in the manifold using only a wrench.

The aluminum manifold body is designed to be manufactured from custom extruded bar stock. The extrusion has all of the manifold's complex outer profiles, minimizing machining and inspection time and complexity to render the final part. Additionally, expensive operations such as radiography (to detect voids and porosity in the metal) can be performend on the entire bar at one time. With this approach, per part radiography cost is minimized and flawed material is discovered prior to machining. The regulator, squib valve actuator, flex circuit assembly, and discharge connectors are then installed in the manifold. Final adjustment is

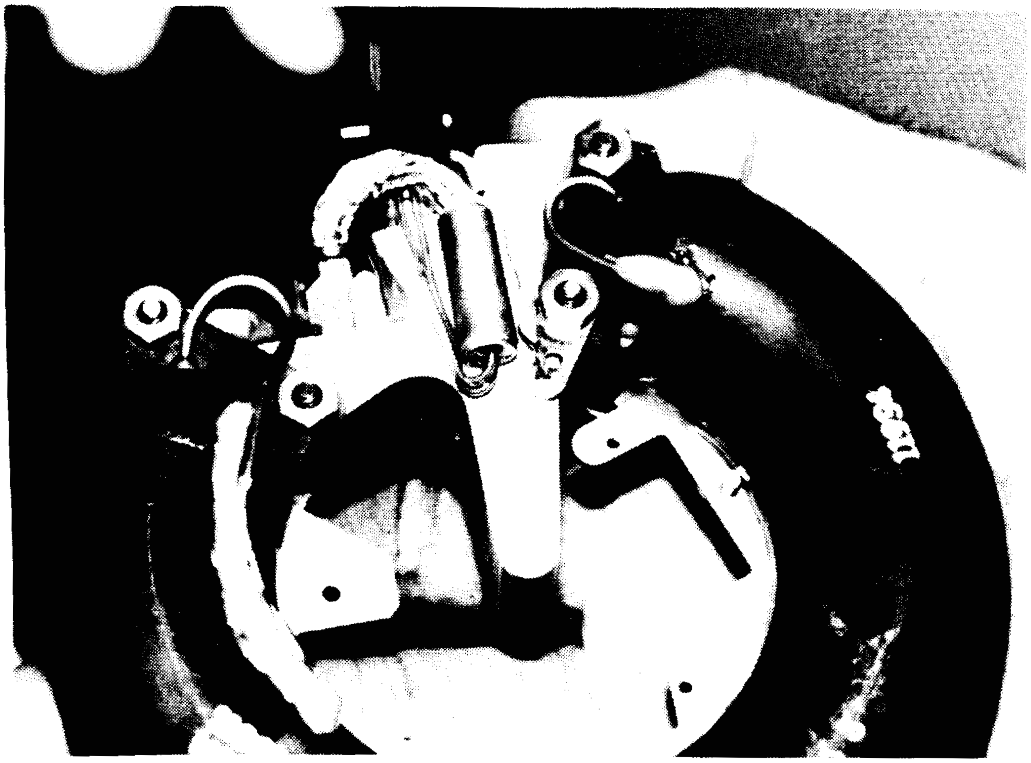


FIGURE 3. Close-up showing the unique manifold design and mounted gas release system.

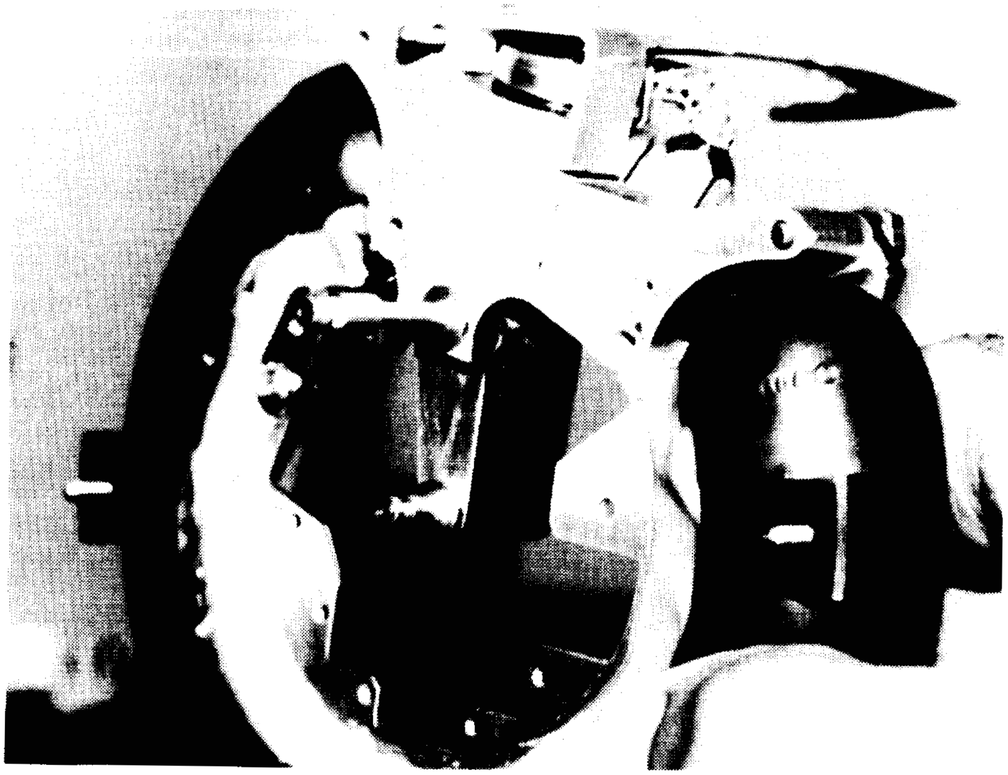


FIGURE 4. Dual output nozzles provide regulated gas pressure at 6700 psi and 100 psi.

then made to the regulator. A fixture is used to accurately locate and epoxy bond many components to the manifold body, in lieu of traditional fasteners, to achieve cost savings, more accurate placement, and reliability enhancement. The fixture is also designed as an in-process inspection device, to insure that all components will meet the final, extremely tight location requirements.

The gas storage vessel employs design approaches similar to the manifold. The thin wall torus ring halves, used to build the main torus ring, are machined from custom extruded specialty stainless steel tubing. This tubing has an outside diameter slightly larger than the outside diameter of the torus and an inside diameter slightly smaller. The part emerges from the machine ready to use without any further proccessing or handling required. The spherical end caps are made in a similar fashion from standard round barstock. Although considerable material is cut away to yield the final part, the extra material cost is more than made up for in reduced handling versus the multiple operations that would be required for forging an initial "blank" then multiple hand loading it for final machining. As in the case of the manifold, radiography can be done on the entire tube and bar prior to final machining.

By machining from solid material rather than forging or casting, the final parts also exhibit far less potential for microfractures, porosity, and residual stresses in the material. The consequences of the above could be leakage and dimensional change after heat treating. Small diameter 304 stainless steel fill and discharge tubes are then furnace brazed all at one time to the end caps. Mounting pads and brackets are precisely located by a fixture and welded. Although presently hand-welded, plans call for performing these operations using automated TIG welding equipment employing specialized assembly fixtures. In contrast to the present hand-welding, the latter automated process will not require the same skill level or specialized training to produce dimensionally accurate storage vessels. After annealing, heat treating, cleaning, proof pressure testing (at 1½ times the final storage pressure) and leak checking, storage vessels are connected in groups, via the fill tubes, to charging manifolds. Each group of vessels is then vacuum baked to eliminate any internal moisture, and charged to final storage pressure.

Final GVA assembly and interconnection of the storage vessel to the manifold is performed in an assembly fixture to insure proper dimensional location of the major assemblies. The same epoxy bonding material previously used on the manifold is used to fill the gaps purposely left between the parts to be mated, and prior to the attachment of fasteners. By the use of this technique, a wider range in manufacturing dimensional tolerences can be allowed for the individual components than are acceptable for the final GVA assembly.

CONCLUSIONS

The technology incorporated in the Gas Vessel Assembly, and the precision manufacturing techniques developed in its support, were developed for specific military applications. They offer multiple opportunities for commercial development of products based on the technology. The manufacturing techniques and process quality control methods were used to satisfy stringent product specifications.

ACKNOWLEDGEMENTS

The authors would like to thank Glenn Haskins, Lidia Savage and Ed Crossen for their contributions in developing the design, manufacturing processes and testing procedures used in this program.

omit

SENSORS AND SIGNAL PROCESSING

ON-LINE PROCESS ANALYSIS INNOVATION: DICOMP™ SHUNTING DIELECTRIC SENSOR TECHNOLOGY

Craig R. Davis
Axiomatics Corporation
3G Gill Street
Woburn, Massachusetts 01801

Frank A. Waldman
Axiomatics Corporation
3G Gill Street
Woburn, Massachusetts 01801

ABSTRACT

The DiComp Shunting Dielectric Sensor (SDS) is a new patent-pending technology developed under the Small Business Innovation Research Program (SBIR) for NASA's Kennedy Space Center. The incorporation of a shunt electrode into a conventional fringing field dielectric sensor makes the SDS uniquely sensitive to changes in material dielectric properties in the kHz to MHz range which were previously detectable only at GHz measurement frequencies. The initial NASA application of the SDS for Nutrient Delivery Control has demonstrated SDS capabilities for thickness and concentration measurement of Hoagland nutrient solutions. The commercial introduction of DiComp SDS technology for concentration and percent solids measurements in dispersions, emulsions and solutions represents a new technology for process measurements for liquids in a variety of industries.

DIELECTRIC PROPERTIES OF MATERIALS

Material dielectrics are physical properties resulting from interaction between an alternating electric field and the ions and dipoles of a material. In dielectric materials, some of the electric field energy is stored while some energy is expended. Energy storage and consumption in the material are interrelated as a complex function of the electric field frequency (f), and are collectively known as the complex permittivity[1]. Complex permittivity is expressed as:

$$\varepsilon^*(f) = \varepsilon'(f) - j\varepsilon''(f) \qquad [1]$$

The real term of the permittivity, ε', is commonly known as the dielectric constant. The dielectric constant value depends strongly on the degree of polarity within the molecular structure of the material. A polar material, such as water, has high relative permittivity while non-polar materials, such as hydrocarbons, have low permittivity values.

The imaginary term of the permittivity, ε'', is commonly known as the loss factor. The loss factor value depends on the energy expended to align dipoles and move ions within the material, so it can be expressed as:

$$\varepsilon''(f) = \varepsilon_R''(f) + \varepsilon_C''(f) \qquad [2]$$

In this expression, ε_R'' is a measure of dipolar relaxation and ε_C'' is a measure of the ionic conductivity of the material. Both terms are a function of the electric field frequency. In general, at lower (kHz) frequencies the conductivity term dominates while at higher frequencies (GHz, or microwave) the dipolar term dominates.

DIELECTRIC MEASUREMENTS IN MATERIALS

Dielectrometry is the measurement and interpretation of the dielectric properties of materials. There are numerous sensors and instruments available whose measurements rely on one of more term of the material dielectrics. Commercially, these include conductivity meters, RF capacitance probes and microwave permittivity analyzers.

In general, each device has a sensor designed to pass an electric field through the material to be tested. Typically, a transmitting electrode and receiving electrode are used for this purpose. The coupling of the electric field between the two electrodes is dependent on the dielectric properties of the material between the electrodes. This coupling results in changes in the amplitude and phase of the original signal, as shown in Figure 1.

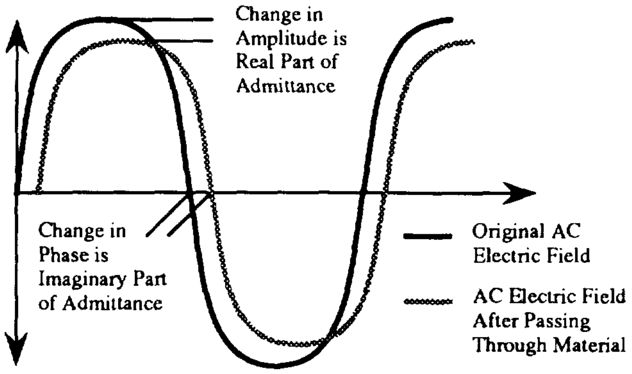


Figure 1: Changes in Measurement Signal Due to Material Under Test

The relative changes in the measurement signal are a unique function of the sensor's electrode configuration and the measurement circuitry. If the electric field can be modeled exactly, it is possible to calculate the material dielectrics by transforming the changes in amplitude and phase.

In the simple case of parallel plate electrodes, a cell constant can be used to characterize the amplitude changes at low frequency and predict the ohmic conductivity of the material. More complex geometries, such as fringing field-type sensors, require the use of more elaborate Fourier transforms to predict material properties from impedance or admittance measurements.

Since the objective of most dielectric measurements is to detect material composition changes, it is not necessary to obtain the absolute value of the material's dielectric properties. It is necessary only that changes in the measurement signals are a function of relative changes in the dielectric properties due to changes in material composition.

LIMITATIONS OF CURRENT DIELECTRIC MEASUREMENTS

Most practical measurement applications involve mixtures of two or more components. The mixture may be relatively stable, as in solutions, emulsions, slurries and dispersions, or may change with time as in a chemical reaction. In each case, the mixture is comprised of components with different contributions to the dielectric properties of the whole.

Solutions, for example, will typically exhibit changes in the ionic conductivity term of the loss factor as the solute concentration changes. Slurries and emulsions will typically exhibit changes in dielectric constant, as their components have lower dielectric constants than the water carrier. Often, a mixture will have both suspended and dissolved components, so both dielectric properties will be affected by changed in composition of the mixture.

If one component has a distinct contribution to the dielectric properties of the mixture at some measurement frequency, then it can be quantified if the instrument system is sufficiently sensitive to the changes in ε' and ε'' at the optimal measurement frequency for a particular component. However, current commercial dielectric measurement systems have limitations in sensitivity.

Conductivity meters are sensitive only to ionic conductivity (ε_C'') and operate a fixed low frequency. RF capacitance probes are limited to measurement in materials with low loss factors. Microwave permittivity systems operate a very high (GHz) frequencies, and therefore cannot detect dielectric relaxations which occur in the low RF (1kHz - 1MHz) region[2]. The ideal dielectric measurement system should have the capability to make measurements sensitive to either ε' or ε'', even at very high loss factors. Such as system should also make measurements at multiple frequencies across the RF spectrum to predict the composition of a wide range of mixtures of materials.

SHUNTING DIELECTRIC MEASUREMENT INNOVATION

Axiomatics has discovered a new dielectric sensor configuration which provides unique sensitivity to changes in both ε' and ε'' across a broad RF spectrum. This innovation is the patent-pending Shunting Dielectric Sensor (SDS).

A conventional fringing field sensor (right) is shown in Figure 2, along with a representation of the coupling of the electric field between the two electrodes.

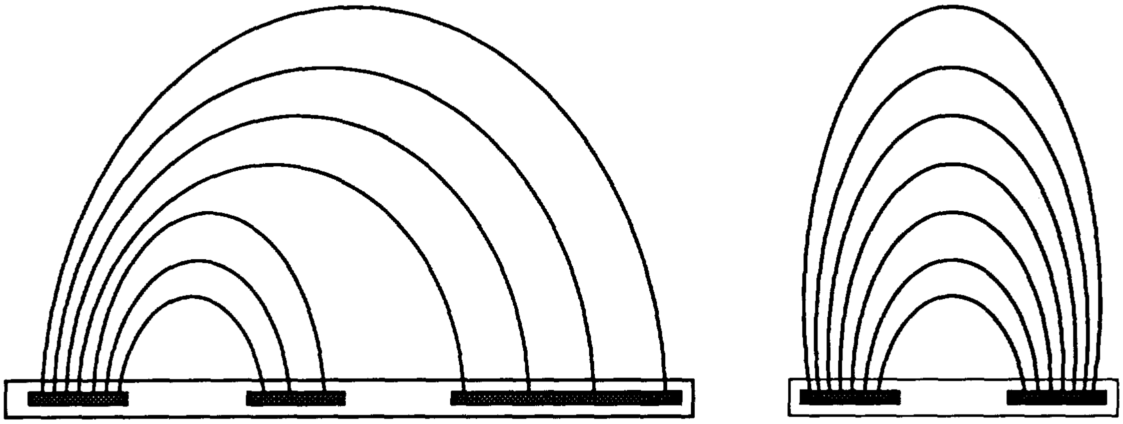


Figure 2: Field Behavior in Shunting (Left) and Conventional Dielectric Sensors

The response of the conventional sensor can be characterized by its RC time constant, which is a function of the geometric configuration of the sensor and the dielectric behavior of any material present in the electric field:

$$RC_C = A * \tau \qquad [3]$$

In this expression, A is a sensor geometry constant and τ is the relaxation time constant of the material in the field. τ is an indication of the time rate at which stationary equilibrium will be reached after the initiation of the electric field. Typically τ is expressed as the ratio of the static permittivity to the steady conductivity. As the conductivity increases, τ and RC_C decrease — this relationship limits capacitance systems to materials with low conductivity.

Axiomatics discovered that the placement of a third electrode in proximity to the conventional electrode pair provides a dominant coupling of the electric field. This shunting of the electric field (left), illustrated in Figure 2, dramatically changes the coupling of the field to the sensing electrode. In effect, the RC time constant for the Shunting Dielectric Sensor (SDS) is multiplied by a term related to the geometry of the shunt:

$$RC_{SHUNT} = B_{SHUNT} * RC_C \qquad [4]$$

As a result, the RC time constant of the SDS sensor is several orders of magnitude higher than the conventional two-electrode sensor. The SDS sensor is therefore significantly more sensitive to changes in τ, regardless of which component of the complex dielectrics is responsible for the change. This sensitivity permits the SDS sensor to be used in measurements of both low and very high conductivity.

More importantly, the SDS sensor has a frequency response which makes it sensitive to non-dispersive dielectric phenomena at much lower frequencies than conventional sensors. Many mixtures, particularly emulsions, have interfacial polarizations which exhibit dielectric relaxations in the MHz spectrum. Complex admittance measurements using the SDS sensor in this frequency range can yield valuable information about the quality and composition of the emulsion.

DICOMP SDS APPLICATIONS

Establishing the response of the DiComp SDS sensor to particular industrial applications has only recently begun. The unique behavior of the SDS makes even the limited published data on conventional dielectric measurements of little value. As with the introduction of any new technology, each application must be approached empirically to establish the measurement response of the SDS sensor for a given material.

In the spirit of the SBIR program, Axiomatics began sales and initial tests with commercial customers in May 1991. The success of this early commercial testing lead to a formal product introduction of our DiComp Analyzer for liquids at the Instrument Society of America ISA/92[3] exhibition in October 1992. The DiComp SDS system has been successful employed in a variety of applications measuring concentration or percent solids in dispersions, emulsions and solutions. The materials detected include acids, alcohols, caustics, halides, hydrocarbons, monomers and water.

In general, the approach to each new application follows a procedure designed to identify the optimal measurement parameters. The first step in this approach is to use the SDS sensor for a frequency scan of a sample with known composition, over the operating range of the Analyzer of 1kHz to 8MHz. By selectively varying constituents in the known sample, a "fingerprint" is generated which establishes the relative contribution of each constituent as a function of frequency. A combination of frequencies may then be used to predict the constituent parts of the sample using a straightforward function of the complex admittance measurements.

For example, Figures 3 and 4 show the real and imaginary components of the complex admittance measurements for two materials as a function of frequency. Material A, which is representative of a material such as deionized water, has a high τ and therefore exhibits a low frequency peak in its imaginary admittance component with a relatively flat response in the real admittance component. Material B, which has the characteristic response of a material with low τ, exhibits peak response at high frequency in both the real and imaginary admittance components.

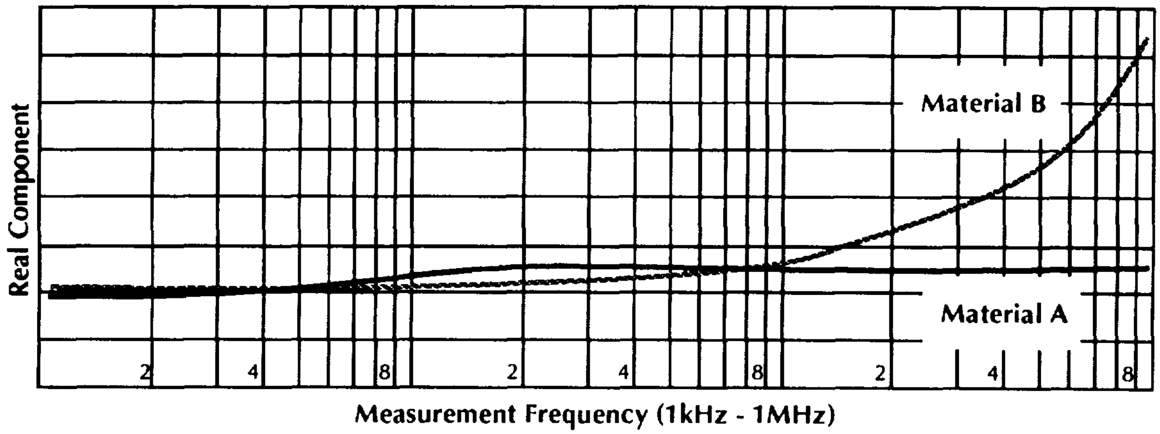


Figure 3: Real Admittance Response for Two Materials as a Function of Frequency

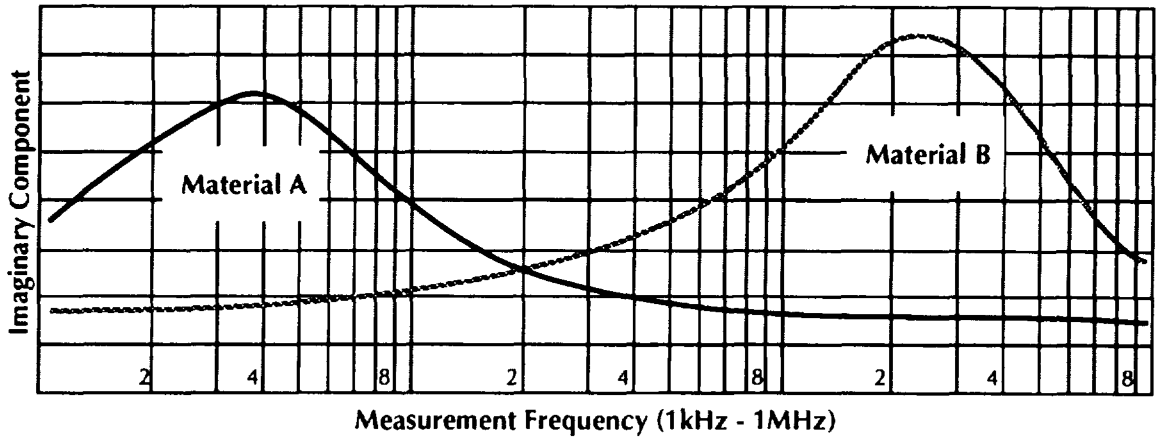


Figure 4: Imaginary Admittance Response for Two Materials as a Function of Frequency

Mixtures of the two constituent materials will result in a shift of the peaks towards the middle of the range, with corresponding drops in the magnitude of the values. A simple polynomial function of the admittance at one frequency can then be used to predict the relative concentrations — low frequency if Material A is the solvent or high frequency if Material B is the solvent.

The addition of a third constituent material, such as a salt or acid, will change the admittance response further as a result of lowering the τ of the mixture. Although this effect may be evident across the frequency range, it will be more pronounced at higher frequencies. A high frequency measurement can therefore be used to determine the salt content, then compensate the low frequency measurement to predict the concentration of the other variable. This technique can be applied to determining hydrocarbon contamination levels in seawater.

Similar strategies have been applied to mixtures such as slurries and emulsions, where the solvent or carrier is typically water with dissolved solids and the other constituent is typically suspended solids or oils with relatively low permittivity. In each case, predictions for each constituent can be made through a calibration procedure in which representative samples are created in the laboratory. The constituents of the samples are selectively varied, and the admittance measurements at various frequencies are recorded. By comparing the sample admittance measurements against actual constituent values determined by laboratory analytical measurement, a predictive function can be established.

In industrial applications involving continuous production, it may not be possible to create representative process samples in the laboratory. However, a pilot or experimental facility often exists where the SDS sensor can be installed in a pilot process, and the calibration procedure is adapted to rely on simple additions or dilutions performed in a flow loop. By comparing process admittance measurements against known addition and dilution values, a predictive function can be established.

Where no pilot facility exists to permit controlled experiments, it is still possible to perform the calibration procedure using a sensor installed in an actual production process. Although the process is under positive control, there will be variations over time in the constituents and these variations are normally recorded as part of routine laboratory monitoring.

For these cases, Axiomatics has developed software to digitally record admittance data in a form compatible with popular spreadsheet packages. The SDS admittance data is then merged with laboratory measurements of process variation, and a calibration is developed using on-line data. Extrapolation of the calibration function outside of the range of actual measured values can generally be used to predict extremely out-of-specification product.

Axiomatics has developed a number of different sensor geometries to accommodate various pilot and production processes as part of an on-going commercialization program.

NASA NUTRIENT DELIVERY SYSTEM APPLICATION

The initial application of the SDS technology in the Controlled Ecological Life Support System (CELSS) Nutrient Delivery System (NDS) at NASA's Kennedy Space Center required the unique capability to predict both the concentration of a solution and the thickness of the solution layer over the sensor.

The CELSS experiment is designed to test new technologies to reliably provide optimized nutrient solution[4] quality and flow rates to sustain crop growth in a micro- or variable-gravity environment. The NDS control system must therefore include sensors for measuring both thickness and concentration of the nutrient solution delivered to one of several alternative NDS plant growth surfaces being evaluated by NASA, including the continuous-flow NDS in the Biomass Production Chamber[5] (BPC) at the Kennedy Space Center Breadboard facility, the porous tube NDS[6] in the Plant Growth Unit (PGU), and the porous stainless steel NDS[7].

Conventional capacitive sensors cannot be used in this application because the nutrient solution has a very high loss factor. Conductivity sensors can measure bulk concentration, but are unsuitable for measurement of nutrient solution thickness, due to electrode interference with plant growth and variances in conductivity of the solution. Infrared detectors are also inappropriate, due to interference from plant roots and biomass.

The SDS provided an innovative and unique solution to the CELSS problem. The SDS sensor could be embedded in the NDS plant growth surface without interference, and its non-invasive measurements do not affect plant growth.

Axiomatics designed an SDS sensor for the NDS[8] with a geometry B_{Shunt} designed to provide the necessary sensitivity for both thickness and concentration independent of the choice of plant growth surface. The NDS sensor design is shown in Figure 5.

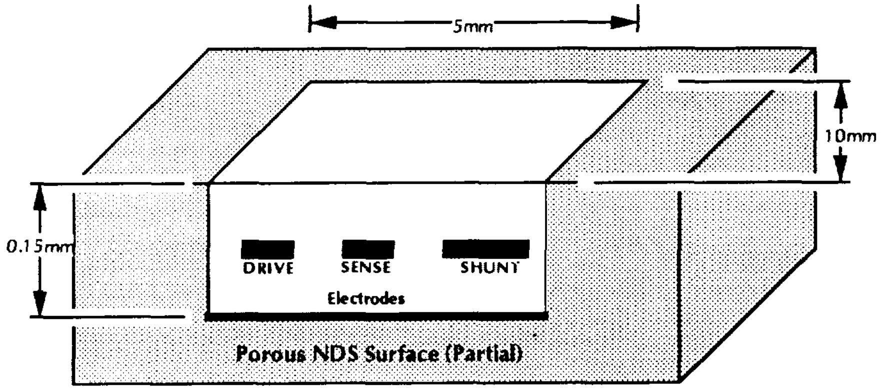


Figure 5: Cutaway View of NDS Sensor

The sensor is constructed of flexible materials to conform to the curved surface of the porous tube growth surface. A wettable surface insures that the thickness of nutrient solution over the sensor is the same as the adjacent area. An architecture which supports multiple NDS sensors is shown in Figure 6. The design employs digital impedance analyzer technology specifically developed to meet the performance and space requirements of the NDS application.

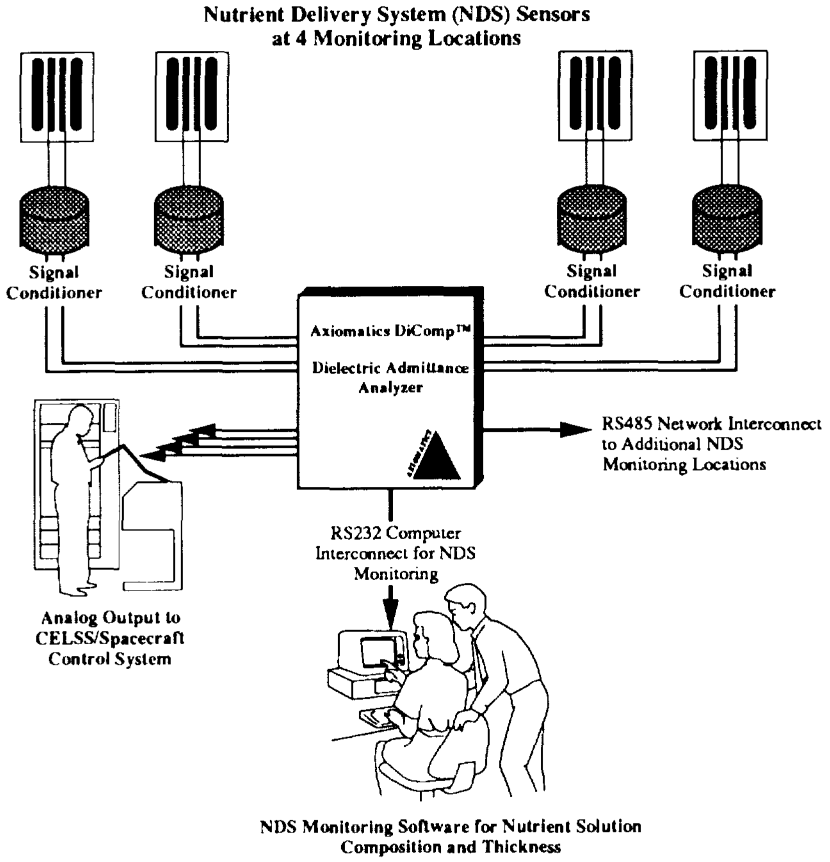

Nutrient Delivery System (NDS) Sensors
at 4 Monitoring Locations
Signal Conditioner
Signal Conditioner
Signal Conditioner
Signal Conditioner
Axiomatics DiComp™
Dielectric Admittance Analyzer
RS485 Network Interconnect to Additional NDS Monitoring Locations
Analog Output to CELSS/Spacecraft Control System
RS232 Computer Interconnect for NDS Monitoring
NDS Monitoring Software for Nutrient Solution Composition and Thickness

Figure 6: NDS Sensor System Architecture

In the NDS application, the target thickness for the nutrient solution at the plant growth surface is 0.5mm. The NDS sensor has been optimized to provide maximum sensitivity over the range of 0 to 0.5mm, although the sensor has been shown to be capable of accurate measurements to approximately 2.0mm. Within the target thickness range, repeatable accuracy of ±0.05mm (0.002 inches) has been demonstrated, as shown in Figure 7.

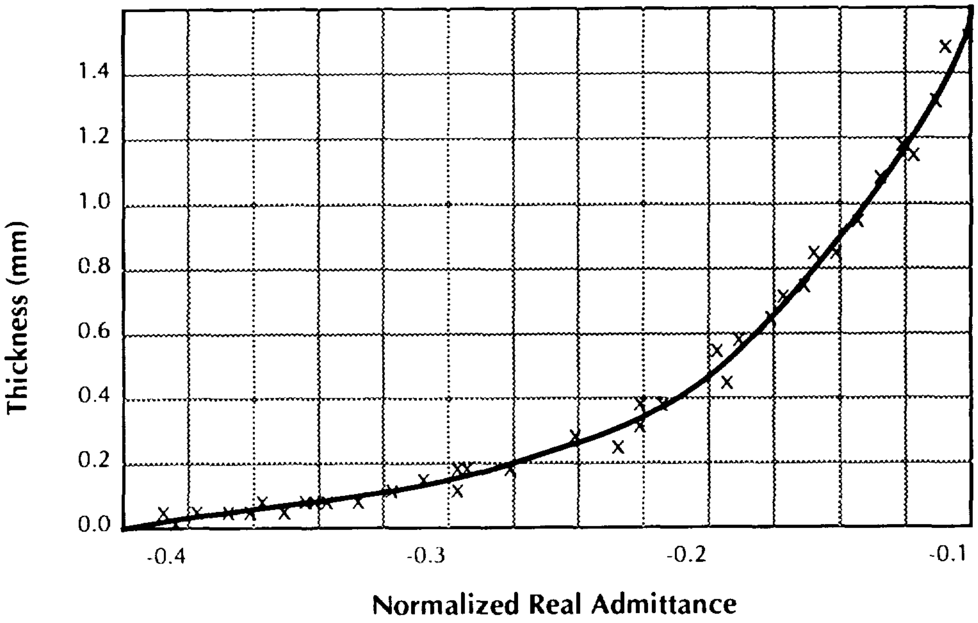


Figure 7: Plot of NDS Sensor Thickness Sensitivity

The development of the NDS sensor system is nearly complete and installation at the CELSS facility at NASA Kennedy Space Center is set for January 1993.

SUMMARY

DiComp SDS technology has numerous industrial applications in single or multi-component liquid streams over a broad range of concentrations from high percent to trace part-per-million. The unique SDS sensor can be flexibly configured to work with nearly any material or process, while the DiComp system retains the stability and low cost of RF frequency operation.

Axiomatics is building on its initial commercial success in measuring concentration or percent solids in dispersions, emulsions and solutions. The broad range of liquid streams which can be used make the DiComp applicable in agricultural productions, chemical processing, foods and beverages, petrochemicals, pharmaceuticals, polymers, pulp and paper and textiles.

REFERENCES

[1] W.M. Siebert, *Circuits, Signals and Systems*, MIT Press, Section 2.5 1989.

[2] M Clausse, Dielectric Properties of Emulsions and Related Systems, *Encyclopedia of Emulsion Technology.*

[3] DiComp, *ISA/Today*, p.21, October 19, 1992.

[4] D.R. Hoagland and D.I. Arnon, The Water Culture Method for Growing Plants Without Soil, *Circular 347,* University of California at Berkeley 1938.

[5] R.P. Prince and W.M. Knott, CELSS Breadboard Project at the Kennedy Space Center, *Lunar Base Agriculture: Soils for Plant Growth*, ASA-CSAA-SSSA, Madison, WI pp. 155-163 (1989).

[6] T.W. Dreschel and J.C. Sager, Control of Water and Nutrients Using a Porous Tube: A Method for Growing Plants in Space, *HortScience*, 24(6), pp. 944-947 (Dec 1989).

[7] H.V. Koontz, R.P. Prince, and W.L. Berry, A Porous Stainless Steel Membrane System for Extraterrestrial Crop Production, *HortScience*, 25(6), p.770 (June 1990).

[8] F.A. Waldman and C.R. Davis, Novel Sensor Technology for Monitoring and Control of Critical Plant Nutrient Parameters, 29th Plenary Meeting of the Committee of Space Research, *Advances in Space Research*, to be published.

A MODULAR, PROGRAMMABLE MEASUREMENT SYSTEM FOR PHYSIOLOGICAL AND SPACEFLIGHT APPLICATIONS

John W. Hines
SENSORS 2000! Program Manager
NASA-Ames Research Center
Mail Stop 213-2
Moffett Field, CA 94035-1000

Robert D. Ricks and Christopher J. Miles
Sensors 2000! Program
Sverdrup Technology-Ames Division
NASA-Ames Research Center

ABSTRACT

The NASA-Ames Sensors 2000! Program has developed a small, compact, modular, programmable, sensor signal conditioning and measurement system, initially targeted for Life Sciences Spaceflight Programs. The system consists of a twelve-slot, multi-layer, distributed function backplane, a digital microcontroller/memory subsystem, conditioned and isolated power supplies, and six application-specific, physiological signal conditioners. Each signal conditioner is capable of being programmed for gains, offsets, calibration and operate modes, and, in some cases, selectable outputs and functional modes. Presently, the system has the capability for measuring ECG, EMG, EEG, Temperature, Respiration, Pressure, Force, and Acceleration parameters, in physiological ranges. The measurement system makes heavy use of surface-mount packaging technology, resulting in plug in modules sized 125x55mm. The complete 12-slot system is contained within a volume of 220x150x70mm. The system's capabilities extend well beyond the specific objectives of NASA's programs. Indeed, the potential commercial uses of the technology are virtually limitless. In addition to applications in medical and biomedical sensing, the system might also be used in process control situations, in clinical or research environments, in general instrumentation systems, factory processing, or any other applications where high quality measurements are required.

SENSORS 2000! PROGRAM

Overview

Sensors 2000! (S2K!) is an Advanced Technology Sensor Systems development initiative based in the NASA-AMES Research Center (ARC) Electronic Systems Branch. The S2K! charter is to research, design, develop, evaluate, and apply biomedical, biosensor, and instrumentation technology for use in NASA Space and Life Sciences Flight programs. The S2K! emphasis applies to all elements of a Sensor System, including transducers, signal conditioners, power and control subsystems, telemetry, packaging, and data systems. To accomplish these objectives, we have implemented a dual strategy, addressing both advanced and enabling sensor technology research, as well as hardware, instrumentation, and systems development, ranging from preliminary engineering breadboards, to ground-based laboratory prototypes, to fully developed, tested, and spaceflight qualified sensor and measurement systems.

Technology Development Emphasis

In keeping with the S2K! strategy outlined above, and recognizing the similarities between sensor and measurement requirements for manifested and planned Life Sciences and Life Support activities, we initially attempted to define and compile discipline specific science measurement parameters, subjects, and test platforms/configurations. The thrust was to converge on a modular, high quality, standardized, hardware architecture which could be applied to multiple scenarios, with the additional capability to be modified, upgraded, programmed, and reconfigured, depending upon the specific requirements and logistics of the application under consideration. Targeted applications and scenarios considered encompassed both research and spaceflight programs.

Research Applications

Research applications included ground-based and near term flight programs, including neurovestibular, cardiovascular, biopotential, musculoskeletal, and life support (water, air quality, and envionmental systems monitoring and control) technology disciplines. Requirements to provide advanced technology biosensors and systems suitable for use in current and planned flight programs involving implantable and external biotelemetry systems over a 5-20 year life cycle have greatly influenced the modular measurement system design.

Spaceflight Program Applications

In addition to the research applications described above, two near-term spaceflight programs have strongly influenced the initial definition of the specific configurations, specifications, and operating parameters of the modular signal conditioning system. These are the US/French Rhesus Research Facility, and the US/Russian Cosmos Biosatellite Program.

The Rhesus Measurement System (RMS) was designed using the modular signal conditioner platform concept to monitor 8-16 channels of physiological data from primate test subjects during an 8-14 day space shuttle mission, planned to be flown aboard the third and fourth dedicated Life Sciences Spacelab missions (SLS-3 and SLS-4). In addition to the primary measurements, the RMS is required to acquire and store digital data during the launch and reentry phases of the flight mission, accept control and command data from the onboard host data management system, and be fully space qualified.

The future requirement to be able to modify, upgrade, or reconfigure the configuration and parameter mix, as well as to double the number of channels from within the same physical space, is a strong justification for the use of a modular, programmable measurement system. Even with the heavy use of surface mount electronics packaging technology, this presents a significant challenge for instrumentation system design.

Bioinstrumentation and sensor systems requirements for the Cosmos '92 Biosatellite mission were even more stringent than those for the Rhesus Project. Logistics of this system required the development of three hybrid integrated circuit function blocks, a manually programmable strain gage signal conditioner card, mode selectable power supply, and an application specific subsystem to measure angular acceleration. An added complication for this system was the requirement to interface and be plug compatible with Russian hardware and signal conditioners, in some cases sharing or splitting responsibilities between functional elements.

Development of modular systems and function blocks for both the Rhesus and Cosmos programs has significantly demonstrated the advantages of using a modular programmable measurement system architecture and approach for Space Life Sciences Instrumentation Development.

MODULAR SIGNAL CONDITIONING SYSTEM

System Configuration(s)

As presently configured, the modular system is configured to accommodate up to 16 channels of input. Physically, it consists of a 12-slot, distributed function backplane, with seven analog signal conditioner slots, two power slots, and three slots allocated for the digital subsystem. Each card measures 125x 55 m (approx. 5 in x 2 in), and is physically contained within a volume of 220x150x70mm (approx. 8.5x6x3 in). Figures 1 - 3 show examples of the backplane and modular cards.

Application specific analog signal conditioner cards have onboard preamplifers, buffers, filter blocks, and programmable gain, offset, and mode selection features. In some cases, preprocessing of the raw analog data is also accomplished. All signal conditioners have onboard calibration circuitry. Although each card is presently application specific, the modular design allows for upgradeability and interchangeability. The signal conditioner design is not limited to biomedical applications, and can be adapted to other measurement requirements with appropriate sensor and logistical considerations.

The microcontroller module uses a Motorola 68HC11 microprocessor and has onboard capability for A/D conversion, parallel digital interface to the signal conditioners, and RS-232 serial communications capabilities. The memory module provides the capability to store up to 3.5 megabytes of 8-bit data storage with battery backup.

In addition to conditioning and level shifting external power (AC or DC), the modular system provides full ground isolation to insure subject safety in biomedical applications. Some power supply configurations which support low noise preamplifiers and signal conditioners use rechargeable batteries to supply the low-noise elements, with a charging circuit activated during periods when data is not being collected. Others operate directly from isolated DC-DC converter power supplies.

Although the primary configuration employs completely self-contained, application specific signal conditioner cards, each having both preamplifiers and output stages onboard a single card, preliminary efforts are underway to separate the preamplifiers from the main amplifiers, and to locate the preamplifers closer to the subject, in an experiment-unique arrangement. This configuration would allow a standard, universal main amplifier block, with application-specific front ends, and may thus reduce the complexity and cost of the signal conditioner module(s), while preserving the specificity of the system. The Cosmos '92 hardware system uses this basic configuration and is being studied for applicability to the overall modular signal conditioner scheme.

General Specifications

For the Rhesus Project, the modular signal conditioner within the RMS is designated the Animal Analog Signal Conditioner (AASC), and is configured as shown in Table 1 for the first (SLS-3) flight mission.

BOARD	INPUT LEVEL (V)	GAIN RANGE	OFFSET RANGE (V)	FILTER RANGE (Hz)
EMG	.1 - 10 mV	400 to 40,000	2.5 nominal	2 to 1000
EEG	.01 - 10 mV	400 to 40,000	2.5 nominal	1 to 100
TEMP	0 - 1 V	0 to 100	2.5 nominal	Low Pass 1 Hz
ECG	1 V P-P	1,2,5,10,20, 50, 100	2.5 nominal	0.05 to 100
Dual RESP	1 V P-P	1,2,5,10,20,50, 100	2.5 nominal	Low Pass 10 Hz
PWR Cond	± 8.5 V	N/A	N/A	N/A
PWR ISO	± 8.5 V	N/A	N/A	N/A
Micro-Controller	68HC11 Microcontroller, provides parallel control ports to cards: 8 ch., 8 bit A/D converter, RS-232 Serial Communications Port			
Micro-Peripheral	Provides control for the memory plus battery backup for memory			
Memory	≥ 2 Mbytes 8 bit SRAM Memory			

Table 1. AASC Board Characteristics.

Module Descriptions

Following is a summary list of some elements, modules, and function blocks currently available or under development. Most have been developed totally inhouse, using the expertise and resources of the Sensors 2000! Program. In some cases, collaborative efforts have been undertaken with commercial sensor/instrumentation/vendors, whereby their technology has been licensed for adaptation and incorporation within the modular signal conditioner architecture. In all cases, a cohesive pathway has been defined, or specified, which can make maximum use of advances in technology, past and current experience, and specific applications,configurations, and logistics concerns.

PRESENTLY AVAILABLE MODULAR SYSTEM ELEMENTS

Backplane

Analog Subsystem(s)

* Electromyogram (EMG) Signal Conditioner
* Universal Strain Gauge Signal Conditioner
* Electroencephalogram (EEG) Signal Conditioner
* Dual Respiration Signal Conditioner

Digital Subsystem(s)

* Microcontroller Module(s)
* Digital Memory Module(s)
* Data Acquisition System

Power Subsystem

* Power Conditioner
* Power Isolator
* Rechargeable Power Supply

Sensors and Interfaces

* ECG/Temperature Biotelemetry Transmitter
* Tendon Force Sensor

Preamplifiers

* Seven Channel Neurovestibular (hybrid)
* Four Channel Neurovestibular (hybrid)
* Four Channel Peripheral (hybrid)

Application-Specific Subsystems

* Biotelemetry Receiver(s)
* Respiration Measurement Systems

PLANNED UPGRADES

Present plans for upgrades and additions to the modular signal conditioning family include further miniaturization and consolidation of existing signal conditioner and function blocks, making further use of surface-mount and hybrid circuit packaging, and semi-custom integrated circuit and multi-chip-module technologies. We intend to further merge the various elements and components into a distributed, integrated system which can be rapidly prototyped and applied to specific applications, including those outside of the primary Space Life Sciences thrust. The family of signal conditioner cards will expand to include other life sciences disciplines including cardiovascular, optical, and biochemical/biological measurements, as well as those parameters particular to environmental, life support, and process control scenarios.

A significant amount of emphasis will be placed on the development, use, and application of implantable and external telemetry systems, in a variety of embodiments, to increase the utility and applicability of the instrumentation in distributed, remote, and unattended situations.

CONCLUSIONS

The modular signal conditioner approach described in this paper represents an effort to consolidate instrumentation and measurement system architectures for Space Life Sciences Sensor Systems. These systems have been and are continually being subjected to rigorous analysis and testing to qualify them for flight on the Space Shuttle and eventually Space Station Freedom. Components have been chosen for the highest reliability and performance standards. Units are built to withstand vibration, thermal variations, ambient pressure fluctuations, and electromagnetic susceptibility, and compatibility. In most cases, the units are designed and built for minimal or unattended operation, in highly critical situations where failure to perform will result in the loss of costly and irreplaceable information.

Although the modular signal measurement system has been developed initially for physiologic measurements onboard Life Sciences Spaceflight missions, the technology lends itself extremely well to any sensor or measurement situation where small, high quality, modular, reconfigurable instruments are needed. Because the device has been developed specifically for spaceflight applications, a great deal of attention has been given to safety, quality and reliability issues. The system lends itself exceptionally well for both general purpose and discipline specific applications.

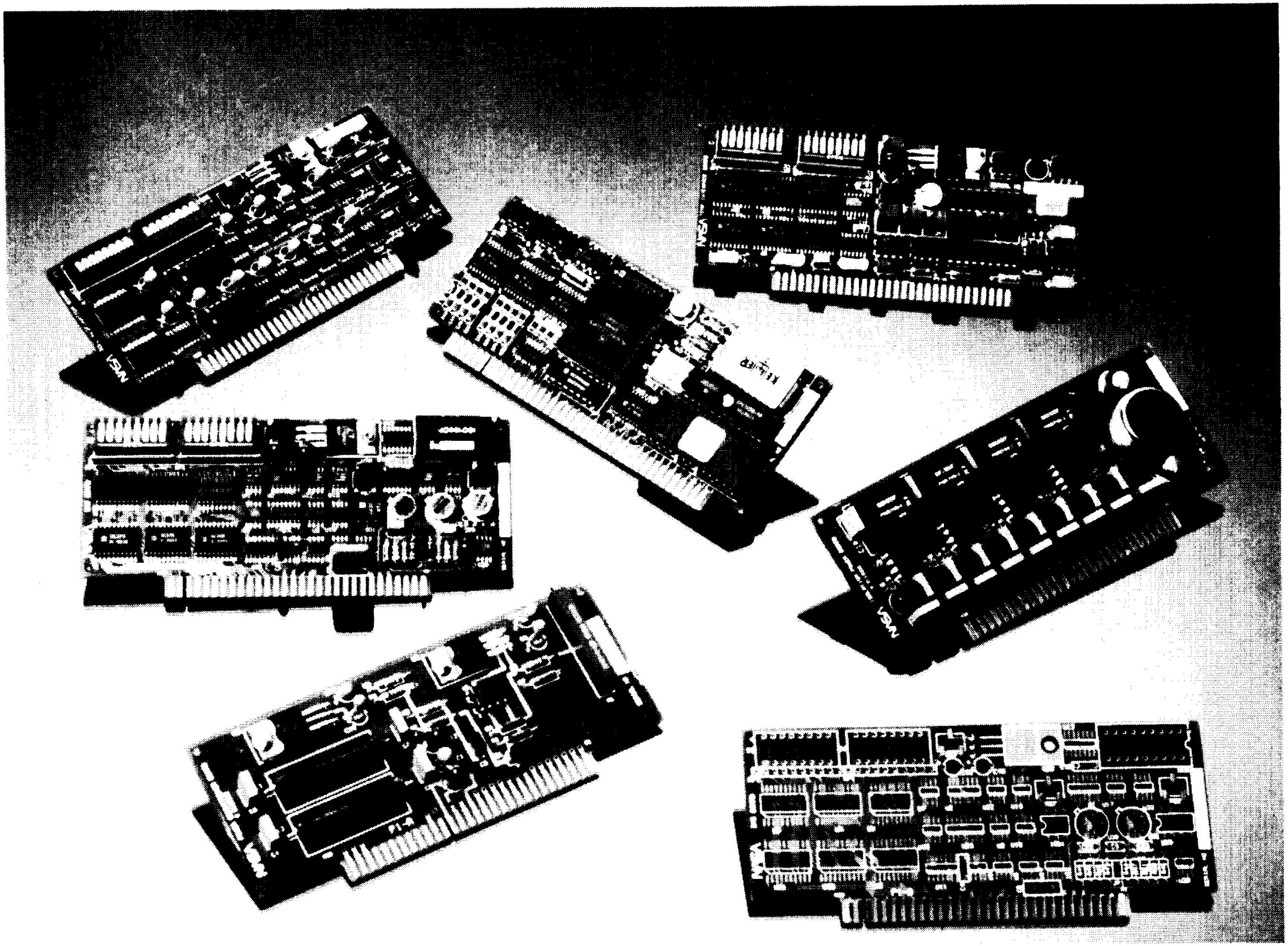


FIGURE 1: Modular Signal Conditioner Card Suite

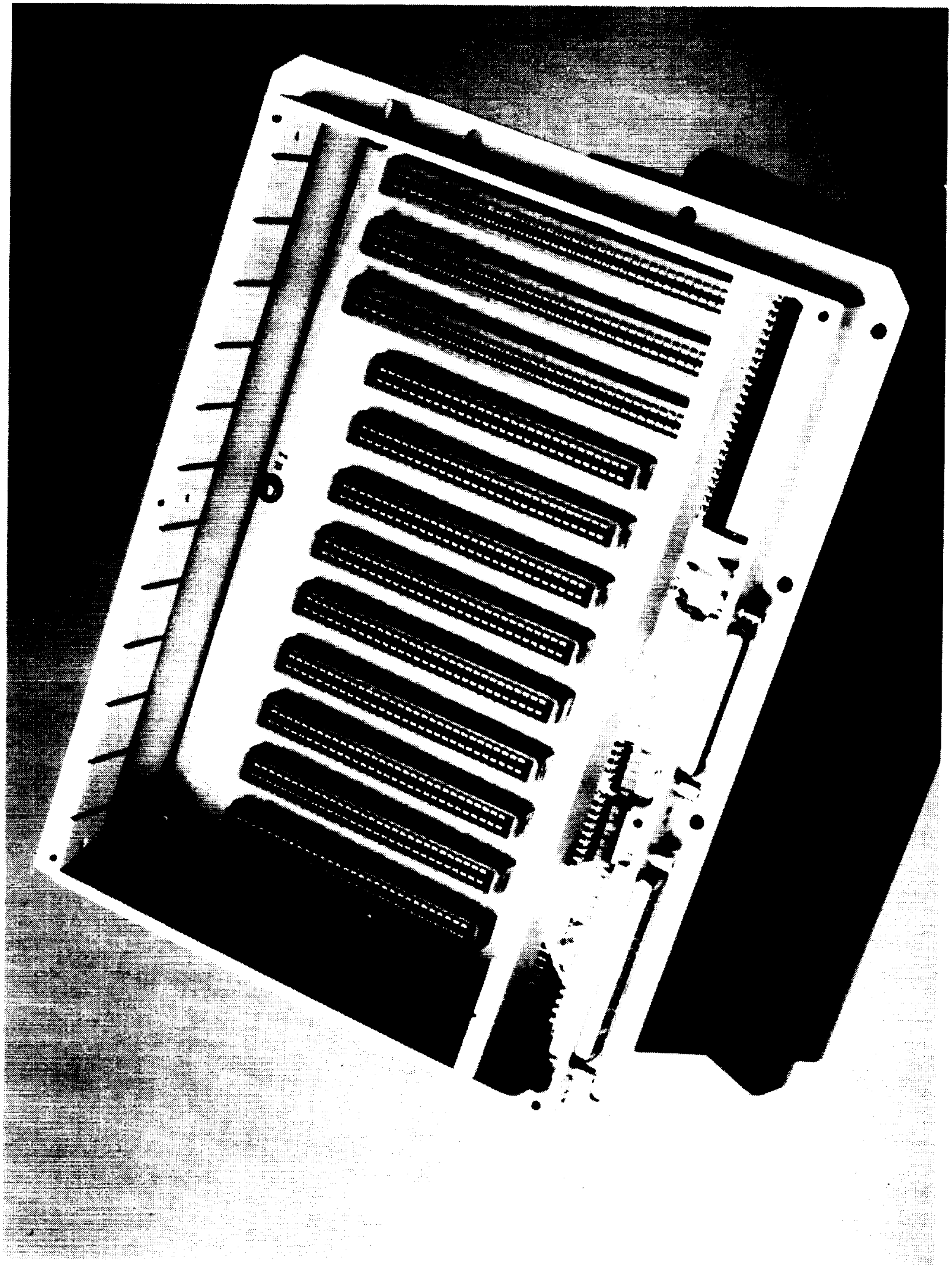


FIGURE 2: Modular Signal Conditioner Backplane

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

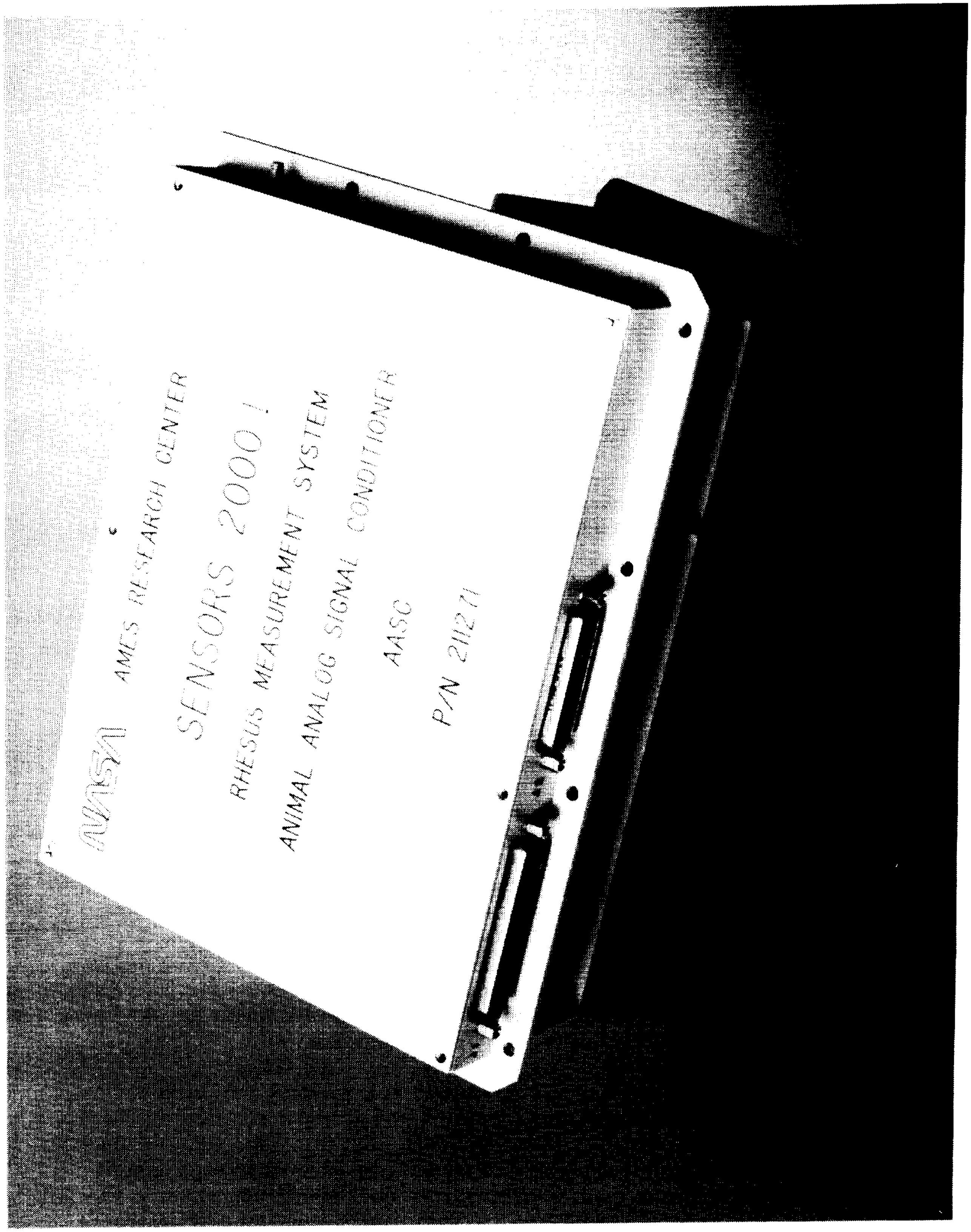


FIGURE 3: Rhesus Measurement System Signal Conditioner

PREDICTIVE SENSOR METHOD AND APPARATUS

Dr. Vivien J. Cambridge, Sr. Engineer
Sverdrup Technology, Inc.
NASA John C. Stennis Space Center
Stennis Space Center, MS

Thomas L. Koger, Engineer III
Sverdrup Technology, Inc.
NASA John C. Stennis Space Center
Stennis Space Center, MS

ABSTRACT

A microprocessor and electronics package employing predictive methodology was developed to accelerate the response time of slowly responding hydrogen sensors. The system developed improved sensor response time from approximately 90 seconds to 8.5 seconds. The microprocessor works in real-time providing accurate hydrogen concentration corrected for fluctuations in sensor output resulting from changes in atmospheric pressure and temperature. Following the successful development of the hydrogen sensor system, the system and predictive methodology was adapted to a commercial medical thermometer probe. Results of the experiment indicate that, with some customization of hardware and software, response time improvements are possible for medical thermometers as well as other slowly responding sensors.

INTRODUCTION

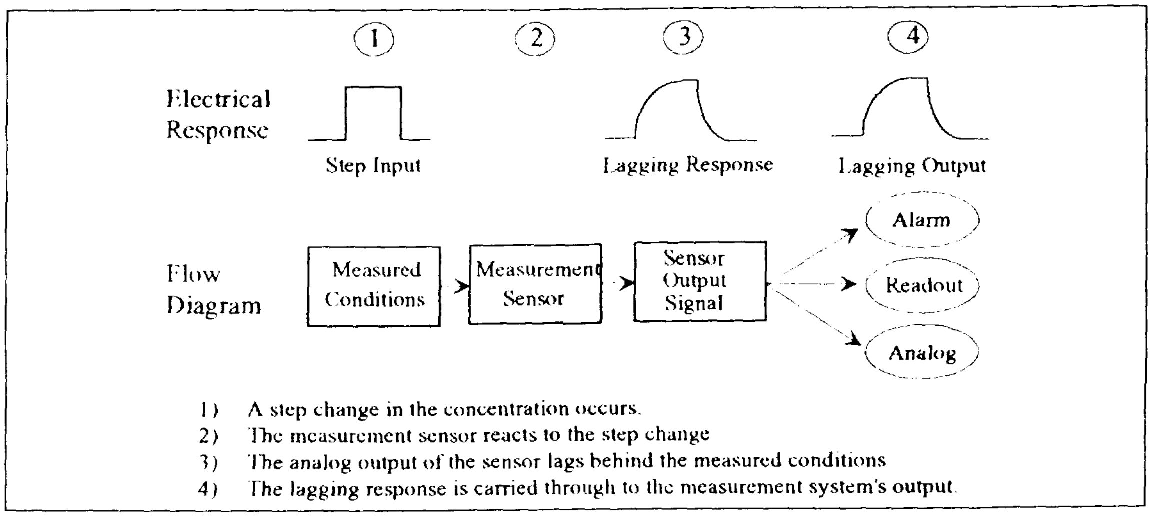
John C. Stennis Space Center (SSC) is NASA's "Center of Excellence" for large rocket engine ground testing. In the course of certifying the Space Shuttle Main Engine (SSME) for flight readiness and conducting engine improvement research and development, SSC consumes 10 million pounds of hydrogen each year. To transport, store, and supply hydrogen for testing, SSC utilizes an extensive system of tank trucks, barges, storage tanks, pumps, and transfer lines. While most of the hydrogen is handled in liquid form through cryogenic storage vessels and vacuum jacketed piping, some hydrogen is also converted to the gaseous state prior to use.

Compounding the variety of hydrogen storage and handling problems found at SSC are a wide range of environmental conditions which make monitoring for leaks particularly difficult. Temperature can vary from rather high, due to the near tropical summers of south Mississippi, to very low, due to leakage of liquid hydrogen. Pressure can vary fairly dramatically in areas near the rocket engines due to overpressure or drawdown effects during engine tests. In some facility areas, inert "purge" gases are used to minimize the possibility of hydrogen ignition in the event of leakage. In the dynamic environment of test operations, dramatic changes in these variable conditions can occur at any time.

The potential for severe damage or injury resulting from the ignition of leaking hydrogen prompted NASA to pursue development of a fast, rugged and reliable hydrogen leak sensor capable of providing accurate results through a wide range of rapidly changing environmental conditions. Although a commercial sensor was available with good ruggedness and immunity to interferences, the sensor responded slowly and exhibited non-linearities with fluctuations in temperature and pressure. A predictive sensor method and apparatus was developed to obtain the fastest possible response time while taking advantage of the slow sensor's more desireable characteristics. The concept was implemented using a commercially available microcontroller to: 1) acquire data from hydrogen, temperature, and pressure sensors, 2) process the predictive algorithm, and 3) linearize the output for the measured fluctuations in temperature and pressure. The success of this method stimulated further investigations of other applications involving slow sensors to determine the suitability of the predictive sensor method and apparatus for commercial development.

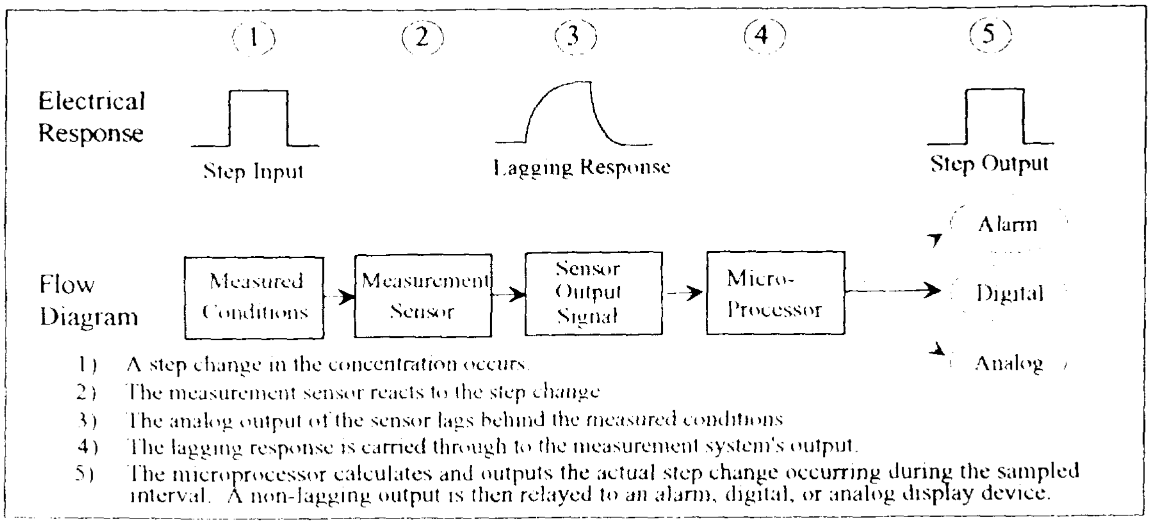
PREDICTIVE SENSOR METHODOLOGY AND SYSTEM DEVELOPMENT

To varying degrees, most sensors exhibit responses which lag behind the input eliciting the response. In conventional measurement systems, this lag is carried through and registered in the systems output device. A diagram of the conventional measurement process is shown in Figure 1. In many applications a lagging response is tolerable. In some cases however, a very rapid or near-real time response may be critical. In such cases, the measurer may be tempted to trade-off other desireable sensor features such as linearity, repeatability, ruggedness, or cost in order to achieve the desired speed of response. With the predictive sensor method and apparatus, the desirable characteristics inherent to the slow sensor are maintained while response time is dramatically improved. An illustration of the predictive sensor method and apparatus is provided in Fig. 2.



1) A step change in the concentration occurs.
2) The measurement sensor reacts to the step change
3) The analog output of the sensor lags behind the measured conditions
4) The lagging response is carried through to the measurement system's output.

Figure 1. Illustration of Lagging Response in Conventional Measurement Systems



1) A step change in the concentration occurs.
2) The measurement sensor reacts to the step change
3) The analog output of the sensor lags behind the measured conditions
4) The lagging response is carried through to the measurement system's output.
5) The microprocessor calculates and outputs the actual step change occurring during the sampled interval. A non-lagging output is then relayed to an alarm, digital, or analog display device.

Figure 2. Illustration of Predictive Sensor Method and Apparatus

The slow responding hydrogen sensor, for which this method was developed, is an electrochemical sensor originally designed for service in nuclear power plant containment vessels. With the exception of response speed the sensor's construction for extreme ruggedness, high reliability, and good accuracy, make the sensor ideally suited for NASA service. The sensor is comprised of a semi-solid electrolyte with a platinum black sensing electrode and platinum reference electrode. The sensing electrode sits behind a polymer membrane which is selectively permeable to hydrogen. The selective permeability of the membrane is the reason for the sensor's good rejection of gases which typically interfere with the accuracy of electrochemical sensors. However, the selective permeability of the membrane is also the primary cause for the sensor's lagging response. Specifically, the hydrogen molecules diffuse through the membrane at a temperature dependent rate thus limiting the rate which the hydrogen enters the electrochemical cell and causes a change in potential between the sensing and reference electrodes.

Hardware and Software Implementation

An analysis of the lagging process led to the development of a mathematical model. After some refinement using spreadsheet analysis, the model was implemented for near-real time estimation of hydrogen concentration and linearization for fluctuations in temperature and pressure. The model was initially coded and tested in the C language. The code was later converted to BASIC and implemented in microcontroller firmware. Ultimately, a compact system, known as the Smart Hydrogen Sensor (SHS) was developed to acquire and process data from the hydrogen sensor, temperature sensor, and pressure sensor. The system configuration is shown in Fig. 3.

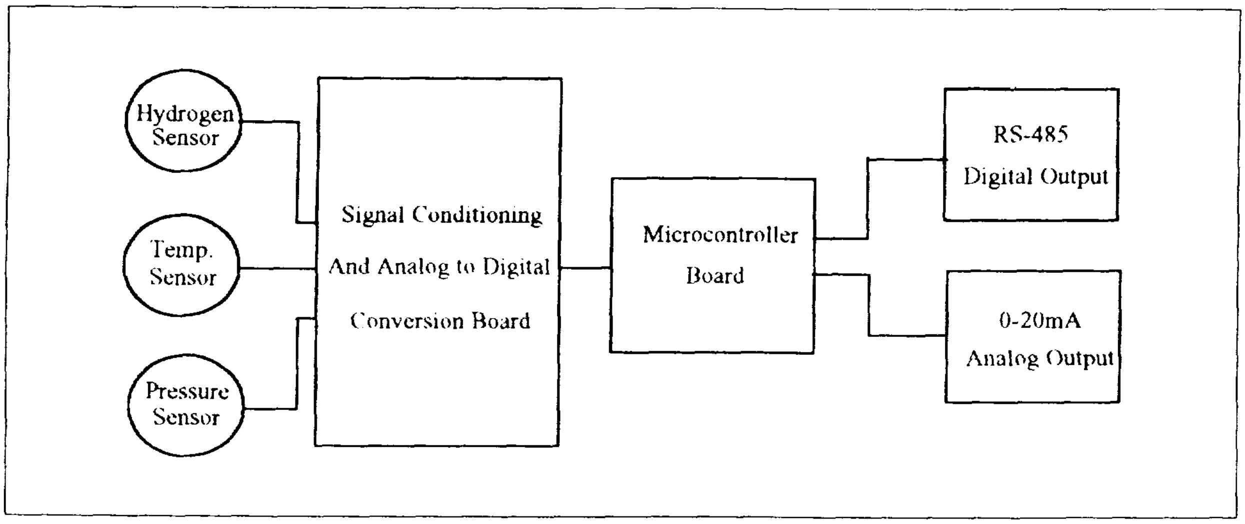


Figure 3. Smart Hydrogen Sensor Hardware Configuration

The approach chosen for the development of the SHS hardware was to use proven, off-the-shelf components. This approach eliminated the need to design new hardware and minimized testing requirements.

The SHS is comprised of three pieces of integrated electronics components. The hardware components are:

(1) Microcontroller Board
(2) Signal Conditioning Board (designed at Stennis Space Center)
(3) RS232C/0-20mA Converter.

Microcontroller Board

The Microcontroller Board is built around an eight-bit microprocessor. It provides three 8-bit parallel I/O ports (24 bits), two asynchronous serial ports (one full-duplex RS-232, one half-duplex RS-485), eight-channel analog-to-digital converter (10-bit resolution at five volts), 96 kb of total on-board memory (EPROM and RAM), and 1024 bits (64 bytes x 16 bits) of EPROM.

As implemented with the SHS, the Microcontroller Board uses a ROM monitor and a multitasking BASIC compiler. Software development was carried out directly on the Microcontroller Board. The board interfaces directly to the Signal Conditioning Board.

Signal Conditioning Board

The Signal Conditioning Board, designed at Stennis Space Center, rides piggy back on the Microcontroller Board. The primary purpose of this board is to (1) convert input power to appropriate levels needed by the SHS unit and (2) condition the signals from the hydrogen, pressure, and temperature sensors prior to the digital conversion by the microcontroller.

The signal conditioning board also features a Computer Operating Properly (COP) Watchdog timer and an eight pole switch (dual-in-line package). The eight pole switch is used to (1) configure the SHS unit for the RS485 network and (2) place the sensor in either NORMAL or CALIBRATION operating mode.

RS-232/0-20mA Converter

The RS232/0-20mA Converter provides a direct means for outputting an analog signal from the SHS unit. The converter uses simple ASCII commands to control a 12-bit digital-to-analog converter. An on board microprocessor provides the communications interface. The RS232/0-20mA Converter receives the ASCII commands from the Microcontroller Board's RS232 serial port.

Software Description

The SHS software, written in a special version of BASIC which is unique to the Microcontroller Board, resides as firmware in EPROM. Although this microcontroller BASIC utilizes commands not standard to BASIC, modifications to the software are relatively simple with an understanding of standard BASIC commands and programming practices.

The main purpose of the firmware is to acquire temperature, pressure, and hydrogen data and use this data to estimate the concentration of hydrogen in the environment. This estimate is updated every second and is sent out to a 0-20mA analog channel. Furthermore, the firmware is responsible for monitoring communications over the RS485 network. If the SHS receives a valid command or request, it is responsible for responding with an appropriate reply.

Testing and Performance

Several iterations of prototype development and tests were completed before arriving at a software and hardware configuration deemed ready for operations in the NASA environment. The prototypes were tested in Stennis' laboratories over a wide range of temperatures and were exposed to a variety of background gases and hydrogen concentrations. Sensors were placed in some of the most severe operational environments imaginable, including the rigors of actual rocket engine firings, and were exposed to cryogenic fluids, saturated oxygen vapors, and heavy water deluge sprays.

Final testing results showed the sensor response to environmental changes was more linear with actual response time increased by a factor of 10. For comparison, sensor response time at 68°F to a 90% step change in H_2 concentration without use of the predictive algorithm was 1.5 minutes. Response time with the predictive method was significantly shortened to 8.5 seconds.

Additional software added to the system enables menu-driven operation, calibration and maintenance of the system's computer which reduces maintenance cost and ensures uniformity in system operations. Benefits of the Smart Hydrogen Sensor are: 1) speed of response, 2) accuracy, 3) reliability, 4) ruggedness, 5) ease of operation, and 6) flexibility.

APPLICATION STUDY - SMART THERMOMETER FEASIBILITY

With the successful application of the predictive hydrogen detection system, it followed that slow responding commercial measurement systems might benefit from this development. Further investigations were thus initiated to determine if the Smart Hydrogen Sensor technology could be readily applied to similar technological difficulties within the commercial sector. To determine viability, a test was developed using a leading brand medical-type electronic thermometer mated with the SHS signal processing hardware and software.

Hardware and Software Design and Assembly

Testing of the commercial thermometer system showed the normal response time to be between 25 and 30 seconds. Since the analog signal of the probe was different from the hydrogen sensor, a custom analog signal conversion circuit was designed and fabricated. The probe and custom circuit were then interfaced with an SHS programmable microcontroller and digital to analog converter. The microcontroller's BASIC software used for the SHS predictive algorithm was modified to process the temperature data from the thermometer probe.

Upon successful test of the software, the analog and digital electronics were assembled and calibrated using a precision calibration water bath. The method is similar to the ASTM standard for calibration of electronic medical thermometers.

Preliminary Test and Evaluation

Since medical practice requires the use of plastic sleeves on the thermometer probe for sanitary purposes, a problem was foreseen with the predictive method in that the individual probe covers could randomly alter the time constant of the temperature measurement assembly. To test the breadboard system and the variable time constant hypothesis, a series of temperature measurements were taken in a controlled temperature bath using different probe covers. Final temperature results obtained with different probe covers were found to be invariable, but, the time of response to reach the final temperature was found to vary by several seconds. The variation in response time necessitated the entry of a new time constant in the software to obtain an accurate prediction of the final temperature. The predictive method would be problematic in a medical application where accuracy is critical without an accurate value for the probe/temperature probe assembly time constant.

Modification of Predictive Method

A concept was therefore developed for solving for a unique time constant of the assembly "on-the-fly" by analyzing the first few seconds of measurement data. The method was developed using the data available from controlled temperature bath experiments. While these predictions from controlled experiments yielded promising results, viability under "clinical" conditions remained in question. Further investigations were then conducted using temperature data from volunteers in the laboratory. Three to four measurements were obtained from each volunteer using a different probe cover for each measurement. Selected data representative of the testing is discussed below.

Results

The data provided in Fig. 4 below shows the raw output of the temperature probe, the standard predictive method (as employed by the SHS), and the modified predictive method. For Fig. 4, the probe was inserted approximately 3 to 4 seconds after starting data acquisition. As can be seen, the standard prediction reaches the final temperature value within approximately 1 second. However, the standard prediction increases above the actual measured condition as a result of an error in the time constant assumed for the probe cover / temperature probe assembly used. This effect may be the result of minor differences in the probe cover or in the way the probe cover is attached to the probe.

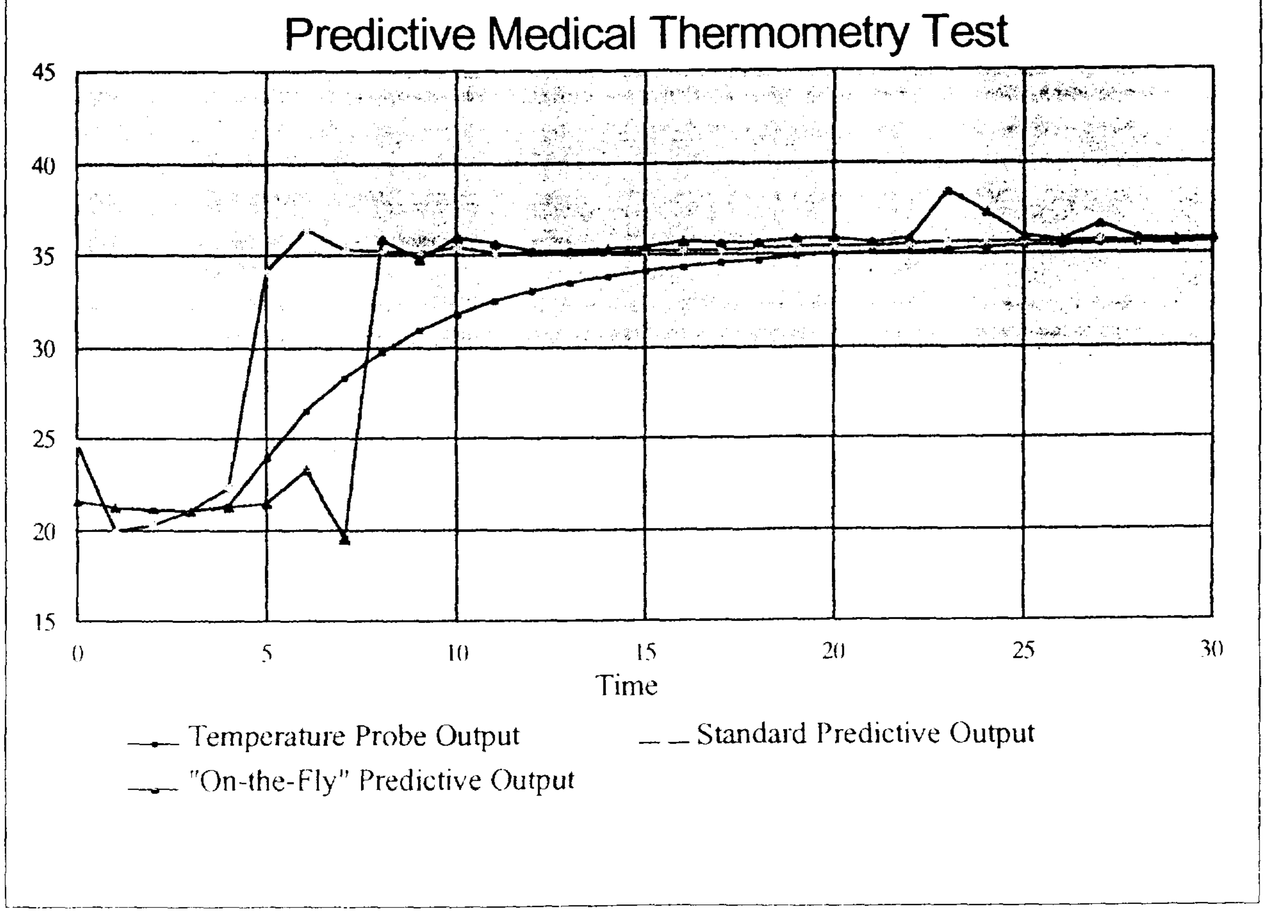


Figure 4. Test of Predictive Method in Medical Thermometry Application

The "on-the-fly" prediction uses temperature data from the first few seconds to calculate a time constant value unique to the probe in use. As seen in Fig. 4, this procedure produces a initial lag until the microcontroller has accumulated sufficient data to compute the time constant and final value, based on the "on-the-fly" time constant.

While a technique was implemented to solve the unique time constant problem associated with the probe covers, the data indicate that the technique does not improve the predictive results as it is highly sensitive to small nuances in the raw data. The use of thinner, more uniform, or more conductive probe covers might reduce the effect of differences in the manufactured probe covers on the time constant. This remains an area for further investigation.

More positively, the predictive method could prove beneficial for less critical applications such as household thermometers which require no sleeve covers. In such applications it is likely that the relationship between cost considerations and accuracy issues would shift, and more emphasis be placed on the design of a low cost chip and software.

PREDICTIVE SENSOR TECHNOLOGY APPLICATIONS

The Smart Hydrogen Sensor was developed to enhance the detection of hydrogen in a variety of gaseous atmospheres. The advanced electronics of the SHS system provide for reliable and rapid estimation of hydrogen concentration along with enough flexibility to function in a variety of environments. No commercial technology has been identified that can out perform the SHS in the areas of speed, accuracy, and reliability. The greatest attribute of this newly developed predictive sensor technology, however, is that it can significantly enhance the speed of response of existing sensor technology. Faster responses can be obtained without developing a faster sensor. Application of the predictive methodology may provide cost effective alternatives for existing sensors that are limited by slow response times. The signal processing algorithm employed can determine in near real time the steady state response of a normally slow sensor.

While a few shortcomings in its use in temperature measurement for critical medical applications remain to be resolved, the technology is readily applicable and adaptable to other types of temperature measurement systems.

ATTIRE

(Analytical Tools For Thermal Infrared Engineering) - A sensor simulation and modeling package

S. Jaggi

Enginnering and Science Department
Sverdrup Technology
Stennis Space Center, MS 39529

ABSTRACT

The Advanced Sensor Development Laboratory (ASDL) at the Stennis Space Center develops, maintains and calibrates remote sensing instruments for the National Aeronautics & Space Administration (NASA). To perform system design trade-offs, analysis, and establish system parameters, ASDL has developed a software package for analytical simulation of sensor systems. This package called " Analytical Tools for Thermal InfraRed Engineering " - ATTIRE, simulates the various components of a sensor system. The software allows each subsystem of the sensor to be analyzed independently for its performance. These performance parameters are then integrated to obtain system level information such as Signal-to-Noise Ratio (SNR), Noise Equivalent Radiance (NER), Noise Equivalent Temperature Difference (NETD) etc. This paper describes the uses of the package and the physics that were used to derive the performance parameters.

In addition, ATTIRE can be used as a tutorial for understanding the distribution of thermal flux or solar irradiance over selected bandwidths of the spectrum. This spectrally distributed incident flux can then be analyzed as it propagates through the subsystems that constitute the entire sensor. ATTIRE provides a variety of functions ranging from plotting black-body curves for varying bandwidths and computing the integral flux, to performing transfer function analysis of the sensor system.

The package runs from a menu-driven interface in a PC-DOS environment. Each sub-system of the sensor is represented by windows and icons. A user-friendly mouse-controlled point-and-click interface allows the user to simulate various aspects of a sensor. Several interactive features allow for data plotting and visualization.

The package can simulate a theoretical sensor system. Trade-off studies can be easily done by changing the appropriate parameters and monitoring the effect on the system performance. The package can provide plots of system performance versus any system parameter. A parameter (such as the entrance aperture of the optics) could be varied and its effect on another parameter (e.g., NETD) could be plotted. A third parameter (e.g., the obscuration) could be varied for each plot and several plots obtained on the same graph. The menu for such " Y vs X plots for different values of Z " contains various such options. The package also allows the user to create customized work-sheets of the simulated system and save the analysis for interface with other packages.

The emissivity, atmospheric transmission and the optical transmission default as constants over the specified spectral bandwidths. There is an option for making these parameters spectrally variable. If more than one of the above-mentioned three parameters are spectrally variable, then it is possible that the upper and lower wavelength values as well as the resolution of the wavelength array may not be the same for all three arrays. The package performs an interpolation of the data to smooth out the curves and then projects them onto a common wavelength array for all the parameters.

INTRODUCTION

The design of sensor systems and analysis of their performance requires the use of several varied aspects of science and engineering. From the understanding of the laws of electromagnetic radiation to the computation of signal-to-ratios of a sensor, the exercise of sensor analysis and design is ideally suited for computer simulation. In this paper one such software package is described. The package called ATTIRE runs on a PC-DOS platform. It can be used for simulating sensor systems as well as understanding the behavior and interaction of various sub-systems that constitute a sensor. Fig. 1 shows the flow-chart for the energy flow that is modeled by ATTIRE. The sensor is represented by its various parameters as shown.

MENU DRIVEN SYSTEM

The user types 'ATTIRE' from the DOS prompt to enter into the package. The screen contains a flowchart for signal propagation of the energy that will be incident on the sensor. Fig. 2 shows the main menu as it may appear on the screen. The energy originates from a flux source. This source could either be the solar energy or thermal energy from a blackbody. An icon is used to represent the parameters that make up the sensor flux source. On clicking at this icon one enters into another window which displays the characteristics of the parameters that make up the source flux. The source flux energy is then propagated through the next subsystem - the atmosphere. The atmosphere attenuates the radiation coming from the source and also adds some radiation from its constituents due to effects such as scattering. This is followed by the optics at the sensor. The optics focus the energy of the source onto a detector. The detector converts the radiation into electrical energy. The optics sub-system and the electro-optics, i.e., the detector sub-system, are the next two icons in the signal flow. The electrical energy can then be amplified and modified in its bandwidth and noise characteristics. This is represented by the electronics icon. The last icon represents the spatial characteristics of the system, i.e., its dwell time, IFOV etc. The user can click onto either of these icons to enter into these sub-systems.

Apart from these six icons the user encounters the main menu. The main menu is made up of a window. The window can be resized or relocated using a mouse. The mouse is pressed at any of the edges of the window for resizing. The window can also be moved anywhere on the screen by pressing the mouse in the topmost line of the window. The outline of the window appears in the form of a rectangle which can then be moved anywhere.

The window is made up of three items - a listing of the system level parameters on a channel-by-channel basis, six buttons with numbers attached to them and a horizontal menu, .

The listing in the window contains information on the current sensor being analyzed. The listing is done on a channel-by-channel basis. Each channel is identified by its bandwidth. For each channel, the main window shows system level parameters. The energy available from each channel after it is attenuated based on its emissivity or albedo, the atmospheric and optical transferrance, FOV, and the nature of the source (thermal or solar) is computed in energy (W/cm^2) and photon units (photons/s/cm^2). The main window also contains the power in watts that is incident on the detector. This is followed by the three system level parameters - SNR, NER, and NETD.

The six buttons are for facilitating the display of a particular channel on the window. This happens when the number of channels is much greater than the window size. Each of these buttons contains a number. When clicking on a button, the channel corresponding to the number in the button is displayed on the top of the window screen. These numbers are programmable by using the item called 'CHAN' in the horizontal menu.

The horizontal menu performs functions to modify or analyze the sensor parameters. The first menu item ‘CHAN’ performs three major functions. It allows the user to modify the number of channels in the sensor being analyzed. The maximum number of channels that currently can be processed is 50. The second function of this menu item is to modify the bandwidths of the channels.

The process of modification of any parameter in ATTIRE follows a standard procedure. The user is first prompted with a menu asking if the modification needs to be done for only one channel, a few channels, or for all channels. For the case of one or a few channels, the user is prompted for the number of the channels that need to be modified. A tick mark appears next to the number of the channel. In the case of more than one channel being prompted for, the user can continue to click on the channel numbers. To delete a number that has been tick marked, the user just needs to click on it again. One continues this process until the required channels have been selected. The user then clicks outside the slection windows to accept the selected channels. Also, all the operations in ATTIRE can be done without a mouse control. In the case of channel selections, the up and down arrow keys are used to move within the selected window. If the number of channels is greater than the size of the selected window, the “Page Up” and “Page Down” keys allow for scrolling to other areas of the index. The tick marks are created by typing the “Enter” key. The end of the selection process is signified by the “Control - Enter” keys. After having selected the number of channels that need to be modified, the appropriate parameter is edited.

If only one channel is selected, the existing value of the parameter appears in a window. This value is appropriately edited. If more than one channel is selected, the modification can be done in two ways. The first method just allows for placing the same parameter value in all channels. This procedure is the same as for one-channel selection. The other manner in which the selected parameter can be modified is by placing a linearly increasing series of values in the selected channels. In this case the user is prompted for the initial and final values of the parameter starting from the first to the last selected channel.

The source emissivity, atmospheric transmission, and optical transmission are initially assumed constant over the specified spectral bandwidths. There is an option for making these parameters spectrally variable. If more than one of the above-mentioned three parameters are spectrally variable, then it is possible that the upper and lower wavelength values and the resolution of the wavelength array may not be the same for all three parameters. The software package performs an interpolation of the data to smooth out the curves and then projects them onto a common wavelength array.

The spectrally distributed incident flux can be analyzed as it propagates through the subsystems. Fig. 1 shows the signal flow and the various parameters that affect the performance of the system. Each of the icons on the main menu represent one of these sub-systems.

SUB-SYSTEMS

The software divided the system into the following sub-systems: 'SOURCE FLUX', 'ATMOSPHERICS', 'OPTICS', 'ELECTRO_OPTICS', 'ELECTRONICS' and 'PIXEL PARAMETERS',

The system can be simulated in a multi-spectral manner, i.e as many as 50 channels can be processed simultaneously for performance analysis. For each channel, the parameters corresponding to these sub-systems can be entered. The parameters can be entered one-by-one for each channel or in a linearly varying manner to analyze the effect of varying one parameter on the system performance. Also, all the channels can be made to assume the same value.

Since each channel is modeled independently of the other channels, the package can be used to simulate one 50-channel sensor or 50 unique sensors or any combination in between.

FLUX SOURCE

For analysis of the system a target has to be simulated as a source for the flux signal. This could be either a black body for thermal sources, the solar irradiance for visible/near IR sources or any customized source depending upon the application. Upon entering the menu the user can choose either of these options.

Blackbody Source

The spectral distribution of blackbody radiant exitance in energy units at a given temperature is obtained from Planck's formula

$$F_n(T) = \int_{\Delta\lambda_n} M_{e,\lambda}(\lambda, T)\, \Re_n(\lambda)\, d\lambda$$

where

$\Delta\lambda_n$ = spectral bandwidth of channel 'n'.

$\Re_n(\lambda)$ = Relative Responsivity of the sensor for channel 'n'.

$$M_{e,\lambda}(\lambda, T) = \frac{c_1}{\lambda^5 \left(e^{\frac{c_2}{\lambda T}} - 1 \right)} \quad \mathrm{W\,cm^{-2}\,\mu m^{-1}}$$

where

$c_1 = 2\pi hc^2 = 3.7483 \times 10^4\ \mathrm{W\,cm^{-2}\,\mu m^4}$

$c_2 = \frac{hc}{k} = 1.4388 \times 10^4\ \mu\mathrm{m\,K}$

h = Planck's constant.

k = Boltzmann constant.

λ = wavelength (μm).

T = temperature of source (K).

(1)

Assuming the source to be Lambertian , the spectral radiance is given by dividing the exitance by π steradians. If the source of the flux is not a blackbody, the spectral radiance is obtained by multiplying by the emissivity of the source as follows

$$L_{e,\lambda}(\lambda,T) = \frac{\varepsilon(\lambda)}{\pi}\ M_{e,\lambda}(\lambda,T)$$

$$= \frac{\varepsilon(\lambda)}{\pi}\left(\frac{c_1}{\lambda^5(\, e^{\frac{c_2}{\lambda T}} - 1\,)}\right)\ \mathrm{W\,cm^{-2}\,\mu m^{-1}\,sr^{-1}}$$

where $\varepsilon(\lambda)$ = emissivity of the source. (2)

For a field-of-view (FOV) Ω_s, the radiant exitance is given as follows:

$$M_{e,\lambda}(\lambda,T) = (\Omega_s)\ L_{e,\lambda}(\lambda,T)\ \ \mathrm{Wcm^{-2}\ \mu m^{-1}}$$

where Ω_s = Field of View. (3)

Solar Source

Solar radiation can be approximated by a 6,000 K blackbody curve. The solar spectrum is modified by atmospheric transmission as it passes through an air mass en route to the Earth's surface. For an airborne remote sensing system, this energy is further affected by the albedo of the surface type and the atmospheric path back to the sensor system. The spectral distribution of the solar irradiance was compiled from various sources[1,2] to arrive at a curve for wavelengths ranging from 0.295 to 2.541 µm. The resolution of this curve is 1 nm. This is the solar curve incorporated in the ATTIRE program, and used to calculate the available solar radiation for each reflecting channel. Fig. 3 shows the solar curve.

Modeling of the Source Flux in ATTIRE

ATTIRE provides an icon in the main menu called 'FLUX'. Fig. 4 shows the flux menu as it appears on the screen. On entering this menu from the main menu, another window appears along with a series of icons. These icons allow the user to change the type of flux (Blackbody or Solar), the temperature of the radiating source, the emissivity or reflectivity of the source and the Field-of-View (FOV) of the source. The window contains a horizontal menu whose items allow the user to perform the same operations as the main menu.

The flux type can be changed by clicking on the icon or typing 't' key. On entering the icon, the user is prompted for identifying the type of the flux. If the type is solar, care should be taken that the bandwidths for that channel are within the solar range. The Solar menu contains items that identify the solar flux; i.e., channel bandwidth and reflectivity of the target. The user enters the appropriate bandwidth values and the program picks up the corresponding irradiance values from a file called SOLAR.BIN. It then multiplies each irradiance value by the reflectivity to obtain the effective spectral flux. Selecting the Other Sources menu, the user can enter the source as a specific lamp or any emitting source whose distribution is known.

To obtain spectral radiant exitance for a source radiating uniformly in all directions, the FOV is made π steradians. To obtain spectral radiant exitance at any other solid angle, the appropriate value is entered.

In order to compute radiance, the FOV should be set equal to one steradian or 57.7° The FOV is modified by clicking on the icon or by typing 'v' key.

If the user has selected a thermal source as the flux, the program uses the temperature assigned for that channel to compute the flux. The temperature icon is used (or typing 't') for that purpose.

The medium between the source and the detector is made up of the atmospheric path and the optical path. The flux passes through the atmosphere before being collected by the optics of the sensor. It then passes through a series of reflective and refractive optical elements before it is incident on the detector.

Depending on the type of flux, the radiation is attenuated by the emissivity for the thermal and the albedo for the solar case. These are modified by clicking on the appropriate menu (or by typing 'e'). This value can be spectrally variable or constant. In the case of spectrally variable values, the user is prompted for a file name. This file name must contain the two arrays - wavelength and emissivity or reflectivity. This is an ASCII file which has these two parameters in each line of the file. The first two lines of the file are ignored and can be used for comments. The data is picked up from the third line. The first number is the wavelength and the second the parameter. The file continues to read until the line it read does not contain two numbers or end-of-file is reached. When a spectrally variable parameter is chosen, the parameters entered in channel bandwidth are not applicable as a wavelength array supplied with the file determines that bandwidth.

ATMOSPHERE

The atmosphere attenuates the amount of flux that enters the sensor. Only a fraction of the flux is incident on the optical aperture. This atmospheric transmission coefficient is denoted by $\tau_a(\lambda)$. It is assumed to be a function of the wavelength.

Also, the atmosphere contributes to the radiance entering the sensor system. This atmospheric path radiance can be divided into two categories, depending on the direction of the flux. The upwelling radiance, $L^{A1}(\lambda)$, is the flux from atmospheric particles that is directly incident on the sensor. The downwelling radiance,$L^{A2}(\lambda)$, is the flux that is incident on the target. This radiance then is reflected from the target back to the sensor.

Thus the effective flux incident on the sensor is given by:

$$L_\lambda(\lambda,T) = \tau_a(\lambda)\ [\ \varepsilon(\lambda)\, L^{BB}_{e,\lambda}(\lambda,T) + L^{A1}_{\lambda}(\lambda) + (1-\varepsilon(\lambda))\, L^{A2}_{\lambda}(\lambda)\ ] \tag{4}$$

ATTIRE Modeling of the Atmosphere

The atmosphere is selected by clicking on the icon for the 'ATMOSPHERE' in the main menu. Fig. 5 shows the menu as it appears on the screen. This menu contains three items for each channel - transmittance, upwelling and downwelling path radiance. The transmittance is modified in a manner similar to that for the emissivity in the flux sub-menu.

OPTICS

The aim of the optics subsystem is to collect, focus, and disperse the radiant flux from the source. The energy is then focused on the individual detectors. The two parameters that are critical to the collection of this energy are the area of the collecting optics and the focal length.

The radiance, limited by the solid angle subtended by the ground pixel, is incident on the entrance aperture and determines the irradiance available to the optics. The area of the entrance aperture minus any obscuration is the other limiting factor.

The solid angle is given by

$$\Omega = \frac{A_{pix}}{a^2} \tag{5}$$

where A_{pix} = area of the ground pixel
a = distance between the source and the entrance aperture (altitude)

This solid angle must equal the solid angle subtended by the detector on the entrance aperture (invariance theorem).

$$\frac{A_{pix}}{a^2} = \frac{A_d}{efl^2}$$

where A_d = area of detector
efl = effective focal length (6)

Depending on the design of the optical system, the entire area of the primary mirror may not be collecting the incident energy. This could be due to an obscuration of the mirror.

A_{eff} = Area of primary mirror - Area of obscuration

The series of reflective and refractive elements that constitute the optics subsystem are modeled by an optical transmission coefficient $\tau_o(\lambda)$, the effective area of the collecting optics (A_{eff}), and the effective focal length (efl).

In the case of several systems , the energy is focused on the field stop so that the beam can be directed to more than one detector. For such systems, the area of the field stop should be substituted for A_d.

ATTIRE Modeling of the Optics

The main menu of ATTIRE contains an icon for the Optics. On clicking this icon, the user encounters a window that contains the optics parameters. Fig. 6 shows the menu as it appears on the screen. These are the diameter of the entrance aperture, the diameter of the obscuration, the effective focal length, and the optical transmittance.

The diameters are modified by entering into the appropriate menu items. The effective focal length is modified through its menu item. This value is also dependent on the area of the detector and the IFOV of the system. For all the flux to be incident on the detector, the following relationship must hold.

$$\alpha^2 = \frac{A_d}{efl^2}$$

where α $= \text{IFOV}$ (7)

Hence, by changing the focal length, the program automatically changes the detector area according to the above equation.

DETECTORS

For a background noise limited infrared photoconductor (BLIP), the theoretical spectral D* is given by

$$D^*(\lambda) = \frac{\lambda}{2hc \sin \theta_{1/2}} \sqrt{\frac{\eta}{M_p^{BB}\ (T_B)}} \ \text{cm Hz}^{1/2}\text{W}^{-1}$$

where

η = Quantum efficiency of the detector

$\theta_{1/2}$ = Cold shield half angle (radians)

M_p^{BB} = Photon background noise flux

= $\int_{pk}^{\lambda} M_{p,\lambda}^{BB} (\lambda, T_B)\, d\lambda$

T_B = Background temperature (^{o}K)

(8)

The peak D* of the detector is obtained by substituting for peak wavelength in the above equation. The D* curve for the detector is now created by linearly extrapolating D*( ) up to the peak wavelength.

$$D^* (\lambda) = \frac{\lambda}{\lambda_{pk}} D^* (\lambda_{pk}) \text{ cm Hz}^{1/2} \text{ W}^{-1} \tag{9}$$

D** is D* normalized to cold shield half angle and is given by:

$$D^{**} = D^* \sin \theta_{1/2} \tag{10}$$

ATTIRE modeling of the Electro-Optics

The main menu of ATTIRE contains an icon for the detectors. On clicking this icon the user enters a screen that contains a window for the parameters that make up the detector sub-system. Fig. 7 shows the menu as it appears on the screen. There is an accompanying icon for each parameter which allows the user to change their values. The user can use the theoretical model to simulate the peak D*. Also, a user-defined peak D* can be input. If a peak D* is input, the program automatically adjusts the other parameters, e.g., Quantum Efficiency, to provide the input D* peak value. The D* spectral curve is created by drawing a straight line from λ=0 (where D*=0) to λ_{pk} (where D^*_{pk} is given). The detector area does not change independently. As discussed, it is dependent on the effective focal length of the optics and the IFOV of the system (Eqn. 7).

PIXEL PARAMETERS

This sub-system models the the pixel geometry as defined by the sensor motion and system spatial IFOV. In the case of airborne scanners, the sensor is mounted on an aircraft, whose velocity and altitude of flight need to be controlled for contiguous scanning.

The pixel size is a function of the altitude (a) and IFOV, and can be computed by

$$\text{Pixel Size} = 2\ \ a \tan\left(\frac{\alpha}{2}\right) \qquad (11)$$

where α = angular IFOV
a = altitude of aircraft in meters

Pixel size can also be determined using the small angle approximation by:

$$\text{Pixel size} = (\text{IFOV})\ (a)$$

For contiguous scanning, the time required for one revolution of the scan mirror is equal to the time required for the aircraft to travel forward a distance equal to one IFOV on the ground. Using the small angle approximation for pixel size

$$\frac{1}{n} = \frac{\alpha a}{v} \qquad (12)$$

where n = scan speed in rps
v = aircraft velocity in meters/sec

ATTIRE Modeling of the Spatial Parameters

The spatial parameters are modeled by entering the menu entitled 'PIXEL PARAMETERS'. Fig. 8 shows the menu as it appears on the screen. Since these parameters are interrelated, changing one may also affect the other. This menu contains the following items:

SCAN-SPEED-HT, SCAN-SPEED-VEL, DWELL-TIME, PIXEL-SIZE, and IFOV.

The scan speed is a function of both altitude and velocity (Eqn. 12).

The dwell time is a function of the scan speed, which in turn is a function of the altitude and velocity. In ATTIRE, changing the dwell time affects the scan speed and the velocity only.

$$t_d = \frac{\alpha}{2\pi n}$$

$$= \frac{\alpha^2}{2\pi\,(v/a)}\ \text{seconds} \qquad (13)$$

The Pixel menu contains the IFOV, pixel size, altitude, scan speed, velocity and dwell time. Several of these parameters are inter-dependent. Therefore changing one might affect some of the others as well.

ELECTRONICS

The signal conditioning and preamplifier electronics are modeled for their bandwidth and noise properties. The aim of the sensor design is to be detector-noise limited. In practice, some noise is added by the electronics. Hence the total noise of the system is represented as follows:

$$\text{Noise}_{\text{total}} = \text{Noise}_{\text{detector}} + \text{Noise}_{\text{electronics}}$$

The electronics noise is input as a noise factor. This is defined as the ratio of the total system noise to the noise of the system prior to the electronics. This makes it simple to define the noise as a multiple of the rest of the noise. Also, it indicates the contribution of the electronics to the total system noise. The detector noise is accounted for in the model where the D* is calculated assuming certain background noise.

The noise of the electronics is defined as follows

$$n_f = \frac{\text{(SNR) input}}{\text{(SNR) output}} \qquad (14)$$

The electronics also determines the sampling interval required to sample the analog output of the amplifiers. In the case of aircraft scanners, as stated earlier, the scan speed 'n', the IFOV 'α', the plane velocity 'v' and the altitude 'a' are related as follows:

$$t_d = \frac{\alpha}{2\pi n}$$

$$= \frac{\alpha^2}{2\pi\,(v/a)} \text{ seconds} \qquad (13)$$

The sampling period is defined by the dwell time of the sensor i.e the time the system spends at each pixel. In the case of a scanner this is given by

The dwell time determines the smallest spatial sampling interval. It is the smallest time interval between which any occurring change can be detected, i.e., any change faster than td will not be detected. It is the time interval between two successive pixels. The spatial sampling frequency is the inverse of the dwell time. This means, based on the Nyquist criterion, the largest spatial frequency component present in the signal is one half of the spatial sampling frequency. Thus in order to collect this signal, the bandwidth of the signal conditioning electronics must, at most, be half the spatial sampling frequency.

$$\text{Dwell Time} = \frac{1}{\text{Spatial Sampling frequency}}$$

$$\text{Bandwidth} = \frac{1}{2}\,\text{Spatial Sampling frequency.}$$

$$\Delta f = \frac{\pi n}{\alpha} \qquad (15)$$

ATTIRE modeling of the Electronics

The user enters into the Electronics sub-system by clicking on the icon in the Main Menu. The noise factor and the Electronics Bandwidth can be modified independently for each channel. Fig. 9 shows the menu as it appears on the screen.

COMPUTING THE BAND-PASS FLUX

The spectral radiance incident on the detector (assuming no path radiance) is given by:

$$L_{e,\lambda}\,(\lambda,T) = \frac{C_1}{\pi}\,\frac{\varepsilon(\lambda)\,\tau_a(\lambda)\,\tau_o(\lambda)}{\lambda^5\,(\exp\,(C_2/\lambda T) - 1)}\ \text{W cm}^{-2}\,\text{sr}^{-1} \qquad (16)$$

The solid angle subtended by the source at the entrance aperture is

$$\frac{A_s}{a^2} = \Omega_s \text{ steradians}$$

$$A_{pix} = \text{Area of the pixel}$$

$$= (\alpha a)^2 \qquad (17)$$

The total energy incident on the detector is given by

$$M_{(e,\pi)}\,(\pi,T) = \int_{\lambda_1}^{\lambda_2} M_{e,\lambda}\,(\lambda,T)\,d\lambda = \int_{\lambda_1}^{\lambda_2} \Omega_s\, L_{e,\lambda}\,(\lambda,T)\,d\lambda\ \text{W cm}^{-2} \qquad (18)$$

where λ_1 and λ_2 are the lower and upper wavelengths of the channel.

For cases where the bandwidth of the channel is defined by a relative spectral responsivity curve, the total energy is given by:

$$\int_{ch\#} R(\lambda)\left(\frac{\Omega_s}{\pi}\right)\frac{C_1\,\varepsilon(\lambda)\,\tau_s(\lambda)\,\tau_o\,(\lambda)}{\lambda^5(\exp\,(C_2/\lambda T) - 1)}\,d\lambda\ \text{W cm}^{-2} \qquad (19)$$

where $R(\lambda)$ is the relative responsivity curve.

Power Incident on the Detector

The power incident on the detector is obtained by integrating the radiance exiting the imaging lens over the channel bandpass and multiplying the integral by the throughput. This is given by

$$P_d = \left(\int_{\Delta\lambda} \varepsilon\,(\lambda)\; \tau_a\,(\lambda)\; \tau_o\,(\lambda)\; L_\lambda(\lambda,T)\; d\lambda \right) (\gamma_d)$$

where γ_d, detector throughput is given by

$$= A_d\,\Omega_d$$

where

A_d = Area of the detector

$$\Omega_d = \text{Solid angle subtended by the detector at lens}$$
$$= \frac{\text{Area of lens}}{(\text{efl}^2)} \qquad (20)$$

Rewriting using the above relationships, we obtain

$$P_d = \left(\int_{\Delta\lambda} \varepsilon\,(\lambda)\; \tau_a\,(\lambda)\; t_o\,(\lambda)\; L_\lambda\,(\lambda,T)\; d\lambda \right) (A_d)\,(\Omega_d) \qquad (21)$$

Noise Equivalent Detector Power

The noise equivalent power (NEP) of the detector is defined as the power incident on the detector such that the signal-to-noise ratio is unity. The NEP can also be obtained if the broad-band D* is available.

In this study, the D* at peak wavelength is modeled. The broad-band D* can be obtained from the peak spectral D* as follows:

$$D^*_{BB}\,(T) = \frac{\int_0^\infty D^*\,(\lambda)\; L_{e,\lambda}\,(\lambda,T)\; d\lambda}{\int_0^\infty L_{e,\lambda}\,(\lambda,T)\; d\lambda} \qquad (22)$$

The NEP is then defined as

$$\text{NEP(T)} = \frac{\sqrt{A_d\,\Delta f}}{D^*_{BB}\,(T)}$$

where Δf = electronic bandwidth (23)

Signal-to-Noise Ratio

The signal-to-noise ratio (SNR) of the detector can be obtained as follows:

$$SNR_d = \frac{P_d}{NEP} \tag{24}$$

Noise-Equivalent Radiance

The noise equivalent radiance (NER) denotes the required bandpass radiance exiting from the target such that the SNR is equal to unity of the system.

$$NER = \frac{\text{Bandpass radiance exiting the target}}{SNR}$$

$$= \frac{\int_{\Delta\lambda} L_\lambda(\lambda, T)\, d\lambda}{SNR} \tag{25}$$

Noise-Equivalent Temperature Difference

The radiometric performance of a thermal sensor is determined by its NEΔT. This is the ability of the sensor to theoretically discriminate between two temperature values. It is defined as the temperature difference required to make the signal-to-noise ratio of the sensor unity.

$$NE\Delta T = \frac{\Delta T}{SNR} = \frac{\partial T}{\partial\left(\frac{V_s}{V_n}\right)} = V_n \frac{\partial T}{\partial V_s}$$

where V_s and V_n are the signal and noise voltages respectively. (26)

$$v_s = (\text{solid angle}) \int_{\lambda_1}^{\lambda_2} \mathcal{R}(\lambda) L_\lambda(\lambda,\ T)\, d\lambda$$

$$\mathcal{R}(\lambda) = \text{Spectral Resposivity} = \frac{v_n}{\sqrt{A_d \Delta f}} D^*(\lambda) \tag{27}$$

Expanding the expression and substituting the appropriate parameters

$$NE\Delta T = \frac{\lambda_{pk}}{D^*\ (\lambda_{pk})}\, n_f \sqrt{\frac{\Delta f}{A_d}} \left(\frac{efl^2}{A_{eff}}\right) \frac{1}{(\text{Integral})} \tag{28}$$

where Integral =

$$\int_{\lambda 1}^{\lambda 2} \tau_a(\lambda)\, \tau_o(\lambda)\ \varepsilon(\lambda)\, \lambda \frac{d\, L_{\lambda}^{BB}(\lambda,T)}{dT}\, d\lambda \qquad (29)$$

GRAPHICS

The program provides extensive graphics capabilities for viewing data parameters. The graphing routines are divided into two categories - PLOT and GRAPH. The GRAPH feature is used exclusively for spectral graphs, i.e., the x-axis is always the wavelength. In the main menu, the GRAPH feature allows for plotting of either the Energy or the Photon Flux versus wavelength. The flux can be plotted for either of the existing channels or for any user-defined set of parameters. The program prompts the user to select one of these two options. If the user desires a spectral plot of a channel, only the channel number is required. In the case of a user-defined graph, the user is prompted for the type of flux (thermal/solar), the temperature (if thermal), the emissivity/albedo, and the FOV.

The graph feature in the FLUX, OPTICS, and ATMOSPHERICS sub-systems are used to graph the spectral emissivity, the optical and atmospheric transmittance respectively. These can be variable or constant depending on the channel. Each of these three sub-systems also allow for an option in their GRAPH Menu to graph all these three spectrally variable parameters on the same graph.

The plot item in the main menu is used to plot non-spectral parameters against each other. One such example is to compute the SNR of the sensor as it varies with the temperature of the source. The user is prompted for the bandwidth of the sensor. All other parameters that enter into computing the SNR are used from a thermal channel. The user is also prompted for that thermal channel number. Next, the range of temperatures over which the data needs to be plotted is asked for.

The other significant feature of the PLOT item is the User-Defined feature. This allows the user to select either of the sensor parameters as it x and y axes. Then for all the existing channels of the sensors the plot is created.

The graphs appear in a resizeable windows form. One window displays the graph and the other window the contents of the array that are being graphed. So the user can not only view the graph, but also view the actual values being plotted. The graphs can be resized by pressing the mouse or the edges and sliding it to the required window size. The top right button of the graphics window is used to make the graphics window fill the entire screen. This can also be done by typing the ‘f’ key (for full screen). On clicking the top right button of the array values window, the user can quit the graphics environment. This can also be done by typing the ‘q’ key. The array values window can be scrolled up or down to view different parts of the array.

The main menu additionally contains an item for Graph menu to plot the energy or photon flux of all the channels. It also allows the user to plot all the spectrally variable parameters - which might include emissivity and atmospheric & optical transmission for a channel - on the same screen. This item allows for obtaining graphs of spectrally varying parameters. Other graphs can be obtained from the plot menu.

An additional option exists for the user to be able to obtain a hardcopy of these graphs on an HP Laserjet II printer. Fig. 10 shows one such graph. These figures are the screen outputs of ATTIRE. This is the result of plotting the energy flux radiated at various temperatures within two prominent thermal bands - 3-5 and 8-12 microns. At low temperatures, the 8-12 channel has more energy.

However as the temperature of the target increases, the 3-5 channel has more energy. Thus to monitor high temperatures (specifically at temperatures greater than 582K), e.g., volcanoes and forest fires, it is better to use 3-5 micron channel and for low temperatures, e.g., normal earth temperatures, at 300 K the 8-12 micron channel gives a better performance.

PLOT Y vs X

The package can provide plots of system performance versus any system parameter. A parameter X e.g the entrance aperture of the optics, could be varied and its effect on another parameter Y e.g NETD, can be plotted. Another parameter Z e.g the obscuration, could be varied for each plot and several plots obtained on the same graph. The menu for such " Y vs X plots for different values of Z " contains various options in each sub-system. Fig. 11 shows the result of plotting the variation in NETD with temperature for three different thermal channels.

SPECTRALLY VARIABLE PARAMETERS

The emissivity, atmospheric transmission and the optical transmission default as constants over the specified spectral bandwidths. These could also be made variable. Indeed, in typical situations these are spectrally variable parameters. Before entering either of these parameters from their respective menus, the program asks the user if they are variable or a constant. If they are a constant, the number as a fraction of 1 is entered in the box. If it is a variable, the box shows the letters VAR. The software then picks up the spectrally variable curves from the disk and defaults the bandwidths for the channels to the corresponding bandwidth and resolution. In the case of atmospheric transmission, for each channel the curve should be stored in ASCII format as a1.dat, a2.dat, ... a6.dat. Similarly for optical transmission the first letter should be 'o' (o1.dat ..) and for the emissivity it should be 'e' (e1.dat ..). The first two lines of the data files are ignored by the reading routine. These shall be for the user to describe the curve for their purposes. The data shall follow from the third line onwards. Each line contains the wavelength (in microns) and the corresponding parameter (in %). Thus the reading routine picks up the values for the wavelength and the parameter, one at a time from each line until it reaches the end of the file.

Interpolation and Projection

If more than one of the above mentioned three parameters are spectrally variable, then it is possible that the upper and lower wavelength values as well as the resolution of the wavelength array may not be the same for any two out of the three arrays. For example, the emissivity data may be available from 3.0 to 5.0 microns in intervals of .1 micron whereas the optical transmission data may be available from 2.2 to 4.85 microns in intervals of .05 microns. ATTIRE contains a methods for resolving such ambiguities.

The package first checks for these discrepancies before deciding on the least common denominator for the resolution and selecting the lower and upper limits for wavelength such that all the existing points available can be plotted as they were obtained. However doing this results in several gaps in the arrays, e.g in the above example, a resolution of .05 microns and a bandwidth of 2.2 to 5.0 microns is chosen. This leaves blanks in the emissivity array at 2.2 - 3.0 microns as well as at 3.05, 3.15 ... microns.

The package performs an interpolation on the data to smooth out the curves and then projects them onto a common wavelength array for all the parameters. The emissivity from 2.2 to 3 microns is not assumed to be zero throughout but starting from the value at 3.0 microns gradually tends to zero at 2.2 microns. Similarly the emissivity at 3.05 microns is the average of the emissivity at 3.0 and 3.1 microns.

SUMMARY

The design of a visible through thermal IR sensor system requires a detailed analysis of how the input signal propagates through the system. The major components in developing a model for the sensor system are the **source**, **atmosphere**, **optics**, **detector**, **spatial parameters**, and **preamplifier electronics**. The final goal of the analysis is to determine the NER for the various spectral channels of a sensor system.

In this paper, a simulation package "ATTIRE", for analyzing sensor systems was introduced. The package runs in a PC-DOS environment and consists of one executable program and several supporting files. The entire package fits on one high-density floppy disk.

ATTIRE is a useful tool for performing design trade offs as it inter relates several aspects of the sensor system to yield performance parameters. It is also useful as a tool for the understanding of the concepts of radiometry.

REFERENCES

1. M. P. Thekaekara, Kruger, and R. Duncan, "Solar Irradiance measurements from a research aircraft," *Applied Optics*, Vol. 8 No. 8, pp. 1713-1732.

2. P. Moon, "Proposed standard Solar-radiation curves for engineering use," *J. Franklin Inst*, Vol. 230 No. 5, pp. 583-617, 1940.

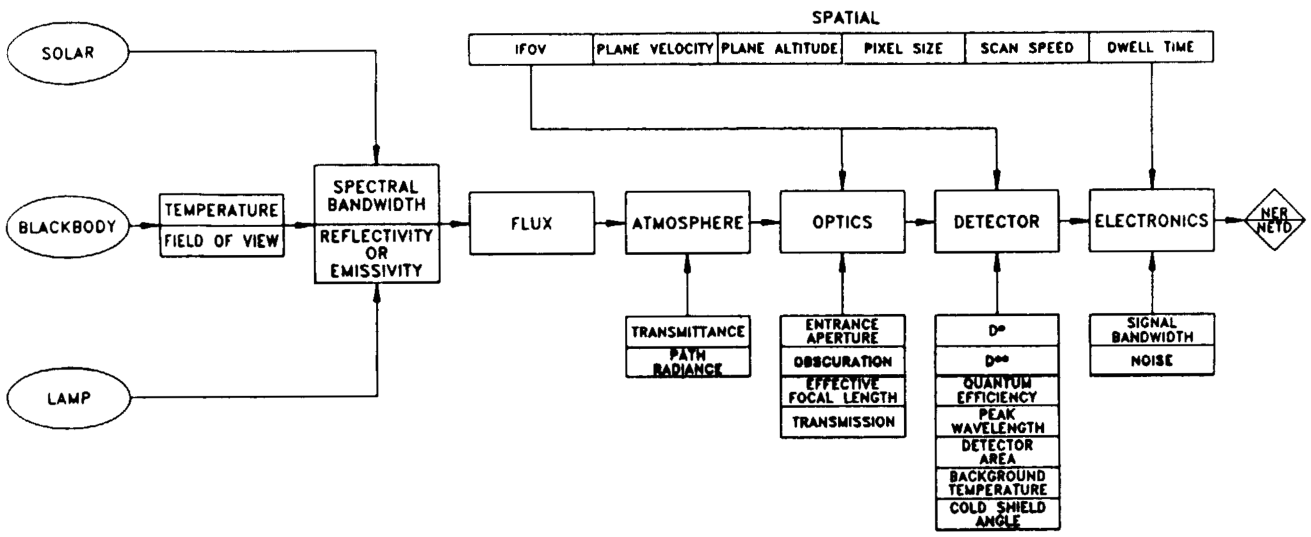
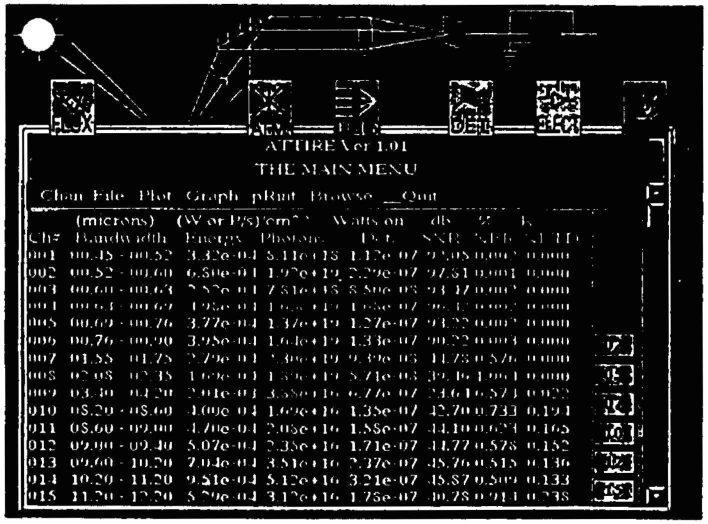


Fig. 1 Flow of energy through a sensor in ATTIRE



```
                    ATTIRE Ver 1.01
                    THE MAIN MENU
 Chan  File  Plot  Graph  pRint  Browse  _Quit
      (microns)    (W or P/s)/cm^2    Watts on    db    %    K
Ch#   Bandwidth    Energy   Photons    Det.      SNR   NEF   NETD
001   00.45 - 00.52  3.32e-04 5.11e+18 1.17e-07 95.05 0.002 0.000
002   00.52 - 00.60  6.80e-04 1.97e+19 2.29e-07 97.61 0.001 0.000
003   00.60 - 00.63  2.52e-04 7.81e+18 8.50e-08 93.47 0.002 0.000
004   00.63 - 00.69  [illegible] [illegible] [illegible] [illegible] [illegible] 0.000
005   00.69 - 00.76  3.77e-04 1.37e+19 1.27e-07 93.22 0.002 0.000
006   00.76 - 00.90  3.95e-04 1.64e+19 1.33e-07 90.22 0.003 0.000
007   01.55 - 01.75  2.79e-04 2.30e+19 9.39e-08 11.78 0.576 0.000
008   02.08 - 02.35  1.69e-04 1.89e+19 5.71e-08 39.16 1.061 0.000
009   03.40 - 04.20  2.01e-03 3.58e+16 6.77e-07 24.61 0.571 0.022
010   08.20 - 08.60  4.00e-04 1.69e+16 1.35e-07 42.70 0.733 0.191
011   08.60 - 09.00  4.70e-04 2.05e+16 1.58e-07 44.10 0.623 0.165
012   09.00 - 09.40  5.07e-04 2.35e+16 1.71e-07 44.77 0.578 0.152
013   09.60 - 10.20  7.04e-04 3.51e+16 2.37e-07 45.76 0.515 0.136
014   10.20 - 11.20  9.51e-04 5.12e+16 3.21e-07 45.87 0.509 0.133
015   11.20 - 12.20  5.29e-04 3.12e+16 1.78e-07 40.78 0.914 0.238
```

Fig. 2 Main menu of ATTIRE

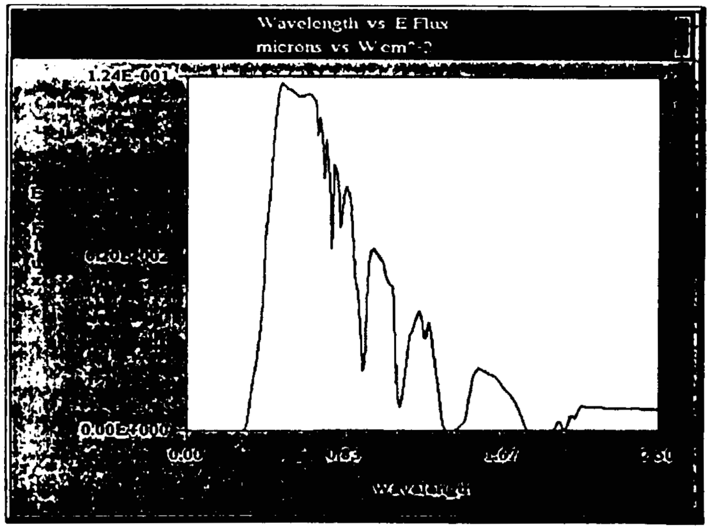
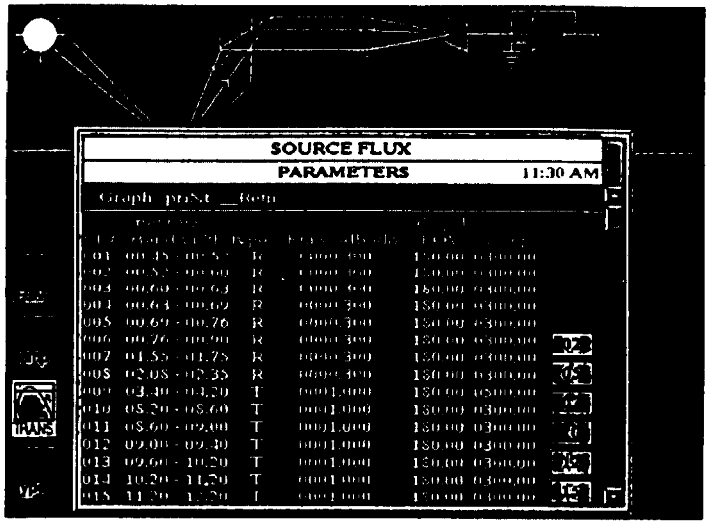
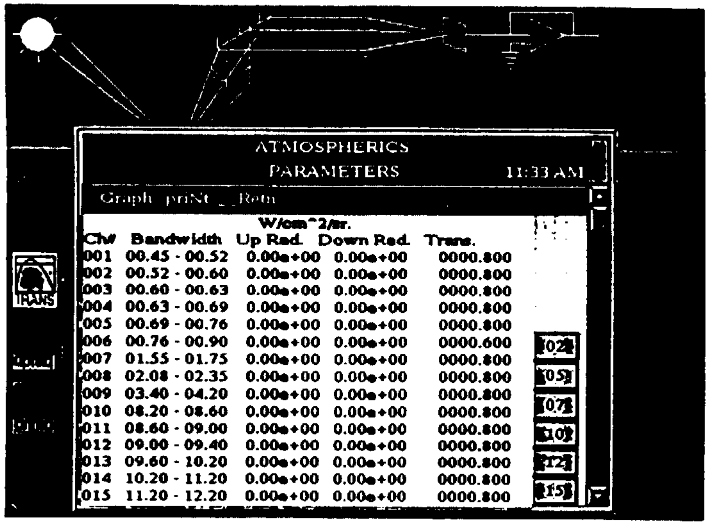


Fig. 3 Solar curve used in ATTIRE



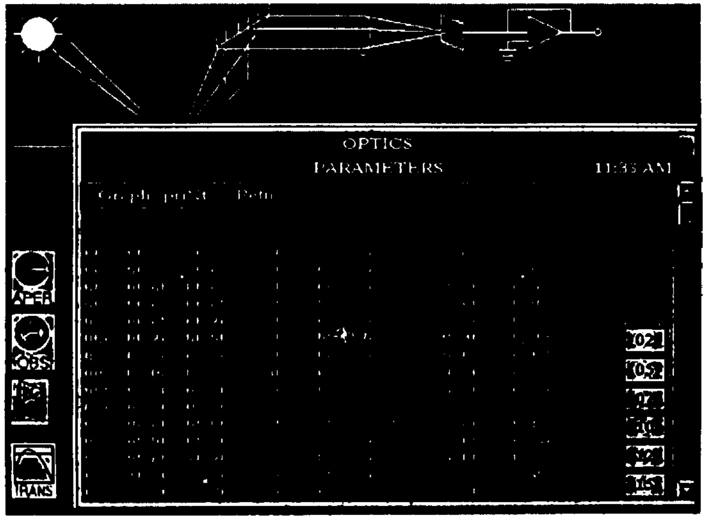
```
SOURCE FLUX
PARAMETERS                         11:30 AM
Graph priNt __Retn
001 00.45 - 00.52   R   0000.400   150.00 0300.00
002 00.52 - 00.60   R   0000.400   150.00 0300.00
003 00.60 - 00.63   R   0000.400   160.00 0300.00
004 00.63 - 00.69   R   0000.300   180.00 0300.00
005 00.69 - 00.76   R   0000.300   180.00 0300.00
006 00.76 - 00.90   R   0000.300   180.00 0300.00
007 01.55 - 01.75   R   0000.300   180.00 0300.00
008 02.08 - 02.35   R   0000.300   180.00 0300.00
009 03.40 - 04.20   T   0001.000   180.00 0500.00
010 08.20 - 08.60   T   0001.000   180.00 0300.00
011 08.60 - 09.00   T   0001.000   180.00 0300.00
012 09.00 - 09.40   T   0001.000   180.00 0300.00
013 09.60 - 10.20   T   0001.000   180.00 0300.00
014 10.20 - 11.20   T   0001.000   180.00 0300.00
015 11.20 - 12.20   T   0001.000   180.00 0300.00
```

Fig. 4 Source flux menu in ATTIRE



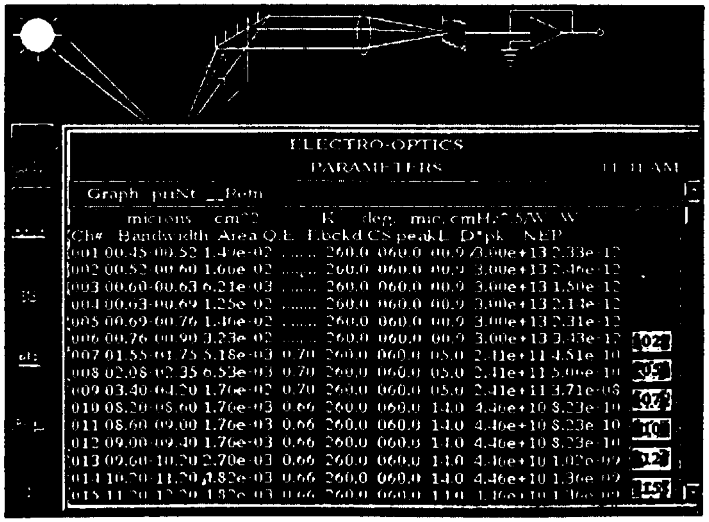
```
ATMOSPHERICS
PARAMETERS                              11:33 AM
Graph priNt __Retn
                          W/cm^2/sr.
Ch#  Bandwidth      Up Rad.    Down Rad.    Trans.
001  00.45 - 00.52  0.00e+00   0.00e+00     0000.800
002  00.52 - 00.60  0.00e+00   0.00e+00     0000.800
003  00.60 - 00.63  0.00e+00   0.00e+00     0000.800
004  00.63 - 00.69  0.00e+00   0.00e+00     0000.800
005  00.69 - 00.76  0.00e+00   0.00e+00     0000.800
006  00.76 - 00.90  0.00e+00   0.00e+00     0000.600
007  01.55 - 01.75  0.00e+00   0.00e+00     0000.800
008  02.08 - 02.35  0.00e+00   0.00e+00     0000.800
009  03.40 - 04.20  0.00e+00   0.00e+00     0000.800
010  08.20 - 08.60  0.00e+00   0.00e+00     0000.800
011  08.60 - 09.00  0.00e+00   0.00e+00     0000.800
012  09.00 - 09.40  0.00e+00   0.00e+00     0000.800
013  09.60 - 10.20  0.00e+00   0.00e+00     0000.800
014  10.20 - 11.20  0.00e+00   0.00e+00     0000.800
015  11.20 - 12.20  0.00e+00   0.00e+00     0000.800
```

Fig. 5 Atmosphere menu in ATTIRE



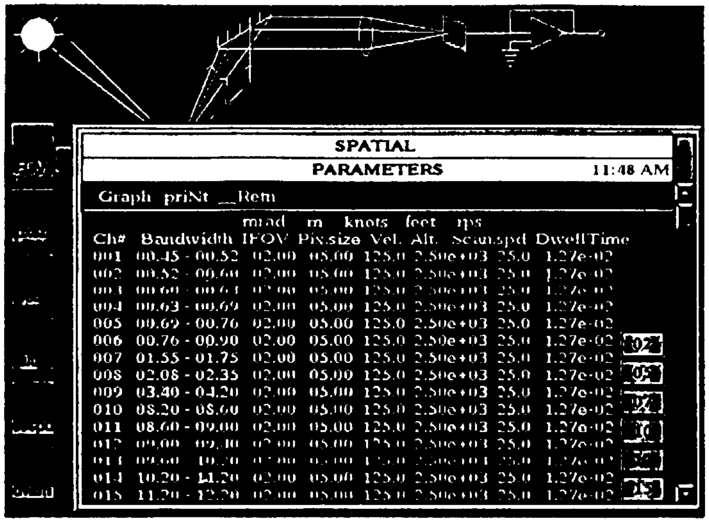
```
OPTICS
PARAMETERS                         11:33 AM
Graph priNt __Retn
```

Fig. 6 Optics menu in ATTIRE



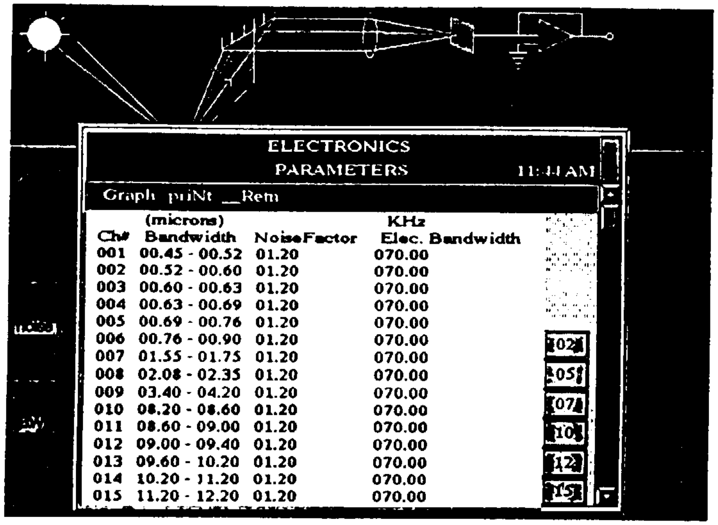
```
ELECTRO-OPTICS
PARAMETERS                                        [illegible]
Graph priNt __Retn
     microns     cm^2          K     deg.  mic. cmHz^.5/W  W
Ch#  Bandwidth   Area    Q.E.  T.bckd CS   peakL D*pk      NEP
001 00.45-00.52 1.47e-02 ...... 260.0 060.0 00.9 3.00e+13 2.33e-12
002 00.52-00.60 1.66e-02 ...... 260.0 060.0 00.9 3.00e+13 2.46e-12
003 00.60-00.63 6.21e-03 ...... 260.0 060.0 00.9 3.00e+13 1.50e-12
004 00.63-00.69 1.25e-02 ...... 260.0 060.0 00.9 3.00e+13 2.14e-12
005 00.69-00.76 1.46e-02 ...... 260.0 060.0 00.9 3.00e+13 2.31e-12
006 00.76-00.90 3.23e-02 ...... 260.0 060.0 00.9 3.00e+13 3.43e-12
007 01.55-01.75 5.18e-03 0.70 260.0 060.0 05.0 2.41e+11 4.51e-10
008 02.08-02.35 6.53e-03 0.70 260.0 060.0 05.0 2.41e+11 5.06e-10
009 03.40-04.20 1.76e-02 0.70 260.0 060.0 05.0 2.41e+11 3.71e-08
010 08.20-08.60 1.76e-03 0.66 260.0 060.0 14.0 4.46e+10 8.23e-10
011 08.60-09.00 1.76e-03 0.66 260.0 060.0 14.0 4.46e+10 8.23e-10
012 09.00-09.40 1.76e-03 0.66 260.0 060.0 14.0 4.46e+10 8.23e-10
013 09.60-10.20 2.70e-03 0.66 260.0 060.0 14.0 4.46e+10 1.02e-09
014 10.20-11.20 1.82e-03 0.66 260.0 060.0 14.0 4.46e+10 1.30e-09
015 11.20-12.20 1.82e-03 0.66 260.0 060.0 14.0 4.46e+10 1.30e-09
```

Fig. 7 Electro-optical menu in ATTIRE



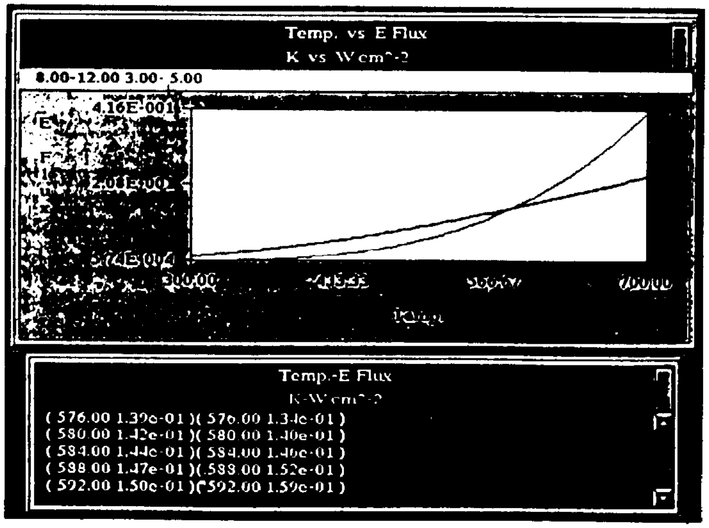
```
                              SPATIAL
                              PARAMETERS                    11:48 AM
Graph  priNt  __Rem
                   mrad    m   knots  feet   rps
Ch#  Bandwidth  IFOV  Pix.size  Vel.  Alt.  Scanspd  DwellTime
001  00.45 - 00.52  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
002  00.52 - 00.60  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
003  00.60 - 00.63  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
004  00.63 - 00.69  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
005  00.69 - 00.76  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
006  00.76 - 00.90  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
007  01.55 - 01.75  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
008  02.08 - 02.35  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
009  03.40 - 04.20  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
010  08.20 - 08.60  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
011  08.60 - 09.00  02.00  05.00  125.0  2.50e+03  25.0  1.27a-02
012  09.00 - 09.40  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
013  09.60 - 10.20  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
014  10.20 - 11.20  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
015  11.20 - 12.20  02.00  05.00  125.0  2.50e+03  25.0  1.27e-02
```

Fig. 8 Spatial parameters menu in ATTIRE



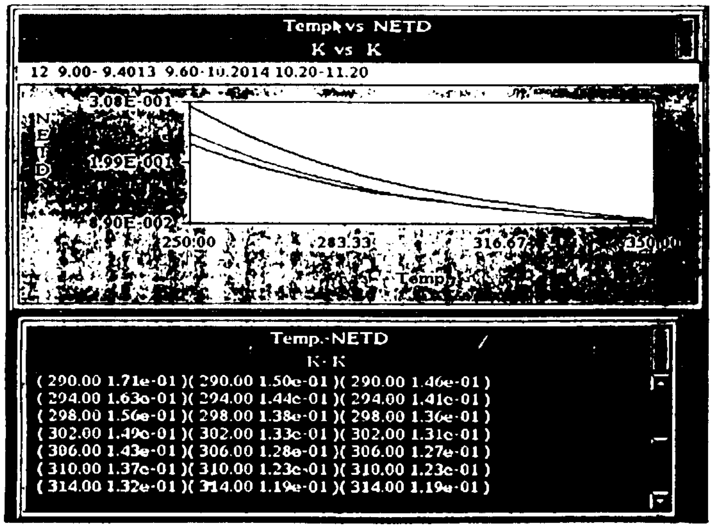
```
                         ELECTRONICS
                         PARAMETERS                11:44 AM
Graph  priNt  __Rem
      (microns)                         KHz
Ch#   Bandwidth      NoiseFactor    Elec. Bandwidth
001   00.45 - 00.52  01.20          070.00
002   00.52 - 00.60  01.20          070.00
003   00.60 - 00.63  01.20          070.00
004   00.63 - 00.69  01.20          070.00
005   00.69 - 00.76  01.20          070.00
006   00.76 - 00.90  01.20          070.00
007   01.55 - 01.75  01.20          070.00
008   02.08 - 02.35  01.20          070.00
009   03.40 - 04.20  01.20          070.00
010   08.20 - 08.60  01.20          070.00
011   08.60 - 09.00  01.20          070.00
012   09.00 - 09.40  01.20          070.00
013   09.60 - 10.20  01.20          070.00
014   10.20 - 11.20  01.20          070.00
015   11.20 - 12.20  01.20          070.00
```

Fig. 9 Electronics menu in ATTIRE



```
               Temp.-E Flux
                 K-W cm^-2
( 576.00 1.39e-01 )( 576.00 1.34e-01 )
( 580.00 1.42e-01 )( 580.00 1.40e-01 )
( 584.00 1.44e-01 )( 584.00 1.46e-01 )
( 588.00 1.47e-01 )(.588.00 1.52e-01 )
( 592.00 1.50e-01 )(592.00 1.59e-01 )
```

Fig. 10 Energy flux vs temperature for 2 channels



```
                    Temp.-NETD
                       K-K
( 290.00 1.71e-01 )( 290.00 1.50e-01 )( 290.00 1.46e-01 )
( 294.00 1.63e-01 )( 294.00 1.44e-01 )( 294.00 1.41e-01 )
( 298.00 1.56e-01 )( 298.00 1.38e-01 )( 298.00 1.36e-01 )
( 302.00 1.49e-01 )( 302.00 1.33e-01 )( 302.00 1.31e-01 )
( 306.00 1.43e-01 )( 306.00 1.28e-01 )( 306.00 1.27e-01 )
( 310.00 1.37e-01 )( 310.00 1.23e-01 )( 310.00 1.23e-01 )
( 314.00 1.32e-01 )( 314.00 1.19e-01 )( 314.00 1.19e-01 )
```

Fig. 11 Plots for variation of NETD vs temperature

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
February 1993

3. REPORT TYPE AND DATES COVERED
CONFERENCE PUBLICATION

4. TITLE AND SUBTITLE
Technology 2002
Volume 2

5. FUNDING NUMBERS

6. AUTHOR(S)
MICHAEL HACKETT, COMPILER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NASA TECHNOLOGY TRANSFER PROGRAM (CODE CU)

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-3189, VOL. II

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 2002 Conference and Exposition, December 1-3, 1992, Baltimore, MD. Volume 2 features 60 papers presented during 30 concurrent sessions.

14. SUBJECT TERMS
technology transfer, computer technology, advanced manufacturing, materials science, biotechnology, electronics

15. NUMBER OF PAGES
524

16. PRICE CODE
A22

17. SECURITY CLASSIFICATION OF REPORT UNCLASS	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASS	19.SECURITY CLASSIFICATION OF ABSTRACT UNCLASS	20. LIMITATION OF ABSTRACT UNLIMITED

*U.S. GOVERNMENT PRINTING OFFICE: 1993-728-150/60021

National Aeronautics and
Space Administration
Code JTT
Washington, D.C.
20546-0001