REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Service, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
25-03-2003	Final Technical Report	9/9/01 – 12/31/02

4. TITLE AND SUBTITLE

The California Central Coast Research Partnership: Building Relationships and Partnership for University Industry Research Collaboration

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-01-1-1049

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Opava, Susan C.; Abler, Fred F.; Bremer, Walt; Smith, Hugh M.; Nico, Phillip; Taufik; Thorncroft, Glen E.; Pascual, Christopher C.; Ahlgren, William L.; DeTurris, Dianne; Chen, Katherine C.; Biezad, Daniel J.; Colvin, Kurt; DePiero, Fred W.; Herter, Roberta J.; Jimenez-Flores, Rafael; Nelson, Yarrow M.; Walsh, Daniel W.; Pohl, Jens G.; Gollery, Steven J.; Saghri, John A.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
California Polytechnic State University Foundation
Sponsored Programs Department
Foundation Administration Bldg. 15
San Luis Obispo, CA 93407

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
George W. Solhan
Office of Naval Research
Ballston Centre Tower One
800 North Quincy Street
Arlington, VA 22217-5660

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release

20030331 057

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The primary focus of this initiative is to forge a strong link between private sector R&D and University applied research, to speed the development of new knowledge and the transfer of technology to the public and private sectors. To this end, communications infrastructure and new R&D facilities have been developed. Relationships with private companies engaged in R&D have been advanced. Important research has been carried out in areas of interest to the Department of Defense and national security. These areas include, computer hardware and software development, aerospace engineering, military field applications, remote sensing, data compression, expert systems, disaster management, photovoltaics and biofilms.

15. SUBJECT TERMS

VLSI systems; aerospace; materiel; computer networks; biofilms; remotes sensing; data compression; expert systems; disaster management; photovoltaics.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Susan C. Opava
U	U	UU	UU	350	19b. TELEPONE NUMBER *(Include area code)* 805-756-1508

The California Central Coast Research Partnership: Building Relationships, Partnerships and Paradigms for University-Industry Research Collaboration.

FINAL REPORT ON ONR GRANT NO. N00014-01-1-1049
PERIOD OF PERFORMANCE: 9/9/01 to 12/31/02

March 25, 2003

Principal Investigator:

Susan C. Opava, Ph.D.
Dean of Research and Graduate Programs
California Polytechnic State University
San Luis Obispo, California

TABLE OF CONTENTS

I. Title of Project 1
II. Summary of Project 1
III. Relevance to ONR Objectives 1-5
A. Relevant partners 1-2
B. Strategic locations 2
C. Relevant R&D focus 2-4
D. University-identity-government partnership 4
E. University strengths 4-5
IV. Summary of Results During the Period of Performance 5-358
A. General 5-6
B. Information Technology Infrastructure, R&D Facilities and Partnership Projects 7-100
1. Development of Information Technology Infrastructure. 7-38
2. New Research & Development Facilities and Partnerships 39-100
a. Computer Networking Research Laboratory. 39-70
b. Photovoltaic facility 71-90
c. Testing facility for semi-conductor processing technologies. 91-97
d. Geographic forecasting collaboration 98-100
3. Fiberoptic Asset Management. 8
(See also Appendix A)
C. Research Projects 101-350
- Design Methodologies for Analog/Mixed Signal VLSI Systems Applied to Infrared Focal Plane Arrays 102-109
- Development of an Autonomous Tactical Reconnaissance Platform 110-162
- Development of an Aerodynamic Table Lookup System and Landing Gear Model for the Cal Poly Flight Stimulator 163-220
- Development of Field Rechargeable Gas Mask Filters 221-226
- Exploitation of Network Bandwidth and the Ethernet/IP Application Layer Standard for Automation Networks 227-241
- Range Sensing and Real-Time Registration 242-259
- When Johnny's Dad Can't Read: Using Technology to Close the Adult Literacy Gap 260-273
- Correlation of Milk Composition and Fouling with Biofilm Formation and Microbial Spore Production in Heat Exchangers 274-289
- The Role of Discovery in Context-Building Decision-Support Systems 290-302
- The TEGRID Semantic Web Application: A Service-Oriented Distributed Approach to Disaster Management Decision-Support Systems 303-316
- Viable Bandwidth Compression for Remote Sensing Applications 317-350

Appendix A. Telecommunication Asset Management in a Global Environment

I. Title of Project

The California Central Coast Research Partnership: Building Relationships, Partnerships and Paradigms for University-Industry Research Collaboration.

II. Summary of Project

The mission of the California Central Coast Research Partnership (C^3RP) is to facilitate the exchange of knowledge and skills between the higher education sector and the private sector in San Luis Obispo County, and to encourage the growth of high-tech companies in the region, thereby enhancing economic development and quality of life. The partnership is a long-term plan to create a dynamic and self-supporting university-industry-government partnership that capitalizes on the strengths and mutual interests of the educational and technology-based business sectors. The plan recognizes the key role of higher education in preparing a highly skilled work force and transferring new knowledge to practical uses. The outcomes of this partnership, when fully realized, will be the expansion of current and the creation of new University technology R&D activities; the development of existing technology-based businesses and the creation of new ones; and the generation of opportunities for job training and research and development activities for University and Community College students and faculty - **all in areas of interest to the Department of Defense and national security**.

The project will eventually lead to the construction (with private financing) of a technology park on the California Polytechnic State University campus that will provide state of the art space for private technology companies engaged in research and development activities, as well as a business incubator that will provide all of the support services needed by start-up, technology-based companies. At this stage, the project involves planning, analysis, relationship-building and pilot research projects related to development of the long-term partnership and its research foci.

III. Relevance to ONR Objectives

A. Relevant partners.

C^3RP represents a coalition of educational institutions, local, state and federal government, and private businesses that have worked together in unprecedented fashion to advance the common goals inherent in the proposed university-industry partnership. The current partners in the project and their contributions include:

California Polytechnic State University

- committed the land for the project, valued at ~$3 million
- provided assistance in financial management of the project
- contributed $90,000 for a pre-feasibility study by Bechtel Corporation

- committed several hundred thousand dollars of in-kind contributions of senior management time and effort over several years; continues to do so
- invested ~$700,000 in efforts to raise additional funds for the project

GEO, International (GEOgraphic Network Affiliates, International; a private company)

- works *pro bono* with C^3RP on communications aspects of the project

CENIC (Corporation for Education Network Initiatives; association of Internet2 universities in CA)

- works with Cal Poly and GEO to further the goals of the IEEAF (see below), goals that will directly benefit C^3RP

IEEAF (International Educational Equal Access Foundation; established by GEO and CENIC)

- is currently securing donations of virtual and physical communications assets in at least 37 countries; some of these fiber-optic assets will directly benefit this ONR project

National Science Foundation

- is working with GEO, IEEAF, CENIC and other universities in the United States to promote the goals of these organizations to develop low-cost fiber-optic networks for the benefit of educational institutions, non-profit organizations and local communities.

Efforts are underway to secure new partners, including:

- National Center for Research Resources of the National Institutes of Health
- National Guard Bureau (federal)
- California National Guard
- National Interagency Civil-Military Institute (federal)
- Governor's Office of Emergency Services of the State of California

B. Strategic location.

New technological developments in San Luis Obispo related to the intersection of major undersea and terrestrial fiber-optic cable networks on the central coast of California have provided an exceptional opportunity to focus the research partnership initially on the emerging technology area of global telecommunications. Linking the research partnership to this burgeoning field has enabled us to recruit many of the important partners listed above and will provide the strong affiliation between the educational and private sectors that is the foundation for success in university-related technology parks. It will also open new opportunities for federally supported research and development projects.

C. Relevant R&D focus.

Global telecommunications, particularly wireless, "over the horizon" communications, is fundamental to many current and developing defense strategies, including the ability to respond effectively to various forms of terrorist activities (including biological and chemical warfare), and in situations that require large-scale humanitarian assistance and disaster relief – areas which have become far more critical in the wake of the September 11, 2001 terrorist attack on American soil. It is also fundamental to the development of the decision-support systems that underlie and are key to these strategies. Cal Poly's Collaborative Agent Design (CAD) Research Center is the architect and lead developer of one of the first such systems: IMMACCS (Integrated Marine Multi-Agent Command and Control System), with JPL, SPAWAR Systems Center and the Stennis Space Center (NRL) as the other principal team members. Because of this, Cal Poly and the CAD Research Center are poised to take the lead in the continued development of complex decision-support systems and to develop a center of excellence in this area that will place Cal Poly at the forefront of the field. The C^3RP project has provided and will continue to provide support for the CAD Research Center to continue, and expand, its work on on-going projects like those completed for the MCWL, such as IMMACS, and the joint MCWL-ONR ELB (Expanded Littoral Battlefield), ACTD (Advanced Concept Technical Demonstration) project.

The scope of the C^3RP project, however, is much broader than simply the continuation of current projects of the kind described above. Command and control technology is developing rapidly and will have nearly universal applications. The basis of this technology is the ability to store "information" rather than "data", a concept that is fundamental to the capacity to utilize that information to support complex decision-making. IMMACS is one of the first of such systems to use this concept, and has led the field, but the bringing together of commercial and University research and development through C^3RP will lead to an explosion of applications and technological advances in this arena.

In addition to telecommunications, many other research areas have been targeted for development through C^3RP. These have been selected because Cal Poly has demonstrated or developing strengths in these areas. They include:

- Agricultural and environmental biotechnology
- Aerospace technologies
- Rendering, animation and modeling
- Software engineering
- GIS and GPS applications in agriculture, biology, architecture and engineering
- Network hardware and software technology
- Data processing
- Remote sensing
- Polymers and coatings
- Electroptics
- Power generation and distribution
- Biological risk assessment and detection
- Protection of the nation's food and water supplies
- Transportation disaster management and control
- Telecommunications technologies
- Photonics
- Graphic communication technologies
- Nano- and micro-technology
- Computer engineering (small systems, peripherals, custom applications)
- Surveillance, logistics, command & control support systems
- e-Learning software development
- Bio-engineering
- Robotics
- Magnetic levitation (transportation)
- Risk assessment and prediction of fire terrorism
- Environmental hazard assessment and detection
- Seismic research

The majority of these research areas mesh with strategic initiatives identified by the Office of Naval Research. Research and advanced development projects in these strategic areas are being developed, often in collaboration with industry partners identified through C^3RP and recruited to join the partnership.

D. University-industry-government partnership.

The primary focus of this initiative is to forge a strong link between private sector R&D and University applied research to speed the development of new knowledge and the transfer of technology to the public and private sectors. San Luis Obispo has become a draw for technology businesses (with a heavy concentration of software development companies) from both the LA Basin and Silicon Valley. For example, SRI (Stanford Research Institute), International has opened a "software center of excellence" in the city. Branches of major corporations are also located nearby, for example, Sun Microsystems, AutoDesk and Sunbay Software. Lockheed-Martin has a research and development group in nearby Santa Maria. At least one small local business is currently working on defense contracts. The company, Visual Purple, develops simulation environments with direct applications to anti-terrorist activities. Also located on the Central Coast are several government entities with whom we are working to advance the development of an Interdisciplinary Center of Excellence in research, education and training relevant to national security that will serve as a national resource. These agencies have been listed on p.2 and include NICI (the National Inter-agency Civil-Military Institute, as well as state entities whose mission includes anti-terrorism training/ response and response to large-scale emergency situations (California National Guard and the Governor's Office of Emergency Services).

E. University strengths.

Cal Poly is a State university that has achieved national distinction as a polytechnic university, with engineering and computer science programs ranked among the very best undergraduate programs in the country. It is moving with vigor into the 21st century and these characteristics have led it to orchestrate the research partnership effort and the consortium of partners proposed herein. Cal Poly also has affiliations with CSA (California Space Alliance) and with Vandenberg Air Force Base, where it has offered an M.S. in Aerospace Engineering by distance learning. Courses in computer science are currently being delivered by distance technology to Keesler Air Force Base in Mississippi. The high bandwidth that will be associated with the eventual physical site selected for the partnership will allow Cal Poly to offer many more academic programs by distance learning to remote locations. In particular, we will have collaborative agreements at cable-head locations around the world (including Asia and Europe) that could make our programs available to military personnel stationed almost anywhere in the world. This could be tied into training programs for ONR and, if desirable, to training directly related to research projects.

In summary, the California Central Coast Research Partnership offers an unprecedented opportunity to take advantage of a confluence of factors, including existing and potential relationships, fortuitous and unique technological and economic developments in the region, the particular strengths and expertise of the CAD Research Center, and a meshing of the research and development interests of the University, the Office of Naval Research, and the private sector. C^3RP is the vehicle for fully realizing the benefits of these common goals and synergies.

IV. Summary of Results During the Period of Performance

A. General.

An overview of accomplishments to date on the overall project follows:

- **Personnel** were hired to lead and staff the project.
- A **potential site** on campus was identified for the technology park. This ~17-acre site is level and close to existing utilities and roads. The land value has been appraised and a preliminary site design done. Phase I environmental studies have been completed, including biological, archeological and geological studies.
- **Project feasibility studies** have been completed. These include:
    - A market and feasibility analysis carried out by a leading national firm (Hammer, Siler, George and Associates) that specializes in University-related research parks.
    - A local commercial real-estate analysis.
- A **web site** (www.c3rp.org) for the project was developed and has been continuously updated and expanded. Data show over 43,000 "hits" on the site in the last three months. The site is attracting technology companies to the region. For example, Datect, Inc., a software development firm currently located in Santa Fe, NM will be relocating to San Luis Obispo, having discovered this area through the C^3RP web site.
- Work has begun on printed materials to promote the project.
- A **database of technology–based companies** that are potential partners in the project was developed and is continuously updated and expanded. New relationships have been developed with companies such as Anritsu, Inc., Datect, Inc., Frontier Technology, Inc., HaveBlue, LLC, and Space Information Laboratories, Inc.
- **New research** has been developed, including the following pilot projects, many with industry collaboration. They include projects highly relevant to defense and national security. Detailed reports of the results of these projects are presented in **Section IV. C. of this report**.

    - Design Methodologies for Analog/Mixed Signal VLSI Systems Applied to Infrared Focal Plane Arrays
    - Development of an Autonomous Tactical Reconnaissance Platform
    - Development of an Aerodynamic Table Lookup System and Landing Gear Model for the Cal Poly Flight Simulator

- Development of Field Rechargeable Gas Mask Filters
- Exploitation of Network Bandwidth and The Ethernet/IP Application Layer Standard for Automation Networks
- Range Sensing and Real-Time Registration
- Using Technology to Close the Adult Literacy Gap
- Correlation of Milk Composition and Fouling with Bio-film Formation and Microbial Spore Production in Heat Exchangers
- The Role of Discovery in Context-Building Decision-Support Systems
- The TEGRID Semantic Web Application: A Service-Oriented Distributed Approach to Disaster Management Decision -Support Systems
- Viable Bandwidth Compression for Remote Sensing Applications

- New **R&D facilities** and partnerships were developed to support current and future research efforts and university-industry collaborations: a computer networking research laboratory, a photovoltaic facility, a testing facility for semi-conductor processing technologies, and a geographic forecasting collaboration. Detailed reports on these activities are presented in Section IV.B.
- **Internet2** connectivity was applied for, approved, and acquired for the campus in November 2001, to support current and future research efforts. Detail is provided in Section IV.B.
- An **industry patent-donation program** was established with SAIC (Scientific Applications International Corporation) and Rockwell Scientific Corporation. These patented technologies can be used to develop new lines of University research.
- Seminal work was completed on **the management of fiber-optic assets as commodities**, for the benefit of non-profit organizations. The full report on this effort is provided in Appendix A and is summarized below.
- The project managers have visited other University-related technology parks and have attended conferences on developing these facilities, as well as on related issues such as technology transfer and start-up companies.
- The project's leaders have continued to work with private and government partners to advance the project and secure additional funding. The particular focus of these efforts has been the future establishment of an **Interdisciplinary Center of Excellence** in areas of relevance to homeland security.
- Efforts are underway to develop industry partners in the **biotechnology sector** for the purpose of developing research and training activities in this field. To this end we have been working with the Central Coast Biotechnology Center in Ventura, CA, with two local community colleges, and with several biotech companies, including Amgen, Baxter, Fziomed, Genentech, Promega BioSciences, and Hardy Diagnostics.

B. Information Technology Infrastructure, R&D Facilities and Partnership Projects.

As noted above, several projects were undertaken to develop infrastructure, facilities and partnerships in support of future research and development. More detail on some of the specific efforts identified in bullet-form above is provided in this section.

1. Development of Information Technology Infrastructure.

Cal Poly, San Luis Obispo is a key node on the higher education network backbone in California. To maintain this position on the backbone and ensure high-bandwidth access to the campus for research and educational purposes, Cal Poly worked on the following initiatives over the past year. These efforts will enhance our ability to support teaching, learning and research programs that require high bandwidth.

a. Internet2

Internet2 is a consortium led by 202 universities working in partnership with industry and government to develop and deploy advanced network applications and technologies, thereby accelerating the creation of tomorrow's Internet. Internet2 is recreating the partnership among academia, industry and government that fostered today's Internet in its infancy. The primary goals of Internet2 are to:

- Create a leading edge network capability for the national research community
- Enable revolutionary Internet applications
- Ensure the rapid transfer of new network services and applications to the broader Internet community

This past year we joined the Internet2 consortium and focused our efforts on promotion and involvement in Internet2-related activities by Cal Poly faculty. To that end, we have:

- Applied for and were accepted into the I2 consortium
- Identified key faculty to be "I2 Champions" in each college
- Installed high-speed (GigE) drops on campus in each College and the Library.

As a result of the effort to stimulate I2 applications, a proposal was developed and submitted to the national Internet2 consortium. The proposal, entitled Objective Networks, is included at the end of this section of the report (pp. 9-38).

b. City Fiber Partnership

To ensure our flexibility to access fiber-optic carrier points of presence and also to promote connectivity to City resources (e.g. City/County Library), Cal Poly has entered into a partnership with the City of San Luis Obispo to install fiber-optic cable within City conduit. This infrastructure will provide Cal Poly with a large degree of flexibility in terms of accessing other network providers and remain a key point of presence on the CENIC CalREN network (www.cenic.org). This partnership and the enhanced access it ensures will be critical for certain research applications, as well as for the viability of the technology park.

2. New Research and Development Facilities and Partnerships.

As noted above, four new facilities and partnerships were developed. Detailed reports on each of these initiatives are included at the end of this section.

a. Computer Networking Research Laboratory (pp. 39-70)
b. Photovoltaic Facility (pp. 71-90)
c. Testing facility for semi-conductor processing technologies (pp. 91- 97)
d. Geographic forecasting collaboration (pp. 98-100)

3. Fiber-optic Asset Management

This effort was directed at performing initial developmental research, investigation, outreach and prototype development of a taxonomy and working software-based model of how to identify, structure and manage certain communications and network assets to enable users to record, track, place into service and dynamically configure key bandwidth-related communications assets. A key feature of this applied research and development was that it created an explicit "performance model" for thinking about and directing the uses of disparate and unrelated assets into time-based or mission-based networks or network connections. Extensive, voluntary assistance or expertise from industry (e.g. Lockheed-Martin), was provided to assist in this effort at no cost to the project.

The essential structures and processes of managing such assets and the metadata to enable additional future applied research were also created and the archetypal design of a potential overall business model was specified. A working demonstration GUI (graphical user interface) and prototype system was created, as well as the thorough documentation thereof. Such documentation has been supplied to ONR (Program Director, George Solhan), and to key U.S. University-led networking consortia on a non-disclosure basis. The full report of these results is provided in Appendix A.

Objective Networks: A Proposal for an Internet2 Advanced Content Delivery Project

Project Investigators:

Fred F. Abler
Project Coordinator Specialist
CAD Research Center

Walt Bremer
Professor
Landscape Architecture Department

Hisham Assal
Software Engineer
CAD Research Center

Objective Networks $_{sm}$ - A Proposal for an Internet2 Advanced Content Delivery Project

Fred Abler, Walt Bremer, and Hisham Assal, The Objective Networks Group, California Polytechnic State University, San Luis Obispo, California.

Proposal Summary

Geographic Management Systems (GMSs)[Abler and Richardson 2003] are increasingly being used for place-based modeling and simulation of various real-time operations management environments (e.g., intelligent transportation systems, intelligent logistics systems, emergency response systems, capital asset management systems, etc.).

GMSs integrate real-time geographic data (GPS/GIS) and other geospatial information into digital *mirror-worlds* [Gelernter 1990] that support distributed, collaborative, and near real-time decision-making. However, the need for greater spatial dimensionality and temporal fidelity in these mirror worlds is increasing rapidly. Traditional static and 2-D modes of representing the physical world are no longer adequate.

One promising approach for addressing the need for high fidelity *world-making* [Abler 2002] are *intelligent object* technologies such as Geometric Description Language (GDL) pioneered by Graphisoft. GDL is capable of representing all the information necessary to completely describe architectural building elements as 2-D CAD symbols, text specifications, and 3-D models.

Broadly applied to geographically-based applications, intelligent object technologies hold considerable promise for addressing the need for additional informational and spatial dimensionality (i.e., 3-D objects) in GMSs. However, these technologies do not yet fully address the needs of real-time integration (i.e., 4-D objects), behavioral constraints, or programmatic accessibility that is needed by intelligent GMS web services.

The purpose of Objective Networks is to provide software tools, a high-bandwidth digital communication infrastructure, and an on-line repository to facilitate the collaborative development of next generation objects that are *virtually embodied autonomous agents*. As an advanced content object-form, these highly accessible and flexibly configurable *digital objects* will enable users of different types to rapidly build *mirror worlds* at multiple levels of detail, and deposit them for viewing and use by others, including web services and intelligent internet agents.

Table of Contents:

Section 1.0	**Objective Networks**	*pages 3 - 7*
Section 2.0	**Objective Networks Architecture**	*pages 8-20*
Section 3.0	**Project Benefits to the Internet2 Community**	*pages 21-26*
Section 4.0	**Project Guidelines**	*pages - 27*
Section 5.0	**References**	*pages 28-29*

Section 1.0 Objective Networks $_{sm}$

1.1 Introduction – *Digital Objects* as Advanced Content Forms

Next generation networks such as Internet2 (I2) offer an unprecedented opportunity for developing advanced content object-forms. I2 is currently approximately 1000 times faster than the commodity Internet, and bandwidth continues to increase rapidly. Gilder's Law states that bandwidth will continue to triple every 12 months for the next decade. Previously, content form and delivery technologies for the commodity Internet (e.g., streaming software, MPEG3, etc.) were driven by the need to conserve limited commercial bandwidth [Lazowska 2001]. However, with next generation networks, compression and bandwidth conservation are no longer the primary drivers of content form, and new advanced content object-forms that take advantage of relatively 'unlimited' bandwidth are now conceivable.

As digital communication infrastructures continue to develop, they enable fundamentally new content forms. These new object-forms are likely to become increasingly holistic, hybrid, and heavy, intelligent, and hyper-dynamic. Objective Networks proposes the unmet needs of emerging real-time technologies (i.e. Geographic Management Systems) may be met by new object-forms where each object is a *virtually embodied autonomous agent.* Virtually embodied autonomous agents, or *digital robots,* may be widely used as world-making primitives for building collaborative models and simulations of the physical environments (*i.e.,* mirror worlds). As an advanced content form, these *digital objects* are crafted collections of heterogeneous content that contain 2-D, 3-D, and 4-D geometry, object attributes, ontological components, and behavioral constraints in a polymorphic content form. Collectively, this dynamically structured collection of geometry and information holistically represents intelligent and autonomous *digital objects,* which in turn, collectively comprise high-fidelity mirror worlds.

1.2 *Digital Objects* as Virtually Embodied Autonomous Agents

As advanced content object-forms, Artifacts are hybrid representations - part autonomous hardware machine (*i.e.,* robot), and part autonomous software machine (*i.e.,* agent). Whereas autonomous agents virtually embody logical machines and robots physically embody mechanical machines, digital objects are *virtually embodied autonomous agents*, or more succinctly, context-aware digital robots. As digital robots, objects have a number of important practical advantages. Because they are virtually embodied, objects can interact *physically* with their virtual environments (e.g. collision detection and avoidance, interference fits, etc.). Because they are also autonomous agents, objects can also interact *contextually* with virtual environments. For example, digital objects may rapidly reconfigure their virtual embodiments and transform their behaviors to suit their respective environmental context (e.g. intelligent assembly, detail-on-demand, information-on-demand, etc.).

Independent context-aware behaviors are useful in their own right, but especially so because mirror worlds are themselves, nothing more than heterogeneous collections of digital objects. Unlike physical robots, virtually embodied autonomous agents are by definition in direct communication with every other object in a virtual world. Thus, by opportunistically sensing and reacting to one another, extremely dynamic and complex emergent behaviors are possible. Virtually embodied autonomous agents extrapolate the *subsumption control architecture* invented by Rodney Brooks at MIT for navigation of physical robots in real world environments [Brooks 1986, 1990] to the navigation of digital objects in virtual environments. Brooks was the first to realize that by quickly coupling 'sensing with acting', real-world robots could navigate almost any physical environment with very little 'planning overhead'. Brooks has considered the use of his subsumption architecture in virtual environments, however, he has previously argued that complex agent behavior is a reflection of a complex environment, and that virtual reality is not a rich enough to provide this level of stimulus [Brooks 91].

However, environmental complexity is readily enabled by the *transitive* nature of objects in digital mirror worlds. Because digital objects are by definition context-aware, the number of their physical and semantic relationships increases factorially with the addition of each new object. Thus, modest initial complexity is rapidly leveraged by the *multidimensional connectedness* [Holm Nelson 1990] of the virtual environment. This complexity from connectivity is further compounded by the flexible reconfiguration, or polymorphism of the digital objects. Each object added to the virtual world may cause other objects to opportunistically reconfigure their relationships to some (but not all) of the objects in the mirror world. Therefore, the number of complex relationships becomes a many-to-many function of the number of objects (and their alternate embodiments) in the virtual world. Thus, even a relatively small number of digital objects can quickly create a rich, dynamic, and adaptive environmental context, which in turn would yield complex agent behaviors (*i.e.,* emergent intelligence).

It is impractical to build real-world objects and robots that have the *virtuality* [Holm Nelson 1990] of their digital counterparts (*i.e.,* that are polymorphic, and in constant contextual communication with one another). Even highly sensored and radio-equipped physical robots would be at a distinct disadvantage because most of the physical objects they would encounter in the real world are not context-aware. Real-world robots cannot ask a physical table 'what are you?" whereas the semantics of a digital *table object* are immediately asserted upon its instantiation in the virtual world. Thus, digital objects possess a *virtuality* that has no real-world analog. For example, autonomous, context-aware, polymorphic, and self-installing (*i.e.,* robotic) architectural windows do not yet exist in the real world, but are readily represented in digital environments. The *'magic world'* of digital environments offers several distinct advantages [Terveen et al. 1995]. In digital worlds, the nature of each object can be cast 'as it seems', rather than actually is. Properly exploited, such *virtuality* can lead to extremely dynamic link grammars, whereby virtually embodied objects opportunistically self-assemble, enact complex associations, or change their relative position based upon GPS data streams for example.

1.3 An Architecture for *World-Making*

Because of their *virtuality*, digital objects present a dynamic architecture for world-making. World-Making is the spontaneous, collaborative, and conceptual processes that organic and artificial intelligences use to make sense of their shared environments or worlds [Abler 2002]. Because digital objects are equally available to both human *and* software-based intelligences, users and next generation web services can synergistically interact by manipulating them in tandem, and, the digital objects themselves become a third form of emergent intelligence that also spontaneously contributes to the *reflective conversation* [Schon 1989]. For example, an Architect may place a *wall object* near the *property line object* in a digital design world. However, in this instance the *wall object* stubbornly resists the architect's placement and automatically retreats from the property line (*i.e.,* haptic feedback). Transparently to the user, the *wall object* has detected an interference fit with a non-displayed *zoning object.* Because all digital objects are context-aware, they obey the constraints embedded in the *zoning object.* Furthermore, because digital objects are also programmatically tractable, a next generation code-compliance web service monitoring the collaboration may intervene with an impromptu explanation, "The lot line setback is 10' unless the wall has a 2-hour fire rating, in which case the setback may be on the lot line".

By virtue of the multidimensional connectedness of virtual environments, the *wall object* has also simultaneously received this information programmatically and, as a *virtually embodied autonomous agent*, the *wall object* may opportunistically transform itself into a 2-hour Fire Rated configuration (via transparent download from Objective Networks), and move to a zero lot line orientation for the Architect's consideration. Thus, context-aware objects can embody a host of useful polymorphic and context-sensitive behaviors. Because they are scalar embodiments and contextually aware of the current display scale, each object is also able to automatically generalize its graphical appearance. For example, transitioning from a nadir-view of a geographical world to an oblique and larger scale planning-view, users could see their objects automatically transform from 2-D symbols to 3-D models with relatively low dimensionality. As users

continue to drill down and navigate their virtual worlds, additional dimensional, graphical, and informational 'detail-on-demand' are provided where needed. Context-aware objects may similarly index their display of attribute information to suit one (or more) professional perspectives (e.g. Geographer, Planner, Architect, Engineer, etc.) as needed. Such context-aware behaviors are not only highly useful and good human factors, but computationally efficient as well.

Polymorphic object behaviors are fundamentally enabled by the multidimensional connectedness and high bandwidth of next generation networks. Should a digital object detect that it needs to reconfigure itself opportunistically based on environmental triggers (i.e. display scale, angle-of-view, interference of objects, etc), user query, or the local enforcement of embedded constraints, the object will introspect and see if it has the ability to spontaneously reconfigure itself as needed. Minor and highly useful transformations will likely be embodied in all objects (e.g. display scales, graphical detail settings, etc). Should the object be unable to reconfigure itself on the local client, it will transparently 'reach back' to Objective Networks and download a *high-fidelity* instance of itself as needed. Thus, depending upon the task at hand, the object population of any given virtual world is likely to be in a constant state of transformation. Thus, digital objects begin to address the dynamic nature of real-time Geographical Management Systems, while conserving computational and bandwidth resources by 'detail-on-demand'.

The *dynamic fidelity* [Kay 1990] of digital objects then, not only supports dynamic and real-time *world-making*, but enables the automatic or purposeful development of alternate mirror worlds. Thus, 'what-if' simulations, and other useful engineering and operations management considerations can be more easily supported in Geographic Management Systems (GMSs). Virtually embodied autonomous agents thus present a dynamic architecture for 'change management' and real-time distribution of geographic information and knowledge on next generation networks. Further, because the architecture of digital objects allows for graduated knowledge representation, the 'just-in-time' delivery of graduated levels of representation is computationally efficient. By caching anticipated levels of detail on local clients, expensive query computation is

largely eliminated and a scaleable architecture is achieved. As Gilders' Law intersects Moore's Law, bandwidth will outstrip computational power to 'push' content through next generation networks [Lazowska 2001]. Distributed backplanes, peer-to-peer, or other forms of 'grid' computing may be able to keep processing power at parity with available bandwidth, however, the 'just-in-time' delivery of detail-on-demand that *virtually embodied autonomous agents* can provide may be a key 'load balancing' technology -even in high powered grid-computing environments.

Section 2.0 The Objective Networks Architecture

All of the useful behaviors of *virtually embodied autonomous agents* discussed in the previous section depend upon the 'context aware' nature of digital objects. However, for digital objects to be context-aware, they will require a semantic architecture that represents knowledge in programmatically tractable, and incrementally scaleable form. Such architectures typically rely on the concepts of ontology construction. Ontology is a structure of knowledge in a given domain. It consists of a description of entities that exist in that domain and the types of relationships that hold among those entities. It also includes all the constraints, assumptions and axioms that define what can exist in the ontology. Using ontology concepts provides easy means for users to define their own digital objects, assigning them any level of detailed description that fits their needs, and for this context to then be automatically available to other virtually embodied autonomous agents, or digital objects.

2.1 Components

Objective Networks consists of a distributed network of I2 servers that share common protocols for exchanging world-making objects. Every server has a backend database that stores the ontological description of the world-making object. It also has another database for storing models that are developed by users. The world-making environment is a client-side software component that includes three main pieces: the display manager, the communications manager and the object editor. The client-side software communicates with the Objective Networks servers through special access

protocols that encapsulate the I2 protocols and add to them more specific protocols to allow the handling of varying levels of ontological details.

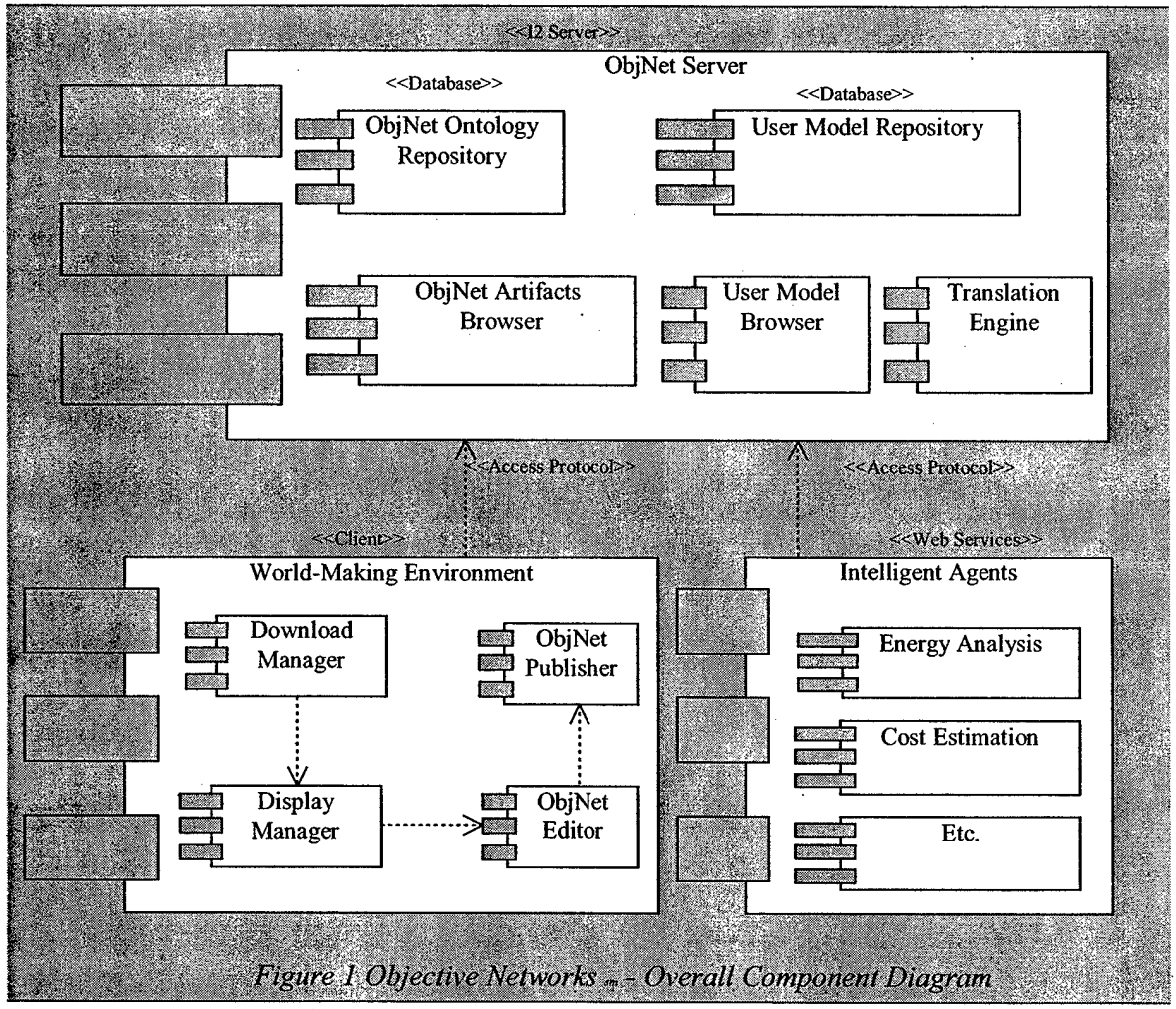


Figure 1 Objective Networks $_{sm}$ - Overall Component Diagram

2.2 The I2 Servers

Objective Networks consist of a distributed network of I2 servers that have similar architecture. They all have 2 types of back-end databases. The first database stores world-making digital objects, which consist of geometric and non-geometric descriptions of objects, and the protocols for joining objects together defined within the ontology of world-making objects. The second database stores models composed by the

users and represents objects in the real world. The user database stores information about the user's interests and the type of access they allow for their models.

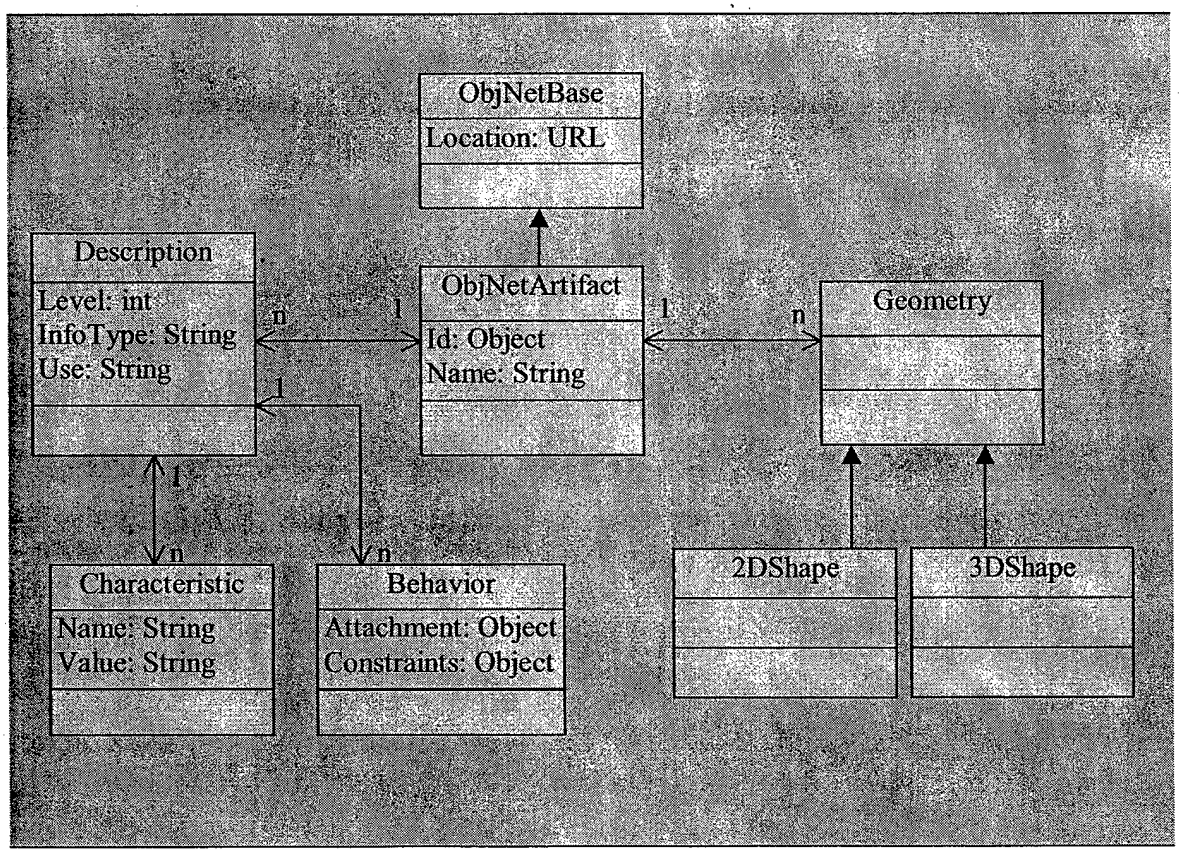
Access to world-Making objects is provided through a specialized browser that understands the ontology structure and communicates with the client-side component using I2 protocols and Objective Networks special access methods. The browser handles the level of detail requested by the user and retrieves the corresponding ontology description from the database. Access to world models is provided through a model browser that has access security levels that can be 'private', 'shared' or 'public'. Private models can only be accessed by the creators of these models. They are stored in the Objective Networks database for further development or until they reach a state that can be communicated with others. Shared models can be accessed by a designated list of users. For example, a user may decide to share world models with a group of associates of collaborators, but not with the rest of the world until it is finished. Public models and worlds can be accessed by any Objective Networks client or web services that implement ObjNet access protocols.

The user-model database will have a space for the user to identify types of information they seek on their models, such as component prices, physical performance, or feedback from other users. Web services and intelligent Internet agents that crawl the user model databases can read this area and provide their services as needed. For example an energy analysis web service can identify building models and make offers to perform energy performance and compliance with federal, state and local regulations.

2.3 Ontological Representation

The representation of digital objects in Objective Networks relies on a base ontology that allows any number of levels of descriptions. The generic ontology also provides basic behaviors and constraints that describe the necessary conditions for interaction among objects. The ontology consists of a number of base classes and base relationships. The 'ObjNetBase' class provides a URL where the object is located. This information is necessary for repeated access to the same object to retrieve the desired

level of description. The 'ObjNetPrimitive' class specializes the base class and defines the type, name and a unique ID for identifying the object. Any object that is defined by the user specializes this class and inherits all its characteristics and relationships. The 'ObjNetPrimitive' class has two main relationships: 'Geometry' and 'Description.' The 'Geometry' class stores the 2D and 3D geometry information such as shapes, polygons, coordinates, and colors. The relationship between 'ObjNetPrimitive' and 'Geometry' is one-to-many to allow the object to have multiple geometric descriptions that correspond to the multiple levels of details. The 'Description' class also has one-to-many relationship to allow the multiple levels of detail. It stores information about the level of detail, the category of information and the type of use for such information. The actual information of one object is stored in multiple instances of the class 'Characteristic.' Each instance stores the name of an attribute and a value for that attribute. This way the description of an object can be built up over time and can be extended later by the same user or by other users.



2.4 Associations

Building a world model is performed by associating objects of different types together into a model. This association is defined in the ontology as behavior attached to each object. Objects with matching behaviors can associate with one another according to that common behavior. The association is also represented in geometry by special handles that can be used to relate two objects together. When users define or extend primitive objects, they can define any number of behaviors and associations to describe the way this object interacts with other objects that are already defined in the repository or to let other users know what behavior they need to have in their objects or models in order to use this object. Basic types of behavior include compositions, translation and constraint enforcement. In composition, the object behavior defines how this object is combined with others and at what levels. This is done by defining methods of attachment among various objects and selecting the applicable ones for the specific case. In translation, the object behavior identifies the different facets of the object in different representations and how the object can map its representation to another. Constraint checking defines how strict or relaxed the constraints of this object can be or should be under different configurations. This allows the user to automate some of the tedious tasks of checking the validity of their compositions by enforcing the properties constraints automatically.

2.5 Translation

The exchange of information between the I2 server and the web services or the client side world-making environment can take place in any common representation format, such as CAD drawings (DXF, DWG, etc.), extensible markup language (XML), portable data format (PDF), simple vector graphics (SVG), or Geometric Description Language (GDL). These common formats allow the users to build their models in a variety of application software systems and apply a variety of analysis tools to the models. This exchange is made possible through a translation engine that resides on the I2 server. The translation engine relies on a set of mappings for all the known formats and a set of tools that allow the user to define mappings to new user-defined ones. The translation engine can work in a transparent manner by recognizing the requested format

when the user attempts to down load a model or an object. It then applies the set of mappings that correspond to the requested format to the model or object.

The result of the down load is the model of the object in the requested format. The translation engine will ideally work in both directions. If the user attempts to upload models or objects in a known format other than the ontology representation, the translation engine applies the set of corresponding mappings to the model before it stores it in the model repository. If the format is unknown to the translation engine, it will allow the user to define or upload a new set of mappings that identify the new format. This facility will allow users to work with models that were created in other software packages, and will build up the number of digital objects available for world-making, much more rapidly than is feasible using a single format alone.

3.6 *World-Making* Environment

The user of Objective Networks interacts with the server through a client-side world-making environment, which is a software component that includes all the required functionality for creating, manipulating and combining Artifacts into models and communicating with the Objective Networks server in both directions: download and upload.

Download Manager

This component handles the download and upload of objects from/to the Objective Networks server using the access protocols. It uses knowledge of the location of the required object and the current level of detail and communicates that to the server to download the required ontological and geometric descriptions. The download manager maintains a cache of downloaded objects and the current level of detail. This information helps the display manager perform its functions in a transparent manner. When the display manager requests another level of detail (higher or lower) the download manager fetches the required information either from its cache area or downloads from the Objective Networks server using the object ID and the URL defined in the object.

Display manager

This software component handles the visual display of objects based on their associated geometry and also the display of ontological, non-geometric information according to the current level of detail, which is set by the user. This component communicates with the download manager to get the required objects at the required level of detail. The display manager has tools to allow the user to manipulate the Objective Networks objects on the screen (e.g. move, scale, rotate, zoom, etc.) and also to connect objects together according to their defined behavior and connectors. The display manager understands the generic ontology structure and performs its functions by manipulating instances of objects in the specific accessed ontology. The user can build a model by downloading any number of artifacts and joining them according to their defined behavior or they can also define their own Artifacts both graphically and ontologically using the Objective Networks editor. When an Artifact or model is finished, it can be uploaded to the user model repository on the ObjNet server using the ObjNet publisher.

Objective Networks Editor

This is a component that may be embedded in the display manager. The editor allows the user to create or edit ObjNet objects in graphical format and associate with them any number of levels of details as they wish. The editor has graphic tools to create any type of geometry at any level of detail. The geometry description and manipulation functions in the editor rely on the geometry description in the ontology. Both 2D and 3D graphic tools are available in the ObjNet editor to allow the user to create objects and models as sophisticated as they wish.

In addition to the graphic description of objects, the ObjNet editor also allows the editing of non-geometric ontological descriptions at any level. The user identifies a level of detail and a simple pop-up menu appears every time the user clicks the right mouse on the defined geometry and the user can select either 'attribute editor' or 'behavior'. The attribute editor allows the user to define attributes and values for these attributes. The behavior editor allows the user to define constraints on what type of objects can attach to

this object and some other basic behavior characteristics. The information defined in this simple menu creates instances in the ontology for the corresponding attributes and behaviors and attaches them to the current object at the currently selected level of detail. The user can then select another level and define more attributes, values and behaviors and attach them to the current object. When the user is done with editing the object they can select to upload it to a common repository using the ObjNet Publisher.

Objective Networks Publisher

This component uploads objects to the ObjNet ontology repository and also uploads finished models into the ObjNet user model repository. The component is integrated into the world-making environment and can be invoked through either the display manager or through the ObjNet editor. The upload function requires knowledge of the server URL to which it should upload the object and the different levels of details that are associated with it. The ontology description that was defined by the user specifies this information and the user can select the server to which the objects or models should be uploaded through the options menu.

2.7 World Making Generalization

The fundamental idea of World-Making Artifacts is to allow the user to describe an object at multiple and progressive levels of detail, and to associate this description with a multiple graphic/geometric representations, which is the means of manipulating the objects.

2.8 Web Services and Intelligent Agents

The World-Making environment forms the basis for web services to perform and run. Models in the Objective Networks server represent the objects and organizations in which users are interested, and web services can interact with such models and provide their analysis, evaluation or information retrieval directly to them. This model provides an intuitive way for product modelers (e.g. manufacturers, designers, architects, etc.) to construct their models and expect services to be provided to them.

Intelligent agents can be viewed as advanced form of web services. Agents roam the network looking for models that fit their understanding of the world (a given

environment) and when they find them they perform their job on them. The functions of intelligent Internet agents can range from search for relevant information (such as real-time geospatial position, prices, engineering properties, manufacturer's information, object status, etc.) to collection of pieces for detailing the model (e.g. finding appropriate hardware for a given window, selecting the paint and finishes for interior spaces, etc.), to preparing bids on completed projects.

Web services and Intelligent agents can work in an autonomous fashion. The user may not be aware of the existence of such services, but they would appreciate any help in information processing on their models. To achieve that, the following process is followed:

- Users deposit their models in a public-access area on the Objective Networks server and define the level of access they permit for them.
- Web services and Intelligent agents that roam the network continuously then access such models up to the allowed level of detail and offer their services.
- The user gets notification of the offered services and the cost for using each of them.
- The user makes a decision on which services to apply to their model and contact the service providers for agreement.
- The user may setup a profile for the accepted and trusted services and may even setup their cost acceptance criteria for any given product.

The use of Intelligent agents in Objective Networks is totally directed and managed by the user. The development of these agents and web services is done by third parties that have access to the structure of the Objective Networks ontology and the development API. This expands the variety of the agents and services and provides an environment for both competing and complementary services.

2.9 Language

The world-making environment will use a language that will be developed to handle ontology manipulation and object compositions. Such a language will be based on previous work that has been done in the area of ontology such as the 'Knowledge

Interchange Format' (KIF) from Stanford University. An ontology language provides higher level of abstraction than regular programming languages such as JAVA and C++ because it handles ontology at a knowledge level. It allows for axioms and assumptions to be part of the definition of a concept and provide mechanisms to define, manipulate, and manage constraints within the ontology. The language will also address the notions of 'ontological concept,' 'ontological theory,' ' constraints' and 'assumptions' as the basic concepts to build world-making objects that are aware of their roles and possible associations in any given environment. This awareness provides context for the world-making objects to apply their and embedded and 'just-in-time' behaviors.

The language will be the underlying building mechanism for the ontology on the server as well as the tool for building artifacts by users. The use of the language will be transparent to the users through user-interface tools that guide the users to the appropriate actions and translate their actions to language constructs. In addition, an application-programming interface (API) will be exposed to allow subscribers of Objective Networks to build applications (e.g. web services, Intelligent Agents, special-purpose applications, etc.)

2.10 Object-Passing Protocol

Current access protocols take into account the bandwidth limitations and attempt to conserve traffic. This leads to simplified protocols that require a lot of re-construction work on the client side. Objective Networks will take advantage of the high bandwidth in I2 to provide access to complex compositions of objects along with all the involved relationships and associations. This will allow complete concepts to be transferred at once in a meaningful way and reduce the amount of work that has to be done on the client side. A special protocol will be developed to allow the passing of complex concepts in an envelope that contains all the required descriptions. The unit of transfer will be a 'concept', which corresponds to an ontological concept with its geometry, relationships, constraints and assumptions. The Artifact-passing protocol will embed the I2 access protocols (IPv6) and provide a layer that describes the structure of concepts within the unit of transfer.

2.11 Delivery Format

To take advantage of the high bandwidth environment of I2, the delivery unit of Objective Networks will be an object. The format for this object delivery will be standardized to handle a wide variety of world-making concepts. As a 'conceptual' unit, the object includes the definitions and all the necessary interfaces for a world-making primitive that will allow its immediate use on the client side environment. The rich description of the 'concept' provides context for use, which defines how this object interacts with other objects that are relevant to it. Since both the server and the client will be using the same concept passing protocol to exchange information, there is no need to convert the complex structure of objects to a standard text-based language such as XML on one end, and then have the other end read and parse the result to re-construct the concepts with all their embedded complexity.

In addition to the specialized concept passing protocol, Objective Networks will also provide a set of translators that will allow the conversion of world-making objects and finished models from their native format to any standard format for use in existing modeling environments (e.g. AutoCAD or ArcGIS).

2.12 Development Plan

The first stage in building Objective Networks includes the following elements (deliverables): the establishment of experimental 'object-making' classes at affiliated institutions, the deployment of the Objective Network servers, the development of world-making software tools, and the selection of communication protocols.

2.13 Objective Networks *I2* Servers

The server will be built by creating a database on an I2 server. The database will preferably be an object-oriented database. The design of this database will have the following requirements;

- Mirror the ontology design so that ontology objects can be stored directly in the database without any conversion.
- Implement an indexing scheme so that artifacts can be found easily and in reasonable time.

- Implement user profiles and access permissions to allow users to store their objects in the database and define the type of access they desire for their objects.
- Implement access and browsing tools for users to navigate the database and search for objects that meet certain criteria.

2.14 World-Making Tools

This is the client side environment and it will implement access methods based on the communications protocols. The requirements for this component are:

- Implement database access methods that include creation, deletion, and modification of objects as well as query capabilities.
- A user-friendly interface that is capable of displaying the objects at the desired level of detail and manipulating the objects by connecting them together according to their defined behavior. The environment will also support the creation of new objects or modifying downloaded ones by providing basic geometry manipulation tools at the local level.
- The ability to verify newly created objects for adherence to the Objective Networks protocols and guidelines for creating objects.

2.15 Communications Protocols

The I2 protocols (mainly IPV6) only address the routing issues of network traffic as well as the increased capacity for name space. Objective Network protocols will address the content of that traffic. The communications protocols have the following requirements:

- Provide a language for composing network messages that encapsulate the rich content of ontology objects (i.e. the objects)
- Provide methods for downloading objects (ontology objects) directly into the world-making environment without the need to convert their description to a standard

language, such as XML. These methods understand the generic information structure of the objects without dealing with their specific content.

- Provide methods for uploading objects that range from simple objects to fully complex models. These methods understand the required format for an ontology entry and can formulate the objects or models into that format for storing in the database.

2.16 Test Case

In order to validate the concept and operations of Objective Networks, a test case will be developed that will carry out the operation from start to end. The test case will involve the following:

- Creation of a small set of objects on the client side (within the world-making environment) that can be related together in a common context.
- Uploading the objects into the I2 server and storing them on the database.
- Through a different client, access the I2 server and browse to find the objects.
- Download the objects into this client.
- Manipulate the digital objects by editing some and connecting them to form a model (or a virtual world).
- Upload the objects and models into the I2 server and store it in the model database.
- Access the stored model via a third client (or maybe the first one) and verify that the model carries all the information that was entered at creation time and in the same structure.

This test case will serve to keep the development process in line and provide means for verifying the validity of the concepts and the correctness of the operations. It will also help to point out any deviations from the original concept at an early stage so that the development process can be guided back to the right track.

Section 3.0 Project Benefits to the Internet2 Community

3.1 Project merit and benefit to the higher education community.

Objective Networks investigates I2 from a theoretical as well as technological perspective. Objective Networks posits that I2 is a shaping technology that will fundamentally redefine academic and professional disciplines by fostering the emergence of transdisciplines. A *transdiscipline* [Kozmetsky 1999] is a discipline or profession that serves other disciplines by providing tools for them (e.g. statistics, logic, design, evaluation, software engineering, etc.), and, that is also a discipline in its own right. For example, the primary disciplines of Geography, City and Regional Planning, Civil and Structural Engineering, Architecture, Landscape Architecture, and Construction Management are unique, but all share common methodologies for modeling and transforming the Earth's surface. These modeling technologies are now also finding unanticipated use in real-time Geographic Management Systems that support a host of operations management environments. Thus, collectively these application areas are served by the emerging transdiscipline of World-Making. World-Making is concerned with knowledge representation and providing the collaborative tools, methodologies, and theory needed for creating, sharing, and managing digital forms of knowledge in network-centric environments.

The project also investigates new hybrid object-forms that will entail the development of a number of technological advances to the state of the art in computer science and knowledge representation. Dynamic ontologies are currently cutting edge research and efforts like the Semantic Web [Berners Lee 2001] and Darpa Agent Markup Language begin to address the need for dynamic ontologies. By distributing and applying ontological concepts via practical World-Making objects, new and potentially significant architectures for dynamic ontology building, distribution, and interchange may be discovered. New protocols for passing conceptually complete objects are also likely to

emerge. For example, instead of currently available protocols for accessing 'simple objects' (i.e. SOAP), complex objects are likely to require new 'Complex Object Access Protocols'.

The distribution of knowledge as digital objects is also likely to have profound import for the teaching and transmission of knowledge in respective disciplines, and a correspondingly large impact on professional practice. New disciplinary models (i.e. transdisciplines) and highly collaborative models for professional practice are likely to emerge from these investigations, and the exploration of such new technologies, epistemologies, and modes of professional practice have traditionally held strong intellectual merit in the higher education community.

3.2 Advancing the Internet2 mission.

Using objects as shared representations for World-Making, several heterogeneous network clients can synergistically collaborate to create virtual worlds exhibiting high degrees of *dynamic fidelity* [Kay 1990]. Dynamic fidelity is the central characteristic of *virtualities*, or virtual computational environments that exploit connectedness to enable the rapid exploration of many different solutions to a given challenge. The classic example of virtuality is the spreadsheet. Spreadsheets support the construction of elaborate mathematical models, and the dynamic enforcement of any number of user-defined mathematical 'truths'. This dynamic fidelity (i.e. *truth maintenance)* was a new form of virtuality that made spreadsheets extremely powerful generalized environments that promoted 'what if' scenarios and other forms of virtual exploration. Consequently spreadsheets, as generalized mathematical modeling environments, served the needs of any number of disciplines (i.e. business, statistics, finance, scientific analysis, construction, etc.).

Early spreadsheets like LOTUS 123 and VisiCalc are widely acknowledged as one of the two 'killer applications' for personal computing, desktop publishing being the other. Curiously, commercial success did not stem from traditional marketing (i.e. software applications for targeted users in existing market segments) but rather from the

virtuality of generalized modeling environments appealing to a broad spectrum of stakeholders. These early *virtualities* were equally useful for balancing your personal checkbook, doing monthly payroll, or the quarterly financial projections of major corporation. Desktop publishing software could similarly make a personal invitation to a New Year's Eve party, your company newsletter, or camera-ready art for professional magazine publications. Ironically then, these 'killer applications' were in fact application-neutral, and may be more accurately described as 'killer virtualities'.

Virtualities are undoubtedly capable of producing a wide range of end products, but it is their versatile 'truth maintenance' and change management facility (i.e. virtuality) that make them relevant to such a broad user base. Virtualities not only introduce new economies of scale (i.e. production capacity), but also *economies of scope* (i.e. version capacity), or what might collectively be called *economies of change.* Technologies that deliver 'economies of change' are extremely significant because they justify their own adoption - *despite* the substantial sunk costs in hardware, training, and communication infrastructures. Because they deliver new and ongoing 'facility for managing change', *virtualities* enjoy a privileged place in an ever-changing world. This bodes well for distributed high-bandwidth computing environments like I2, which are intrinsically more dynamic, connected, and collaborative than stand-alone or even narrow-band client-server applications. As a technology platform, I2 has an unprecedented advantage in supporting virtual capacity for managing change; however, commercially significant virtualities that exploit network-centric architectures have yet to emerge.

The project proposes that I2 fundamentally enables powerful new network-centric *virtualities*, which will serve the emerging transdiscipline of World-Making. By using I2 as a test bed for this potential 'killer virtuality', the project intends to establish pre-commercial traffic-ways between Objective Network Affiliates, establish new and more meaningful uses of high bandwidth environments, and collaboratively develop the semantic capital needed for subsequent development of next generation web services and intelligent Internet agents. Because environmental disciplines are also connected to large communities of professional practice that routinely use advanced technologies and

software applications, and because these professions are in turn connected with multi-trillion dollar finance and sales industries, the project may also initiate a ready migration path for commercializing I2 as a platform of distributed high-bandwidth computing.

3.3 Proposed time frame and project stages.

The Objective Networks project is a long-term collaborative research and development undertaking. Its fundamental goal is to develop World-Making as a proof-of-concept network application (i.e. 'killer virtuality') for Internet2 that may lead to commercialization of I2 as a distributed platform for collaborative computing. In this respect, the project is intended to be ongoing, commercially supported, and based on a 'Network Affiliate' business model.

3.4 Cost Recovery Arrangements – Network Affiliation Fees

The Objective Network Project will employ a network affiliate model for project participation. Organizations wishing to become Objective Networks Affiliates are welcome to contact Fred Abler abler@calpoly.edu. Academic Affiliates will be expected to offer experimental classes on 'object-making' and contribute the digital objects and models developed in the classroom to Objective Networks. Commercial affiliates are welcome to contribute in-kind services or software, as well as financially to the ongoing research and development of Objective Networks.

For example, a commercial organization may wish to engage Objective Networks for the purposes of creating a product catalog on Objective Networks. The project will provide a quote for collaborative research services, and if accepted, The Objective Networks Group will manage the collaborative research and development for the manufacturer. All research and development funds will be expensed according to the following overhead schedule:

1) 10% of all research funds will go to the Objective Networks Information Utility for ongoing development of World-Making $_{sm}$ tools and services, contract coordination and administration, and general network maintenance.

2) 10% of all research funds awarded to Network Affiliates (i.e. Universities and non-profit organizations) will be allotted (in proportion to the allocation of funds for research and development) to supporting I2 physical infrastructure related to contract research at that location. No awarded funds shall be used for overhead or contract administration provided by non-profit, or not-for-profit, or for-profit University Foundations and Sponsored Programs unless collaborative research awards exceed $250,000 dollars annually, in which case project overhead and administration fees will be negotiated by the Objective Networks Group and local research faculty and the home university.

3)

4) 80% of all research funds will go directly to research and development of World-Making technologies, software tools, Artifacts, world-models and/or web services as directed by the awarded research contracts.

Alternately, those commercial organizations wishing to use World-Making $_{sm}$ technologies to develop product catalogs, digital objects, models and worlds internally, and then publish them on Objective Networks may do so for a negotiated contribution to the Objective Networks Project. Such arrangements will be possible via the Objective Networks Affiliate Agreement, and negotiated fees are to be paid directly to the Objective Networks Information Utility. Such fees will be used for ongoing research, development, coordination and other project costs, and all fees and expenditures will be made available annually to Objective Networks Affiliates.

The Objective Networks Group is currently seeking supplemental funding for the continued development of World-Making toolkits and services that will be distributed via the Internet2 platform. These toolkits will be freely available to Objective Networks Affiliates and sponsoring industry partners for the collaborative development of environmental and other types of Artifacts. All objects will then be served by a distributed network of I2 servers hosted by distributed Objective Networks Affiliates. Objective Network affiliation will be defined in the forthcoming Objective Networks Affiliate Agreement, which will base network affiliation on active participation and performance levels.

3.5 Aspects relevant to Internet2 Working Groups.

Geospatial Working Group – Using advanced networking to enable rapid and ubiquitous access to geospatial Artifacts may be of interest to this group.

Internet2 Commons – creating a large scale distributed collaborative environment for research and education for Objective Networks may be interest to this group.

Distributed Storage Infrastructure- using Objective Network Affiliates sm as a distributed network of I2 Reality Engines sm for serving Artifacts to the I2 community may be of interest to this group.

Internet2 P2P and Distributed Computing WG – Terrestrial bandwidth will eventually outstrip existing processing power, thus requiring new forms of grid computing that will be able to move Artifacts across the network in real and hyper time frames.

Ipv6 Working Group – Objective Networks will require an addressing space that may exceed commercial Internet applications. The need for wide and highly dynamic addressing spaces may be of interest to this group.

3.6 Objective Networks sm Contact Information.

The Objective Networks I2 Project is currently operating with an initial collaboration with Architecture faculty at Penn State University. There are several charter members of the project noted in the next section. Additional membership is by invitation. Interested parties should contact Fred Abler abler@calpoly.edu .

3.7 Objective Networks sm Founding Members

Frederick Abler	Project Director & PI, Objective Networks Group, Cal Poly,
Walt Bremer	Co-PI, Objective Networks Group, Professor LARCH, Cal Poly
Hisham Assal	Senior Software Engineer and Research Associate, Objective Networks Group, Cal Poly
Madis Pihlak	Professor ARCH, Department, Penn State University
Marc Fredrickson	Director of 3-D Art and Level Design, Angel Studios, Carlsbad CA

Section 4.0 Objective Networks Project Guidelines

- All collected network performance data and metrics will be freely available to the Internet2 community via http://www.objectivenetworks.com.
- Any content providers collaborating on the project must be Internet2 members, and all existing policies governing Abilene and Objective Networks (sm) project participation apply.
- The project duration will be a one-year, subject to extensions at the request of the Objective Networks Group and Fred Abler.
- Participation in subsequent phases of the project will follow guidelines established by the Objective Networks $_{sm}$ Group. Project participation will be based on active participation, goals and metrics. Interested parties are welcome to contact the Fred Abler, via email at abler@calpoly.edu.
- Internet2 and Objective Networks $_{sm}$ Affiliates will make no claim on the intellectual property rights for the content produced and distributed through these collaborations.
- Distribution of World-making software tools and use of third party proprietary software will be governed by the Internet2 Intellectual Property Framework, V.2.0. http://www.internet2.edu/members/html/intellectualproperty.html
- Objective Networks $_{sm}$, ObjectNet $_{sm}$, and World-Making $_{sm}$, are service marks, brand assets, and intellectual property of the Objective Networks Information Utility, a not-for-profit corporation promoting the commercialization of Internet2.
- Service marks may not be used commercially or by Objective Network Affiliates without prior written permission. Non-commercial use must acknowledge organizational ownership (i.e. Objective Networks Information Utility).
- Status reports will be made accessible to the Internet2 community via the http://www.objectivenetworks.com twice a year. More frequent or ongoing reporting, including project presentations and demonstrations during Internet2 meetings, may also be periodically made available.
- This proposal and all contents are © 2002 of the Objective Networks Information Utility.

Section 5.0 References

[Abler 2002] Fred Abler. "Objective Networks: How High Bandwidth Environments Will Enable *Context* to Become *Content* on Next Generation Networks. : InterSymp 2001, 14th International Conference on Systems Research, Informatics and Cybernetics, Baden-Baden, Germany, August 2002.

[Abler and Richardson 2003]. Real-time Geographic Management Systems for Homeland Security. The Geographic Dimensions of Terrorism (a book forthcoming from XXX press.)

[Berners-Lee 2001] Tim Berners-Lee, J. Hendler, O. Lassila. "The Semantic Web" Scientific American, May 2001.

[Brooks 1986] R. A. Brooks (1986) "A Robust Layered Control System For A Mobile Robot", *IEEE Journal Of Robotics And Automation*, RA-2, April. pp. 14-23.

[Brooks 1990] R. A. Brooks (1990) "The Behavior Language; User's Guide", *M. I. T. Artificial Intelligence Laboratory, AI Memo 1227*, April.

[Brooks 1991] Brooks, R., "Artificial Life and Real Robots", Towards a Practice of Autonomous Systems: European Conference on Artificial Life, MIT Press, Paris, France, pp. 3-10, 1991

[Chanrasekaran et al. 1999] Chandrasekaran, B., Josephson, John R. and Benjamins, V. Richard. *What are Ontologies, and Why Do We Need Them?* IEEE Intelligent Systems, Vol. 14 No. 1, January/February 1999.

[Fridman –Noy et al 1997] Fridman-Noy, Natalya, Hafner, Carole D. *The State of the Art in Ontology Design: A Survey and Comparative Review.* AI Magazine, Vol. 18, No. 3, Fall 1997.

[Holm Nelson 1990] Ted Holm Nelson. 'The Right Way to Think About Software Design'. The Art of Human-Computer Interface Design 1990 (Brenda Laurel ed.) pp 235-243. Addison Wesley Publishing Company.

[Kay 1990] Alan Kay. 'User Interface: A Personal View'. The Art of Human Computer Interface Design, Brenda Laurel, ed. Addison-Wesley, New York, 1990

[Kozmetsky 1999] (*Transdiscipline* 1999) George Kozmetsky. Knowledge and Innovation for the 21st Century: Perspectives on Creating Knowledge through Global

Knowledge Partnerships. Proceedings. 3rd International Conference on Technology Policy and Innovation. http://www.ic2.org/08_30_99.doc .

[Lazowska 2001] Ed Lazowska. Bill & Melinda Gates Endowed Chair, Computer Science and Engineering, University of Washington. Computer Science: Still Crazy After All These Years. Science Forum, University of Washington. May 2, 2001. http://programs.researchchannel.com/displayevent.asp?rid=1043

[McIlraith 2001] S. McIlraith, T. Son, H. Zeng. "Semantic Web Services". IEEE Intelligent Systems, March/April 2001.

[Schon 1983] Donald A. Schon. The Reflective Practitioner, How Professionals Think in Action. Basic Books.

[Tervereen et al. 1995] Loren Terveen , Markus Stolze, and Will Hill. "From 'Model World' to 'Magical World': Making Graphical Objects the Medium for Intelligent Design Assistance." A workshop performed at the CHI '95 Conference (Computer Human Interface), a special interest group of the ACM.

[Uschold and Gruninger 1996] Uschold, Mike and Gruninger, Michael. *Ontologies: Principles, Methods and Applications*. Knowledge Engineering Review, Vol. 11, No. 2, June 1996.

[Uschold and Martin 1995] Uschold, Mike and King, Martin. *Towards a Methodology for Building Ontologies*. Workshop on 'Basic Ontological Issues in Knowledge Sharing' held in conjunction with IJCAI-95.

[UCAID 1997] UCAID. *Internet2 Preliminary Engineering Report.* http://www.internet2.edu/html/97engineering.html January 1997

Computer Networking Research Laboratory

Project Investigators:

Hugh M. Smith, Ph.D.
Associate Professor
Computer Science Department

Phillip Nico, Ph.D.
Associate Professor
Computer Science Department

Computer Networking Research Laboratory
C3RP Project #54410
Final Report

Hugh Smith and Phillip Nico

Department of Computer Science
California Polytechnic State University
San Luis Obispo, CA 93407
email: husmith@calpoly.edu
pnico@acm.org

28 January 2003

We proposed to use the C3RP funds to establish the Cal Poly Network Research and Performance Laboratory (CPNRPL), a research laboratory dedicated to the investigation of networking technology in conjunction with external partners from both industry and academia.

Since the inception of the project, we proceeded along three major tracks: the establishment of the laboratory itself, investigation into network performance (latency and throughput), and investigation into network support for quality of service (QoS).

The following sections of this document describe our progress in each of these areas during the period of the grant as well as a brief discussion of our current activities under the continuation grant, project #55380 for the 2002–2003 year.

1 Laboratory Development

At the close of the project in September, 2002, CPNRPL, now renamed the Cal Poly Network Performance Research Laboratory (NetPRL), consisted of three faculty and four full-time undergraduate and graduate researchers. The faculty concentrated on developing the relationships and infrastructire necessary for the lab to succeed and directing the students who concentrated on learning the technology and performing initial experiments.

Second to recruiting researchers, the most pressing need for a research lab is a place to work and equipment to work with. We generated a request for space to provide for a larger work area, but in the interim, we have been able to co-locate with another research project currently under way. Sharing space, while not suitable for the long term, has allowed us to get the students working right away.

Student advising is conducted via twice-weekly meetings as well as ad-hoc consultations in the lab. Because a significant portion of the work at this point is familiarization with the technologies involved, the weekly meetings consist of a seminar segment followed by detailed discussions of current work. The meetings themselves are often conducted via videoconferencing, an important application area for network quality of service work.

The majority of the faculty time has been spent developing relationships with potential industrial partners. We developed a prospectus for the lab and are currently working with three companies—IBM, Intel, and Brocade—to secure support for the future of the lab.

For IBM: During the period of the project, we visited two IBM sites and met with contacts there to determine the appropriate avenue for funding requests. We plan to return to IBM in the future.

For Intel: we met with a representative of Intel Corporation and submitted a written grant proposal to evaluate the performance of their Internet eXchange Architecture (IXA) as well as the performance of our current CiNIC platform in conjunction with the next-generation PCI Express bus technology.

For Brocade: during the summer we met with representatives to make a presentation on our work planned future meetings that took place at Brocade the end of the summer.

In addition to pursuing industrial partners, we also investigated avenues of support available through the National Science Foundation or DARPA. While industrial partners are our primary goal, some small academic sponsorship would open up opportunities for collaboration with other labs in the academic community as well. In addition, the prestige of NSF or DARPA support is, itself, helpful when seeking industrial support.

2 Network Performance Progress

Our efforts in network performance covered three separate areas: development of a testbed network, development of an experimental architecture for evaluation of network co-processors, and measurement of firewall performance.

First, we set up a testbed network in the lab to allow us to study the effects of different latency and bandwidth constraints on network performance. The system consists of two independent subnets connected by a bridge under our control. This bridge is running NistNET, a tool that allows us to simulate network delays, congestion, loss-rates, etc. Now that the network is set up, we intend to use it to investigate the performance of Web100, a package of tools under development to help systems tune their network performance to utilize all the available bandwidth in their current network connections.

Second, we had a student working to develop an experimental platform for evaluating network co-processors. This network processor is implemented using field programmable gate array (FPGA) technology that allows it to easily evaluate the effects of different hardware and software configurations. The full paper is included in Appendix A.

We are also preparing a paper for publication in the Linux Journal, "Performance analysis of the Linux firewall `iptables` in a host," which describes our work to instrument the Linux kernel to monitor the progress of network packets through the network stack. This

instrumentation has allowed us to empirically measure he behavior of the operating system and to call into question an industrial rule of thumb about the number of packet handling rules a firewall can efficiently implement.

3 Network Quality of Service Progress

During the period of the grant, we had one student exclusively devoted to working on network quality of service, focused on videoconferencing technology. Videoconferencing strains best-effort networks, because in order to provide acceptable performance, it requires both high bandwidth and low latency. In addition to studying the relevant literature and setting up a conference testbed in the lab, he developed a Linux implementation if the IETF standard Session Initiation Protocol (SIP) and has studied the underlying Real Time Transfer Protocol to allow him to use is to study QoS issues.

In addition, our semi-weekly video meetings gave first-hand experience with the conditions that generate unacceptable quality of service.

4 Current and Future Developments

In September, 2002, project #54410 ended, but in November it was continued by project #55380. Since that time, the laboratory has continued to develop and currently has five participating faculty members and nine undergraduate and graduate students drawn from the Computer Science and Electrical Engineering Departments. Work continues toward development of the laboratory itself, developing intelligent network interfaces, monitoring network performance, and improving network quality of service.

A Publications

The work under project #54410 resulted in the following two attached papers.
Refereed Publications:

J. Hatashita, J. Harris, H. Smith, and P. Nico, An evaluation architecture for a network coprocessor. In *Proceedings of the 2002 IASTED International Conference on Parallel and Distributed Computing and Systems (PDCS),* Cambridge, Massachusetts, November, 2002.

In Preparation:

A. Melara, J. Harris, H. Smith, and P. Nico, Performance analysis of the Linux firewall `iptables` in a host.
In preparation for submission to *Linux Journal.*

Performance Analysis of Linux Firewall iptables in a Host

Project Investigators:

Hugh M. Smith, Ph.D.
Associate Professor
Computer Science Department

Phillip Nico, Ph.D.
Associate Professor
Computer Science Department

James G. Harris, Ph.D.
Professor
Electrical Engineering Department

Americo J. Melara
Electrical Engineering Department

In preparation for submission to the *Linux Journal*

Performance Analysis of Linux Firewall iptables in a Host

James Harris and Americo J. Melara, Electrical Engineering Department
Hugh Smith and Phillip Nico, Computer Science Department
California Polytechnic State University

July 29, 2002

ABSTRACT

This paper presents a performance analysis of the Linux firewall `iptables` for a single host. As part of the analysis, the sensitivity of the performance to the parameters that determine the functionality of the firewall are determined.

In order to be able to measure the performance and the sensitivity of the firewall, we designed and instrumented each layer of the Linux TCP/IP stack. This instrumentation was used to test the host's firewall under two scenarios: In the first scenario, we captured the path and the latency of one single packet; in the second scenario, we captured the latency of multiple packets sent to the host at various transmission rates.

Our measurement results indicate that the firewall is sensitive to the number of rules and the type of filtering. The results demonstrate that for each type of filtering, latency increases linearly as the number of rules increase. The results also show that the percentage overhead generated by a firewall when a single packet travels the TCP/IP stack ranges from 6% for a rule-set of zero and up to 75% for a rule set of 100 rules.

INTRODUCTION

Firewalls are the first front line defense mechanism against intruders. There are two different goals for testing them. The first goal is to analyze and test the firewall policies, in other words, to model and test how secure a firewall is in a "real-world" environment. The second goal is to test the performance impact generated by the firewall. We decided to analyze the performance cost of having a firewall in the host. After searching for conference papers that addressed the firewall performance on single hosts, we found that very little research had been done on the topic. In our end-of-the-year meeting with 3Com in December 2001, we were told that a third party vendor discovered that the firewall would degrade the performance of a system tremendously after 30 rules. In view of the lack of research and the uncertainty on what the firewall performance cost might be, we decided to study the performance of the Linux firewall `iptables`. This paper presents a study of the sensitivity and the performance impact produced by the Linux firewall `iptables` in a host.

We decided to test the performance of the firewall under two test scenarios changing various parameters. The first test scenario traced one single packet in order to measure the sensitivity of the firewall to: INPUT policy, number of rules, type of filtering, payload size, and transmission

protocol. The second test scenario validated the single packet results by varying the throughput by sending a stream of packets at 5 and 10 Mbps. The first test results document that the performance is only sensitive to the number of rules and the type of filtering. The measurement results obtained in the throughput tests confirm that the single-packet test measurements are valid, and that they serve as conservative estimates for finding the performance impact generated by the firewall.

The remainder of this document is organized as follows: First, previous research and methods used to test firewalls is presented. Then, the background and experimental approach for this work on iptables is presented. The next section presents the results of out testing. Finally, the conclusions are presented.

PREVIOUS RESEARCH

The literature was reviewed for previous work in analyzing the performance of firewalls. Previous work can be partitioned into two areas of interest: performance in firewall routers, and performance in single-host firewalls. Firewall router performance is presented first.

Firewall routers are in charge of filtering and forwarding packets destined for a specific network. Various studies have been made on these types of firewalls, for example, Patton, Doss, and Yurcik [2] compared the performance of open source versus commercial firewalls. They compared the old Linux ipchains included in RedHat version 6.0 against CISCO's IOS firewall, the latter consisting of hardware and software. At the time, the older Linux netfiter (ipchains) had the disadvantage of lacking functionality; it was not a *stateful* firewall while IOS was. In their results they showed that "the Linux firewall has consistently higher transaction throughput rates than the Cisco's stateful firewall for rule sets varying from 0 to 200 rules and for packet sizes of 1 and 128 bytes [2]." No specifics were given on the rule set used.

Other studies measured and compared the *latency* and *total transaction time* to download small and large HTTP and FTP files [3]. The tests setup included several clients inside a LAN connecting to a server outside the LAN and a router firewall sitting in between the networks. The firewall would be configured to 7 different policies, one for each HTTP and FTP test. The clients would run a script to establish the connections. The tests for HTTP and FTP were performed independent from one another. For HTTP tests, the clients made connections to download small sizes of data. On the other hand, for FTP tests the client would make small or large number of connections and download files of either 1MB files during one test or 5MB files in another. Those tests were also independent from one another. The results implied that "the performance difference among security levels due to the overhead of packet filtering for more security is negligible when compared with the outside traffic interface [3]." In other words, performance decreases as the number of connections increase, and is not affected by the security policy. Unfortunately, no specifics were given on the rule-set.

Other tests have been performed to compare commercial router firewalls such as [4] [5] and [6], but the results are not presented in this paper because our work focused on single-host firewall performance.

Different from edge firewall routers, there has not been much research done to analyze the performance or the processing overhead produced by single-host firewalls. One paper presented the results on the throughput and CPU utilization of two machines connected through a 10Mb hub [7]. The purpose of the tests was to measure `iptables` on a single-host. The CPU utilization was measured using "vmstat 3." The sending box sent a byte stream of 187,000,000 bytes. The payload size per packet was 3,700 bytes. The throughput was measured by dividing the size of the bit stream by the time (in seconds) to receive the stream. Finally, the input policies (i.e. INPUT/ OUTPUT/ FORWARD) were set to ACCEPT. The results of the four tests are described below.

On the first test of four tests discussed, without a firewall and with one single connection, the throughput was 9.09 Mbits/sec. The CPU utilization was not provided. The second test running "real-world" `iptables` rules and one single connection showed a 9.10 Mbps and a CPU utilization of 19-23% on the sender and 16-20% on the receiver. The third test included establishing five TCP connections and no rules, in order to measure the CPU impact by TCP/IP traffic. The sum of the throughput was 9.13 Mbps, and the CPU utilization varied from 19-20% on the sender and 15-18% in the receiver. For the fourth and last test, the intention was to "measure real-world stress on the iptables rule-set. Five connections were used: two open TCP ports, a TCP port rejected with a TCP reset, a closed TCP port, and an open UDP port." The CPU utilization on the receiver was 15-20% and 23-30% on the sender. For the UDP component the throughput yielded 10.57 Mbps. For the two non-blocked TCP connections the throughput yielded 8.14 Mbps. For TCP in the latter test, it is understandable that as the amount of filtering and connections increase the throughput might decrease. However, for UDP, having a 10.57Mbps throughput on a 10 Mbps hub is suspect.

BACKGROUND AND EXPERIMENTAL APPROACH

Firewalls have been tested by modifying parameters such as the number of rules, the number of connections, the number of bytes and the transmission rate. Our investigation focuses on analyzing and testing the sensitivity of the firewall, and the performance impact generated by it, by varying a set of *external* and *internal* parameters presented in Figure 1. External parameters are those that cannot be controlled by the firewall such as *transmission protocol, transmission speed,* and *payload size.* Internal parameters are those that can be controlled by the firewall such as *Input policy, filtering type,* and *number of rules.*

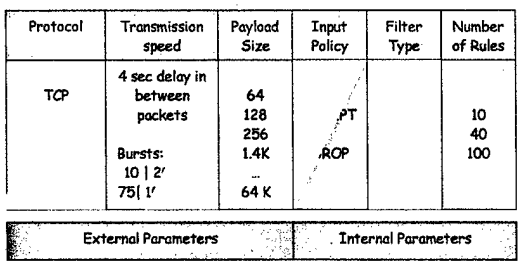


Figure 1 Parameters to determine the sensitivity of the firewall

A brief description of the `iptables` algorithm will help to understand the definition of the tests performed. A flow chart of the `iptables` algorithm is presented in Figure 2. Notice what

happens when the netfilter's framework `NF_IP_LOCAL_IN` hook is called in Figure 2. `Iptables` executes `ipt_do_table`, which then executes `ipt_packet_match`, Figure 3.

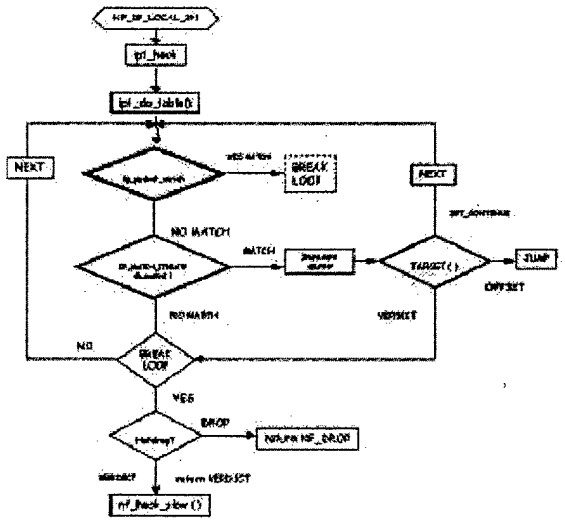


Figure 2 Firewall Algorithm – ipt_do_table( ) checks for matches in the rule set

Observe in Figure 3 that the firewall checks the source and destination IP address first. If no match is found, it tries to find a match for the input device, then check for the output interface device, then the protocol, and finally it checks if the packet is a fragment. If a match is not found, `ip_packet_match` will return 0, continue to the next rule, or break out of the loop.

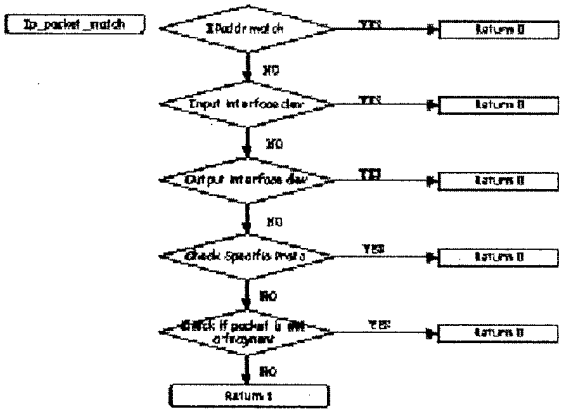


Figure 3 ip_packet_match function – IP addresses are always checked first regardless of the type of filtering

After passing through the `ip_packet_match` and finding no matches, the next step is to execute the `IP_MATCH_ITERATE`, in Figure 3. Here, the firewall calls the `do_match` function pertaining to the specific rule. Every module has a specific do match function. In other words, if we are filtering/matching a MAC address the `iptables` algorithm will call the `do_match` function specific to MAC addresses. If a match is found, the chain breaks to perform

a TARGET check. Targets can be ACCEPT, DROP, QUEUE, STOLEN, REPEAT or "JUMP" to another chain when a chain has been added to the rule-set. If `IP_MATCH_ITERATE` does not find a match it will either continue to the next rule or exit the loop. Iptables breaks out of the loop when it finds a match, when the packet is a fragment, or when all the rules have been checked.

In summary, the firewall will always go through the `ip_packet_match` function regardless of the type of matching. For example, every rule that filters TCP ports includes checking for IPs, interfaces, protocol, fragments, and at last matching the TCP port.

The tests were performed for two different scenarios: a single-packet test, and a throughput test. The first scenario consisted on sending a single packet every 4 seconds and analyzing the latency as it traversed the stack. The second test consisted on sending a stream of identical packets at different rates to the host using the SmartBits network testing system. The purpose of the throughput test scenario was to verify the results of the single-packet test for the case of multiple packet inputs with different input rates. In order to take the measurements the Linux TCP/IP stack was modified, adding timestamping instrumentation from the Data Link to the Socket layer.

In order to implement any timestamp instrumentation it became necessary to study the path that an incoming packet follows in the stack. This path, shown in Figure 4, was traced from the data link layer up to the point where the socket layer hands the data to the application. The symbols in Figure 4 represent the following:

< > enter and exit function ( )
>>> enter function ( )
<<< exit function ( )

The path depicted in Figure 4 served as the basis for our analysis because all TCP and UDP packets destined for the host follow the path outlined in this figure.

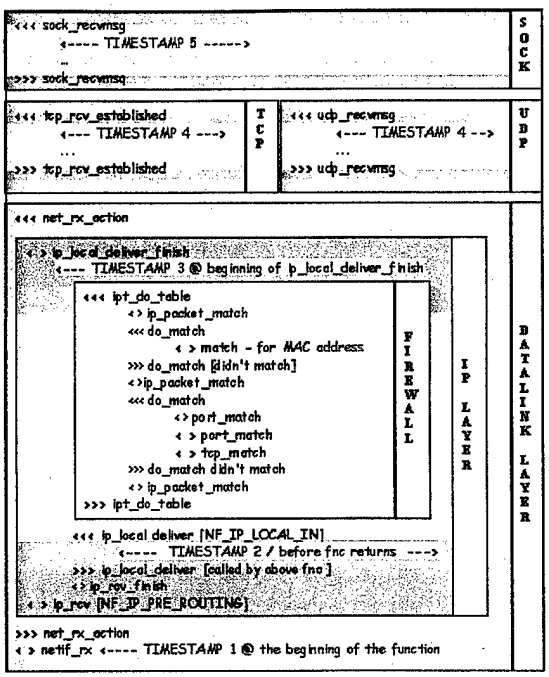


Figure 4 Instrumentation in the Linux network stack - from the Data Link up to the Socket layer

Initially, as a packet comes in from the physical layer it causes the Ethernet device to "fire up" an interrupt. Interrupts are handled by top-halves and bottom-halves [8]. The top-half is handled by the network adapter's device driver (e.g. 3c59x.c). The device driver calls the `eth_type_trans()` function located in the *eth.c* file. This function organizes the first part of the packet header (i.e. MAC header) inside an `sk_buff` structure.

When the `eth_type_trans` function returns the device driver calls the device controller (i.e. `netif_rx`) located inside the *dev.c* file. This file controls all the network device drivers and it is located in the `usr/src/linux/net/core/` directory. Two main functions separate the top-half from the bottom-half: `netif_rx` and `net_rx_action`, respectively.

After the top-half executes, the *swapper* will be in charge of running the bottom-half. Note that it is the *swapper* and not the *scheduler* who handles this operation. The difference between the *swapper* and the *scheduler* is that the swapper is in charge of completing the execution of the pending bottom-halves [8] and the latter is in charge of handling processes.

The `netif_rx` function takes a timestamp by calling the `get_fast_time(&skb->stamp)` function. This timestamp serves as a unique ID for each packet. This packet ID is transferred throughout the entire stack inside the `skb` structure, serving as a mean to match the measurements of a specific packet at the data link layer, with the other measurements. After the top-half executes, the swapper schedules the bottom-half which starts with `net_rx_action`.

Single-packet tests are performed using two PCs. Volans, our Device Under Test (DUT), is a dual 450MHz Intel Pentium processor with a modified 2.4.7 kernel running the server application. One of the CPUs is turned off in the SMP option of the kernel configuration. The version of `iptables` used was `1.2.4`. The files with the timestamp instrumentation are: *dev.c*, *ip_input.c*, *tcp_input.c*, *udp.c*, and *socket.c*. At boot time, both machines will start in run-level 3. During the tests, all the services are shutdown. Libra, the client, is a 233MHz Pentium II processor. Both machines are isolated from any outside traffic and connected through a 100 Mbps 3Com switch. Both machines used a 10/100 Mbps 3Com NIC, model 3C905C. Refer to Figure 5 to see the test bed.

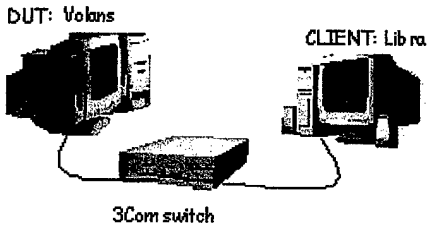


Figure 5 Test setup to measure the latency when a single packet is sent every 4 seconds

The parameters under test for the single-packet tests environment are shown in Table I. These parameters included the transmission protocol, connection speed, payload size, number of rules, type of filtering, and the INPUT policy.

Table I Parameters under test for the single-packet test scenario

Generic Test Setup	Parameters under test	
Transmission Protocol	TCP	UDP
Type of filtering/matching	TCP, IP, MAC	UDP, IP, MAC
INPUT policy	ACCEPT & DROP	DROP
Connection speed	100Mbps	
Payload size	64 &1400 bytes	
Number of rules	No firewall,10, 40, 100	

Tests were performed for both TCP and UDP protocols. The connection speed was 100 Mbps in an isolated system, sending one packet every 4 seconds. The payload size varied between 64 and 1400 bytes. The type of filtering was IP and MAC for both protocols, and TCP and UDP for each respective transmission protocol. The number of rules under each type of filtering was zero (or No Firewall), 10, 40, and 100 rules. Both INPUT policies ACCEPT and DROP were tested for TCP, but only the INPUT policy DROP was tested for UDP because, as it will be shown, the INPUT policy does not impact the performance of the firewall.

A total of 40 packets or samples on one single test were sent to the host. The results were accessed via the `/proc/` file system. Three tests were performed for each type of filtering, from which we took the median of the total samples to exclude any outliers. The medians were averaged for a final result.

Table II Difference in the total processing time for INPUT policies ACCEPT and DROP – the firewall is not sensitive to the INPUT policy

Number of rules - (payload size)			[units: μs]
	INPUT policy - [T5 - T1]		
IP matching	Accept	Drop	Acc - Drop
10 rules - (64 bytes)	29.94	29.92	0.02
10 rules - (1400)	41.14	40.77	0.38
40 rules - (64)	34.12	34.00	0.12
40 rules - (1400)	45.00	44.90	0.10
	INPUT policy - [T5 - T1]		
MAC matching	Accept	Drop	Acc - Drop
10 rules - (64)	35.66	35.23	0.43
10 rules - (1400)	47.00	46.57	0.43
40 rules - (64)	57.13	55.41	1.72
40 rules - (1400)	68.67	66.84	1.83
	INPUT policy - [T5 - T1]		
TCP matching	Accept	Drop	Acc - Drop
10 rules - (64)	35.93	36.00	0.07
10 rules - (1400)	47.02	47.30	0.28
40 rules - (64)	54.48	54.73	0.25
40 rules - (1400)	65.66	65.92	0.26

Timing the network stack

To assess the impact of the firewall on the overall performance of the network stack, it is useful to observe the overall latency time using a plot of the data obtained for all the timestamps in every layer. Figure 5 shows the timestamps of T2 to T5 with respect to T1 for a particular test case with a firewall with no rules imposed. Referring to Figure 4 for reference, T2-T1 or T3-T1 without a firewall can be considered the time spent in the IP layer, the time difference between T3-T1 and T2-T1 represents the time that it takes for the firewall to execute, the time difference T4-T3 represents the TCP or UDP layer latency time, and the time difference T5-T4 represents the socket layer latency time. The time difference T5-T1 is the overall network stack latency time. Note that the data in Figure 5 shows that even when there are no rules imposed for the firewall, that there is a small penalty in the overall network stack performance with the firewall installed – approximately 1.6 us compared to 28.58 and 39.64 us, or less than 6% of the total latency time (an explanation for this is given later). We now discuss each of these latencies and then conclude with a discussion of the overall latency.

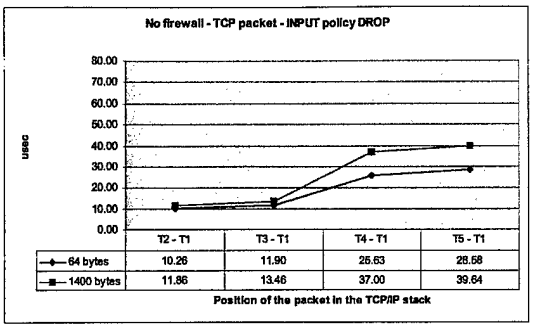


Figure 5 Latency increases as the payload size increases

T2 – T1

The results for TCP and UDP in Table III show that the difference between T2 – T1 increases as the payload size increases. For example, compare the averages for 64 bytes with the averages for 1400 bytes. The reason for this is because the packet is copied from the network into kernel space.

Table III Payload impact in T2-T1 – time increases as the payload size increases

T2 - T1 (units ... us)

	IP	TCP	TCP		IP	UDP	UDP
64 bytes	T2 - T1	T2 - T1	T2 - T1	64 bytes	T2 - T1	T2 - T1	T2 - T1
No firewall	10.26	10.26	10.36	No firewall	10.48	10.48	10.48
10 rules	10.46	10.23	10.54	10 rules	10.45	10.50	10.45
40 rules	10.58	10.59	10.26	40 rules	10.51	10.60	10.61
100 rules	10.72	10.64	10.56	100 rules	10.66	10.55	10.65
Average	10.51	10.43	10.43	Average	10.52	10.53	10.55
1400 bytes				1400 bytes			
No firewall	11.86	11.86	11.86	No firewall	12.34	12.34	12.34
10 rules	12.02	11.90	11.94	10 rules	12.38	12.27	12.36
40 rules	12.05	12.10	11.92	40 rules	12.39	12.43	12.53
100 rules	12.31	12.15	12.30	100 rules	12.42	12.48	12.54
Average	12.06	12.00	12.06	Average	12.38	12.38	12.44

T4 – T3

At the TCP and UDP layers, the latter is processed faster than the TCP layer because of the nature of the complexity of their algorithm. However, the time to process the layers is influenced by the payload size because the data is copied from kernel space to user space. For example, the results in Table IV demonstrate that the average time to process 64 bytes of payload is shorter than 1400 bytes of payload.

Table IV Impact of the payload size in T4 – T3 – time increases as the payload size increases

		T4 - T3					
		TCP					UDP
64 bytes	ID	Max	UDP	64 bytes	ID	Max	UDP
No firewall		18.74	18.74	18.74	No firewall	9.43	9.43	9.43
	10	15.89	14.14	14.20	10	9.48	9.40	9.42
	40	14.92	14.68	14.72	40	9.63	9.95	9.89
	100	14.57	15.02	14.81	100	10.09	10.25	10.10
Average		14.13	14.39	14.37	Average	9.65	9.75	9.72
		TCP					UDP
1400 bytes	ID	Max	UDP	1400 bytes	ID	Max	UDP
No firewall		23.54	23.54	23.54	No firewall	19.46	19.46	19.46
	10	23.65	24.07	24.22	10	19.46	19.35	19.23
	40	24.00	24.86	24.67	40	19.48	19.83	19.75
	100	24.39	25.07	24.62	100	19.90	20.02	20.05
Average		23.89	24.38	24.26	Average	19.57	19.66	19.62

T5 – T4

Different from T2 – T1 and T4 – T3, the socket layer latency time exhibits random values independent of the payload size. Two tests were performed in order to study the time to process the socket layer with respect to the size of the payload. The tests did not include a firewall. Figure 6 shows that the time to process this layer is not dependent on the payload size. The socket layer is a process controlled by the scheduler, and therefore is executed at times independent of the payload size.

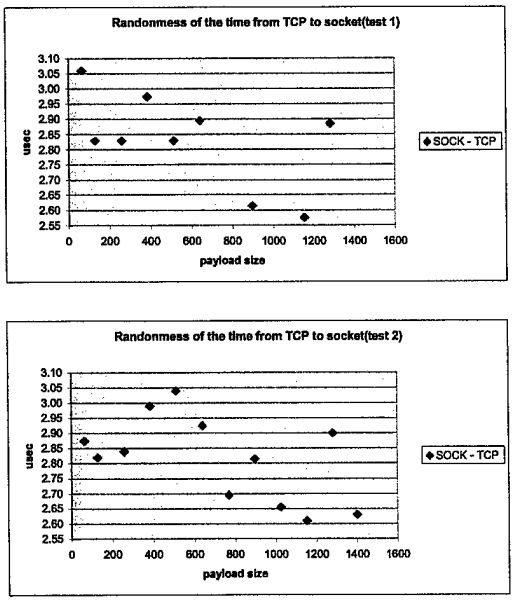


Figure 6 Randomness at the socket layer –socket layer is called randomly regardless of the payload size

Firewall Performance T3-T2

We now study the firewall. First, we focus on the impact that the payload size has on the values of the latency time for the firewall, T3-T2. Reviewing the data in Tables V through VI, it is observed

that the difference in the latency times between the 64 byte and the 1400 byte payloads is less than 0.27 us. Therefore, this leads us to conclude that the impact of the size of the payload on the latency timing for the firewall is negligible.

Next, we focus on the effect that the number of rules has on the performance of the host. Notice, in the Tables V through VII that as the number of rules increases the difference between T3 – T2 also increases. Subsequently, the firewall is sensitive to the number of rules.

Notice in the tables that for "No firewall," the results for T3 – T2 is non-zero. This can be explained with the `iptables` algorithm because, as depicted in Figure 2, when the netfilter hook NF_IP_LOCAL_IN is called, the function `ipt_hook` is executed. This latter will return a call to `ipt_do_table` which checks the iptables rule-set. If no rules are found the function will exit `ipt_do_table` and `ipt_hook` and make a final call to `ip_local_deliver_finish`. This process will take between 1.60 to 1.90 microseconds.

Table V Matching IP – time increases as the rules increase

TCP packets	IP matching [units: µs]			UDP packets	IP matching [units: µs]		
64 bytes	T2 - T1	T3 - T1	T3 - T2	64 bytes	T2 - T1	T3 - T1	T3 - T2
No firewall	11.26	11.90	1.64	No firewall	10.48	12.32	1.84
10 rules	11.46	13.10	2.64	10 rules	10.45	12.29	1.84
40 rules	11.58	16.31	6.22	40 rules	10.51	17.91	6.39
100 rules	11.72	24.66	13.94	100 rules	10.66	24.72	14.06
1480 bytes	T2 - T1	T3 - T1	T3 - T2	1480 bytes	T2 - T1	T3 - T1	T3 - T2
No firewall	11.86	13.46	1.60	No firewall	12.34	14.21	1.87
10 rules	11.92	14.66	2.57	10 rules	12.38	15.57	1.18
40 rules	11.85	18.45	6.60	40 rules	12.39	19.12	6.72
100 rules	12.31	26.27	13.94	100 rules	12.42	27.10	14.68

Table VI Matching MAC addresses – time increases as the rules increase

TCP packets	MAC matching [units: µs]			UDP packets	MAC matching [units: µs]		
64 bytes	T2 - T1	T3 - T1	T3 - T2	64 bytes	T2 - T1	T3 - T1	T3 - T2
No firewall	10.26	11.90	1.64	No firewall	10.48	12.32	1.84
10 rules	10.23	18.19	7.96	10 rules	10.58	19.59	9.01
40 rules	10.39	37.94	27.55	40 rules	10.68	39.41	28.73
100 rules	10.44	80.36	69.92	100 rules	10.55	80.65	70.10
1480 bytes	T2 - T1	T3 - T1	T3 - T2	1480 bytes	T2 - T1	T3 - T1	T3 - T2
No firewall	11.86	13.46	1.60	No firewall	12.34	14.21	1.87
10 rules	11.90	19.95	8.06	10 rules	12.27	21.49	9.22
40 rules	12.10	39.49	27.40	40 rules	12.43	41.75	29.32
100 rules	12.15	82.08	69.93	100 rules	12.48	82.86	70.38

Table VII Matching TCP ports – time increases as the rules increase

TCP packets	TCP filtering [units: µs]			UDP packets	UDP matching [units: µs]		
64 bytes	T2 - T1	T3 - T1	T3 - T2	64 bytes	T2 - T1	T3 - T1	T3 - T2
No firewall	10.26	11.90	1.64	No firewall	10.48	12.32	1.84
10 rules	10.54	19.04	8.51	10 rules	10.45	19.16	8.71
40 rules	10.26	37.19	26.93	40 rules	10.61	38.67	28.06
100 rules	10.56	78.46	67.90	100 rules	10.68	78.56	67.89
1480 bytes	T2 - T1	T3 - T1	T3 - T2	1480 bytes	T2 - T1	T3 - T1	T3 - T2
No firewall	11.86	13.46	1.60	No firewall	12.34	14.21	1.87
10 rules	11.94	20.53	8.59	10 rules	12.36	21.33	8.97
40 rules	11.92	38.81	26.89	40 rules	12.52	41.30	28.78
100 rules	12.30	80.28	67.98	100 rules	12.54	80.69	68.16

It should be noted from the data in Tables V, VI, and VII that the latency values T3-T2 for the TCP and the UDP protocol test cases differ by at most 2us. Thus, one can conclude that the

performance of the firewall is not very sensitive to the protocol of the packet (TCP or UDP) that is being processed.

The plot of the data just presented shows a linear relationship between the performance impact and the number of rules. Figures 7 and 8 present the T3 – T2 time latencies for TCP and UDP using the data obtained for packets of 64 bytes of payload. They also present a set of equations that may serve to estimate the time to process T3 – T2 up to 100 rules. It also is noted that there is not a significant difference between matching on the TCP and the MAC parameters when compared to matching on the IP parameter.

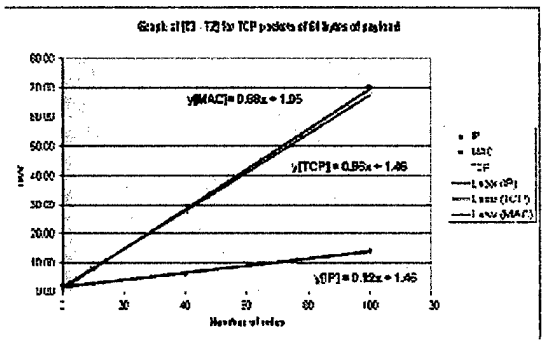


Figure 7 TCP connection – [T3 – T2] – linear relationship between the number of rules and the time to process the firewall

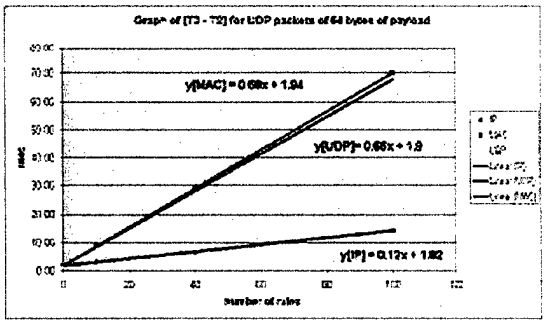


Figure 8 UDP connection – [T3 – T2] - linear relationship between the number of rules and the time to process the firewall

Impact of T3 – T2 on the total processing time T5 – T1

The total processing time (T5 –T1) is expected to be smaller for TCP packets than for UDP packets; refer to the T4 – T3 section presented earlier. The results in the Table VIII confirm that the number of rules directly affects the total processing time. In addition, it shows that the payload size also impacts the total processing time (e.g. compare "No firewall" for 64 and 1400 bytes.)

Table VIII TCP and UDP – Difference in Total processing time [T5 – T1] for three types of filtering rules

		T5 - T1		[units: µs]			
TCP PACKETS				UDP PACKETS			
	IP	MAC	TCP		IP	MAC	UDP
64 bytes	T5 - T1	T5 - T1	T5 - T1	64 bytes	T5 - T1	T5 - T1	T5 - T1
No firewall	28.58	28.58	28.58	No firewall	24.46	24.46	24.46
10 rules	29.92	35.23	35.00	10 rules	25.81	31.81	31.42
40 rules	34.00	55.41	54.73	40 rules	29.46	52.25	51.48
100 rules	41.99	98.28	96.90	100 rules	37.64	93.87	91.75
1400 bytes				1400 bytes			
No firewall	39.64	39.64	39.64	No firewall	36.90	36.90	36.90
10 rules	40.77	46.87	47.30	10 rules	38.37	44.17	43.78
40 rules	44.90	66.84	65.92	40 rules	41.91	64.97	64.45
100 rules	53.09	109.47	107.38	100 rules	50.37	106.32	103.99

The impact of the firewall latency T3 – T2 on the overall network stack latency T5-T1 can be expressed in terms of percentage overhead produced by the firewall:

$$\text{Firewall's \% overhead} = (T3 - T2)/(T5 - T1) * 100$$

The results in Tables IX through XI show that: (1) as the number of rules increases the percentage overhead increases - up to 75% for UDP and 71% for TCP; (2) as the payload size increases for a specific number of rules, the percentage overhead decreases – this is because the firewall is not sensitive to the payload size, consequently, an increase in payload will increase T5–T1 while T3–T2 will remain the same; and (3) MAC and TCP matching percentage overhead are larger than IP matching overhead, and there is little difference between MAC and TCP matching percentage overhead.

Table IX Percentage overhead of IP matching over the T5 – T1 – overhead increases as the number of rules increase

TCP PACKETS				UDP PACKETS			
IP matching				IP matching			
64 bytes	T5 - T1	T3 - T2	% overhead	64 bytes	T5 - T1	T3 - T2	% overhead
No firewall	28.58	1.64	6%	No firewall	24.46	1.84	8%
10 rules	29.92	2.44	9%	10 rules	25.81	3.13	12%
40 rules	34.00	6.32	19%	40 rules	29.46	6.38	22%
100 rules	41.99	13.94	33%	100 rules	37.64	14.36	37%
1400 bytes				1400 bytes			
No firewall	39.64	1.60	4%	No firewall	36.90	1.97	5%
10 rules	40.77	2.63	6%	10 rules	38.37	2.18	6%
40 rules	44.90	6.30	14%	40 rules	41.91	6.73	16%
100 rules	53.09	13.96	26%	100 rules	50.37	14.63	29%

Table X Percentage overhead of MAC matching over the T5 – T1 - overhead increases as the number of rules increase

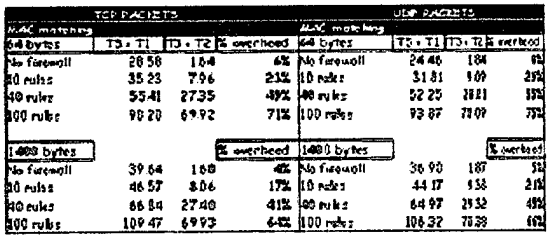
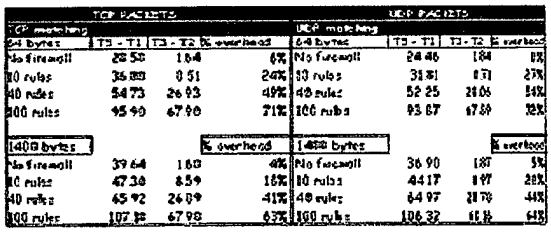


Table XI Percentage overhead of TCP matching over T5 – T1 - overhead increases as the number of rules increase



Throughput Tests to Validate Single Packet Results

The above conclusions are based upon single packet test results. To determine the impact of the incoming throughput rate on these conclusions, throughput tests were performed using the Spirent's Network Tester "SmartBits". These tests show the latency of the network stack when multiple packets are sent at different transmission rates. The tests were performed in the Cal Poly Cisco lab. A Windows 95 PC is connected via a Patch panel to control the SmartBits 2000. The SmartBits cards, model ML-7710, were connected via the patch panel to a Cisco 2900 XL switch to communicate with our DUT, Volans. The test setup is shown in Figure 9.

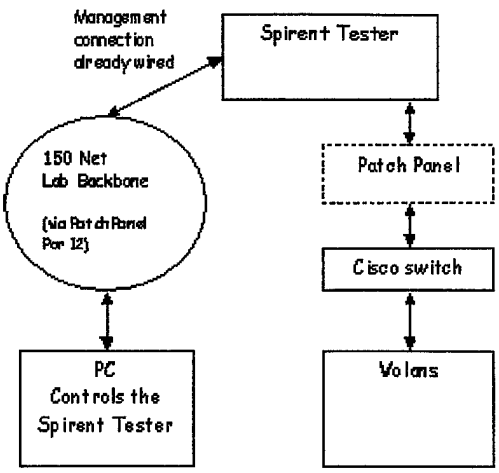


Figure 9 SmartBits testbed

Two SmartBit cards were connected to the switch, one is to send the stream of test packets at different rates, and the other is used to send 2 packets to port 6789 which serves to reset the memory buffers where the timestamps are stored. During the tests, the SmartBit's "Smart Window" was used to allow the test to run for one minute before the timestamp measurements were taken so that the system could reach steady-state before taking any data. The DUT would count the number of incoming packets until a minute had elapsed, and then a total of 4000 timestamps were stored in the pre-allocated memory buffers. The test scenario addressed two limitations: SmartBits only supported UDP packets, and the timestamp instrumentation overhead prevented input rates greater than 12 Mbps before the system lost interrupts and failed. In spite of these two limitations, we felt that the conclusions of the single packet testing presented above supported valid throughput testing. The parameters under test are shown in Table XII. The tests provided for two types of filtering, IP addresses and MAC addresses (the two basically distinct cases for matching). The number of rules used was zero, or no firewall, and 100 rules (relying on the linearity of the performance to the number of rules). Even though SmartBits supported 100 Mbps throughput, only tests for 5 and 10 Mbps were performed because of the timestamping instrumentation limitation. Tests were performed without the instrumentation and 100% link utilization at 100 Mbps could be reached without any loss of interrupts even when filtering 100 MAC addresses. This latter test is very important because it eliminates the possibility that the firewall is the cause of the interrupt loss; rather, the loss is due to the instrumentation overhead. Nevertheless, it is felt that the data results from the two cases of 5 and 10 Mbps are sufficient to validate the conclusions derived from the single packet test cases.

Table XII Parameters under test for the Throughput tests

Generic Test Setup	Parameters under test
Transmission Protocol	UDP
Type of filtering/matching	IP, MAC
INPUT policy	DROP
Throughput / transmission rates	5 & 10 Mbps
Payload size	64 bytes
Number of rules	No firewall & 100

The results in Table XIII showed that as the input throughput rate increased from 5 to 10 Mbps, the network latency time T5-T1 decreased.. In other words, as the throughput increases, the latency decreases.

Table XIII Difference in the latency for various throughput – latency decreases as the throughput increases

SINGLE-PACKET every 4 seconds [units: us]				
64 BYTES	T2 - T1	T3 - T1	T4 - T1	T5 - T1
No firewall	10.48	12.32	21.73	24.46
100 rules IP	10.66	24.72	34.81	37.64
100 rules MAC	10.55	80.65	90.90	93.87
MULTIPLE PACKETS - 5 Mbps				
64 BYTES	T2 - T1	T3 - T1	T4 - T1	T5 - T1
No firewall	11.26	12.46	18.52	20.40
100 rules IP	11.61	20.27	27.07	29.15
100 rules MAC	12.35	77.08	84.76	87.14
MULTIPLE PACKETS - 10 Mbps				
64 BYTES	T2 - T1	T3 - T1	T4 - T1	T5 - T1
No firewall	11.06	12.15	17.67	19.51
100 rules IP	11.84	20.03	26.27	28.29
100 rules MAC	12.03	76.30	82.66	84.95

The data in Table XIII was plotted in Figure 10. Note that the single-packet tests show the highest latency as measured by T5-T1, and that its smallest latency occurs when the input throughput rate is 10 Mbps. Consequently, this data shows that the single-packet measurement results may serve as a conservative upper bound to estimate the time to process the packets by the stack.

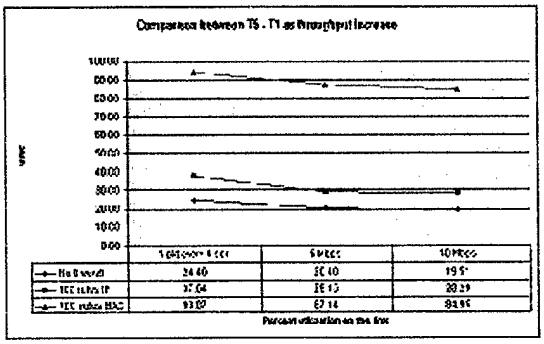


Figure 10 Comparison between T5-T1 for different transmission rates – latency decreases as the throughput increases

The time that a packet is processed by the firewall, the UDP layer, and the Socket layer during the throughput test is shown in Table IXV.. Notice in the table that between T3-T2 (i.e. the firewall) and T4-T3 (i.e. the UDP layer) the packet is processed faster as the throughput increases. On the other hand, this is not the case for T5-T4 (i.e. the socket layer) where the time to process this layer is a random constant value, lying between 2 and 3 μs regardless of the input throughput rate.

Table IXV Time that a packet is held on each layer

SINGLE-PACKET every 4 seconds [units: us]			
64 BYTES	T3 - T2	T4 - T3	T5 - T4
No firewall	1.84	9.42	2.73
100 rules IP	14.06	10.09	2.83
100 rules MAC	70.09	10.25	2.97
MULTIPLE PACKETS 5 Mbps			
64 BYTES	T3 - T2	T4 - T3	T5 - T4
No firewall	1.20	6.06	1.88
100 rules IP	8.67	6.80	2.08
100 rules MAC	64.73	7.69	2.37
MULTIPLE PACKETS 10 Mbps			
64 BYTES	T3 - T2	T4 - T3	T5 - T4
No firewall	1.09	5.52	1.85
100 rules IP	8.19	6.24	2.02
100 rules MAC	64.28	6.35	2.30

CONCLUSIONS

The goal of this research was to study the sensitivities and the performance impact of the Linux firewall `iptables` in a host. We placed timestamps throughout the TCP/IP stack of a host PC running Linux version 2.4.7. To collect accurate data from our instrumentation, we analyzed the path that an incoming packet follows in the stack. With each timestamp, we were able to measure the latency of a packet as it traversed the entire network stack. We used a single packet test scenario to find the sensitivities of the firewall, and then used a input throughput rate test to validate the single packet test conclusions. The results obtained showed the firewall processing latency was not sensitive to the INPUT policy (ACCEPT or DROP), was not sensitive to the payload size (40 or 1400 bytes), was slightly sensitive to the transmission protocol (maximum difference of 2 us between TCP and UDP). We found the firewall to be sensitive to the type of filtering and the number of rules. When filtering IP addresses, TCP/UDP ports, and MAC addresses the cost per rule increases linearly and its cost is approximately 0.12, 0.66, and 0.68 µs/rule, respectively. We were able to explain the difference in the performance cost between IP and the other types of filtering through the `iptables` algorithm. The linear relationship between the latency time for the firewall and the number of rules for the linux `iptables` netfilter firewall indicates that it does not exhibit a threshold effect with the number of rules such as reported by 3Com for another vendor's firewall product. Also, our results showed that the percentage overhead generated by a firewall for the latency time of a single packet traveling the network stack will vary between 6% to 75% as the number of rules vary from 0 to 100, respectively.

Acknowledgement: This work was performed at Cal Poly within the Network Performance Laboratory (NetPrL), and was supported by 3Com; please visit www.ee.calpoly.edu/3comproject for more information.

REFERENCES

[1] James Fischer. "CiNIC – Calpoly Intelligent NIC," California Polytechnic State University, Master's Thesis, San Luis Obispo.

[2] Samuel Patton, David Doss, William Yurcik. 2000. "Open Source versus Commercial Firewalls: Functional Comparison." Proceedings of the 25th Annual IEEE Conference on Local Computer Networks (LCN '00)

[3] Michael R. Lyu, Lorrien K.Y. Lau. 2000. "Firewall Security: Policies, Testing and Performance." Proceedings of the 24th Annual International Computer Software and Applications Conference (COMPSAC '00).

[4] Molitor, Andrew. (n.d.). "Measuring Firewall Performance." Network Systems Corporation. <http://web.ranum.com/pubs/fwperf/molitor.htm>.

[5] E Testing Labs. October 2001. "P-Cube SE1000: Layer 4 and Layer 7 Performance Tests." <http://www.etestinglabs.com/main/reports/pcube10_01.pdf>. Accessed May 2002.

[6] [Anonymous]. July 1999. Network World Fusion: "Performance tests turn up big differences." <http://www.nwfusion.com/reviews/0719fireperf.html>. Accessed April 2002.

[7] [Anonymous]. (n.d.). "Iptables Performance." <http://industrial-linux.org/mlug/2001-10-13/iptables_thruput.txt>. Accessed May 2002.

[8] D. P. Bovet and M. Cesati. 2001. "Understanding the Linux Kernel." O'Reilly. ISBN 0-596-00002-02.

[9] James P. Anderson, Sheila Brand, Li Gong, Thomas Haiigh. September/October 1997. "Firewalls: An Expert Roundtable." September/October 1997. IEEE Software Magazine Vol 14, No. 5: 60-66.

An Evaluation Architecture for a Network Coprocessor

Project Investigators:

Hugh M. Smith, Ph.D.
Associate Professor
Computer Science Department

Phillip Nico, Ph.D.
Associate Professor
Computer Science Department

James G. Harris, Ph.D.
Professor
Electrical Engineering Department

Jason Hatashita
Electrical Engineering Department

Appearing in: *Proceedings of the 2002 IASTED International Conference on Parallel and Distributed Computing and Systems (PDCS)*, Cambridge, Massachusetts, November, 2002.

An Evaluation Architecture for a Network Coprocessor*

Jason Hatashita and James Harris
Department of Electrical Engineering
California Polytechnic State University
San Luis Obispo, CA 93407
email: jhatashi@alumni.calpoly.edu
jharris@calpoly.edu

Hugh Smith and Phillip L. Nico
Department of Computer Science
California Polytechnic State University
San Luis Obispo, CA 93407
email: husmith@calpoly.edu
pnico@acm.org

Abstract

This paper outlines research currently being conducted by the Cal Poly Intelligent Network Interface Card (CiNIC) project to develop a network coprocessor. The purpose of this coprocessor is to free the host machine from its network processing duties as well as to allow for additional functionality such as hardware-based firewalling or quality of service (QoS) support.

We provide an overview of the current CiNIC architecture as well as an introduction to and evaluation of the next generation CiNIC architecture. Our evaluation consisted of analyzing the performance and capabilities of an FPGA processor in order to determine whether it will meet our future development requirements. The FPGA's performance was tested by timing the execution of the uClinux TCP/IP stack during send operations. The processor's capabilities were tested by adding custom logic to the system and interfacing it with the uClinux operating system. We determined that both the performance and flexibility of the FPGA make it an ideal next generation CiNIC architecture.

Keywords: Network Interfaces, Network Performance, Reconfigurable Architecture, Network Coprocessor

1 Introduction

In recent years, computer systems have become more and more reliant on networks in the performance of their most common tasks. The increase in network communications has brought with it a corresponding increase in the computational burden on the processor to support them. At the same time, microprocessors have become less expensive and more powerful leading many to consider the potential benefits of adding a dedicated network coprocessor to free the host machine's resources for more important work. This goal applies equally to true parallel architectures[8] and networked workstations[7].

The goal of the Cal Poly Intelligent Network Interface Card (CiNIC) project is to offload the network functions from a host machine onto a dedicated network coprocessor. In this case, we define network functions to be those processes that manage the movement of data to and from the network, including TCP/IP stack processing, firewalls, routing, and Quality of Service (QoS) features. Our current development architecture consists of a x86 Linux host machine and a STRONG-ARM Linux coprocessor which sits in the host's PCI bus and appears to the host to be a simple PCI device. The system architecture is shown schematically in Figure 1.

In spite of its simple appearance, the CiNIC coprocessor is a complete linux system in itself. Running a full operating system on our coprocessing board provides developers with a familiar, easy-to-use, environment and affords access to industry-proven software for the coprocessor's network functions. The flexibility of a microprocessor and full operating system on the coprocessor system allows for experimentation with any of the host's software routines. Rapid software prototyping and easy access for performance evaluation make this architecture a powerful and flexible platform for exploring the capabilities of a network coprocessor.

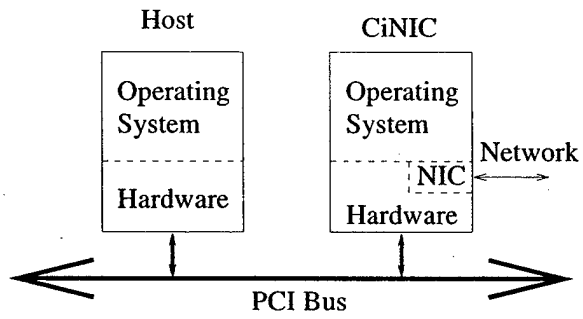


Figure 1. Logical Diagram of CiNIC Architecture

The only major limitation of the current architecture is that its hardware configuration is fixed. Because the demands placed on a network processor may differ significantly from those on a general purpose processor, design decisions based on the expected task-set of one may not be applicable to the other. That is, there might be significant opportunities to improve performance or functionality of the CiNIC if it were possible to tune the hardware architecture.

The next generation architecture for CiNIC intends to incorporate programmable logic into the design in or-

*The work described in this paper was partially supported by the Department of the Navy, Office of Naval Research.

der to explore the benefits of architectural flexibility. One such technology that is well suited to the task is Altera's NIOS processor[1, 2]. This is a 16 or 32 bit microprocessor and peripherals that can be programmed onto a Field Programmable Gate Array (FPGA). This technology provides a processor, peripherals, and custom logic all on one chip. Another reason we chose to evaluate the NIOS processor was the availability of the Microtronix Linux Development Kit[6]. The Microtronix LDK contains hardware expansions for an existing NIOS development board and a port of the uClinux operating system. This provided us an FPGA development board running uClinux that we used to test the performance and features of the NIOS technology.

This paper is an overview of the current CiNIC architecture and discusses our experience evaluating a new programmable logic-based hardware architecture. The first section is a brief overview of the hardware and software architecture of the current CiNIC system. The second section is a discussion of the goals for the next generation CiNIC architecture. The third is brief overview of the NIOS technology and Microtronix uClinux. The final section presents the results of some experiments used to evaluate the capabilities and performance of the NIOS system.

2 Overview of CiNIC Architecture

The goal of the CiNIC architecture is to build a network coprocessing engine that is completely transparent to the user. The current CiNIC architecture consists of a coprocessing system mounted in a host computer. The coprocessing system is running a full Linux 2.4 operating system which provides a software development environment, a TCP/IP stack, and services such as a firewall and QoS support. It is this complete Linux implementation that makes our coprocessing scheme different from others. Instead of the host operating system processing a network function, the parameters for that function call are packaged up and shipped to a dedicated coprocessor for execution. The results are propagated back to the host system after the network function is completed[4].

The intent of offloading the entire TCP/IP stack is to save valuable host processing cycles by allowing network functions to be executed in a coprocessor. In addition to processing savings, one can add network functionality such as firewalling, QoS, and traffic shaping without adding a processing load to the host. Similar to the way that video cards began to offload video processing, the CiNIC architecture takes steps to providing offloaded network functions.

The original CiNIC architecture began with just a simple coprocessor board, but evolved when it was realized that running a second operating system within the coprocessing system provided many benefits. The key to running two operating systems within one system is the use of a non-transparent PCI-to-PCI bridge to separate the devices of the host system and the devices of the coprocessor system. Communication between the two operating systems is accomplished via shared memory scheme in which the host maps some of the coprocessor memory into its own address space.

2.1 Hardware Overview

The current CiNIC coprocessor is an Intel EBSA-285 board comprising the following components:

- SA-110 microprocessor running at 233 MHz,
- 21285 core logic chip set,
- 148 MB of system memory for development purposes,
- a set of flash memory for BIOS images,
- and a serial port to use as a console.

Although a real commercial network coprocessor would not need all of these components, they are included to provide for our development environment with the greatest possible flexibility. This coprocessor card is plugged into an Intel 21554 EB backplane[5]. The bridge is a non-transparent PCI-to-PCI bridge that separates the host PCI devices from the co-host PCI devices, effectively creating primary and a secondary PCI bus. The host processor controls devices on the primary bus and the coprocessor controls devices on the secondary bus. Figure 2 is a detailed diagram of this configuration.

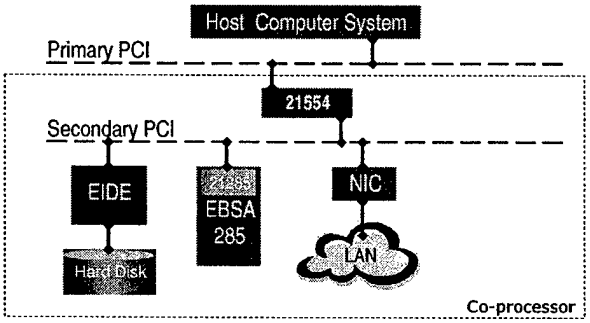


Figure 2. Detailed Diagram of CiNIC Architecture

The addition of a secondary PCI bus, allows any devices having Linux drivers to be easily added to the coprocessing system. In the case of Figure 2, an IDE disk has been added to allow for web caching. What are not shown in Figure 2 are potential devices on the host computer system.

The piece which allows the two computer systems to co-exist is the non-transparent PCI-to-PCI bridge. From both the host computer system and the coprocessor viewpoint, the bridge appears to be a single PCI device, but is responsible for moving data between the two buses. To do this the bridge performs address mappings and address translations. An address range in the primary bus can be

mapped to an address range in the secondary and vice-versa. This allows the bridge to respond to PCI transactions destined for the other side by performing address translation as shown in Figure 3.

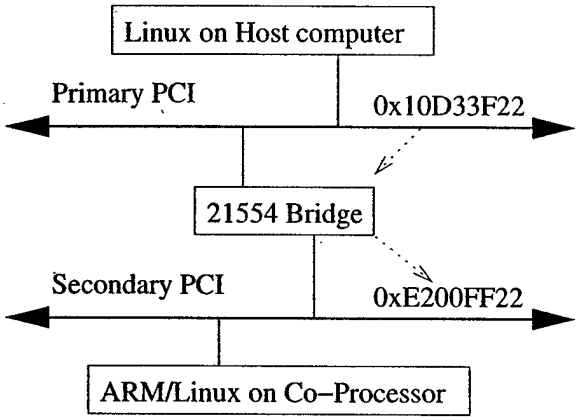


Figure 3. Address mapping through the 21554 bridge

In the example shown in Figure 3, a transaction is started on the primary bus at address 0x10D33F22. The bridge sees this address is in the range of addresses it is translating. Using an internal translation base register, it maps the host address of 0x10D33F22 into the co-processor's address space on the secondary PCI bus at 0xE200FF22. It then carries out the transaction on the secondary PCI bus using this translated address. If this were a read, then the bridge would return the read data on the primary PCI bus. In the case of a write, the bridge would execute the write on the secondary PCI bus using the data passed to it by the primary bus [9].

2.2 Software Overview

The goal of the system software is to make the coprocessor invisible to applications while providing all TCP/IP state processing. The entire data transfer process between the two computing systems is based on a shared memory scheme. This shared memory is located in the coprocessor's memory and is mapped to the PCI space of the secondary bus via registers in the 21285 core logic. This secondary PCI space is in turn, mapped to the primary PCI space via the bridge. It is in this way that the host system can access the SDRAM of the coprocessor operating system.

To execute the host's network functions, the coprocessor must intercept and redirect the TCP/IP stack calls to the network coprocessor. In the CiNIC architecture this is done at the socket layer. When an application makes a socket system call, a kernel module takes over the processing of the system call. It marshals the parameters for the call, places them in shared memory and invokes the coprocessor to complete the function call. When the coprocessor has completed the function call, it passes the parameters back to the host. The host saves process cycles by scheduling another process while waiting for the socket call to return.

The example above is the ideal operation of each system call. Many situations are much more complex. Functions such as select() which can execute on local file descriptors and sockets, needed to be coded with care. Once the functions were coded, the end result is a library of socket system calls that to the applications and user behave exactly like the standard system calls. Standard applications such as FTP were used to test the functionality of this coprocessing architecture[4].

3 Goals for the Next Generation CiNIC Architecture

The current CiNIC architecture is based on a commercially available microprocessor. Our goal is to move to an architecture using programmable logic which will make more parts of the system customizable. While there is nothing fundamentally wrong with the old CiNIC architecture, a completely customizable system will allow us to experiment with the hardware aspects of offloaded network functions and give us the ability to adapt the hardware architecture to match the needs of the software. This new architecture will no longer be static, but can be changed to enhance the functionality and performance of the coprocessing system. The new architecture that was tested is based on Altera's Excalibur technology: RISC processors synthesized and running on programmable logic. The processor provides the power to run an operating system and the flexibility of programmable logic to implement custom hardware on the chip. It is a classic example of hardware/software co-design: designing hardware and software concurrently to meet system-level objectives[3].

This architecture allows us to conveniently decide what to implement in hardware and what to implement in software. With the added capability of programmable logic, much more of the system is under our control. It would allow us to use a physical shared memory, rather than just mapping the SDRAM of the coprocessor to the host. Since we utilize very few of the PCI bridge's capabilities, this is another component that could be replaced with a more elegant, custom solution. The original CiNIC architecture was developed to aid in network performance analysis and experimentation. Moving the entire platform to programmable logic will provide extra features and allow further analysis and experimentation.

4 NIOS and Microtronix uClinux

The core piece of this project is the Altera NIOS development board. The Altera board is part of a technology they call Excalibur. It is based on running RISC microprocessors on a FPGA and providing room for user-defined logic. The NIOS technology is a soft-core processor, mean-

ing that the hardware for the processor is implemented in a Hardware Description Language. This code is compiled and programmed into the FPGA, making part of the FPGA into a microprocessor. One advantage of this architecture is the ability to use leftover logic on the FPGA for processor peripherals. This allows peripherals like a UART, Ethernet controller, PCI Logic, or IDE controller to be implemented on the same chip as the microprocessor. In addition to peripherals, the architecture lends itself to easy implementation of user-defined logic on the chip. This allows the functionality of a microprocessor, peripherals, and ASIC all on the same chip (see Figure 4). The result is an architecture with extremely flexible software and hardware, perfect for embedded systems work[1].

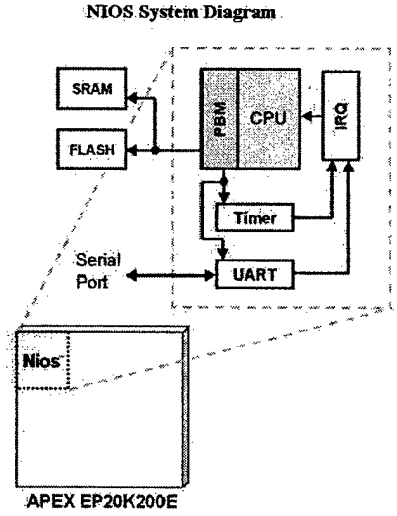


Figure 4. NIOS System Diagram[2]

The Microtronix Linux Development Kit (LDK) is a hardware and software expansion to the NIOS board. It contains extra memory, in the form of 16 Mbytes of SDRAM, a real time CMOS clock, and an 10BaseT Ethernet controller chip plus transceiver. The software expansion is a port of uClinux, a real-time, embedded version of Linux. The embedded aspect refers to the small memory footprint of the operating system kernel. Also ported were the standard Linux tools and commands, cp, ls, etc. With the addition of this kit, the NIOS processor had a fully functional operating system. For more information on the contents of the Microtronix LDK see [6].

5 NIOS System Evaluation

The evaluation of this technology to use as a next generation architecture comprises two aspects: performance and capabilities. First, we needed to determine whether the FPGA processing system could handle the Linux network processing code. Next experimentation needed to be done in order to determine how easy it is to add custom logic to the system and how that custom logic integrates with the Linux operating system. Once these two aspects were analyzed, it was shown that the NIOS system was a good target for moving the next generation CiNIC architecture.

5.1 Performance

Since the purpose of the CiNIC architecture is to improve performance related to network processing, we conducted a series of timing measurements to determine network processing latencies on the new coprocessor architecture. The experiment consisted of running a test application under uClinux on the NIOS development board and measuring the time required to traverse the stack. The test application sent marked packets using the `send()` function. A timer was added to the hardware of the FPGA specifically for counting how long a `send()` function took to execute on the NIOS system. The server that was receiving the packets simply discarded them, since the timing was all done while the data traversed the TCP/IP stack of the NIOS system.

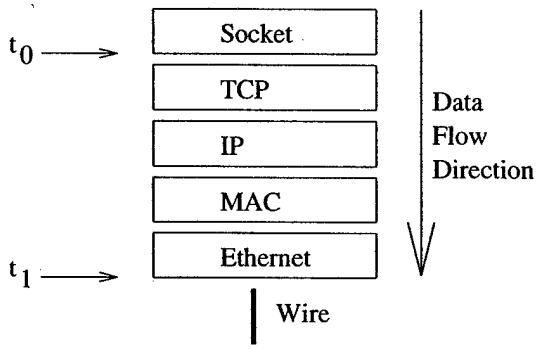


Figure 5. Model of uClinux TCP/IP stack

The timing model is relatively straight-forward. The timings for each measurement were based on the buffer size of the `send()` function. These buffer sizes range from 64 bytes to 64,000 bytes. The timer for the system is started when the `send()` function enters kernel space. The timer is stopped after all the data for the send has been written to Ethernet controller. Figure 5 shows the approximate placement of the timing hooks in the stack.

The hooks themselves are snippets placed into the kernel stack that start, read, and stop the timer. In order to make the ends of the buffers recognizable, each buffer was marked with 6 0xAA characters at the front and 6 0xEE characters at the end. On each buffer size, 2 timestamps were recorded: entry and exit times. *Entry time* is defined to be the time when the first packet of the buffer (the one with the 0xAA's) hits the t_0 marker (see Figure 5). The *exit time* is defined to be the time when the last packet of the buffer (the one with the 0xEE's) hits the t_1 marker.

This test was run 25 times for each buffer size. From there we took the median value of the 25 runs and calcu-

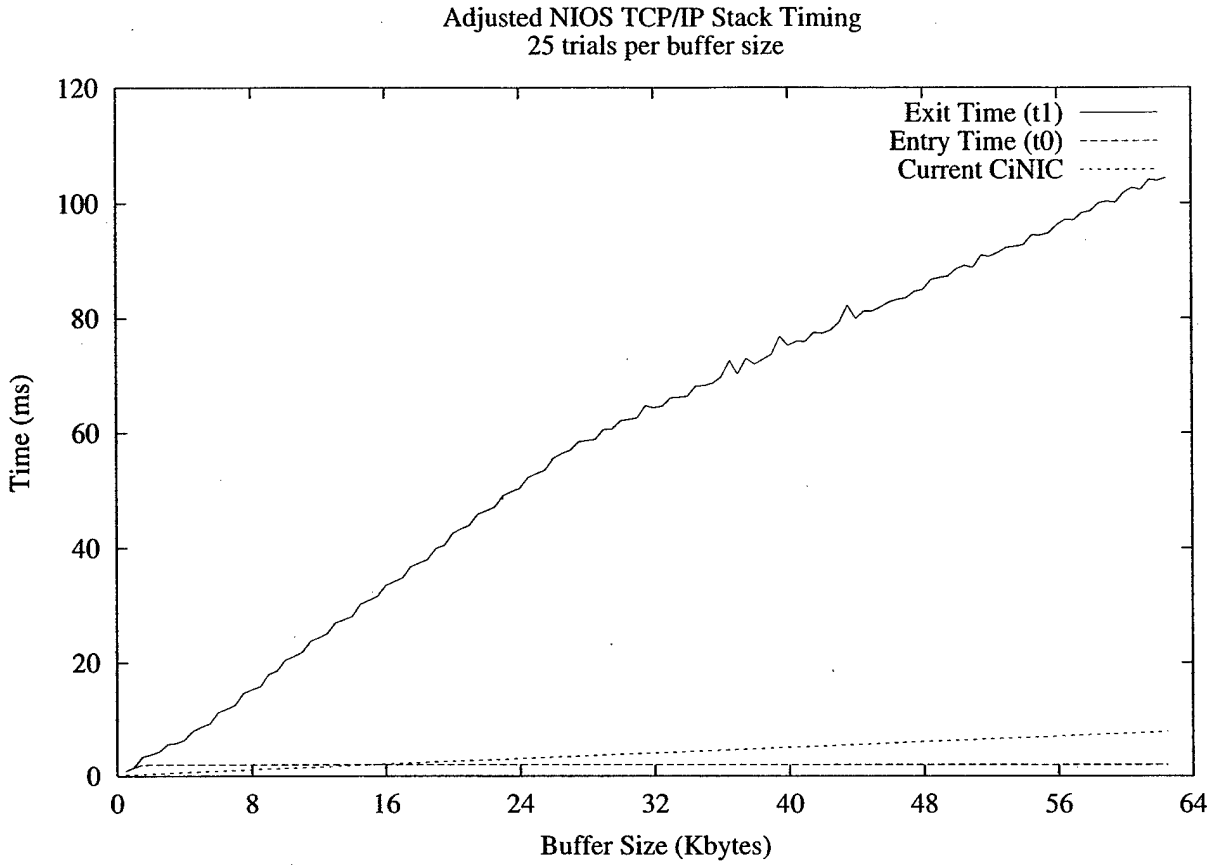


Figure 6. Adjusted NIOS TCP/IP Stack Timing

lated the standard deviation of each of the measurements. The timings showed that 99.5% (3 standard deviations) of the measurements were within 3% of the median value. This means that the uClinux stack on the NIOS system has a consistent processing time. Figure 6 shows the results—adjusted to eliminate the cost of the timestamping itself—and compares them to the results of the current CiNIC TCP/IP stack timings.

The entry time stayed constant as was expected. Whether sending a buffer of 64,000 bytes, or 1500 bytes, the time it takes for the first packet to hit the Ethernet controller is the same. The exit time is proportional to the buffer size that was sent. That is, the processing time is linearly dependent on how much data needs to be sent.

The TCP/IP stack measurements for the NIOS system give us an idea of the network performance of uClinux running on a NIOS processor. When comparing this performance measurement to the measurements from the CiNIC system[4], the NIOS processing appears to be approximately an order of magnitude slower. The reasoning for this, however, can be attributed to three major differences in the hardware and software between the two systems.

First, the architectures and processor speeds of the two systems are different. The current CiNIC using a StrongARM processor runs at 233MHz while the NIOS processor runs at 33.333 MHz. That is a difference of 7x. Second, there is a difference in the Linux kernel versions. The CiNIC's kernel version was 2.4 and the NIOS has version 2.0. There are large differences in the TCP/IP stack portion of these two kernels. Many of the modifications between the 2.2 and 2.4 kernels were improvements to network processing. Further analysis in this area is needed, but when comparing these results it is important to remember that the NIOS and CiNIC are running similar (but different) stacks.

The third difference is the fact that the NIOS has a 10 BaseT Ethernet interface, while the tests for the CiNIC were done with a 100 BaseT Ethernet interface. This issue suggests a difference of 10x if the network interface is the limiting factor. This hypothesis was easily tested since the driver for the controller chip contained a "chip busy" state. There is a finite amount of memory for packets in the chip, so when this is full the driver will wait and try again. A single debug statement proved that the driver was entering the "chip busy" state during multiple packet sends. This shows the NIOS processor and the uClinux stack are capable of

generating packets faster than 10 Mbps. The addition of 100 Mbps Ethernet would allow us to determine if the processor is capable of 100Mbps speeds.

5.2 Capabilities

This section discusses the use of programmable logic within the NIOS FPGA. The idea of adding custom hardware is one of the main advantages of this programmable system. The process of modifying or adding hardware can be done quickly, giving us flexibility not present in the current CiNIC architecture. The process of examining the capabilities of custom hardware, was implementation of a very simple custom hardware device and integrating it with uClinux.

The simple hardware device was a two register device which used control values in register A to decide when to perform a one's complement on the value in register B. This was implemented in Verilog code, along with the interfacing logic necessary for tapping into the NIOS system bus. From the NIOS system side, a user-defined interface was setup on the bus and assigned a region in the memory-mapped space. This allowed all software to see the registers of the custom hardware at a particular physical address. Since NIOS does not contain an MMU, it only has to dereference a pointer to that physical address to access the registers.

The process of adding custom hardware to the system was found to be very painless. Once the initial busing structure was understood. Integration with uClinux was also simple due to the lack of a virtual addressing for the NIOS processor. Our custom hardware wrote data to register B, flipped a bit in register A, waited for the hardware to flip the A bit back, and read out the "processed" value. This allowed us to implement a simple scenario, where the processor waits for the custom hardware to finish[1].

6 Conclusion

This paper presented a network coprocessing architecture based on an FPGA running a modified version of the Linux operating system. The architecture will allow us to implement changes and analyze the hardware/software trade-off very quickly. We studied both the performance and capabilities of this architecture. In the performance aspect, the NIOS under-performed the Strong Arm based CiNIC. But, it was shown that this may have been due to the NIOS's 10 Mbps Ethernet versus the CiNIC's 100 Mbps Ethernet. We feel confident that the NIOS, FPGA based architecture can adequately perform network functions. Also, if Moore's law holds true, the performance of the FPGA based architecture will continue to improve rapidly, providing more performance in the near future. In the capabilities arena, NIOS was found to be very flexible and fit our needs from an ease of use and development standpoint. The uClinux operating system provided us with the needed software support and the ability to add custom hardware to the system.

[1] NIOS 2.0 has the capability to provide up to five custom instructions. I.e., there now is a direct implementation for custom hardware instructions.

References

[1] ALTERA CORP. *Excalibur Backgrounder, Version 1.0*, 1.0 ed., June 2000.

[2] ALTERA CORP. *NIOS Embedded Processor—Getting Started, Version 1.1*, Mar. 2001.

[3] DE MICHELI, G., AND GUPTA, R. Hardware/software co-design. *Proc. IEEE 85*, 3 (Mar. 1997), 349–365.

[4] FISCHER, J. CiNIC—CalPoly intelligent NIC. Master's thesis, California Polytechnic State University, June 2001.

[5] INTEL CORP. *21554 PCI-to-PCI Non-Transparent Bridge Evaluation Board User's Guide, Version 1.0*, Jan. 2001.

[6] MICROTRONIX CORP. *LDK Getting Started Guide, Version 1.5*, Sept. 2001.

[7] ROSU, M., SCHWAN, K., AND FUJIMOTO, R. Supporting parallel applications on clusters of workstations: the intelligent network interface approach. In *Proceedings of the Sixth IEEE International Symposium on High Performance Distributed Computing* (Portland, OR, Aug. 1997), pp. 159–168.

[8] STEENKISTE, P. Analysis of the Nectar commuinication processor. In *IEEE Workshop on the Architecture and Implementation of High Performance Communication Subsystems* (Feb. 1992), pp. 1–3.

Building Photovoltaic Facility

Project Investigator:

Taufik, Dr.E.
Associate Professor
Electrical Engineering Department

Executive Summary

Solar energy is a promise for the future. It is a clean and sustainable source of energy that can provide a significant share of our energy needs. It is not only from an environmental point of view that solar energy has a future: from an economic perspective prospects abound. Large multinationals such as Shell, BP and Siemens are focusing their efforts in the field of sustainable energy, most particularly on photovoltaic solar energy. They are doing this primarily because they expect that solar energy will offer their companies good economic prospects, rather than through environmental concern.

However, solar energy appears to be developing into a perennial promise. The big breakthrough is long in coming. The predominant reason for this is price. Solar energy is much more expensive than conventional energy, and as long as this remains the case, solar energy will remain an unrealized promise.

As part of the effort to see fossil fuels phased out in favor of renewable sources of energy, it is very important to educate people about solar energy such that it becomes widely accepted and used. The development of Photovoltaic Facility at Cal Poly State

University, as described in this report, is aimed to increase public awareness on Solar Energy to campus community in particular, and to the central coast community in general. Furthermore, Cal Poly as the leading academic community in the central coast should be the logical choice to host such facility to educate not only its students but also overall central coast community about Solar Energy as an alternative source of energy.

Another aspect of solar energy which will help in its global acceptance is its easiness of use and its compatibility to utility grid. Solar energy users should experience minimum hassles when it comes to installing solar panels to power their houses, either as a stand-alone energy source or when connected to grid. The enabling technology that plays the main role in bridging the solar panels to customer loads and to utility grid is what we know as Power Electronics. The use of this technology ensures that fairly high efficiency is achieved during the conversion process from the solar energy collected to a battery to any electrical load. On the other hand, the use of power electronics in solar energy conversion circuitries may concern utility companies because of the amount of electrical pollutant or harmonics being injected by any power electronic circuitry to the utility grids. The presence of excessive harmonics has been known to cause harmful effects to transmission and distribution line equipment. For instance, the neutral conductor in a harmonic congested three-phase system has to be resized to handle the third harmonic current. A distribution transformer which is connected to predominantly power electronic driven loads will have to be derated to handle the heat caused by the harmonics reflected back from the load, or otherwise the K-rated transformer which is specially made to withstand harmonics will have to be installed. The establishment of the Photovoltaic facility at Cal Poly is aimed to be the place where students, faculty, local community and industries may share interests and conduct any study, project, or research in power electronics for Solar Energy. Particularly to Cal Poly's students, the facility is yet another addition to Cal Poly's research/project labs wherein students will experience the learn-by-doing education, in line with Cal Poly's philosophy.

By far, the Photovoltaic facility has attracted numerous undergraduate and graduate student projects. Three projects are currently planned to start in Winter 2003, and more are expected in the next academic year. In addition, involvement of several industries has

also been secured and is still on going to support the student projects. Future industry commitments are being sought to help strengthen the facility.

In the 2001 – 2002 academic year, the C3RP committee decided to allocate $ 30,000 to fund the establishment of the Photovoltaic Facility at Cal Poly. Most of the funding received had been used for acquiring the equipment essential to the facility. With the help of two EE students, a movable 450W complete solar panel system has been established, along with the necessary metering and measuring equipment. In order to avoid a large start-up cost, the existing Power Electronics laboratory (Building 20 Room 104) was chosen to host the facility. The choice should come in naturally since one of the objectives of building the facility is, as mentioned previously, to study the power electronics associated with Solar Panels.

The "Building of Photovoltaic Facility" has been quite a challenging and educative experience for the many individuals involved in the project. However, it does not stop here. Future funding is currently being sought to extend its capability and more importantly to expand beyond solely the solar energy, but also other sources of alternative energy. In the long-term plan, the facility will be part of the Renewable Energy Center at Cal Poly and of the central coast community where all sorts of renewable energy will be studied and exhibited.

Introduction

Based on latest projection, it is thought that PV solar could produce one quarter of the world's energy by 2040. The Photovoltaic (PV) power generation market is currently experiencing rapid growth with worldwide PV module shipments increased 38% in 1997 and 29% in 1998 [1]. This rapid growth is expected to continue. The international photovoltaic industry is projected to grow at a rate of around 20 to 25% per year over the next 15 years. It is anticipated that by the year 2010 annual PV shipments could reach 1,600 MW [1]. Industry analysts estimate that solar power is a $1.5 billion business today. The vast majority of present photovoltaic sales are for applications such as navigational signals, call boxes, telecommunication centers, consumer products, and off-grid electrification projects. More recently, small grid-interactive rooftop installations have started contributing to the demand for PV products.

Several factors support the growing interest in building photovoltaic systems. Increased concerns over global warming, ex-President Clinton's Million Solar Roofs program, legislation that requires utilities to buy excess energy generated by on-site, distributed

power sources, and the fact that 40% of U.S. energy consumption is attributed to buildings are all providing incentives to incorporate photovoltaic into building. In addition, all of these factors are coupled into the energy crisis that is central to California including the central coast which in turn triggers the growth and the popularity of distributed generation such as microturbines and photovoltaic. The Photovoltaic Facility at Cal Poly has been completed to serve as a facility where projects, ideas, research, and any other public interest activities pertinent to photovoltaic technology will be accommodated. The facility will also provide an avenue for the facility staff's professional development, and a means of helping students increase their technical skills while working on problems on interest to industry.

A survey of 900 building professionals in the United Kingdom found that 88% would consider the use of photovoltaic products if there was greater evidence of the performance and reliability of these products [2]. Forty-nine percent of the survey respondents noted that they would only consider building PV products after they had seen them utilized in demonstration sites [2]. Although a similar survey has not been conducted within the U.S., it is anticipated that the results would be comparable. The Photovoltaic Facility at Cal Poly also serves as a demonstration site where residential customers or other interested entities, for example, can come to visit the facility and learn and witness a successful demonstration of the photovoltaic module. Hence, the facility will help increase public awareness in the solar energy alternative to ease the ever-increasing demand of California's fossil-based energy consumption in particular, and to promote the use of renewable energy sources in today's energy conservative world in general.

Finally it was also reported that almost two billion people in developing countries have no access to electricity, but a recently study indicates that the use of solar energy could help bring electricity to many of these people. The report, "Solar Photovoltaic for Sustainable Agriculture and Rural Development," by Food and Agriculture Organization states, "The time is ripe now to advance towards a new phase of solar energy beyond the light bulb. We should not only use solar systems for household lighting, but also for

pumping water, irrigation, cattle watering, small cottage and agro industries, facilitating educational radio and TV programs and health services". According to the report, the dominant use of solar energy could be in agriculture including solar pumping for drip irrigation. The uses of solar energy are not limited to agriculture says the report, but could be used to operate such small tools as drills, blenders, mobile phones, and televisions. FAO is calling upon governments to promote the use of solar energy to meet the needs of the rural poor. The establishment of the Photovoltaic Facility at Cal Poly is significant to show our contribution to this global effort in solar energy, especially in the state of California, where agriculture is a big part of the state and where every summer people get nervous about having their electricity service interrupted for rolling black out or even a brown-out.

Facility Capability

At the start of the development, the proposed list of equipment to be acquired was modified due to the reduced amount of funding received. To follow is the list of equipment, along with its capability, obtained during the course of the project; both from the funding and industry support. Figure 1 shows a partial snapshot of the facility.

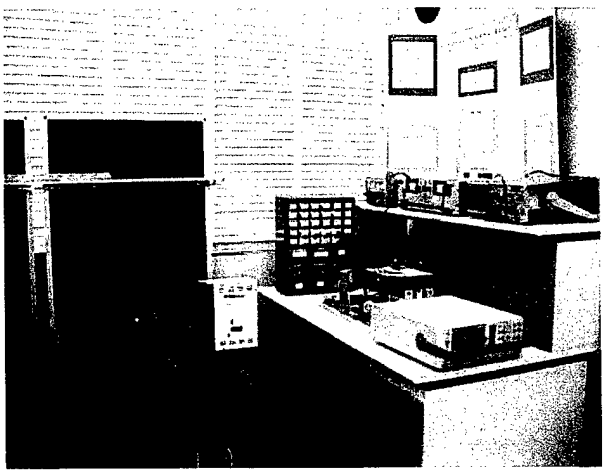


Figure 1. Workbench for PV related student project

a) Solar Panels

After a very extensive research on solar panel manufacturers, BP was finally selected due to its competitive price and significant educational discount that they provided. It was also decided that a 450-W PV system would be the optimum capacity given the limited space of the facility and the cost of the associated accessories to set up a complete system. As can be seen from Figure 2, there are three solar panels, each of 150W capacity. The three solar panels are structured on a movable frame which will make it convenient to move the panels around.



Figure 2. 450-W BP Solar Panels

b) Battery System

Once the solar panel was selected, a battery would then be needed to store the solar energy collected by the panel. Also, to ensure that the battery is being charged properly, a charge controller will be required. We were able to allocate money to acquire two different types of charge controller: the standard charge controller and the MPPT (Maximum Power Point Tracker) charge controller. The standard charge controller will

not provide the optimum collection of energy from the panels to the battery, while the MPPT will always look for the maximum possible energy to be stored in the battery. MPPT is of course a more expensive controller than the standard type. By having the two types in the facility, we will now be able to study the trade-offs between the two types, and study is currently on going to improve the MPPT technology. Figure 3 shows the battery system along with the two types of the aforementioned charge controllers. Each component in this system was obtained through generous educational discount from their manufacturer.

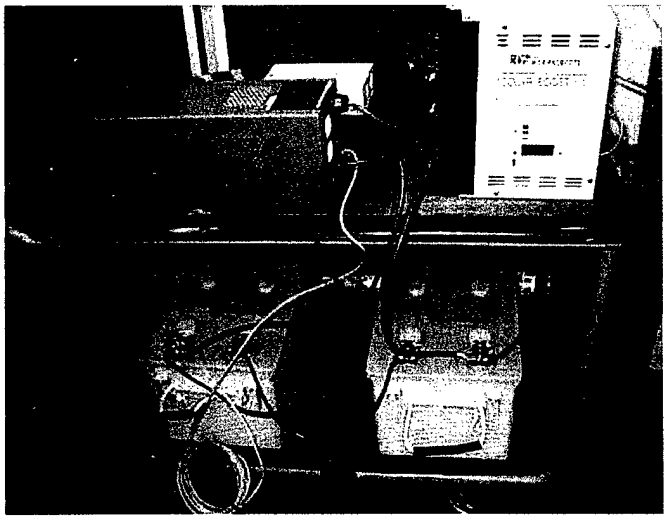


Figure 3. Battery System and Charge Controllers

c) Electronic Load

Electronic was needed to study dynamic response of the solar panels upon sudden load changes. The electronic load also provides quick and convenient way for testing and measurement of electrical system at different level of load. Generally speaking, the electronic load is a fairly expensive equipment. However, we were able to find one, as shown in Figure 4, which provides most of necessary functions at a very economical price (under $800). The electronic load can function as a constant current source both in AC or DC from, and it can also have the option to provide constant resistance.

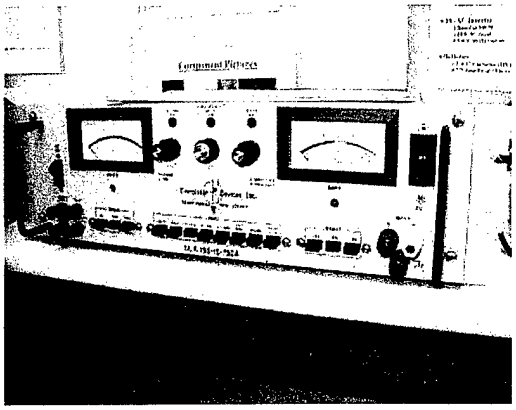


Figure 4. Electronic Load

d) Computers

To perform all the computing and data gathering tasks, several computers were obtained, two of which were donated. Another computer (a laptop) was purchased for power quality on-site measurement (see next section). With the help and support from the EE department, the PC's are now connected to the Internet.

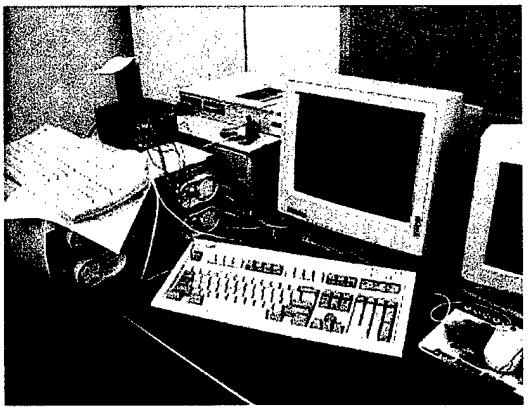


Figure 5. Personal Computers with Internet Connection

e) Power Analyzer

This particular equipment is the heart of any Power Quality study. Since power electronics is being used in the conversion from DC energy stored in the battery to many of the AC loads that we are using, it is then critical to study the power quality and its effect on the solar panels. There are two power analyzers obtained in this project, each

has its advantages and disadvantages. One power analyzer (Fluke 43B), as shown in Figure 6 on the left, is a stand-alone and portable tool. The current and voltage waveforms can be measured and captured, along with other power parameters such as THD, power factor, VAR, phase angle.

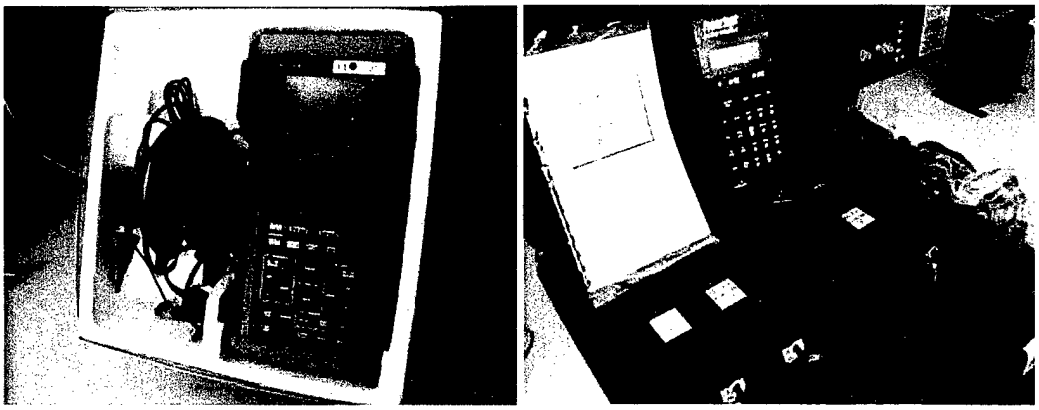


Figure 6. Power Analyzer: Fluke 43B (Left) and Power Sight (Right)

The other power analyzer is the Power-Sight has all the functions that the Fluke 43B does and more. The drawback is that it has to be interfaced with a laptop to further process the data gathered by the device. It can handle more power and more versatile than the Fluke. The two power analyzers, especially the Power Sight, were acquired with significant educational discount.

Industry Involvement

In line with the spirit of C3RP wherein industry involvement into undergraduate research and projects is an important part in the C3RP program, several companies were contacted and invited to join the effort and get involved with the "Building Photovoltaic Facility" project. Here is the summary of these companies and their involvement in this project.

a) Venable Industries

This company designs and manufactures Spectrum Analyzer. This type of equipment is useful when analyzing stability of a system by looking into the frequency response of the system (gain margin, phase margin). It is also a very useful tool when designing a loop compensator for any system which needs to be stabilized. The equipment is known to be very expensive, somewhere around $30,000. Hence, it was very excited to be able to get the support from Venable Industry who was generous in loaning their Spectrum Analyzer free of charge for the duration of this project. The Venable spectrum analyzer was delivered in March 2002 and returned in June 2002. During this period, many students

had shown their interest and used the equipment for their projects. Figure 7 and 8 are two pictures showing students working on the Venable spectrum analyzer.

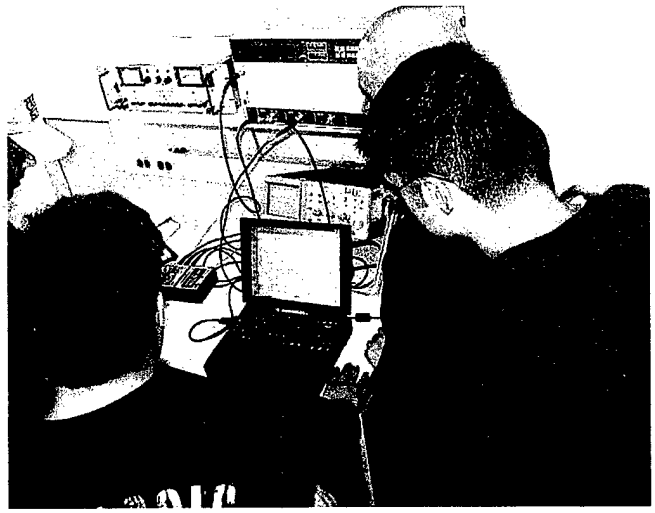


Figure 7. Students using Spectrum Analyzer for their project



Figure 8. EE Master Student using Spectrum Analyzer for his project

b) Linear Technology

This company manufactures ICs such as those used for switching regulators for energy conversion which is one of the areas of interest in this project. The company donated various switching regulator ICs which will be used by any student interested in doing undergraduate projects on solar energy conversion circuits.

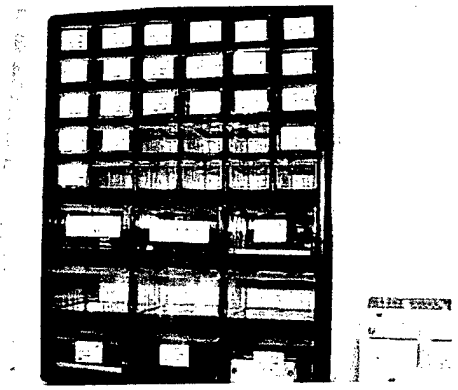


Figure 9. Linear Technology IC's

c) QuickSilver Control

In order to achieve the maximum solar energy collected by the panels, it is necessary to track the movement of the sun to get the optimum angle of incident. One method to track the sun is through the use of motor. QuickSilver Control is a company that makes servomotors. Their sales office is located in Atascadero and when contacted they expressed their interest in supporting this project. As many as 14 servomotors were donated for use by students interested in servomotor control projects. Figure 10 shows a prototype of two-axis servomotor controller which may be used for solar panels. The project was undertaken by an EE student for his senior design project.

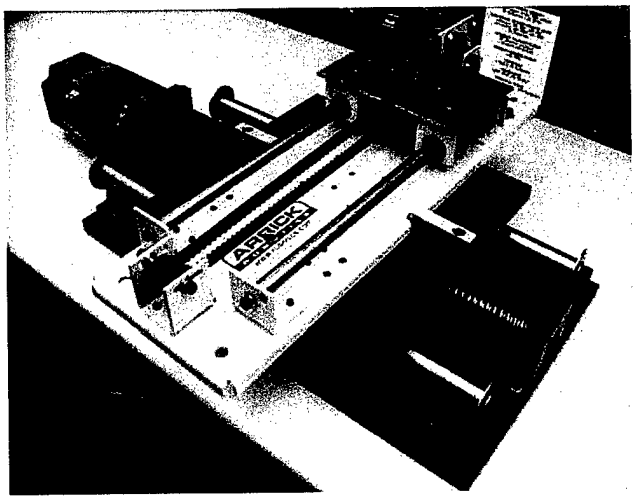


Figure 10. Two-axis Servomotor Controller

d) Enerpro, Inc.

When connected to utility grid, the solar panel would have to provide bi-directional power flow capability such that during excess of energy, the battery can provide power to the grid. One way to do this is to use a three-phase thyristor controlled rectifier. The rectifier when operated in its inverting mode will provide the path for the real power flow from the battery to the utility grid. Enerpro, Inc. in Goleta is a company that designs and builds the controller for the thyristor rectifier. The company was interested in supporting the program and they donated 6 units of controller board. Figure 11 depicts the ready to use controller module designed and built by an EE student for his senior design project.

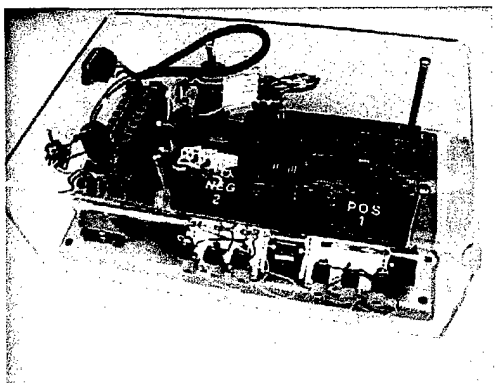


Figure 11. Enerpro's Thyristor Rectifier Controller Module

e) Alltech, Inc.

Once the solar energy has been transported and stored to the battery, it may be necessary to use a dc-to-dc converter to supply a dc load at a different voltage level. Alltech, Inc. is a company in San Luis Obispo that deals with telecom equipment. They contributed to this project by donating two of their dc-dc converters and two three-phase rectifiers.

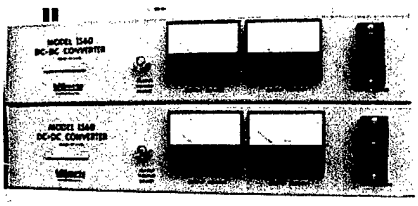


Figure 12. Alltech's dc-dc converter

List of Student Projects

One important entity in this project and the C3RP program in general is none other than our own students. Each of C3RP project should generate interests to our students and invite their involvement to stimulate the learn-by-doing and at the same time real world educational experience. The "Building Photovoltaic Facility" has attracted several undergraduate students who were interested in getting involved in the solar energy related project while fulfilling their senior design requirement. The facility was also being used as exhibition tool for one graduate course in Switching Power Supply Design. Furthermore, work is currently under plan for one graduate student who has chosen his master's thesis topic in improving the efficiency of the MPPT charge controller for the Solar Panel. Future senior design projects using the facility are also currently under plan to start in Winter 2003. In all, the projects that were completed benefited from the facility and took advantage of the equipment acquired for the facility. Here is the complete list of student projects generated from the facility:

List of Past Projects:

Name(s)	Topics
Kevin Quetano and Karl Buckman	Development of Photovoltaic Facility
Augusto DeCastro	Internet-Based Servomotor Control for Solar Panel Remote Operation
Daniel Fritz	Thyristor-based Rectifier and Inverter Control
HongDiem Dinh	Internet-Based DC Motor Control for Solar Panel Remote Operation
Edward Surber	Two-axis Servomotor Controller for Solar

List of Future Projects:

Name(s)	Topics
Angel Martinez	Resonant Boost for MPPT Charge Controller
Victor Velasquez	Efficiency Improvement using Synchronous Rectification Topology
Matthew Hartman	Power Factor Correction Circuit Using Boost Converter
Gernard Ferril and Marvin Camillon	Sinusoidal PWM Inverter for Solar Panel

Future Plans

This project does not stop here. Work is currently on going to attract more companies to be involved in the facility. Since the use of solar energy may directly impact the utility companies, an effort will be put forward to get the support from California's utility companies such San Diego Gas and Electric, Pacific Gas and Electric, Southern California Edison, and Los Angeles Water and Power, to mention a few. Moreover, since on of the focus for the facility is to study energy conversion, industry contacts will also be planned to get the involvement of energy conversion companies such as Rantec, Capstone Turbine, etc.

Finally, external funding both from the government and industrial sectors will be sought to extend the capability of the facility and to accomplish the long-term goal of having a Renewable Energy Facility at Cal Poly State University, San Luis Obispo.

References

[1] "Markets: Worldwide PV Module Output Heading for Record 190-204 MW range in 1999," Photovoltaic Insider's Report, XVIII No. 8, pp. 1 and 6.
[2] Schoen, T.J., "Information", Renewable Energy World, 2, No. 5, p.84, 1999.

Development of Technologies for Semiconductor Processing: A Partnership with Applied Materials Corporation

Project Investigators:

Glen E. Thorncroft, Ph.D.
Associate Professor
Mechanical Engineering Department

Christopher C. Pascual, Ph.D.
Associate Professor
Mechanical Engineering Department

DEVELOPMENT OF TECHNOLOGIES FOR SEMICONDUCTOR PROCESSING: A PARTNERSHIP WITH APPLIED MATERIALS CORPORATION

C^{3}RP Project Update: April-December 2002

Glen E. Thorncroft and Christopher C. Pascual
Department of Mechanical Engineering
California Polytechnic State University
San Luis Obispo, CA 93407

Abstract

The work outlined in this report describes development of a research partnership between the Cal Poly Mechanical Engineering Department and Applied Materials Corporation, inaugurated under a grant from the California Central Coast Research Partnership (C^{3}RP). In this work, the partnership was established, and a test facility was designed and built at Cal Poly to test throttling isolation valves for vacuum processing of semiconductors. Preliminary testing of the system and software is complete, and further funding was sought from C^{3}RP to obtain additional equipment and to qualify the test facility. Additional funding from C^{3}RP was not granted; therefore, additional funds have been sought from the National Science Foundation. Pending additional funds, qualification of the test facility will allow accelerated-life testing to be performed to measure degradation of valve performance. A research study will also be initiated to measure and model the transient behavior of the vacuum chamber when the chamber pressure setpoint or the mass flow rate through the chamber is changed. The result of this research will be the establishment of a sustained program of qualification testing and of fundamental research in the semiconductor industry.

1 Background

Microchips are manufactured by layering thin films of conducting, semiconducting, and insulating materials onto silicon wafers. Techniques such as Chemical Vapor Deposition (CVD) and Physical Vapor Deposition (PVD) are commonly used to create these layers; these processes take place in the presence of a gas, and often take place at low pressure (as low as 0.1 mTorr). Conversely, the removal of layers by chemical etching may also take place under a vacuum. Whether layering or removing material, the vacuum pressure and the flow rate of reacting gases must be regulated precisely to ensure that the layers are uniform; this is especially critical given that the layers can be as thin as a few atoms.

Applied Materials Corporation, the leading manufacturer of microchip manufacturing equipment, is seeking ways to improve the performance of their vacuum pressure control equipment used in vapor deposition and etching processes. To this end, they have identified a number of new technologies that may improve and streamline the way the vacuum and gas flow is controlled and monitored throughout the process. One of these technologies is throttling isolation valves (TIVs), which are used to control the vacuum pressure in the chamber, as well as to isolate the chamber from the vacuum pumps. Previously, the tasks of *throttling* and *isolation* were performed by separate devices; combining these functions into one device will streamline the design of the vacuum chamber, reducing both the cost and the space requirements of the

product. This change represents a significant improvement to the next generation of vacuum processing equipment at Applied Materials.

Extensive testing must be performed on any new technology before it can be integrated at the production level; such testing occurs in two stages. First, the validity of the design modification must be tested to ensure that it improves the performance of the overall system. Second, once satisfactory performance is verified, the new devices purchased from suppliers must meet the performance requirements specified by Applied Materials. Therefore, a program of qualification testing must be established.

Applied Materials is also interested in developing a deeper understanding of the parameters that influence the accuracy and control of the vacuum, as well as the flow rates of gases through the chamber. For example, as the flow rate of a reacting gas is varied, or the setpoint of the chamber pressure changed, the system must be allowed to reach steady state before the process can continue. This "settling time" affects production throughput, and reducing this effect can have a significant and direct benefit to the performance of the system. If a mechanistic model of the behavior can be developed, analysis could lead to modifications to the facility or to control algorithms that could significantly reduce the settling time for the system.

2 Project Overview

Through a grant from the California Central Coast Research Partnership (C^3RP), Cal Poly and Applied Materials have begun a technology partnership to develop a state-of-the-art testing and research laboratory for semiconductor processing. The overall scope of the project is separated into immediate and long-term goals, which are described below.

Immediate Goals

The first goal of the project is to establish a premier qualification testing program for original equipment manufacturer (OEM) components used in vacuum processing of semiconductors. The laboratory at Cal Poly is to be built to the testing specifications set forth by Applied Materials, and is similar to a test facility being used by them for component testing. Initially, the test facility will be used to qualify throttling isolation valves (TIVs) for use in Applied Materials products. Once the facility is established at Cal Poly, manufacturers of TIVs, mass flow controllers, and other components will qualify their parts through the Cal Poly laboratory. The laboratory will perform testing initially in two areas: performance testing (accuracy, transient behavior), and accelerated life testing (time to failure).

In conjunction with the testing program, the second goal of the partnership is to initiate a program of research in vacuum technologies for semiconductor processing, beginning with a parametric study of vacuum control. Initially, Applied Materials will be a partner in this research, though it is anticipated that the scope will expand to development work for component manufacturers.

A critical function that this project will allow is opportunities for student involvement in research and industry. Opportunities are anticipated for undergraduate students in the areas of testing and data analysis, while at the graduate level, experiment design and research will be

conducted to study fundamental behaviors in vacuum systems. In addition, Applied Materials has committed to seek opportunities for student internships and co-ops on site in support of the partnership.

Long-Term Goals

The initial work in this research program is considered a first step in developing a broader technology partnership. Once a strong reputation for research is established with Applied Materials in the area of vacuum processing, extending the scope to other semiconductor processing technologies will be explored. For example, one notable area of work is in chemical mechanical planarization (CMP), a process whereby the silicon wafer is ground flat prior to deposition of conducting layers.

It is expected that this program will foster research in other disciplines as well. Such areas of collaboration include the optimization of automated/robotic processes and improvement of control algorithms and technologies. Additional topics include the exploration of new materials in semiconductor conductor processing, as well as semiconductors themselves. Finally, improvements to manufacturing methods are being sought in all processes. One of the goals of this project is to determine what potential exists to grow the partnership into other areas of research.

3 Project Status

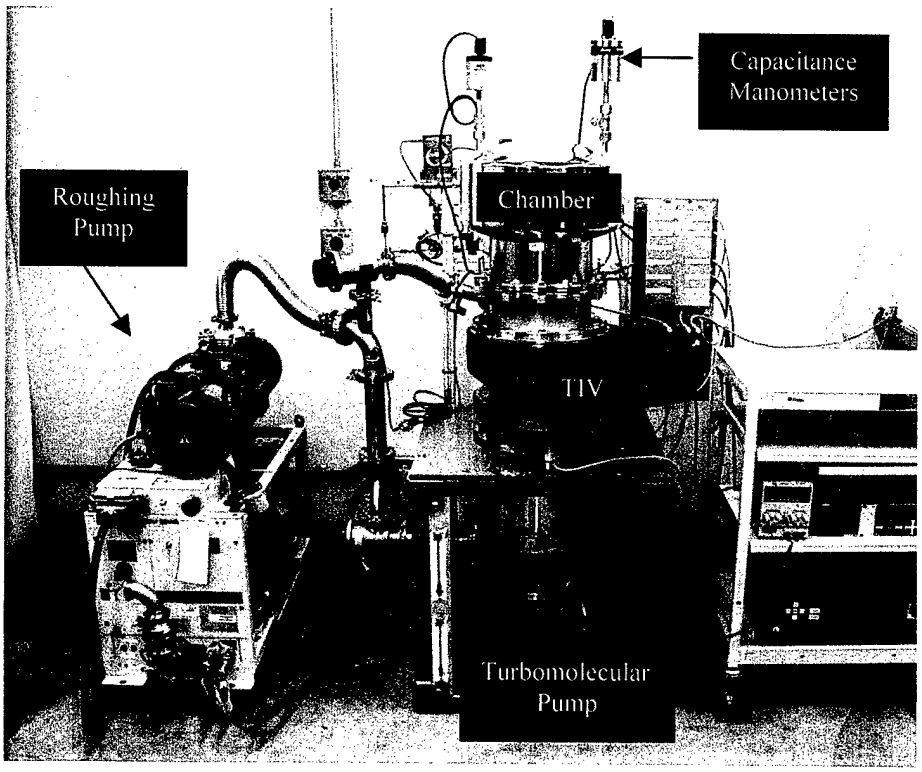


Figure 1. Vacuum processing test facility at Cal Poly.

The work performed during the initial phase of the program (January to August 2002) was divided into three tasks: (1) defining the roles of Cal Poly and Applied Materials in the

technology partnership, and establishing the administrative and legal structure of the program within Cal Poly; (2) designing the facility, obtaining and setting up laboratory space on campus; and (3) fabricating and operationally testing the facility. Completing fabrication and operational testing of the facility became the primary focus of the project from August-December 2002.

The Cal Poly test facility was constructed with $50,000 of donated equipment from Applied Materials and the initial $27,000 grant from C^3RP. This test facility, shown in Figure 1, will initially be used to qualify vendor-supplied throttling valves for Applied Materials. This qualification test will ensure that the valve meets the rigorous requirements of the manufacturing process as well as the reliability necessary to ensure the required production rate. Two undergraduate students will be hired to conduct these tests. Research will also continue with testing of mass flow controllers. Two graduate students will be used to explore the fundamental nature of the fluid flow in a vacuum system. These graduate students will develop a model based on molecular theory to predict the response time or "settling time" of the system to changes in throttling valve position and/or mass flow. The graduate students will also investigate the possibility of using pump speed to control the chamber pressure instead of a throttling valve.

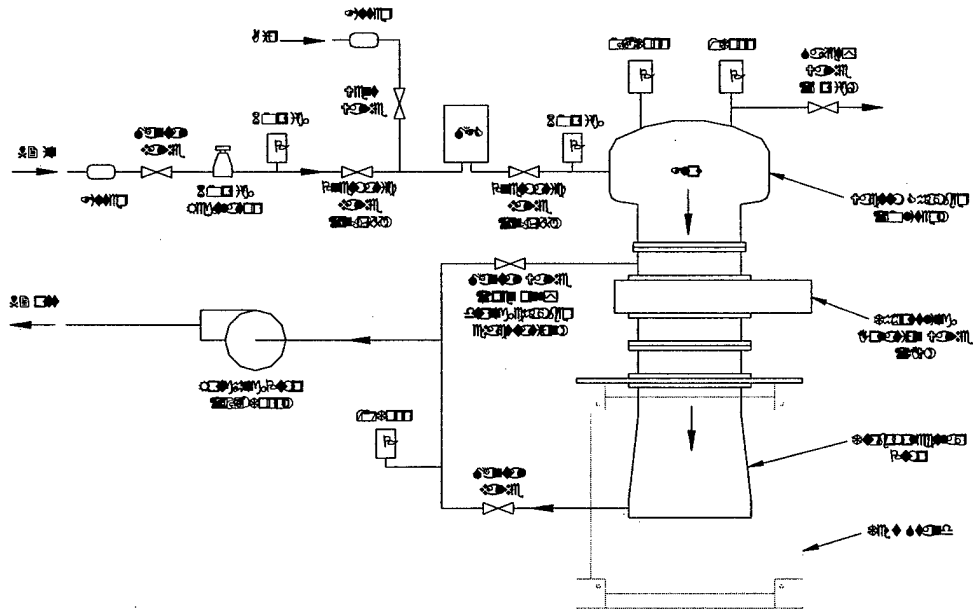


Figure 2. Schematic diagram of facility.

A schematic diagram of the test facility is depicted in Figure 2. The facility consists of a 30-liter vacuum chamber (approximately 40 cm diameter). A pneumatically controlled throttling isolation valve (TIV) is mounted below the chamber, and in typical systems, is adjusted to vary the pressure in the chamber during operation. The chamber and TIV are attached to a turbomolecular pump, which maintains the system operating pressure. A roughing pump is attached to both the chamber and the exit of the turbomolecular pump; the purpose of the roughing pump is to lower the system pressure to below 1.2 Torr, thus allowing the turbomolecular pump to be started. The manual valve from the chamber is only open during initial start-up. Nitrogen is used as the test gas for the facility, which enters the chamber at the top through a mass flow controller. The controller is isolated from the chamber by two pneumatic valves located at the inlet and exit of the mass flow controller. The pressure in the

chamber is measured using two capacitance manometers. One manometer has a maximum range of 0.1 Torr, while the other manometer has a maximum range of 1.0 Torr. Attached to the 1.0 Torr manometer is a pressure relief valve, which opens at 3 psig to prevent overpressurization of the vacuum chamber. Three additional pressure transducers are used to measure the pressure going to the roughing pump, the pressure of the nitrogen flow at the inlet of the chamber, and the pressure of the nitrogen flow upstream of the mass flow controller.

4 Critical Needs

To complete the testing and qualification of the facility, additional equipment is required. While a second proposal to C^3RP was not funded, the authors have submitted a proposal to NSF for an essential piece of equipment: a Helium Leak Detector. To ensure consistent, quantifiable, and repeatable results, the leak rate into the chamber must be minimized. Leak testing in this type of system is accomplished with a helium leak detector through a *spraying test*. The leak detector is attached to a test port on the facility, and the detector's onboard vacuum pumps create a vacuum in the system. Seals and parts are sprayed with helium; any leakage into the system will be measured as it flows through the detector.

Cal Poly does not have helium leak detection capability. The authors are, therefore, requesting funding for a helium leak detector to commission the test facility. With the leak detector, the test facility will be able to conduct qualification testing at industry-level standards and fundamental molecular flow research at the level needed for peer-reviewed publications. Without leak detection capability, the facility will not be industry-standard, and data produced from experiments will be of no commercial or academic value. Additionally, because life-cycle testing involves frequent leak testing over a period of months, rental of leak detection equipment is prohibitively expensive.

Semiconductor Equipment and Materials International (SEMI) publishes standards for vacuum processing, and in particular specifies the requirements and test methods for ensuring leak integrity of a vacuum system. Applied Materials references this specification in their processes, which dictates that the inboard leak integrity of the system and all components must be less than or equal to 10^{-9} atm-cc/s of helium. This requires the use of a mass spectrometer-based helium leak detector. Dry pumping is also required in the detector (no fluid seals) to prevent backflow of contaminants into the system. In addition to these requirements, the detector should have an internal leak standard for automatic self-calibration. In addition, the leak detector should be large enough to evacuate the facility without operating the system's pumps; this simplifies the leak detection process by eliminating "split flow" through the system. Finally, it is desirable for the system to be portable in order to be shared among faculty and research labs across disciplines.

5 Future Work

Once the facility has been qualified and approved for testing by Applied Materials, the first task of the project will be to assess the performance characteristics of new technologies used in vacuum processing. The first components to be tested using this system are throttling isolation valves, and will be tested using two protocols currently being developed by Applied Materials. The first protocol is to test basic performance of the device, which will include the following characteristics:

- Leak integrity across the closed valve, which requires removing the valve from the facility and testing separately
- Steady state performance in pressure control:
  - Accuracy of maintained pressure
  - Repeatability
  - Hysteresis
  - Stability
- Perturbation tests (response to sudden change in setpoint pressure)
  - Response time
  - Stabilizing time
  - Overshoot/undershoot
  - Settling time

A major concern to Applied is ensuring the consistent accuracy and performance of the throttling isolation valve over repeated cycles. Degradation of the valve affects its ability to maintain an accurate, constant vacuum in the system. The extent of this degradation has never been measured, and is necessary not only to establish performance of the valve, but also to aid Applied Materials in establishing a performance specification. Pending the outcome of this data, accelerated life testing may become an on-going function of the laboratory when performing qualification testing of vendor components. Life cycle testing will consist of operating the valve through a series of mass flow rates and pressures (and thus valve positions), and repeating this cycle 400,000 times. After every 20,000 to 40,000 cycles, the above performance tests will be repeated, and any degradation on performance will be tracked.

In summary, a test facility has been fabricated and operationally tested for studies on vacuum technologies for semiconductor processing. Test protocols have been developed, and software has been and continues to be refined for operating the facility. Acquisition of a helium leak detector is necessary to verify the proper operation of the facility and to meet industry standards for leak integrity, and though continued C^3RP funding was not awarded, an NSF grant was submitted in January 2003 to obtain funding for further development. Pending further support, the system and software will be operational by the second quarter 2003, and qualification testing will commence.

Geographic Forecasting: Simulation and Analysis of Fire Patterns

Project Investigator:

Max A. Moritz, Ph.D.
Social Sciences Department

Goal of the Project.

The goal of this project was to initiate a collaborative research effort based on spatially explicit geographic analysis and prediction. For an initial application to begin the project, we aimed to parameterize a new simulation model of fire spread (HFire) for the region encompassing the Main Division of Los Padres National Forest. This work will aid in the analysis and prediction of hazard due to wildfire, and it should support the protection of human life and property in the study area. It is also relevant to national security concerns. There are several days each year, during the hot and dry Santa Ana winds of California, when a small number of well-placed intentional ignitions could result in massive wildfires across the state. Little is known about the spatial distribution of extreme fire weather or how to predict rapidly spreading fires under these conditions.

Results.

We have begun to acquire the fuels data (i.e., digital maps of vegetation types and fuels characteristics), topographic data (i.e., digital elevation models), and weather data (i.e., historical archives from local weather stations) that are necessary inputs for the fire simulation model. For the front range of the Santa Ynez Mountains (located behind Santa Barbara), some of these inputs are available through www.hfire.net. Although this has been a helpful start, many of the digital inputs are at the wrong spatial resolution and will need to be re-sampled from 30m to 100m; in addition, the front range of the Santa Ynez Mountains is only a small fraction of the study area. Finding and/or creating vegetation and fuels characteristics for the remainder of the study area will be a challenge, but one we have already faced in parameterizing the model for the Santa Monica Mountains National Recreation Area. At least two manuscripts are in progress that discuss aspects of this work. We had mixed results in our attempt to extend the simulation domain to include the region surrounding Santa Barbara. Although the model will perform long-term simulations (e.g., 500-year runs, as in the Santa Monica Mountains), output does not appear realistic. This is an area for future work, as most of the inputs (topography, weather, fuels) are quite similar and should not cause the fire model to generate novel patterns/behaviors. Regardless, it was instructive to parameterize the model in a new environment.

A second goal of this work was to coordinate a group of collaborators under the theme of spatial prediction, or "geographic forecasting," under an administrative unit that would eventually generate enough recognition and funding to be self-sustaining. Because many Cal Poly faculty, from a variety of disciplines, do research that has both a predictive and spatial component, a "Center for Geographic Forecasting" could serve as an administrative vehicle to bring them together and generate research in this area. Several potential collaborators, both within and outside the University, were pursued. These included Cal Poly Professor Walt Bremer in Landscape Architecture, Fred Abler, CAD Research Center staff, James Petroni of the California Safety Training Institute, and Professor Cort Willmott, a renowned physical geographer from the University of Delaware.

We also made contact with other hazard-related researchers and began the design of a WWWeb-based consortium utilizing spatially explicit geographic analysis and prediction. Our primary partner at this point is the Southern California Wildfire Hazard Center (SCWHC) at U.C. Santa Barbara (see www.icess.ucsb.edu). This center is funded by NASA and is currently at the forefront of meso-scale weather modeling and acquisition of remotely sensed fuels data. Additional opportunities for collaboration in climate-related hazards research, such as drought and flood prediction, have also been identified at U.C. Santa Barbara.

Several proposal ideas were discussed by collaborators. Two proposals were written, one to NSF and one to NASA, with Co-PIs from the University of California at Santa Barbara in both cases. At this early stage, much of our effort has focused on establishing a "web presence" to disseminate preliminary results and provide a virtual entity to which other interested parties could link. This site is now more or less stable, and is available to academic researchers, fire managers, or those that may be interested in hazard forecasting as a marketable service (see http://cla.calpoly.edu/~mmoritz/firesim.html). Parts of this site may also be used for an "online supplement to an upcoming publication.

SECTION IV. (Continued from P. 8)

IV. C. Research Projects.

An important part of the work undertaken this year was the support of research projects. Broadening and strengthening the applied research base at the University is key to both the proposed technology park and to the development of critical areas of research excellence at Cal Poly. As noted above, eleven research projects were carried out during the grant period. Detailed individual reports on these research projects follow.

(SPACE INTENTIONALLY LEFT BLANK)

Design Methodologies for Analog/Mixed Signal VLSI Systems Applied to Infrared Focal Plane Arrays

Project Investigator:

William L. Ahlgren, Ph.D.
Electrical Engineering Department

DESIGN METHODOLOGIES FOR ANALOG/MIXED SIGNAL VLSI SYSTEMS APPLIED TO INFRARED FOCAL PLANE ARRAYS

Report for the period January-June 2002

by

William Ahlgren
Electrical Engineering Department

Abstract

This project is developing a design methodology for analog/mixed signal (AMS) systems, based on emerging CAD tools. Our example system is a high-performance image sensor for infrared wavelengths, a so-called infrared focal plane array, employing 0.18-micron CMOS technology. We use a suite of tools from Cadence Design Systems which includes an analog hardware description language and new signal integrity analysis tools for system/chip-level design, as well as traditional transistor-level simulation, layout, and verification tools. We review the current state-of-the-art of CMOS image sensors, and describe the impact of next-generation deep sub-micron fabrication technology in sensor design. We propose a pixel design to exploit the opportunities presented by the continuing advance of CMOS technology, following Moore's Law. Finally, an image sensor architecture incorporating this pixel design is proposed.

Computational Infrastructure and CAD Tools

Through partnership with Cadence Design Systems, Cal Poly has access to state-of-the art analog/mixed signal system-on-chip design computer-aided design tools. However, the infrastructure required for utilizing these tools was lacking. Further, due to the newness of the tools, Cadence had inadequate documentation for their installation and use. We undertook the task to develop, on campus, a facility including client/server hardware platforms for the tools, that would provide convenient 24-hour access for both faculty and students doing project work. Then, working with Cadence, we developed and documented the required installation and start-up procedures.

The Cal Poly Computer Science Department and Information Technology Services organization provided us with space in their Java Projects Lab, as well as giving us the loan of a Sun E220 server and two Sun Ultra 10 workstations. We were also supported by the College of Engineering, who loaned us five additional Sun Ultra 10 workstations. The E220 server hosts all the Cadence tools as well as user's home directories, and also is the license server. Four of the Sun Ultra 10's reside in the Java Projects Lab, a facility available by card access 24-hours a day, hence very suitable for both faculty and student project work. However, this facility will get increasingly heavy use for teaching classes.

To insure adequate access to the Cadence tools at all times for project work, three Ultra 10 workstations will be placed in an alternative location, probably the Electrical Engineering Department's Digital Projects Lab. The installation of the tools on the server was supported by Cadence, who sent a field engineer to work with us on the project for two days. We, in turn, created documentation that Cadence will be able to provide to other customers needing similar support.

CMOS Image Sensors and Infrared Focal Plane Arrays: Current State-of-the-Art

Both visible and infrared image sensors were based on charge-coupled devices (CCD's) during their early development in the 1970's and 80's. Infrared focal plane arrays (IRFPA's) transitioned from CCD to CMOS technology in the late 80's, and visible imagers have increasingly adopted the same technology [Refs. 1-3]. Part of the advantage of CMOS technology is the possibility of integrating an entire camera system on a single chip, enabling higher performance, smaller size and lower cost. An example of such a system is provided in Refs. 4 and 5. The leading characteristic of CMOS technology is reflected in Moore's Law, which states that transistor sizes decrease with time in an exponential manner. The minimum feature size decreases by 0.7x every 3 years [Ref. 6]. As technology advances, more transistors, and hence more complex signal processing functions, can be fit into smaller areas. The consequence of this is that with time, signal processing functions are migrating up the signal chain, toward the pixel.

The most common current architecture can be described as follows. Pixels are arranged in an N columns by M rows matrix, where N might be on the order of 1200, M on the order of 900. Each pixel would ideally contain 1-3 detectors sensitive to different colors, or spectral regions. In current practice, different colors are usually obtained on adjacent pixels, thus we can think of pixels as containing several sub-pixels with different spectal responses. Light is converted to current in each sub-pixel by a photovoltaic mechanism that can be modeled as a current source proportional to light intensity. This is followed by a current-to-voltage converter, invariably a capacitor in one configuration or another. In IRFPA's, it is most often incorporated into a capacitive transimpedance amplifier (CTIA). Light-generated current is integrated on this capacitor, providing a voltage output. The integration operation is one of the most important signal-processing operations in the overall signal chain, providing an important noise-reduction. The integration time is a key system parameter. A sample-and-hold mechanism is included, enabling the output to be stored at the end of the integration time, until it can be transferred to an output buffer register. After the transfer has taken place, the pixel storage capacitors are reset, and a new frame begins. Meanwhile the output buffer register is continually being emptied and re-filled from other pixels, until an entire frame has been read out, after which the process begins again. The output of this register is the video signal that the image sensor array provides to the next processing stage in the system.. This is an analog signal, deriving from the current-to-voltage conversion capacitor at the pixel. A high-speed analog-to-digital converter (ADC) is usually inserted at this point in the system.

Imager performance is measured in such terms as resolution, speed, noise, and dynamic range. The latter performance factors can be improved by moving the ADC up the signal chain, as close to the point of signal generation as possible. Placing an ADC at every column or even in every pixel has been explored [Refs. 9-18]. The space available is the limiting factor. Typical pixel sizes for IRFPA's are 20 - 50 μm square. For visible imagers, 5 - 20 μm pixels are typical. Pixel sizes are ideally determined by the limit of resolution of the light being imaged, which is proportional to wavelength. The larger size of IRFPA pixels comes from the longer wavelength of the light being imaged. IRFPA pixels also have more room available for signal processing than visible pixels, because visible imagers typically use part of the pixel area for the photosensor, whereas IRFPA photosensors are incorporated in a separate non-silicon semiconductor chip that is "hybridized" (flip-chip bonded) to the silicon chip with an indium-bump interconnect in every pixel. References 9 and 10 describe an approach to incorporating ADC in each pixel using a bit-serial approach with 0.35 μm CMOS technology and 10 μm pixels. The layout permitted 6 transistors per pixel, after 25% of the pixel space was allocated to the photodetector. In order to implement ADC at the pixel level, four pixels had to share ADC circuitry, and also ADC had to be implemented in bit-serial fashion. Therefore, a current-to-voltage conversion capacitor is still required in each pixel, capable of sample-and-hold operation, to preserve the analog value after each integration time until it can be digitized and read out in serial fashion.

Pixel Design for Advanced Technologies

IRFPA pixels can use of 100% of the CMOS area for signal processing, because the photodetector is on a separate chip. Based on the experience of Ref. 9, this would allow about 8 transistors per 10 μm pixel, using 0.35 μm technology. The number of transistors per pixel can be assumed to scale as (P2/P1)^2*(L1/L2)^2, where P is pixel linear dimension and L is technology minimum feature size. Using Ref. 9 data as a benchmark, the number of transistors per cell is 8*(P/10)^2*(0.35/L)^2, with P and L in microns. Using current-generation 0.18 μm technology, this becomes 128 transistors per 20 μm pixel. This is a far different constraint than the authors of References 9 and 10 had to contend with, and suggests a different approach, utilizing more processing per pixel. Arguably, we should be designing not for the current 0.18 mm technology, but for the next technology generation, at 90 nm, which will be available in about 2005. Then we will have available 512 transistors per 20 μm pixel. However, we will design with a 128 transistor per 20 μm pixel target, since the additional capability made possible by 90 nm technology may be better used for implementing smaller and/or multicolor pixels. Note that for many IRFPA applications, optical resolution will not be improved for pixels smaller than 40 μm, so that with 90 nm technology we can expect to have available 2048 transistors per pixel!

The availability of hundreds or even thousands of transistors per pixel, rather than the current 6, suggests using a completely different pixel design. Perhaps the simplest design

is as follows: the current-source output of the photosensor is first converted to frequency, and then fed to a digital counter. That is, the current source is first converted to a sequence of pulses, the number in a given time interval proportional to the light intensity. By counting these pulses, a digital representation of the light intensity is obtained. This is a form of full-parallel ADC per pixel. The inherent simplicity of this approach is appealing. Similar designs have been discussed in the literature [Refs. 19-25]. Two key components are required: the current-to-frequency converter [Refs. 19-22] and the counter [Refs. 26-28].

Current efforts are directed toward the design and layout of these two components. For initial planning purposes, we reserve 32 transistors for the current-to-voltage converter and pixel control circuitry, leaving 96 transistors for the counter. We further assume an 8-bit counter, hence need a counter design that can be realized with 12 transistors per bit, including control of the shift operation needed to read out the data.

Image Sensor Architectures

An image sensor can be thought of as an N x M pixel array. Each pixel may need K sensors of different spectral response (color), resulting in NMK analog data values per frame. Each of these values may be obtained at different sensitivities (light-to-electrical conversion factors), specified by S sensitivity bits, and may be converted to digital form with R resoluiton bits, resulting in Q = R + S quantization bits. The output of the sensor may be on P parallel lines. The integration time is Ti and frame time is Tf. In coventional expose/readout mode, Tr = Tf-Ti is the time available for readout of all the values before a new frame begins. Date acquisition can be pipelined if simultaneity within a given frame is not mandatory, resulting in Tr almost equal to Tf, if needed. Alternatively (or additionally), increasing the number of parallel outputs P will reduce data rate. This is important because power consumption is proportional to data rate, so strategies for keeping the data rate low should be considered, even if not essential to meet other design objectives.

The sensitivity bits (S) deserve further discussion. These implement the floating-point data concept described in Reference 10, and will be included in a sensor design when very wide dynamic range is required. Pixel data is represented as the sum over s of r*2^s, where s is an S-bit binary number (0 </= s < S) and r is an R-bit binary number (0 </= r(s) < R). It is equivalent to q, a Q-bit binary number where Q = R + S. However, however the S-bits are derived from different photosensor sensitivities, rather than increased ADC resolution. This is significant for image sensor arrays, because dynamic range is limited by noise. The ADC resolution cannot be made indefinitely fine; there is no further gain once the resolution goes below the noise. Instead, the range must be extended upward, without increasing power supply voltages and currents, since power consumption must almost always be kept as low as possible. The way to do this is with selectable pixel sensitivities. After the the pixel has saturated at high sensitivity, it is switched to a lower sensitivity mode. For a conventional current-to-voltage converter, this might mean

switching to a higher value of integrating capacitor. For a current-to-frequency converter, this means switching to fewer pulses/charge, which might also be implemented with switchable capacitors.

Our initial architecture will be as follows. We assume single-color operation (K = 1). Each of the M rows will be selected in sequence, and N columns will be read in parallel into a N x Q memory cache (an output buffer "shift register" having Q bits in parallel rather than just one). The data will then be multipelxed out onto P parallel lines. Initial design will assume S = 0 (no sensitivity adjustment in the pixel), R = 8 (8-bit ADC resolution), and P = 1.

Sub-systems ot be designed include the row select module, column amplifiers, output cache, output amplifier(s), and control module.

Further Work

One of the next tasks is the system-level simulation of the proposed chip architecture using the Verilog-AMS hardware description language. Noise modeling and simulation will be incorporated, including noise due to fundamental processes, and noise due to unintentional signals. Design for manufacturability, including design for test, will be emphasized. We will assess our design using Monte Carlo methods to simulate the random variation during manufacturing of key parameters, and their effect on yield. Test is a major cost driver, to be addressed by creating software test benches for our design. A software test bench is an HDL program used to verify ("test") another HDL program. This is a standard design technique for digital VLSI systems, which we expect to be able to extend to analog/mixed signal systems.

Another task is to continue the detailed design of sub-systems, especially the pixel. This task entails the transistor-level simulation of alternative designs for the current-to-frequency converter and counter, and physical layout and verification. Other sub-systems include control, clock generation, row select, and column amplification and multiplexing functions.

The final task is to integrate the subsystems into the complete SoC, and perform chip-level physical verification (parasitic extraction, simulation, signal integrity analysis).

References

1. E. R. Fossum, "Active pixel sensors: are CCD's dinosaurs?" in *Charge-Coupled Devices and Solid State Optical Sensors III*, M. M. Blouke, Editor, Proc. SPIE 1900, pp. 2-14 (1993).
2. L. J. Kozlowski, J. Luo, and A. Tomasini, "Performance limits in visible and infrared image sensors." IEEE International Electron Devices Meeting, 1999, pp. 867-870.

3. L. J. Kozlowski et al., "Theoretical basis and experimental confirmation: why a CMOS imager is superior to a CCD," in Infrared Technology and Applications XXV, B. F. Andresen and M. S. Scholl, Editors, Proc. SPIE 3698, pp. 388-396 (1999).
4. M. J. Loinaz et al., "A 200-mW, 3.3-V, CMOS color camera IC producing 352 x 288 24-b video at 30 frames/s." *IEEE J. Solid-State Circuits* 33, 2092-2103 (1998).
5. M. Loinaz and B. Ackland, "Video cameras: CMOS technology provides on-chip processing." *Sensor Review* 19(1), 19-26 (1999).
6. H.-S. Wong, "Technology and device scaling considerations for CMOS imagers." *IEEE Trans. Electron Devices* 43, 2131-2142 (1996).
7. C. Shen et al., "Low voltage CMOS active pixel sensor design methodology with device scaling considerations." *IEEE Hong Kong Electron Devices Meeting*, 2001, pp. 21-24.
8. A. Afzalian et al., "Comparison of bulk vs SOI for low power low voltage CMOS imager." *IEEE International SOI Conference*, 2001, pp. 133-134.
9. D. X. D. Yang, B. Fowler, and A. El Gamal, "A Nyquist-rate pixel-level ADC for CMOS image sensors." *IEEE J. Solid-State Circuits* 34, 348-356 (1999).
10. D. X. D. Yang, A. El Gamal, B. Fowler, and H. Tian, "A 640 x 512 CMOS image sensor with ultrawide dynamic range floating-point pixel-level ADC." *IEEE J. Solid-State Circuits* 34, 1821-1834 (1999).
11. W. Lian et al., "Sensor array with A/D conversion based on flip-flops." *Sensors and Actuators A* 21-23, 592-597 (1990).
12. G. de Graaf, F. R. Riedijk, and R. F. Wolffenbuttel, "Colour-sensor system with a frequency output and an ISS or I2C bus interface." *Sensors and Actuators A* 61, 441-445 (1997).
13. W. Mandl and C. Rutschow, "All digital monolithic scanning readout based on sigma-delta analog to digital conversion," in *Infrared Readout Electronics*, E. R. Fossum, Editor, Proc. SPIE 1684, pp. 239-244 (1992).
14. W. J. Mandl, "Focal plane analog to digital conversion development," in *Smart Focal Plane Arrays and Focal Plane Array Testing*, M. Wigdor and M. A. Massie, Editors, Proc. SPIE 2474, pp. 63-71 (1995).
15. W. Mandl and R. Fedors, "Direct digital conversion detector technology," in *Photonic Device Engineering for Dual-Use Applications*, A. R. Pirich, Editor, Proc. SPIE 2481, pp. 170-184 (1995).
16. W. Mandl, "Visible light imaging sensor with A/D conversion at the pixel," in *Sensors, Cameras, and Systems for Scientific/Industrial Applications*, M. M. Blouke and G. M. Williams, Jr., Editors, Proc. SPIE 3649, pp. 2-13 (1999).
17. C. Jansson, U. Ringh, and K. Liddiard, "On-chip analog-to-digital conversion suitable for uncooled focal plane detector arrays employed in smart IR sensors," in *Smart Focal Plane Arrays and Focal Plane Array Testing*, M. Wigdor and M. A. Massie, Editors, Proc. SPIE 2474, pp. 72-87 (1995).
18. U. Ringh et al., "CMOS analog to digital conversion for uncooled bolometer and infrared detector arrays," in *Smart Focal Plane Arrays and Focal Plane Array Testing*, M. Wigdor and M. A. Massie, Editors, Proc. SPIE 2474, pp. 88-97 (1995).

19. Y. Kuroda, A. Hyogo, and K. Sekine, "A current-to-frequency converter for switched-current circuits." *Analog Integrated Circuits and Signal Processing* 20, 145-148 (1999).
20. J. Ohta et al., "Pulsed vision chip with inhibitory interconnections," in *Optics in Computing 2000*, J. Smith, Editor, Proc. SPIE 4089, pp. 488-495 (2000).
21. J. Ohta et al., "Low-voltage operation of a CMOS image sensor based on pulse frequency modulation," in *Sensors, Cameras, and Systems for Scientific, Industrial, and Digital Photography Applications II,* M. M. Blouke, J. Canosa, and N. Sampat, Editors, Proc. SPIE 4306, pp. 319-326 (2001).
22. A. Bermak, A. Bouzerdoum, and K. Eshraghian, "A digital vision sensor with pixel level analog-to-digital converter," in *Electronics and Sensors for MEMS II*, N. W. Bergman, Editor, Proc. SPIE 4591, pp. 353-358 (2001).
23. K. Eshraghian and S. Lachowicz, "Fully digital pixel readout architecture with a current-mode A/D converter," in *Design, Characterization, and Packaging for MEMS and Microelectronics II*, P. D. Franzon, Editor, Proc. SPIE 4593, pp. 228-233 (2001).
24. M. L. Simpson et al., "An integrated CMOS microluminometer for low-level luminescence sensing in the bioluminescent bioreporter integrated circuit." *Sensors and Actuators B* 72, 134-140 (2001).
25. E. K. Bolton et al., "Integrated CMOS photdetectors and signal processing for very low-level chemical sensing in the bioluminescent bioreporter integrated circuit." *Sensors and Actuators B* 85, 179-185 (2002).
26. J. Mavor, M. A. Jack, and P. B. Denyer, *Introduction to MOS LSI Design*, Addison-Wesley (1983).
27. J.-R. Yuan, "Efficient CMOS counter circuits." *Electronics Letters* 24, 1311-1313 (1988).
28. J. Yuan and C. Svensson, "New single-clock CMOS latches and flipflops with improved speed and power savings." *IEEE J. Solid-State Circuits* 32, 62-69 (1997).

Development of an Autonomous Tactical Reconnaissance Platform

Project Investigators:

Sema E. Alptekin, Ph.D.
Professor and Chair
Industrial and Manufacturing Engineering Department

Dianne DeTurris, Ph.D.
Assistant Professor
Aerospace Engineering Department

Development of an Autonomous Tactical Reconnaissance Platform

Abstract

This research effort led to the invention and prototyping of a man-portable, autonomously controlled aerial surveillance platform. The ultimate goal in this development effort is to provide a method for obtaining remote sensing data using an inexpensive, expendable device. There is a wide array of potential uses for this device in both military and civilian applications. The baseline scenario used to define our performance objectives is the design of a device weighing less than 1.5 lbs. that is launched to an altitude of 300 meters, where it deploys and flies a user selected program of autonomous flight maneuvers. In this baseline scenario, telemetry from the ATRP provides the user with real time output from onboard sensor instrumentation.

During the first phase of this ongoing research, an Autonomous Tactical Reconnaissance Platform (ATRP) was developed consisting of a standard parafoil that carries below it instrumentation for autonomous flight and remote sensing capabilities. In initial prototype test flights, the controller board, associated battery and guidance sensors added about 1 lb. to the total system weight. At the same time the cost of these components was less than $1,000 USD. The technology used in the ATRP is commercial-off-the-shelf (COTS) components, easily acquirable and inexpensive to manufacture.

A range of launch mechanisms including rocket, compressed air, artillery shell and tethered flight (flown like a kite) are possible with this device. It was decided to use a rocket launch mechanism to build on earlier success with a rocket launched parafoil. A 4-ft tall rocket was designed and prototyped. Extensive ground testing was conducted to characterize the flight performance and load carrying capacity of the smaller parafoils used in this project.

An autonomous, fuzzy logic flight controller was developed which pilots the ATRP. A fuzzy logic architecture was developed and optimized over a range of simulated flight conditions. The fuzzy logic flight control algorithm for the ATRP provided a simple yet robust means of controlling the vehicle, thus enabling the use of low cost, light weight components. Prior to launch the user selects one of a finite set of pre-programmed flight scenarios. After launch and deployment, the autonomous controller executes the pre-selected program of flight maneuvers as it glides in the descent phase. A fuzzy logic control method is employed because of advantages it has in fault tolerance and graceful response to missing or noisy sensor input.

All original design and development goals were either met or exceeded, however, many hurdles remain before the demonstration of a successful baseline product will be achieved. Follow-on efforts will concentrate on developing the real-time link between sensors on board the ATRP and ground instrumentation. A number of improvements in payload weight reduction, instrumentation, and improved instrumentation accuracy will also be explored with follow-on funding. Highlights and additional information on this project including video of a rocket launch can be accessed at the web site,

http://www.ime.calpoly.edu/salpteki/Parafoil/Parafoil-Index.htm.

1. Introduction

The ATRP device is a parafoil with a payload that is instrumented for autonomous flight and remote sensing capabilities. A parafoil is a flying wing made of some flexible material, supported by lines that maintain its shape by virtue of the air flowing over and through it. The purpose of this project is to prototype a system that can be used as a personal reconnaissance device. The ATRP would have many military and civilian uses that are not adequately addressed by remote sensing systems currently available. This project has potential for military and civilian applications by greatly enhancing the ability to obtain information inexpensively.

Reconnaissance data from various sources have proven to be a significant contributor to the success of most battle campaigns, particularly in recent years. Today our military employs a number of reconnaissance assets including satellites, manned aircraft, unmanned aircraft and human infiltration. All of these assets can provide vital information to commanders at home and in the field, however they all have drawbacks when it comes to supporting the individual soldier in real-time, hostile environments.

In its military version the ATRP would provide the soldier with a close range, real-time view of the immediate surroundings. This capability would be extremely valuable in rugged or jungle terrain and during nighttime operations. An infrared and/or visual sensor aboard the ATRP would pinpoint enemy location and numbers during such exercises. The need for real-time reconnaissance data is not limited to the battlefield.

Non-military uses include providing fire location and egress points for forest fire fighting crews and aiding search and rescue teams in locating victims of avalanche and lost hikers. This device also shows great promise in helping farmers and land management professionals to determine areas of plant stress due to drought, pests, salt intrusion and disease. Another potential use for the ATRP is to provide an enhanced capability for long-range communications in obstructed or rugged terrain. Other life saving and communications scenarios can be implemented for this device when outfitted with the appropriate instrumentation.

One of the more promising non-military applications for the ATRP is to provide inexpensive and good quality remote sensing data to farmers and land management professionals. A number of studies have shown the value of remote sensing data in discovering areas in fields where crops are under stress due to disease, lack of nutrients or pests [5,6]. Current collection methods such as satellites and manned aircraft are expensive, geographically imprecise, and frequently unavailable at critical times in the growing season. The ATRP equipped with a Global Positioning System and the appropriate sensors would overcome all of these obstacles and would allow farmers to perform their own personalized data collection and evaluation.

The ATRP is designed to provide personnel in the field with a means of expanding their knowledge of the near environment in a safe, timely and cost effective manner. The ATRP

is launched to low altitude (typically <1000 feet) and then flies a pre-selected course while transmitting sensor data to the user on the ground. During flight, the device is fully autonomous allowing the user to direct their concentration on the sensor data (i.e. IR imaging, visual imaging, radio communication and potentially laser designation) being communicated to a hand held receiving device. Prior to launch, the user can select one of several pre-programmed flight patterns for the device to follow and then launch, all in a matter of seconds.

The reason a parafoil design was chosen is that it has many advantages. A parafoil is lightweight, easily compacted and deployed, and very stable in flight. An autonomous, fuzzy logic flight controller pilots the parafoil for the ATRP. Prior to launch the user will select one of a finite set of pre-programmed flight scenarios. After launch and deployment, the autonomous controller executes the pre-selected program of flight maneuvers as it glides in the descent phase. A fuzzy logic control method is employed because of advantages it has in fault tolerance, graceful response to missing or noisy sensor input and compact memory requirements.

The ATRP and its launch mechanism are man-portable, with set-up and launch operations achieved in a matter of seconds. A range of launch mechanisms including rocket, compressed air, artillery shell and tethered flight (flown like a kite) are possible with this device. The ATRP is intended to be an inexpensive, expendable device, suitable for a wide variety of mission scenarios. The technology used in the ATRP is commercial-off-the-shelf (COTS) components, easily acquirable and inexpensive to manufacture.

Grants previously received by the PI have provided an excellent foundation for this project. A team of faculty members and undergraduate students developed a smart product as part of this project. This product made use of various sensor devices and a microprocessor based control unit as it tracked a human target. A byproduct of this project was the development of a Mechatronics Laboratory that will serve as one of the facilities for this research. Publications that resulted from this work are cited in the references section of this proposal [10-17].

The research collaboration between the Aero Department and the IME Department goes back to 1995 when a three year grant was funded by NASA's Multidisciplinary Design and Analysis Program. Several graduate student research projects were funded by this project. Joseph Fournell was one of these graduate students who investigated the utilization of Fuzzy Logic in the control of an autonomous device: a Fly-by-Wire System. Ervin and Alptekin subsequently used MatLab and simulated flight data as they developed a Fuzzy Logic Controller as part of the Fly-by-Wire System. Publications resulting from this grant are provided in the references section of this proposal [4,7,8,9].

Development of the rocket launch mechanism is provided in Section 2. Design and development of electronic hardware is summarized in Section 3. An overview of the development of the Fuzzy Logic Algorithm and its optimization through simulation is provided in Sections 4 and 5. Results of flight tests can be found in Section 6, followed by conclusions in Section 7, and recommendations for future work in Section 8.

2. Rocket Development and Construction

In order to support the goal of creating a man-portable device, it was decided that an operational ATRP should not exceed 3 feet in length. Alternative launch mechanisms (e.g. compressed air launch, artillery shell, tethered launch and deploy) may prove to be even more compact than the current rocket launched design. However, a rocket length of 3 feet or less should prove to be easily manageable for one person to carry even in rugged terrain. This constraint in length also helps to determine various other constraints such as; system weight, rocket motor size, parafoil sizing, and maximum altitude of flight and deploy.

A standard motor burnout delayed ejection charge was used to separate the nose from the rocket body and release the parafoil. In keeping with the goal to control costs and enhance manufacturability, only commercially available solid rocket motors were purchased and used in test flights. These motors were sized to allow the rocket reach to apogee and deploy the parafoil at an altitude of between 500 and 2000 ft., depending on the goals of each particular test flight.

Prior experience with a much larger rocket and parafoil, as picture in Figure RDC-1, demonstrated that the parafoil was relatively stable and easy to fly in radio control mode. This 10-ft tall rocket and parafoil were developed under an earlier project supervised by Dr. DeTurris. Extensive ground testing was needed to help characterize the flight performance and load carrying capacity of the much smaller parafoils and rockets used in this project. The first parafoils purchased for this project were commercially available kites and there was some concern that they would not be optimal for free flight. One of these kite parafoils is pictured in Figure RDC-2 during early ground test evaluation.

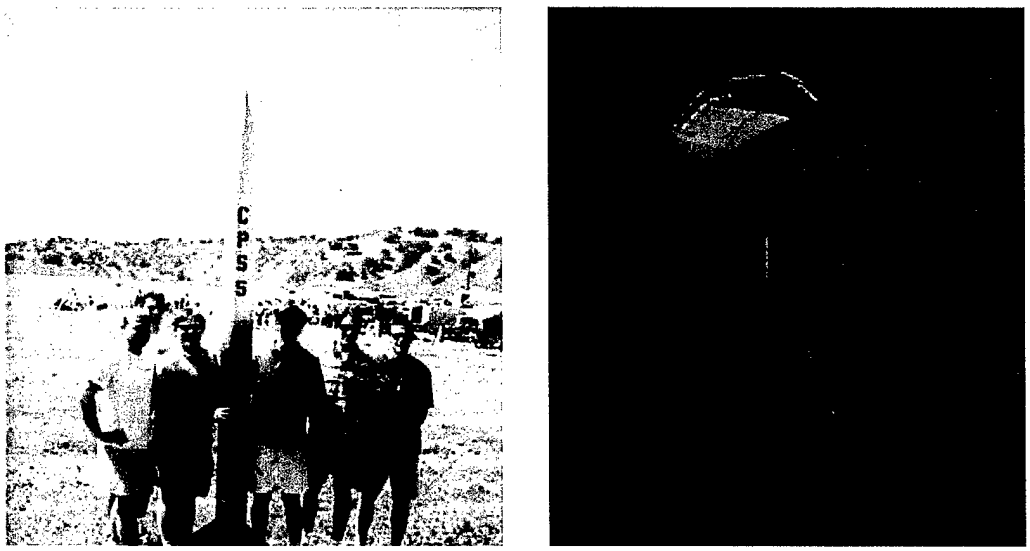


Figure RDC-1 Cal Poly High Power Rocket with Parafoil Recovery System

Another concern with these kite parafoils was related to the geometry of the two control lines. When flown as a kite, these control lines are spaced about 2 feet apart up near the kite body as shown in figure RDC-2. It was suspected and soon confirmed that this

separation of control lines was necessary for stable flight. Since the control package would be much narrower than 2 feet, a mechanism had to be constructed to separate these lines after the parafoil had been deployed from the rocket. This was accomplished using two attached rods with a spring loaded core carried aloft on the outside surface of the rocket. During deploy the rods straightened at the bend and the core rod was driven through the joint to provide a rigid structure. Control lines were threaded through the ends of this rod providing the control line separation needed for stable flight. Figure RDC-3 shows the control line separation rod in the stowed position on the exterior of the rocket body.



Figure RDC-2 Early Ground Testing of a Commercial Kite Parafoil



Figure RDC-3 Rocket Being Prepared for Launch (note folded separator rod stowed on the side of the rocket)

It soon became apparent that there were several drawbacks with this design of parafoil with deployable separator rod. Having the parafoil deployment occurring at the same time as the separator deploy proved to be a complicated event with a high probability of tangles and breakage of control lines. In addition, the parafoil pictured in Figure RDC-2 and the larger version of the same model pictured in Figure RDC-4 were designs with a large number of control and support lines, which added to the potential for tangles during deploy. Ultimately, it was determined that fixing the drawbacks of these parafoils would consume too much time and energy of the design team and a new parafoil solution was sought as a replacement.

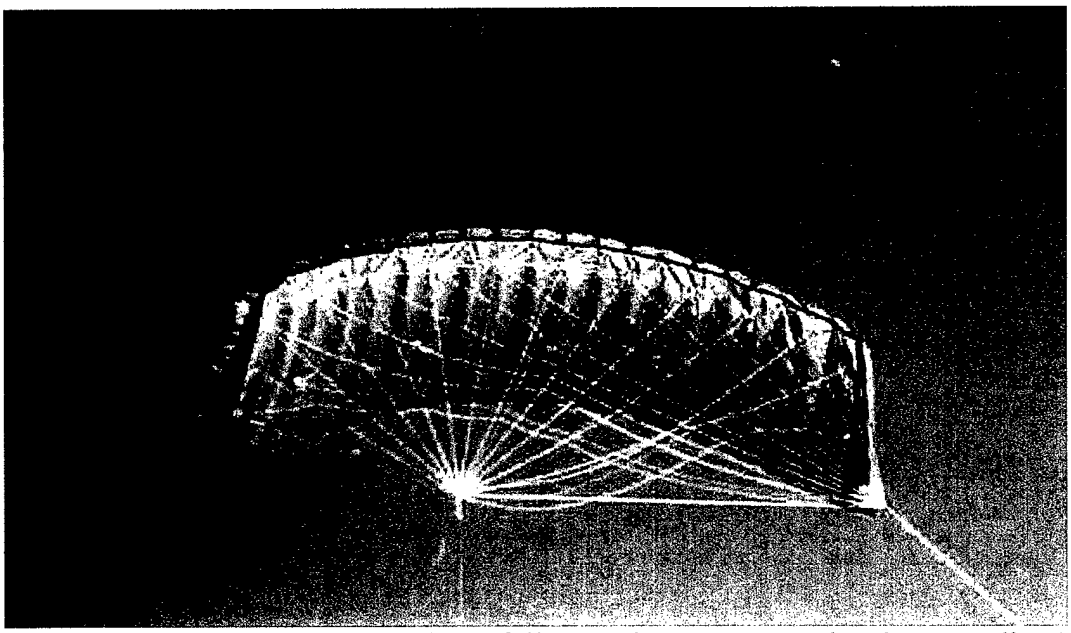


Figure RDC-4 Large Early model Parafoil (note the many control and support lines)

The search for a new parafoil resulted in the purchase of a radio controlled powered parafoil. This product consists of a small radio controlled, gas powered engine that is suspended below a parafoil canopy. The gas motor was not used for this project, but the parafoil proved to be far superior to earlier models in several important respects. The new parafoil was designed for free flight and therefore had the correct line lengths to give the proper angle of attack while flying. This model also had fewer lines attached to the parafoil, thus reducing the possibility of tangles. It also was designed for the control lines to angle much closer together, which reduced the length of the separator rod needed for stable flight. This new parafoil with the gas engine is shown below in Figure RDC-5.

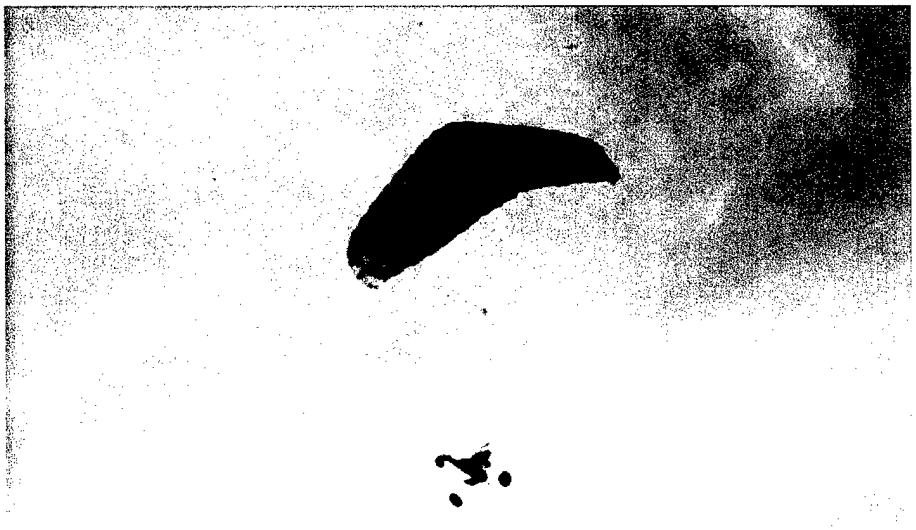


Figure RD-5 The New Parafoil as delivered with Suspended Motor

The much smaller separation distance required for the control lines made it possible to launch the rocket with the separator rod already deployed. The rocket with the separator rod in position is pictured below in Figure RDC-6. There was very little drag evident with the rod in this position and this design proved to be very reliable in deploying properly.

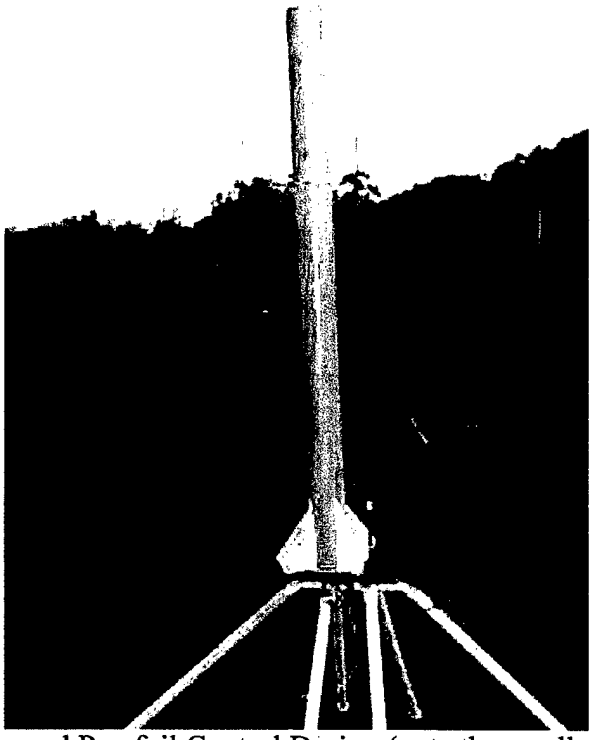


Figure RDC-6 Improved Parafoil Control Design (note the small control line separator mounted sideways through the rocket)

Another improvement made from early designs was the significant reduction in weight achieved by using lighter materials and better construction techniques. In one sense it would seem that keeping the rocket low in weight would not be important in a prototype vehicle since larger motors are easily purchased that can launch the rocket to the desired deploy altitude. However, reducing the rocket weight was a significant improvement from a practical stand point, because larger rocket motors can only be used in strictly controlled designated areas. This meant that the first launches of the heavier rockets were performed as far away as Fresno and the Lucerne Valley. This placed a practical limit on the frequency of test launches and significantly slowed development. The lighter rocket allowed for the use of less powerful rocket motors resulting in a much less restricted launch

regimen. Launches could be performed in virtually any open field, which allowed for an increase in the frequency of launches and the development pace.

3. Electronic Hardware

The control function is provided by a Motorola M68HC11 microcontroller on a Handy Board with a system clock speed of 2 MHz. This unit also has 32 kilobytes of battery backed RAM memory for user programming. For this project we also purchased the optional extension, the Expansion Board. The Handy Board/Expansion Board is equipped with one RS-232 serial port, 21 analog input ports, 8 digital input ports, 9 digital output ports and 6 servo motor output ports. Programming is performed on a PC and the resulting code is downloaded over the RS-232 port via a separate Interface/Charger board (Figure EH1) to the Handy Board micro controller.

A Precision Navigation Inc. TCM2-50 digital compass was used to provide directional input to the Handy Board (Figure EH2). An RS-232 communications channel was also needed for the Handy Board to communicate with this compass. Fortunately the Handy Board manufacturers provide a small amount of board space to accommodate additional, user developed, circuitry. A circuit design incorporating a Maxim MAX232CPE interface chip and associated electronics was developed and soldered into place on the board by project team members. This circuitry was configured to allow the Handy Board to communicate with the compass. Various pieces of the software that control the acceptance of input data from the compass were obtained as freeware, from various sources on the internet [23-27]. These pieces were combined and modified by team members to perform the necessary communications task.

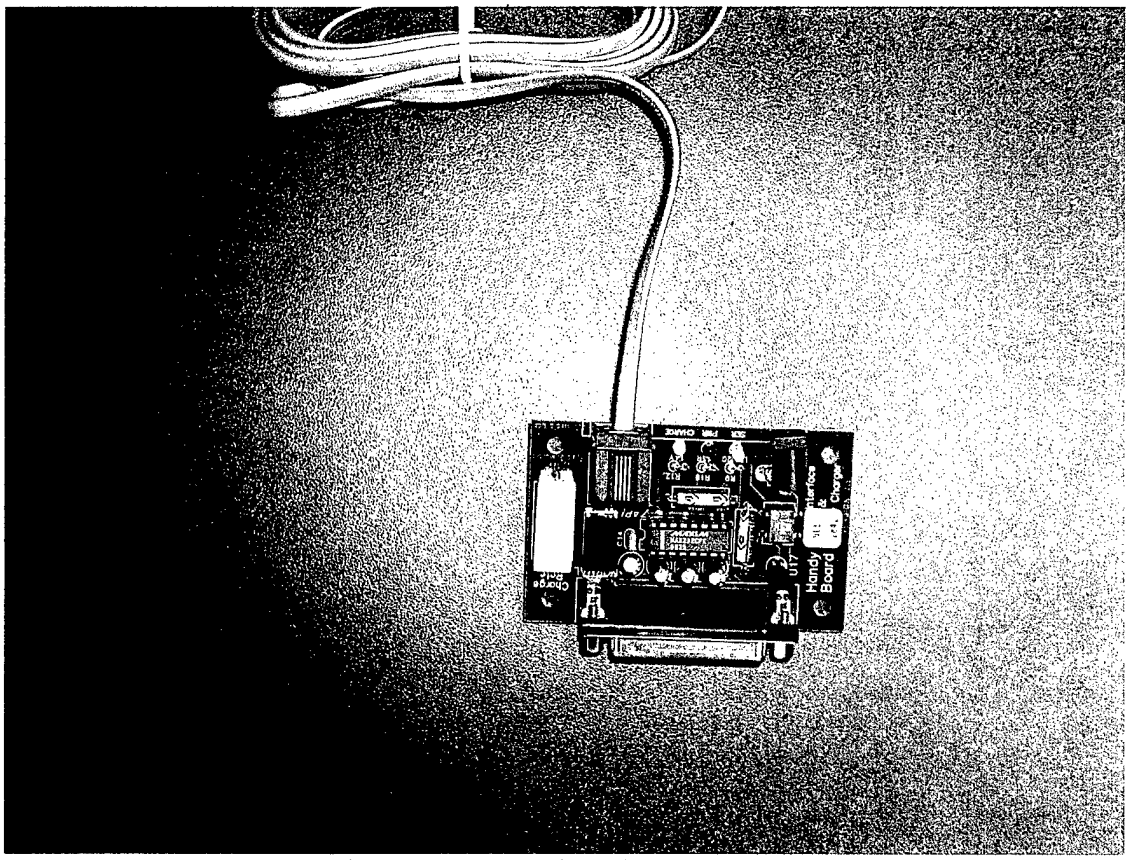


Figure EH1 Interface/Charger Board

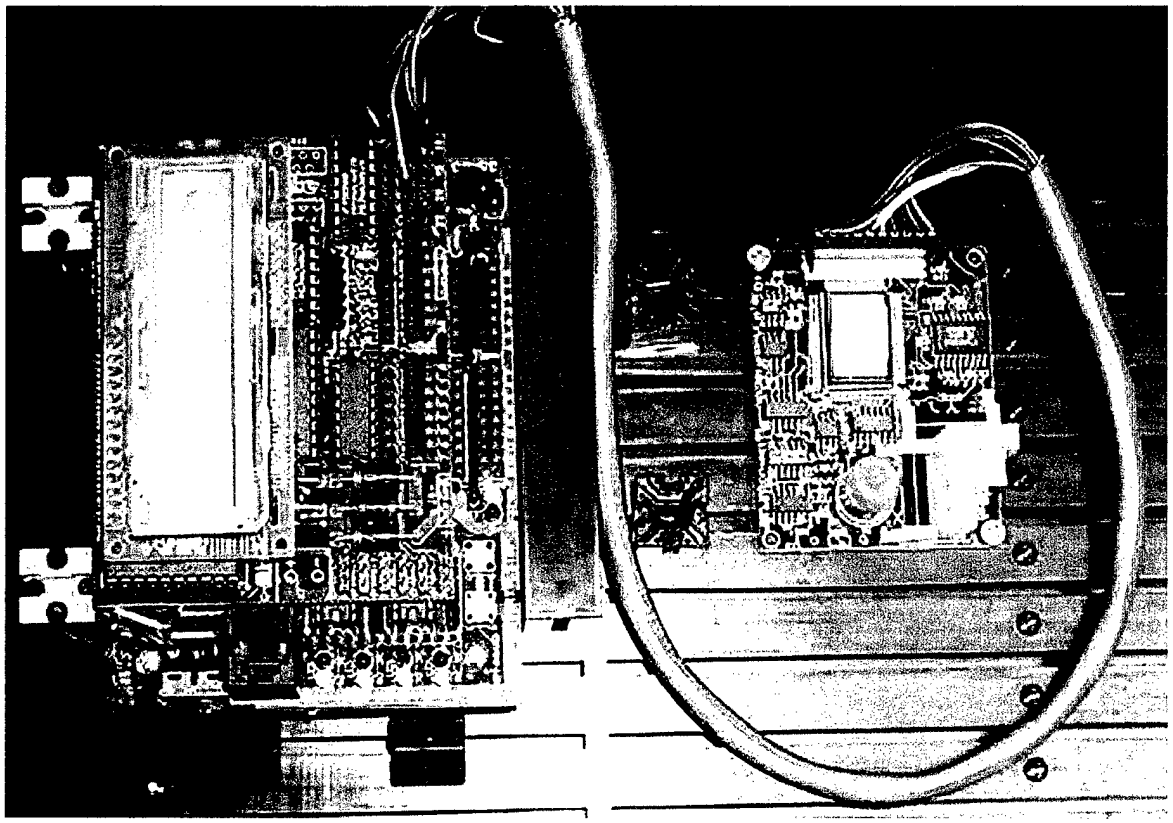


Figure EH2 Handy Board (left) and Compass (right)

The TCM2-50 compass was the single most expensive electronics component purchased for this project. The TCM2-50 spec sheet claims an accuracy of ±1.5° RMS even when tilted from level by as much as ±50°. The compass is capable of providing both digital and analog output, but is considered to be more accurate in the digital mode. It also has several auxiliary outputs including; temperature, pitch and roll angle, and three axis measurements of the local magnetic field.

Once the compass has been positioned in the electronics assembly package, it must be calibrated to eliminate local magnetic distortions. These distortions arise from close proximity to ferrous materials and magnetic fields surrounding various electronic components and support structures. In calibration mode, the compass is tilted and rotated through one or more full circles thus allowing the compass to collect magnetic field data in three dimensions. The three dimensional map of the distorting fields is computed and stored in the compass electronics.

The accuracy of the compass even when tilted to extreme angles was considered very important during the initial design phase. There was a great deal of uncertainty in how much tilt motion would be experienced by the instrument package during flight. Video of subsequent parafoil flight tests suggests that these tilt angles are not as severe as expected. In Phase II of this project an array of up to 5 cheaper compasses, with only limited tilt compensation capability, will be tested and compared to the performance of the TCM2-50.

These cheaper compasses would have advantages with regard to cost, ruggedness, and redundancy.

An Analog Devices, Inc. ADXL190 accelerometer was also included in the electronics package (Figure EH3). This device is capable of measuring forces up to ±100 g's along a single axis. Although output from the accelerometer was not used directly as input to the guidance package, data collected from this instrument was stored on the Handy Board to aid in future design efforts. Early in the ATRP design, the accelerometer was considered as a means of establishing when the deployment charge had been fired. This concept was discarded and a simple magnetic switch located at the separation seam was used to signal that the deploy charge had fired.

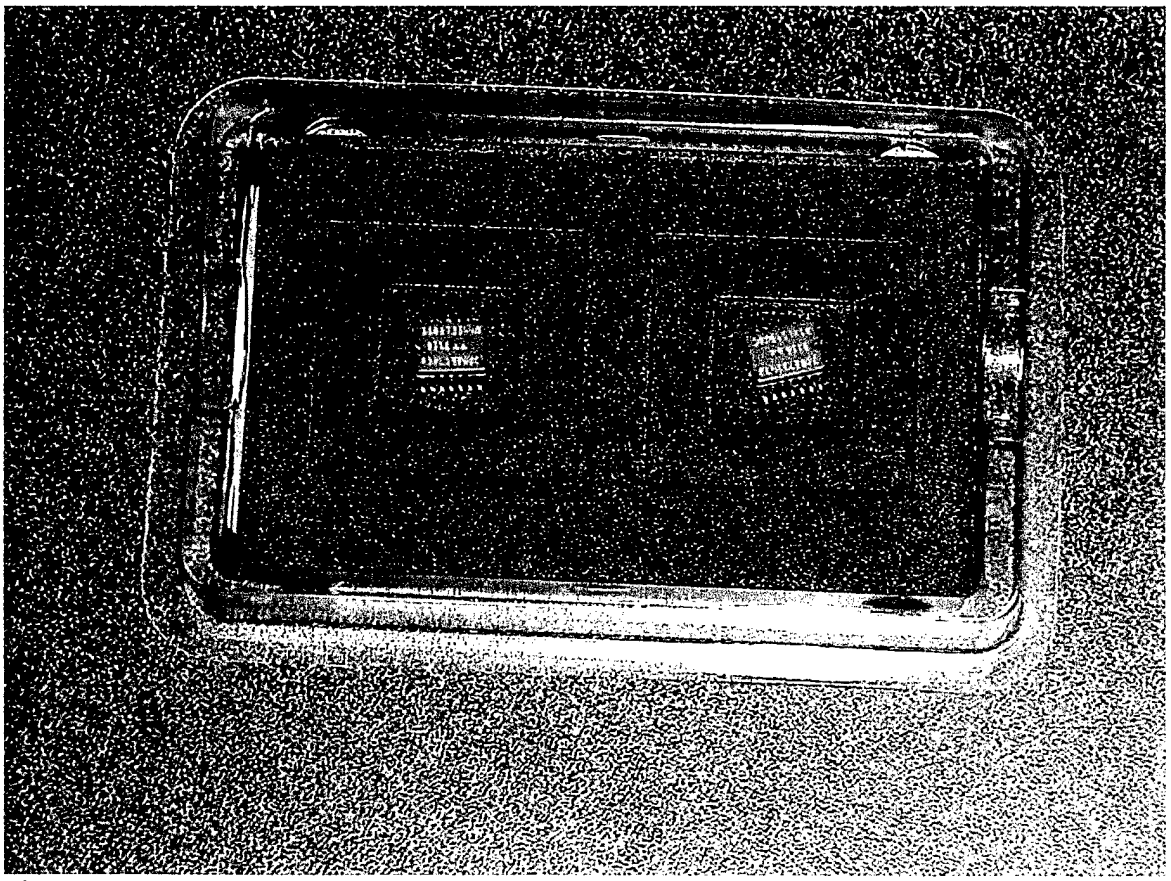


Figure EH3 A pair of ADXL190 accelerometers provided courtesy of Analog Devices, Inc.

Communication checks between the various electronic components were carried out on the bench top. Figure EH4 shows all of the various flight components connected for testing and software download. Simulations of the hardware, performed in the MATLAB Simulink software environment were verified by simulated flights performed on the breadboard configuration pictured below. This synergy between computer simulation and breadboard verification played a vital role in the overall optimization process.

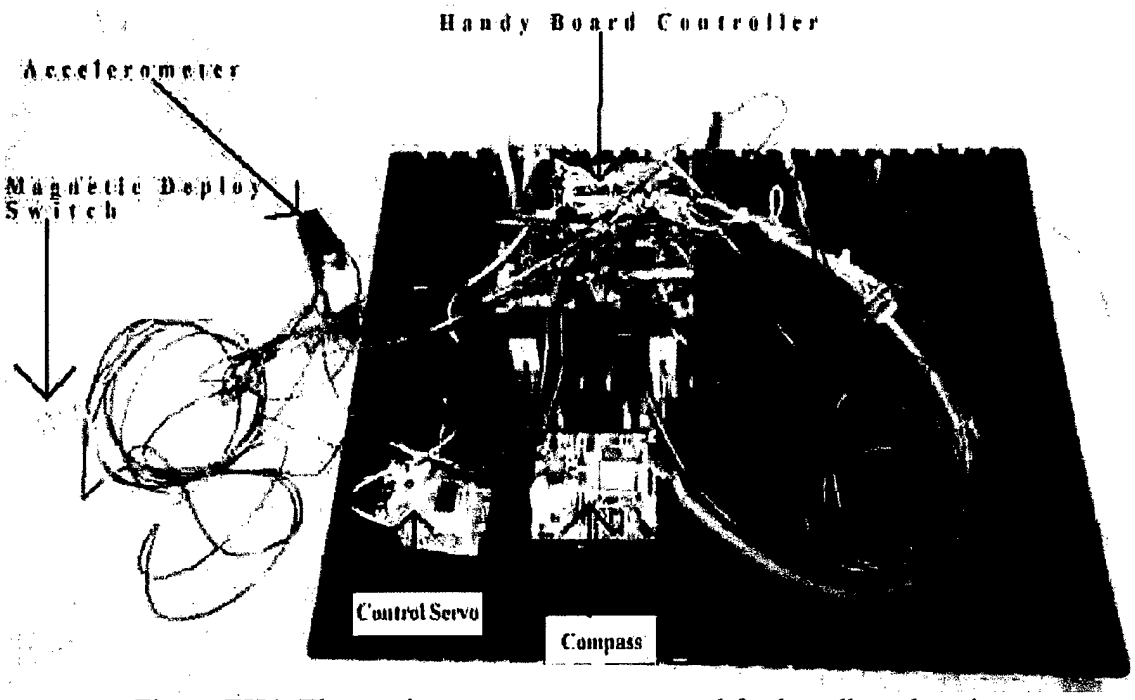


Figure EH4 Electronic components connected for breadboard testing

Assembling the various components into a rugged package that would fit within the confines of the rocket body was a subject of significant development effort. The eventual package design consisted of several lightweight bulkheads joined together by 4 threaded metal rods. The original steel joining rods had to be replaced with brass when it was determined that the steel interfered with the magnetic compass calibration. Figure EH5 shows the electronics package prior to insertion in the rocket.

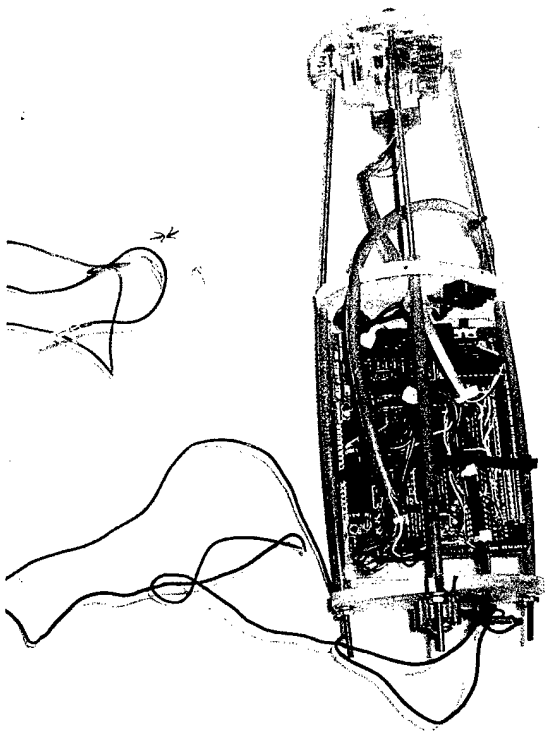


Figure EH5 The electronics package prior to insertion into the rocket

Access to the electronics package was provided by a pair of holes drilled through the rocket body. One hole provided access to the power switch which allowed us to power up the electronics on the launch platform, thus conserving battery power just prior to launch. The other hole provided access to a switch that initiated a two minute countdown. A green LED was lit when this switch was activated and then went out at the end of the two minute countdown. Launch was manually initiated from a safe distance at the end of this electronic countdown.

4. Development of the Fuzzy Logic Algorithm

A fuzzy logic control algorithm provides the decision making strategy for the ATRP. There are a number of other excellent control methodologies available and indeed proportional integral derivative (PID) algorithms of one type or another are used in 90% of current controllers. PID controllers have been enormously successful in a variety of applications and in most cases have proven to be both very efficient and accurate. However, the fuzzy logic controller does have distinct advantages in our particular application.

The credit of the invention of Fuzzy logic is generally given to Dr. Lotfi Zadeh. Dr. Zadeh and others noted that most circumstances in the real world are not binary in nature. For instance, the temperature of the bath water may be hot or it may be cold, but it also might be warm. Indeed, the definition of what is hot, cold and warm will vary among individuals. Fuzzy logic was invented in part to deal with a continuum of possible input states in a variable while still providing a single, crisp output. Advantages of fuzzy logic control

include; robust handling of noisy and/or missing input data while providing crisp output commands, and a flexible and easily modified control architecture. Fuzzy logic controllers are seeing much more widespread use in a wide range of industrial and consumer devices (i.e. washing machines, automobiles and microwave ovens).

Attributes of fuzzy logic that make it appealing for this project are the ability to model nonlinear functions, robustness in the face of imprecise input and ease of code generation. Fuzzy logic algorithms are intuitively easy to understand and allow the user to encapsulate the experience of experts in an efficient manner. In the following paragraphs a brief description of the development of the fuzzy logic algorithm is provided, many of the references cited at the end of this report can give the reader a much more comprehensive insight into fuzzy logic fundamentals.

We began the process of developing the fuzzy logic architecture by determining what inputs and what instruments will be used to provide those inputs, in order to provide directional control of the ATRP. A number of quantifiable factors including cost, weight, size and instrument accuracy as well as more subjective factors such as ease of use and compatibility were part decision of the process. Ultimately a digital compass was selected to provide heading information in this first prototype effort.

The compass provides heading data with respect to true magnetic North and by differentiating this heading by time we are able to determine angular velocity. These two parameters, absolute heading and angular velocity, were considered to be the minimum inputs necessary for directional control. A third input, angular acceleration, was also considered with many simulation runs performed to determine whether it would be required. Angular acceleration was obtained by differentiating angular velocity over time. There was a trade-off in including angular acceleration between improved directional performance versus processing speed and memory requirements and ultimately it was determined to be unnecessary to include it.

The next step in the process was to develop membership functions for the three inputs (heading, angular velocity and angular acceleration) and the one output (control line pull). The magnitude and sign of the output would determine the magnitude and direction of the pull on the control lines, which in turn determined the magnitude and direction of turns of the ATRP. Figure DFL-1 shows the membership functions for the Heading input as displayed in the MATLAB Fuzzy Logic Toolbox development software. In this figure one of the membership functions, positive small (psmall), is highlighted in red.

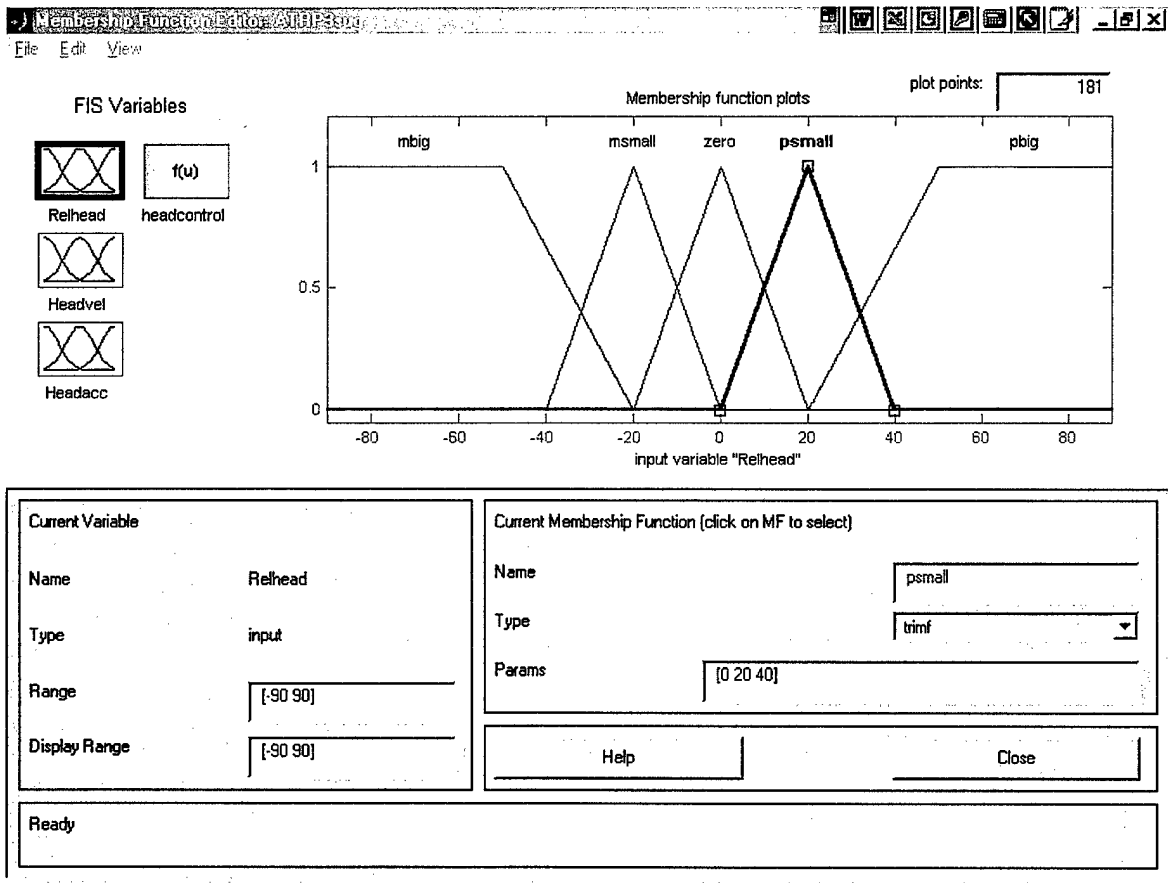


Figure DFL-1 Membership functions for Heading input

The three membership functions in middle of figure DFL-1 are triangular in shape while the outside two are trapezoidal. Each vertex of these 5 functions can be relocated to control the spacing and the amount of overlap in the pursuit of optimizing performance. Each of the inputs and the output have a similar degree of freedom in the arrangement and number of membership functions, thus providing an infinite variety of possible combinations. This extreme flexibility is something of a double-edged sword in that it provides the means of optimizing the system but also makes the optimization process a very daunting task.

After an initial number and shape are determined for the membership functions for each input and output, then rules are developed to relate the various potential input scenarios to the desired output result. Figure DFL-2 shows a screen shot of the rule editor from the MATLAB Fuzzy Logic Toolbox software program. The maximum number of possible unique rules is calculated by multiplying the number of membership functions for each input. For instance if there are 5 membership functions for heading and 5 for angular velocity, then there will be 5x5 = 25 possible, unambiguous rules. By adding a third input with 5 membership functions, such as angular acceleration, the number of possible rules increases to 125. The potential for the rule base to expand rapidly makes it imperative to judiciously limit the number of input parameters.

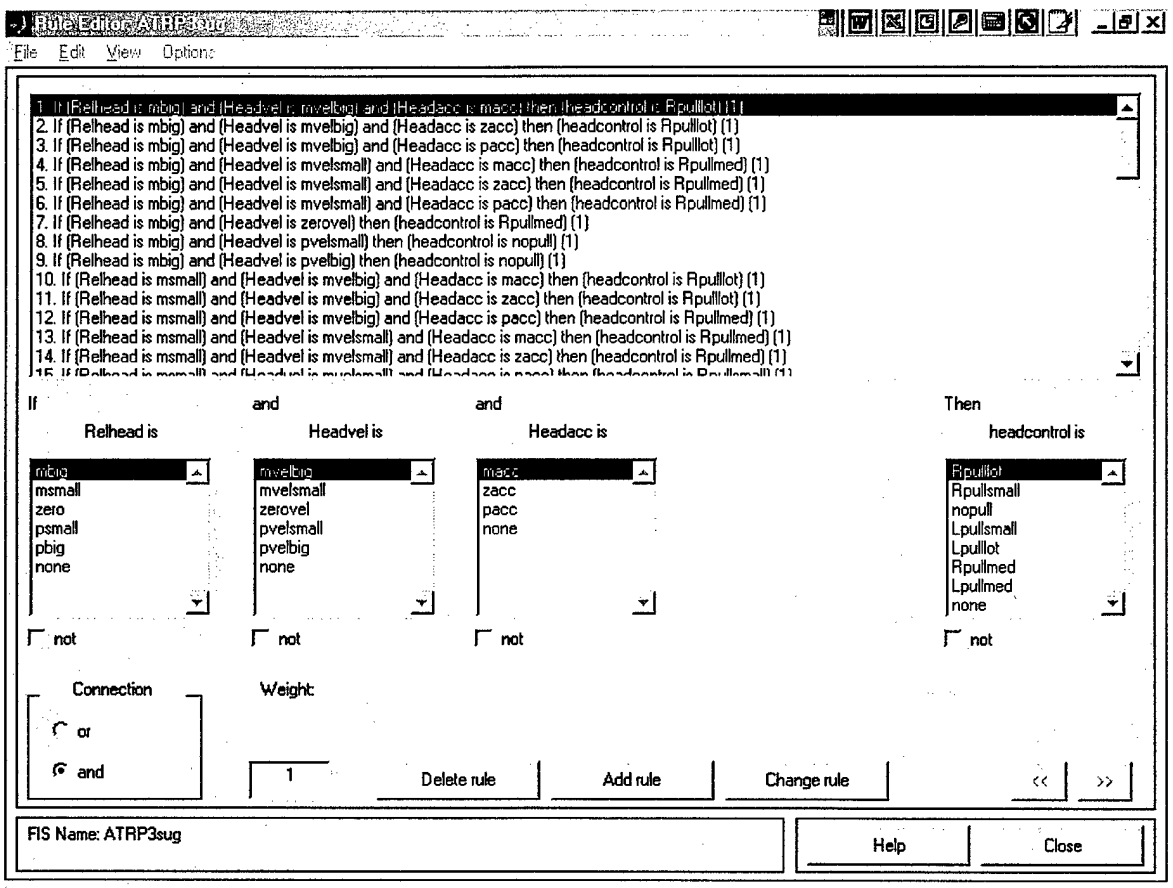


Figure DFL-2 Screen shot of rule base editor from MATLAB Fuzzy Logic Toolbox

In practice, the number of rules can be trimmed by reducing the number of membership functions for each input. In the example shown in the figure above, the angular acceleration has been limited to just three membership functions, reducing the possible number of rules to 75. A further reduction in the rule base can be achieved by eliminating rules that will never be activated or "fired" in the real world mission of the ATRP. Additional rule pruning can be achieved by eliminating rules that are redundant. For instance output from rule 7 in Figure DFL-2 is always the same no matter what value is given for angular acceleration and therefore the single rule 7 replaces 3 possible combinations of rules involving three different values for angular acceleration. Still there is a great danger of rapidly expanding the rule base by adding input parameters, this condition is sometimes described in the literature as the "curse of dimensionality".

The MATLAB Fuzzy Logic Toolbox provides a convenient graphic for visualizing the various rules and showing which rule will fire for a given set of input values as shown in Figure DFL-3 below. In this example it can be seen that rules 5,7,14, and 17 fire based on the given inputs of –29.7 for the heading, -3.42 for the angular velocity and 0 for the angular acceleration. The output levels for each of the rules that fire are averaged in order to obtain a singular or "crisp" output value. In the example below the crisp output value is shown to be –0.64.

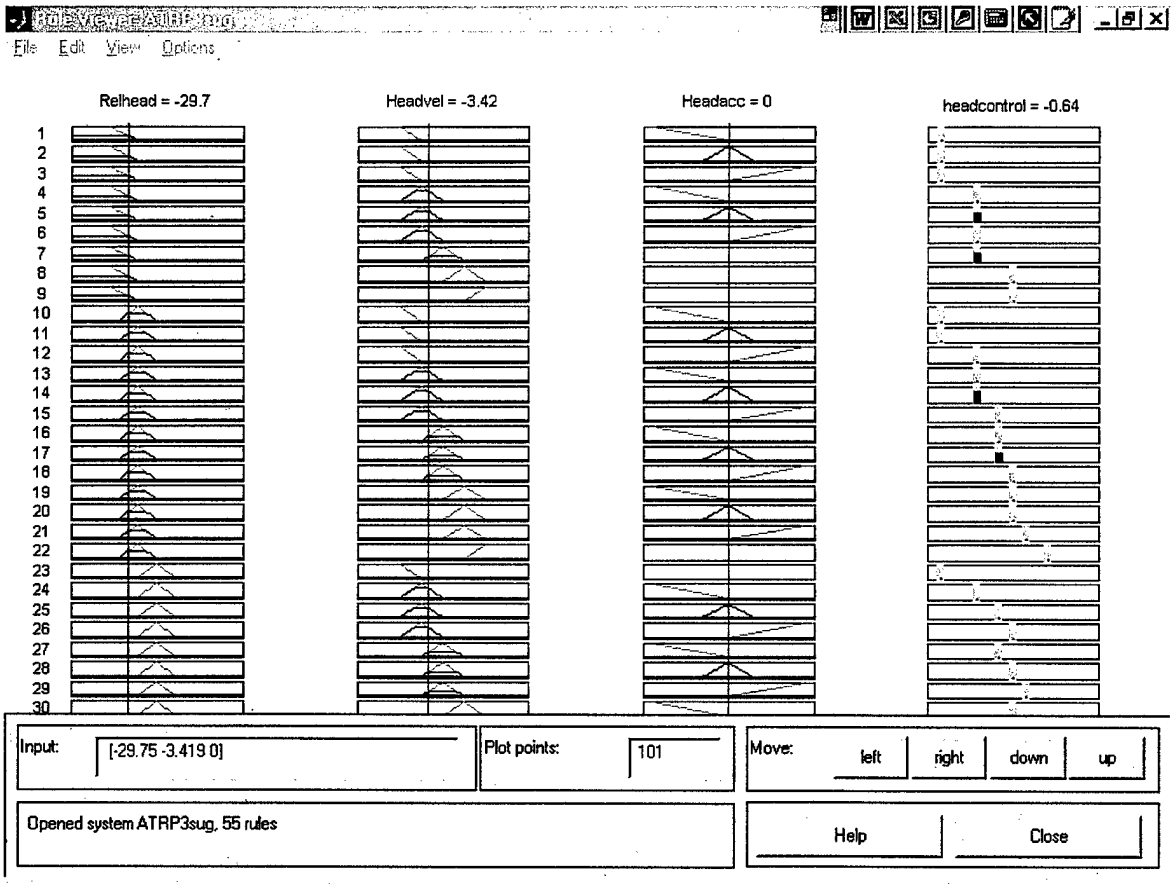


Figure DFL-3 Screen shot of the rule viewer in MATLAB Fuzzy Logic Toolbox

It should be noted that in Figure DFL-3, the membership functions for the output variable look very different from those of the three input variables. In general, the membership functions for the output variable could be similar in form to those of the input variables, but instead we have chosen them to be simple impulse functions. This type of fuzzy logic architecture with impulse output membership functions is of the Sugeno type. This type of architecture was chosen because it is somewhat simpler and more compact to convert to computer code and it has been proven to be as robust and encompassing as more traditional architectures.

An additional advantage of using the MATLAB Fuzzy Logic Toolbox is that it can be directly linked to the Matlab Simulink software to provide a full simulation of an ATRP flight scenario. It is this simulation capability that made it possible to refine the control software and thus drastically reduce the number of test flights necessary for control optimization.

5. Flight Simulation and Fuzzy Algorithm Optimization

The parafoil design is inherently a very simple and stable flight vehicle that does not require a vertical control surface. Flight maneuvers are simple in the extreme, directional control is effected by pulling on one of two control lines connected to either side of the

parafoil wing. One simplifying assumption made in the simulations was that the parafoil flies at a constant speed with no need for pitch axis control. This assumption is at least valid to a first degree of approximation and was not a limiting factor in optimizing the control algorithm.

It was not the task of the simulation to optimize the aerodynamic qualities of the parafoil platform and therefore only a rough degree verisimilitude was necessary to achieve the goals we were striving to achieve. The goals of the simulation were to: 1) determine the minimum number of sensor inputs to achieve satisfactory, stable control, 2) determine the most effective sensor inputs for achieving control, 3) optimize various fuzzy logic components (i.e. membership functions, rule base, etc.) under ideal conditions and, 4) test and re-optimize the algorithm under simulated sensor and environmental sources of imprecision (i.e. sensor electronic noise, wind gust disturbance, wind drift). Figure FS-1 shows a block diagram of the various software components used in the simulation and optimization process.

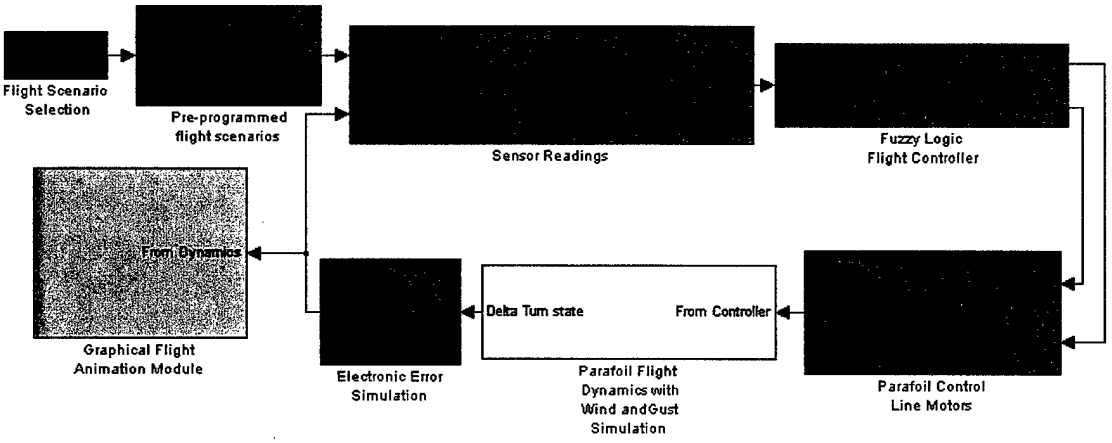


Figure FS-1 Parafoil Flight Simulation/Optimization Block Diagram

The simulation blocks shown in green in Figure FS-1 represent functions that occur in hardware and/or software on the parafoil. The yellow block contains simulations of parafoil interactions with the environment and the blue block provides output for the user. Following the block diagram above, the simulation begins with the selection of a desired flight scenario which is compared to the current heading. The difference between the desired and actual heading along with the first derivative or change over time of this difference is supplied to the fuzzy logic controller. The controller uses these two inputs to produce a control line pull command. The effect of the control line pull command on the aerodynamic performance of the parafoil is computed and a incremental parafoil heading vector is determined. Any simulated electronic errors are combined with the heading vector to complete the loop and to provide input to the graphical output module. The simulation runs for the length of a single simulated parafoil flight.

In the following paragraphs the referenced figures are supplied in appendix A for hard copy versions of this report and are in the accompanying zip file for electronic copies. Each set of figures represents a single run of the simulation shown in the block diagram. Only selected results have been included in this report as they represent the culmination of countless optimization runs. As mentioned above, a variety of parameters can be altered in order to optimize the fuzzy logic control algorithm. In fact this wealth of options presents a formidable challenge to the researcher in the optimization process. Various authors have attacked the optimization problem using genetic algorithms, neural networks and other even more exotic methods, they all tend to be very complicated and fragile. Therefore, we chose a more brute force approach of running a large number of simulations and comparing graphs of a few key outputs from each run. This approach was labor intensive and somewhat subjective, but it didn't require a lot of time or money in development, which were both key constraints in this project.

Ultimately it was determined that a continuously updated compass heading and a delta compass heading, or angular velocity, would provide sufficient input to the fuzzy controller for the ATRP to accomplish its mission. One additional input, angular acceleration, was considered and provided improved results in certain scenarios. It was determined however that the price in processing speed, complexity, and memory requirements did not justify the modest improvement gained by adding this input.

Three types of graphs were used to evaluate the performance of a fuzzy logic algorithm in any given simulation run. The first type, as shown in Figure A1, is a ground track plot or view from above looking down on the simulated flight path of the parafoil. Each simulation run started at x-y coordinates of (0,0), and in this case proceeded straight until making a single 180 degree turn to return near the starting point. This plot gave an overall view of how capable the fuzzy logic algorithm was in following the desired path. This type of plot was especially useful in observing how the algorithm responded to various input perturbations.

The second type of evaluation graph (Figure A2) plotted the difference between the ideal flight heading and the "actual" flight heading as a function of time. This provided a means of comparing the efficacy of each fuzzy logic control architecture in maintaining heading control. This type of plot was given the greatest weight in the evaluation process.

The third type of graph (Figure A3) is a plot of the magnitude and direction of control commands over time. These control commands are sent to the motors controlling the direction and amount of pull applied to the control lines. An overly aggressive control sequence with large fluctuations would quickly exhaust batteries and produce flight instabilities. This graph was used as a "sanity check" to ensure that a theoretically optimized but physically unworkable solution was not produced.

Figures A1 – A3 show a flight simulation under ideal conditions, where the ATRP initially deploys flying in the exact desired direction, there is no electronic noise nor are there any wind gusts to drive the vehicle off course. As can be seen in these figures, the fuzzy controller performs very well under these ideal conditions. The course is very straight, the

difference between the actual and desired course is small and controller commands are well within hardware limits. Unfortunately, these ideal conditions are not very close to reality and so the majority of the analysis effort was expended on evaluating various fuzzy architectures under non-ideal conditions. Thus not only did the fuzzy logic architecture require optimization, but it also required optimization over a range of simulated flight conditions.

Figures A4 – A5 demonstrate what happens when the ATRP deploy with an initial heading that is exactly opposite to the desired heading. The ATRP quickly executes a 180 degree turn to get on track, but there is some residual oscillation that causes it to over compensate when the programmed turn is executed. The resulting flight track is not ideal, but would seem to be reasonably well executed for most applications.

As mentioned above the addition of a third input, angular acceleration, to the fuzzy controller gave improved results under some flight scenarios. Figures A6 –A7 show the same flight scenario as in Figure A4 but with the addition of the angular acceleration input. Here the ATRP recovers from the bad initial heading much more quickly and the difference plot (Figure A7) shows much smaller excursions from the desired heading. There were occasions however, when the addition of the angular acceleration actually hurt performance.

Figures A8 through A11 show the ATRP ground track and associated heading errors with and without angular acceleration input to the fuzzy logic algorithm. These simulated flights were made with an initial heading error of 90 degrees which is the average error that can be expected from the random deployment of the ATRP. As can be seen in these plots, despite some ripple in the ground track, the ATRP generally responds quite well under these conditions. As a result of this type of analysis for a variety of initial headings, it was determined that an initial random heading error alone would not significantly degrade the performance of the ATRP flight mission.

Another potential source of error for a real world system is the effect of gusts of wind pushing on the parafoil and inducing a change in heading. While in flight, the ATRP must be able to recover from these gusts and maintain the required heading. Figures A13 and A14 illustrate the response of the ATRP in the presence of gusts of wind. In the simulation, gusts were constructed as relatively slow varying and pseudo random events in both direction and intensity.

The effect of a third type of error source, random noise, is shown in Figures A14 through A19. Real world sources of noise come from the error tolerances of the compass and various components of the ATRP electronics. In the simulations, it is assumed that the effect of this noise is to increase the uncertainty in the heading information provide to the fuzzy logic algorithm. This noise is assumed to be random in nature and therefore will have a mean of zero degrees. The variance of this noise term is chosen to be 2.0 degrees, which is in line with the published uncertainty in the compass specifications with an additional 0.5 degree added in to account for other electronic sources.

The effect of the noise is most easily discerned in the ragged nature of the "heading error" plots A15 and A18. The difference between these two plots being, that the former does not have angular acceleration as an additional input to the controller and the latter does. A comparison of these two plots shows that the addition of angular acceleration information actually tends to degrade the performance of the controller. Despite the fact that some averaging of the angular acceleration data is being performed in the simulation, it would appear that this additional input is producing a sort of whipsaw effect on the fuzzy logic control mechanism.

Another effect of the random noise is to significantly increase the direction and magnitude of the output commands. This effect can be most easily observed by comparing Figures A3 and A15. Despite this increase in command output, it was determined that the required motions were well within the capabilities of the motors and batteries used in controlling the ATRP.

Figures A20 through A22 show the combined effects of all the sources of error described above. In these plots the assumption is made that the ATRP is deployed with an initial heading error of 90 degrees. Figures A24 and A25 show a worst case scenario where the ATRP is deployed with an initial heading 180 degrees off from the desired heading. Of course while every mission scenario will have it's own criteria for what is an acceptable deviation from the planned flight profile, it is our belief that the flight simulations shown are within acceptable limits for most applications.

6. Flight Tests

Flight tests performed early on in the project were designed to test the rocket design, the deploy mechanism, and gain some information on the aerodynamic characteristics of the parafoil. A standard motor burnout delayed ejection charge was used in conjunction with a piston to separate the nose from the rocket body and release the parafoil. The rocket was launched on commercially available solid rocket motors, of varying sizes depending on vehicle weight, to allow the rocket to reach apogee at an acceptable test altitude.

In the initial test flights, the parafoil was controlled from the ground by means of a radio control mechanism designed and built for this project. As noted above, this first rocket design needed a powerful motor to achieve useful deploy altitudes, which forced us to launch from some fairly remote locations (Fresno, Lucerne Valley, etc.) during specific, scheduled event days. This restriction made it difficult to achieve the high turn around needed to optimize the system design. Figure FT-1 below shows the launch of the rocket at one of these early events.

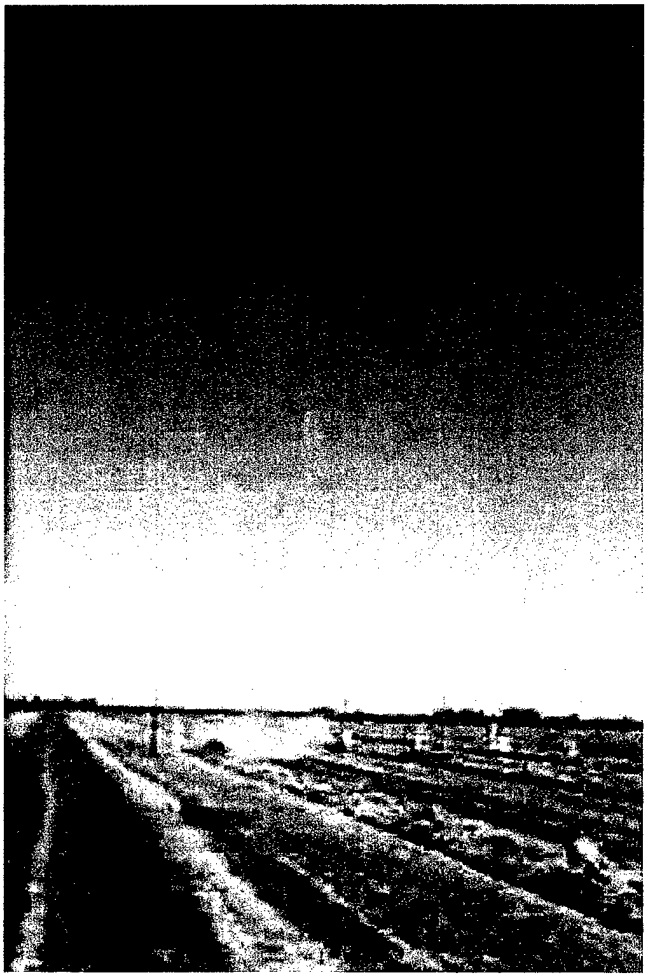


Figure FT-1 Rocket Launch Outside Fresno, CA

The rockets and deploy mechanisms were designed and built by students in the Aerospace Engineering department, under the guidance of Dr. DeTurris, and every one of these flight tests resulted in a successful launch and deploy separation. However, there were problems encountered during the deploy event. Parafoil lines were often getting tangled or broken during deploy, resulting in an uncontrollable descent. Incremental improvements were made in the design, but it became apparent that a new approach would be required to achieve a high percentage of successful deploy and flight events. The solution was to use a different parafoil model that had fewer support lines and did not require the deployment of a control line separator rod.

The rocket design was also overhauled to create a new vehicle that was much lighter and could be launched to the required altitude using a much smaller motor. The regulations governing the firing of these smaller motors is much less restrictive, allowing us to perform tests with greater frequency and in locations much closer to Cal Poly. The very first launch of this new design resulted in an unqualified success in launch, deploy and radio controlled landing. See Figure FT-2 below for this first launch of the redesigned rocket.



Figure FT-2 First Launch of the New Rocket Design

The final launch, which was the first under autonomous control, was performed at Camp Roberts Army base. Camp Roberts is located on the Central Coast, just a few miles north of the town of Paso Robles. In addition to the ATRP team members, the launch was witnessed by several officials based at Camp Roberts. A successful liftoff and ascent was once again achieved, however, the delay charge was too long and deployment occurred 3-4 seconds late. The late deploy event occurred as the rocket was well into its descent and at high speed. The stress of the high speed deploy of the parafoil caused one of the support line attachment points on the payload section to fail. The lack of support lines on one side made it impossible for the autonomous controller to fly the parafoil during the rest of the descent. Figure FT-3 shows the rocket lifting off at Camp Roberts. However, despite the resulting hard landing, the electronics package survived and data from the flight was down loaded to computers in the laboratory. At the time of this writing another launch of the autonomous control package is scheduled to occur at Camp Roberts on January 2-3 of 2003.

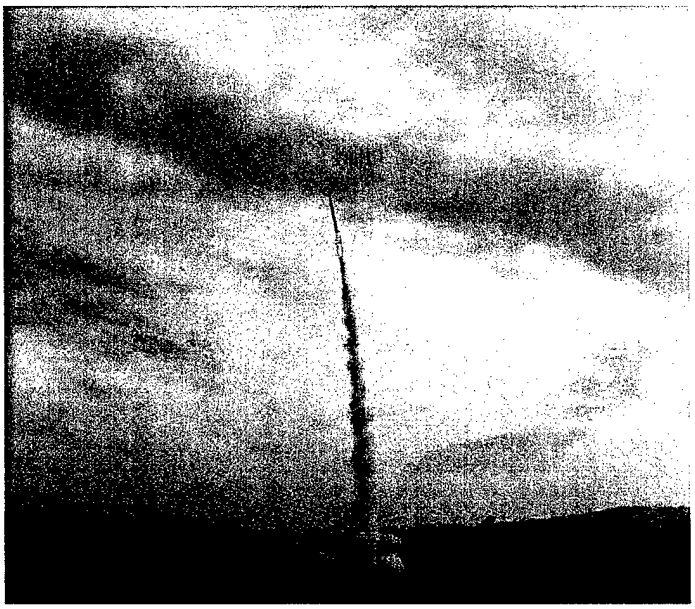


Figure FT-3 Camp Roberts Launch

7. Conclusions

There are a number of applications where the ATRP would provide a unique capability that is not adequately satisfied by other devices currently available. The ATRP is superior in terms of cost, ruggedness/durability, ease of use, portability, time to activate and reusability

when compared with other competing technologies in a variety of mission scenarios. In the following paragraphs we provide some conclusions from the efforts just completed, a brief projection of the potential uses for the ATRP and our goals for a follow-on project.

The ATRP can play an important role in many civilian applications where there is a time critical need for reconnaissance information. Non-military uses would include providing fire location and egress points for smoke jumpers and forest fire fighting crews. Tragedy has occurred in this profession when shifting winds and/or miscommunication has allowed crews to become encircled or overrun by wild fires. An ATRP equipped with an IR sensor could be rapidly deployed to provide ground crews with a birds eye view of fire hot zones in the immediate vicinity. In rugged terrain, the ATRP could provide a low cost, real-time communications boost and perhaps save lives in unpredictable conditions.

Search and rescue teams would use the ATRP to help locate avalanche victims, lost skiers and hikers or victims of other types of natural disaster. The sooner a device can be deployed, the more likely a successful rescue can be effected. If a search team can launch a reconnaissance device with an IR sensor quickly, a lost hiker has less time to expand the required area of search coverage. With the ATRP in tethered mode (flown like a kite) a constant surveillance can be maintained to help direct search teams and again act as a communications boost in rugged terrain.

Perhaps the most promising civilian application of the ATRP is in the area of farming and land management. As an affordable device, the ATRP would allow farmers to evaluate crop health over vast fields on a regular basis. Studies have shown that data collected in the IR wave bands can be extremely helpful in identifying trouble spots even before they become evident through visual inspections. Environmentalists would use similar types of data to monitor threatened ecosystems. Evidence of salt intrusion disease and pollution could be identified much more quickly and at very little cost with the ATRP.

There are also a number of military uses for the ATRP other than the baseline scenario described in this report. A scaled up ATRP could be airdropped and used to deliver critical provisions to troops and civilians in the field to precise, GPS guided, landing zones. This precision guidance feature would eliminate the danger to personnel on the ground and ensure that materiel was recovered by friendly forces.

An ATRP deployed prior to terminal approach by a smart weapon, such as a cruise missile, or by an attacking aircraft would provide images for instant assessment of battle damage. This inexpensive means of verifying that a facility has been denied to hostile forces would allow planners to concentrate on other targets or re-target the same facility if required. Multiple sorties could be made against the same facility successively without having to wait for satellite imagery for damage assessment.

Looking farther in the future, the ATRP would be a precise cost effective means of distributing ground sensors in militarily contested areas. Communications among sensors and control personnel could be enhanced through the use of launched or tethered ATRP devices. A scaled up ATRP could also be used to deliver future robotic mechanisms, such

as the Unmanned Ground Autonomous Vehicle (TUGAV), into hostile territory and thus extend the range of operations for these devices.

One typical scenario for the device would be for the soldier in a small unit, in potentially hostile surroundings, to launch the ATRP with an IR sensor to detect the presence of enemy forces in the immediate area. Launch can be effected day or night and the flight pattern selected can call for the device to be return or not depending on the degree of concern over revealing the location of the launch point. In either case, the device can be equipped with a beacon to allow for easy retrieval, but it is also low cost and can be left behind if circumstances are unfavorable for recovery.

8. Recommendations for Future Work

The successful completion of the current Autonomous Tactical Reconnaissance Platform (ATRP) project has produced a prototype rocket launched parafoil capable of autonomous flight. The goal of future efforts will be to build on the capabilities of the current configuration by exploring new techniques and technologies. Some of the technologies are quite unique and would result in quantum leaps in capability rather than simple incremental steps.

In the follow-on project we proposed to provide a number of improvements and new technologies that will be necessary to field a system ready for operational activity. The size and weight of the prototype system electronics and motor control package will be reduced by more than half by using more basic electronic elements in the design, a lighter and more efficient motor and battery and lighter instrument packaging. Reducing weight in these payload areas will allow for both a smaller rocket body and parafoil requirement, thus further reducing weight and increasing portability with improved performance. Another goal will be to investigate and demonstrate one or more alternative, silent and/or simplified launch mechanisms. Potential alternatives will include but not necessarily be limited to; compressed air, artillery shell, tethered launch and release and electric motorized launch and flight.

Other enhancements to be incorporated in the follow-on proposal are collision avoidance sensors, and a remote sensing device (e.g. IR camera) with telemetry transmitted to a ground station. The remote sensing device will communicate to a handheld wireless device with the user on the ground in real-time. The onboard sensing platform will be modular in order to accommodate a variety of sensors including but not limited to; IR, visual, and long-range communications capabilities. The ATRP would also be enhanced with a tethered flight mode when long dwell times are required for communications and surveillance missions.

And finally, a very exciting opportunity for at least doubling the parafoil flight performance can be achieved by adopting a technique pioneered by a recent Cal Poly graduate. The technique involves dynamically changing the parafoil geometry while in flight. The span or wing tip to wing tip width of the parafoil is increased or decreased as required throughout the flight. It has been demonstrated that decreasing the span allows for faster flight without significant loss of L/D or glide ratio. This capability also allows the parafoil to maintain

course in much higher wind conditions. The parafoil can fly rapidly from point to point with slow dwell times over areas of interest. The optimum parafoil geometry would be regulated throughout the flight by the onboard fuzzy logic controller.

References

1. Passino, K.M. and Yurkovitch, S., Fuzzy Control, Addison-Wesley, CA, 1998.
2. Sugeno, M., "Development of an Intelligent Unmanned Helicopter", presentation at the World Automation Congress, May 10-14, 1998, Anchorage, Alaska.
3. Swanson, S., Fuzzy Control of the Shuttle Training Aircraft, Applications of Fuzzy Logic Towards High Machine Intelligence Quotient Systems, Prentice Hall, NJ, 1997, pg. 387.
4. Fournell J. and S. E. Alptekin, "Fuzzy Logic Fly-by-Wire System", Proceedings of World Automation Congress, May 10-14, 1998, Anchorage, Alaska.
5. GopalaPillai, S., Tian, L., Beal, J., "Detection of Nitrogen Stress in Corn Using Digital Aerial Imaging", 1998.
6. Johannsen, C.J., "Agricultural Applications of Remote Sensing", 2001.
7. Fournell, Joseph A., Mechatronic Philosophy: A Fly-by-Wire System, Masters Thesis, Industrial and Manufacturing Engineering Department, Cal Poly San Luis Obispo, June 11, 1997.
8. Ervin J. and S. E. Alptekin, "Fuzzy Logic Control of a Model Airplane", Proceedings of the 1998 IEEE International Conference on Systems, Man, and Cybernetics, October 11-14, 1998, San Diego, California.
9. Alptekin, S.E., Fournell, J. and Ervin, J. "An Investigation of Fuzzy Logic Control of a Complex Mechatronic Device", Proceedings of the 2nd International Conference on Recent Advances in Mechatronics, ICRAM'99, May 24-26, 1999, Istanbul, Turkey.
10. Alptekin, S.E., "Smart Products - A Tool for Mechatronics Education", Proceedings of International Conference on Recent Advances in Mechatronics - ICRAM'95, Volume I, pp: 288-292.
11. Alptekin, S.E. and Menon, U., "Use of Mechatronics in Integrated Product Development", with Proceedings of American Society for Engineering Management, pp:307-310, October 23-26, 1997, Virginia Beach, VA.
12. Stedman, B., Menon, U. and Alptekin, S.E., "Video Tracker Application for Boeing", Proceedings of Mechatronics'96, San Francisco, June 13-15, 1996, pp:135-150.
13. Mason A., Towne, D. M. and Alptekin, S.E., "Simulation in Mechatronics Instruction", Proceedings of Mechatronics'96, San Francisco, June 13-15, 1996, pp: 84-91.
14. Alptekin, S. E., "Development of a Mechatronics Design Studio", Proceedings of ASEE Conference, Milwaukee, June 15-18, 1997.
15. Alptekin, S. E., "Preparing the Leaders for Mechatronics Education", Proceedings of 1966 Frontiers in Education Conference, Salt Lake City, Utah, Nov. 6-9, 1996.
16. Alptekin, S.E. and Freeman, J. H. "Mechatronics Education: Model and Implementation", Proceedings of Mechatronics'96, San Francisco, June 13-15, 1996, pp:1-8.
17. Alptekin, S. E., "Mechatronics Design Studio: Sample Student Projects", Proceedings of Mechatronics'96, San Francisco, June 13-15, 1996, pp:180-185.
18. Martin, F., "The MiniBoard 2.0 Technical Reference", Media Laboratory, MIT, 1994.
19. MATLAB, http://www.mathworks.com.
20. Wind Dance Parafoil Kite http://www.seattleairgear.com/
21. X-38 Crew Return Vehicle, http://www.dfrc.nasa.gov/Projects/X38/intro.html

22. Guided Parafoil Air Delivery System – Light, http://www.fas.org/man/dod-101/sys/ac/equip/gpads-l.htm
23. Clark, Kevin, "Sampcode.c" http://www.egr.msu.edu/classes/ece482/Teams/99spr/design5/deliverables/code/main.c
24. Martin, Fred, "Serialio.c" http://216.239.33.100/search?q=cache:Y3ohOvJ5bI4C:www-isl.ece.arizona.edu/~soccer/Fall2002/ai/serialio.pdf+serialio.c+fred+martin&hl=en&ie=UTF-8
25. Mulders, Symen "Comm.c" http://clubs.plattsburgh.edu/csc/handy_car/source/comm.c
26. Drushel, Richard F., "Conio.c" http://handyboard.com/software/contrib/drushel/conio.c
27. Lang, Kam, "RF Serial Communcation for the MIT Handy Board or Motorola 68HC11 MCU"

Contact People:
Dianne DeTurris, Ph.D., Aerospace Engineering Department, Cal Poly, ddeturri@calpoly.edu
Sema E. Alptekin, Industrial and Manufacturing Engineering Department, Cal Poly, salpteki@calpoly.edu
Jon Ervin, Apogee Research Group, apogee@alumni.calpoly.edu

This project is sponsored by the Department of the Navy, Office of Naval Research (N00014-01-1-1049)

Appendix A

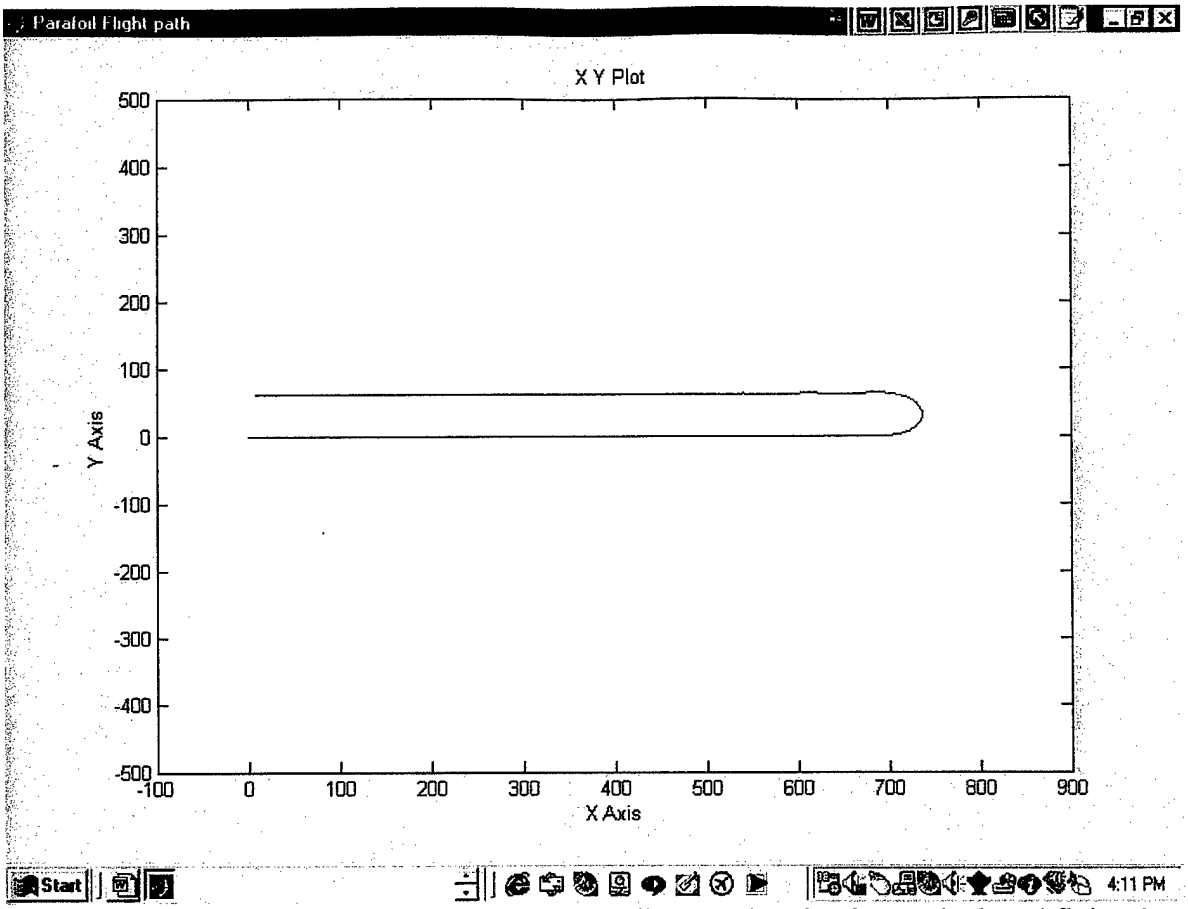


Figure A1 - A single 180 degree turn without heading acceleration input, deployed flying 0 degrees from desired heading

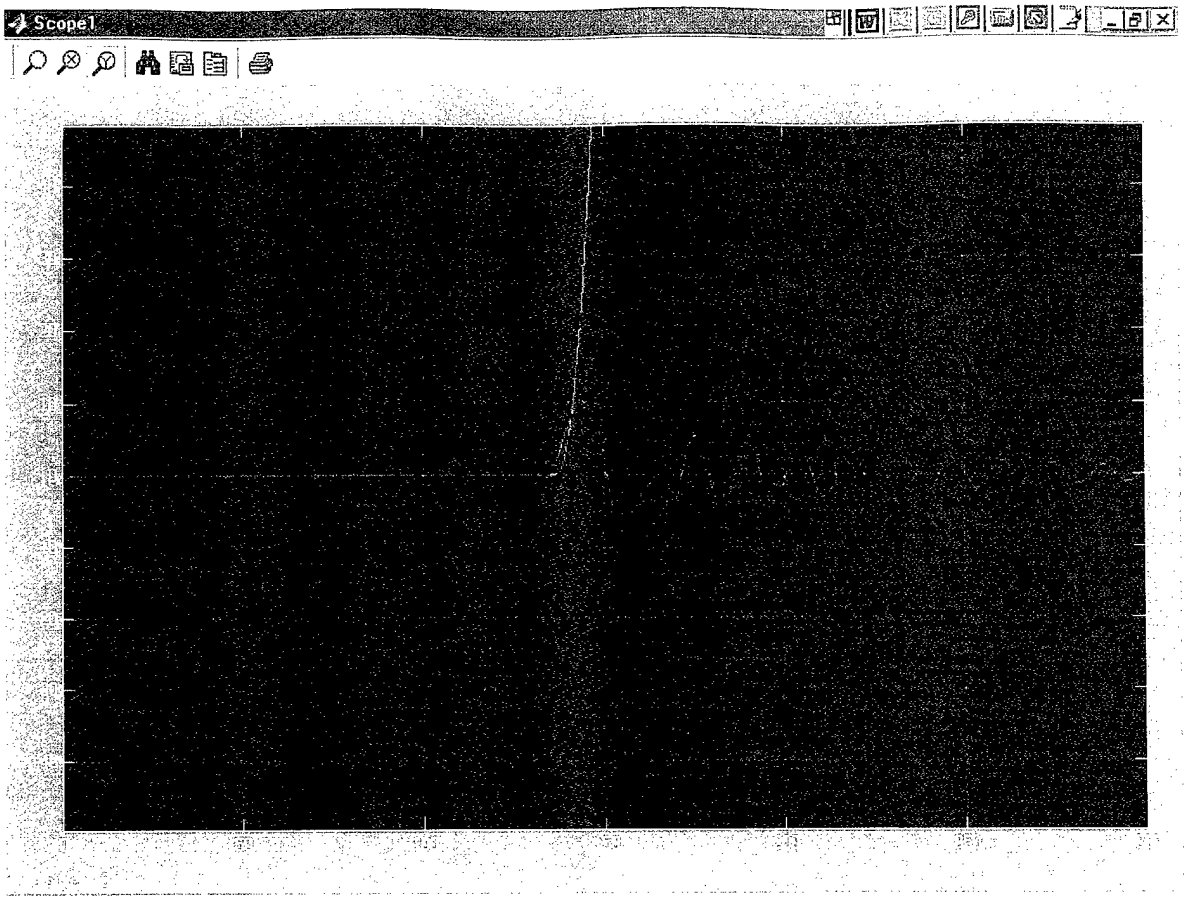


Figure A2 A single 180 degree turn without heading acceleration input, deployed flying 0 degrees from desired heading

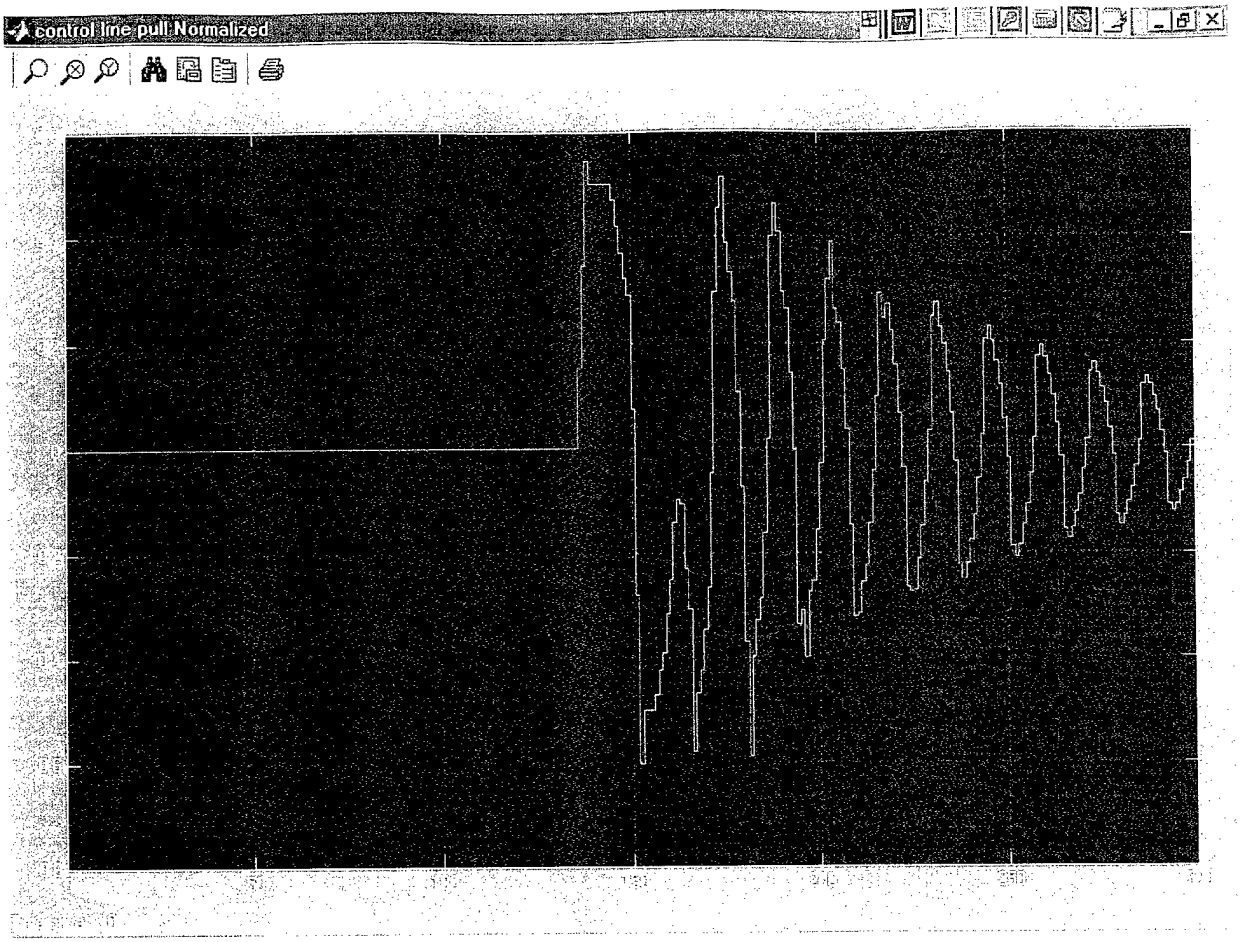


Figure A3 A single 180 degree turn without heading acceleration input, deployed flying 0 degrees from desired heading

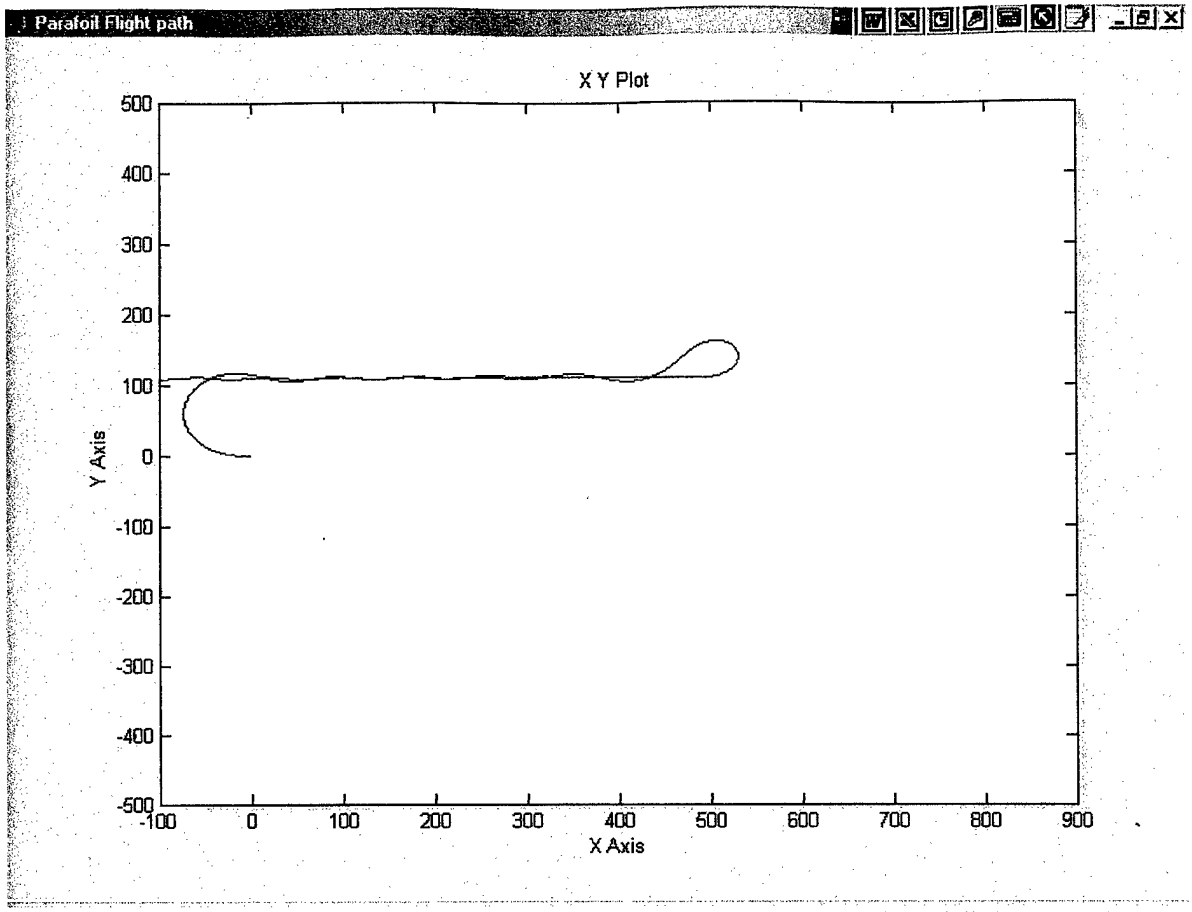


Figure A4 1 180 degree turn without heading acceleration input, deployed flying 180 degrees from desired heading

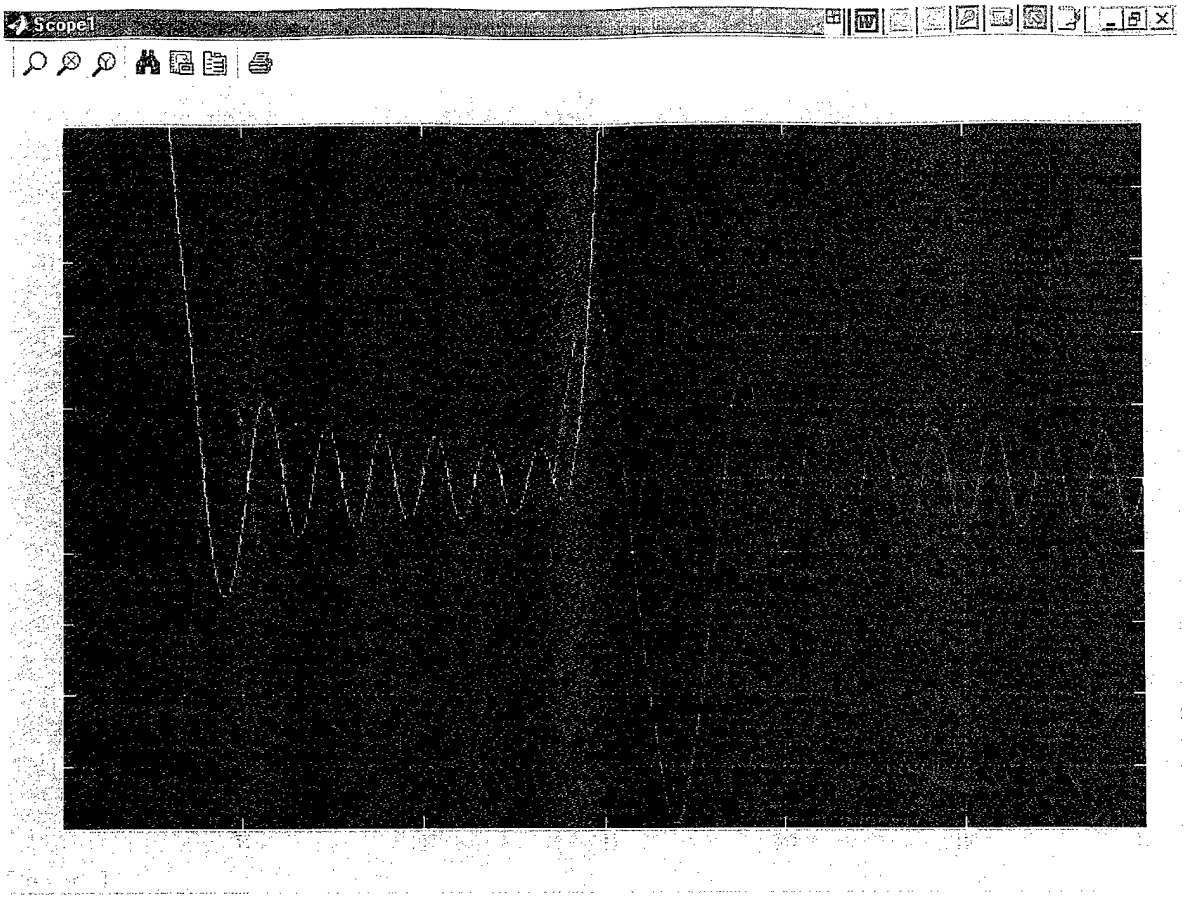


Figure A5 1 180 degree turn without heading acceleration input, deployed flying 180 degrees from desired heading.

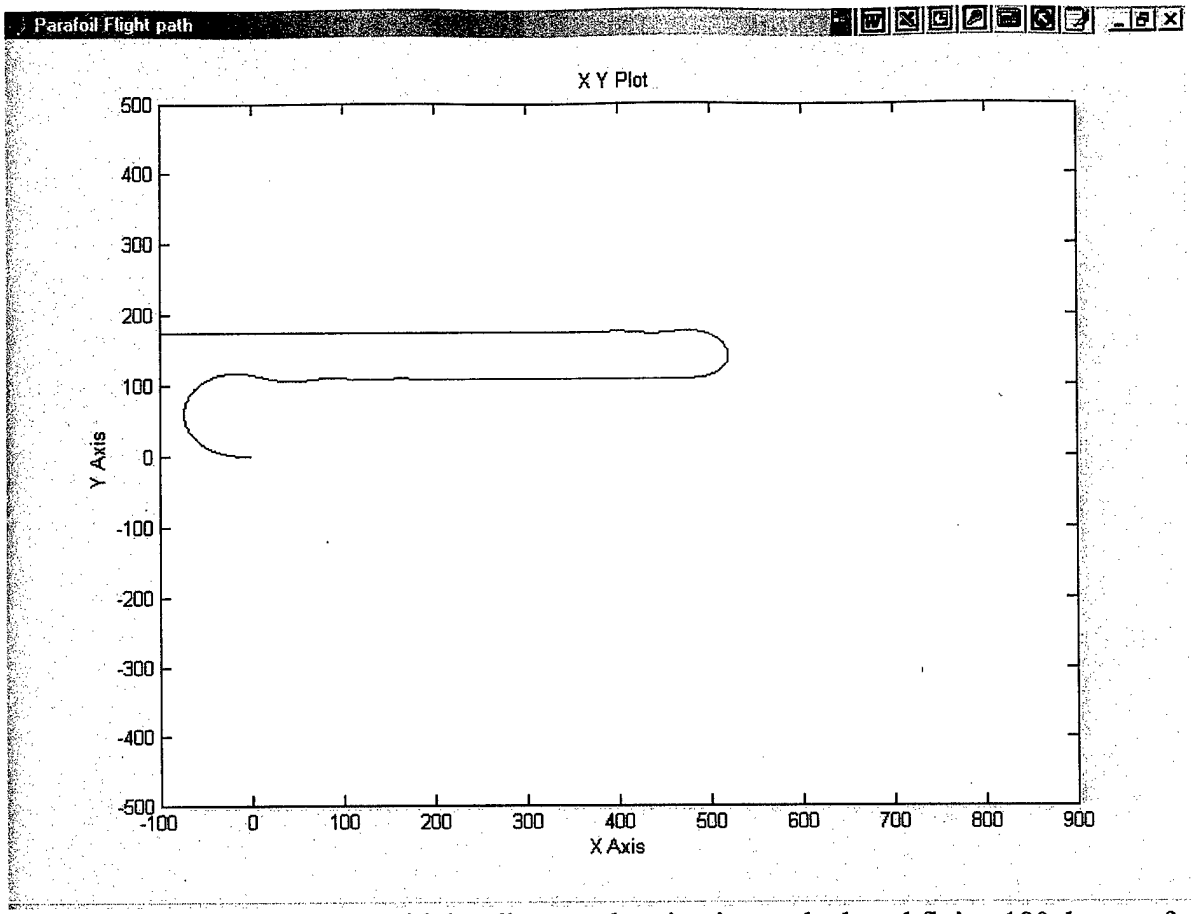


Figure A6 1 180 degree turn with heading acceleration input, deployed flying 180 degrees from desired heading



Figure A7 1 180 degree turn with heading acceleration input, deployed flying 180 degrees from desired heading

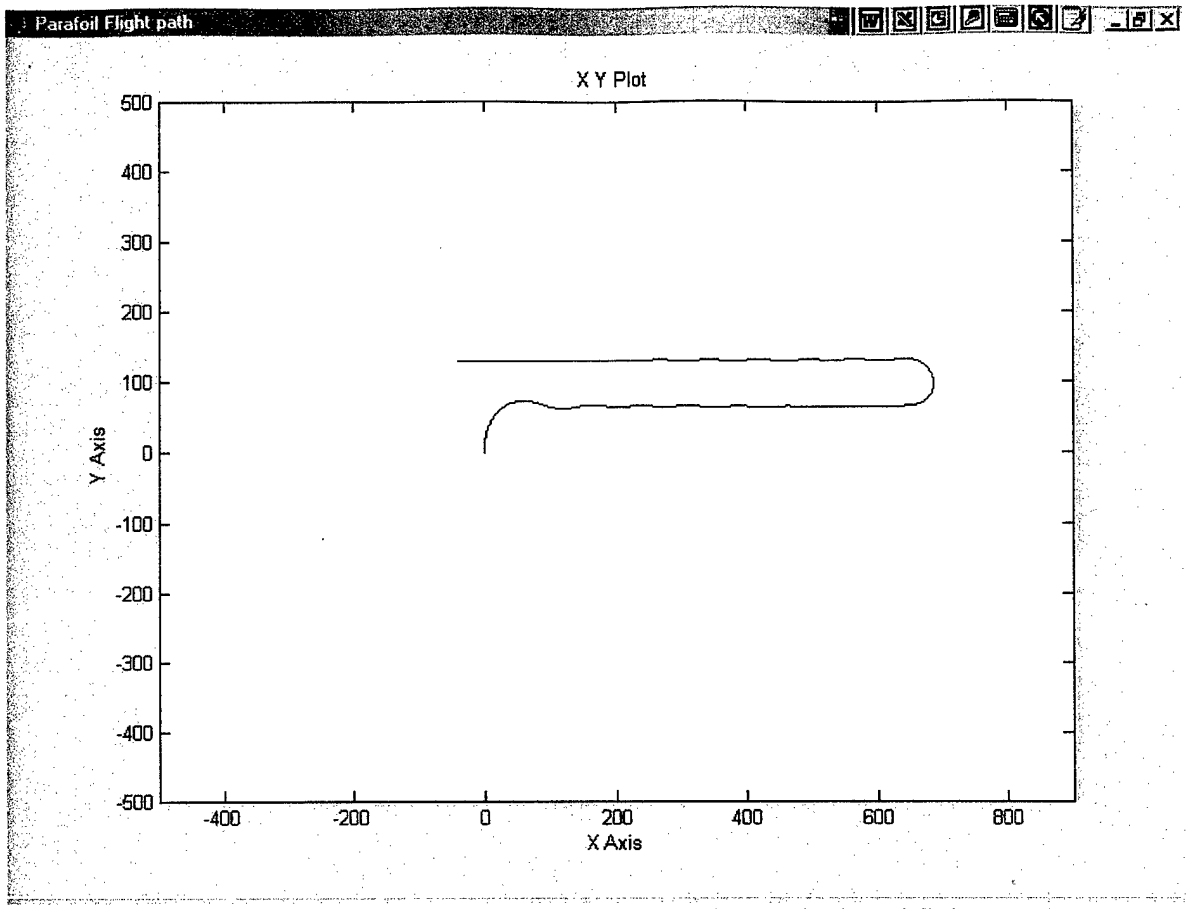


Figure A8 1 180 degree turn with heading acceleration input, deployed flying 90 degrees from desired heading

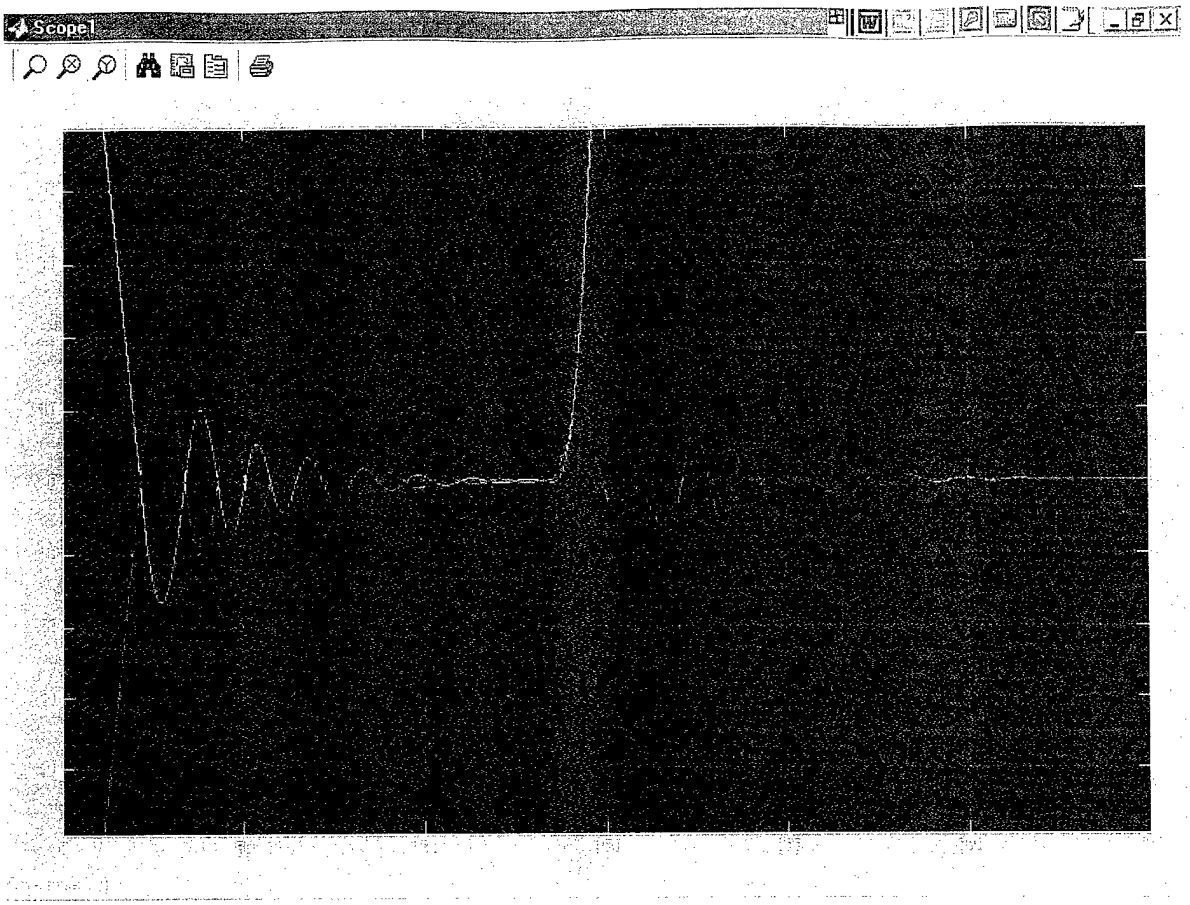


Figure A9 1 180 degree turn with heading acceleration input, deployed flying 90 degrees from desired heading

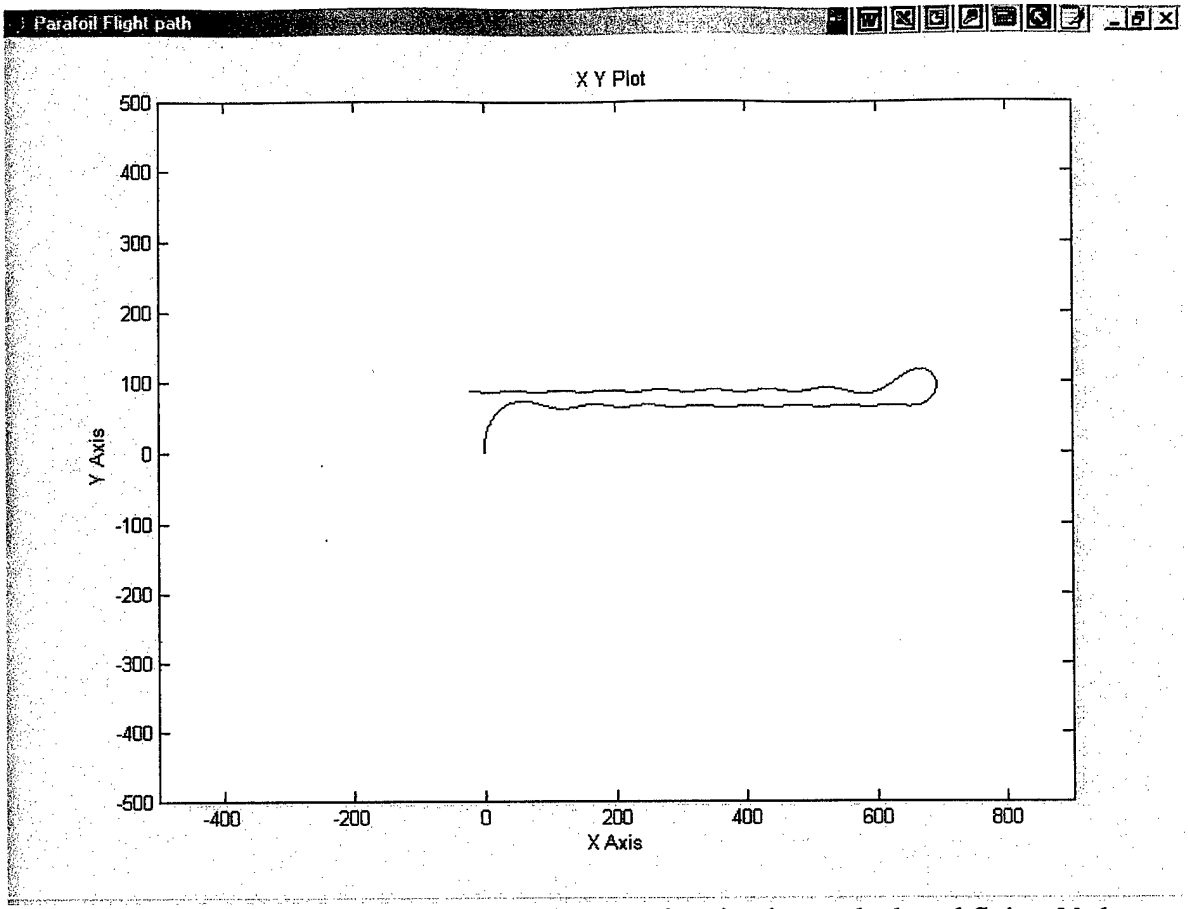


Figure A10 1 180 degree turn without heading acceleration input, deployed flying 90 degrees from desired heading

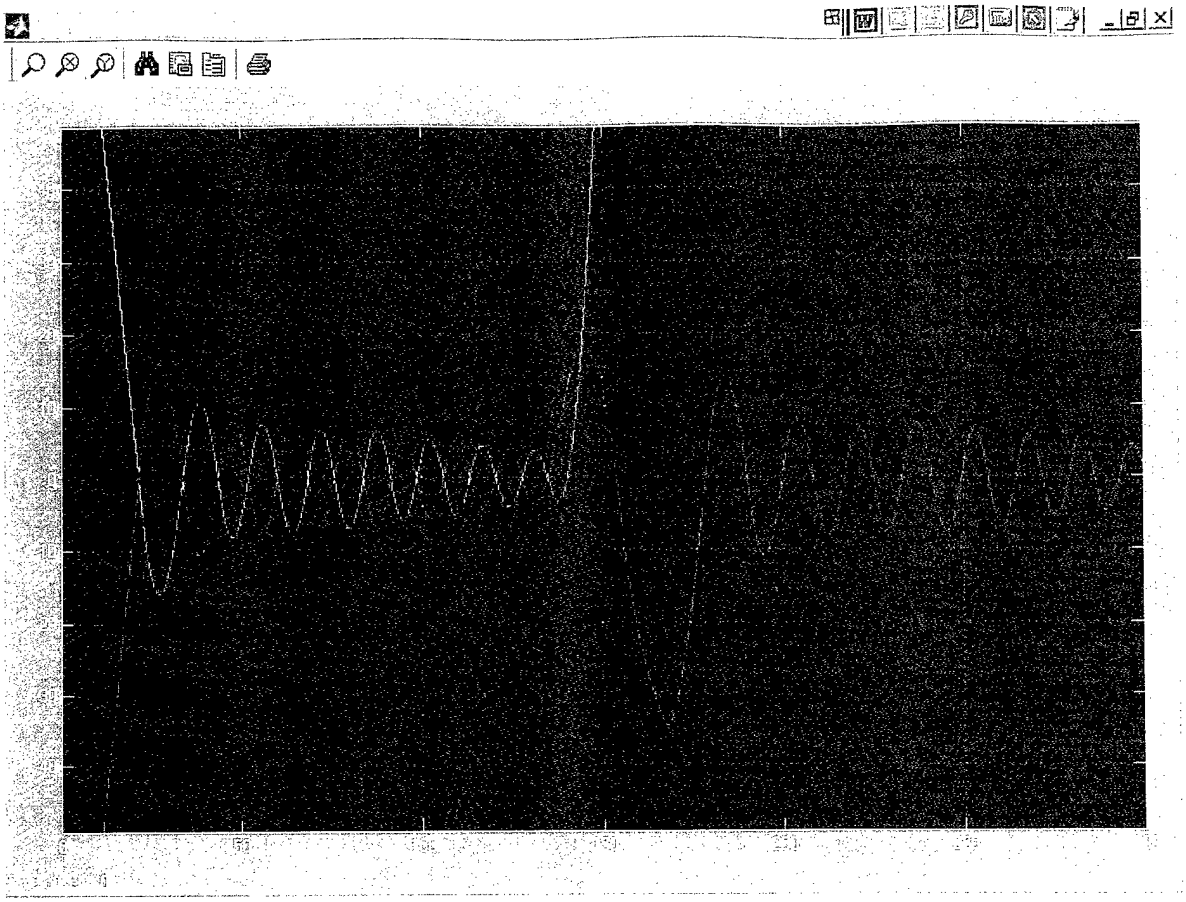


Figure A11 1 180 degree turn without heading acceleration input, deployed flying 90 degrees from desired heading

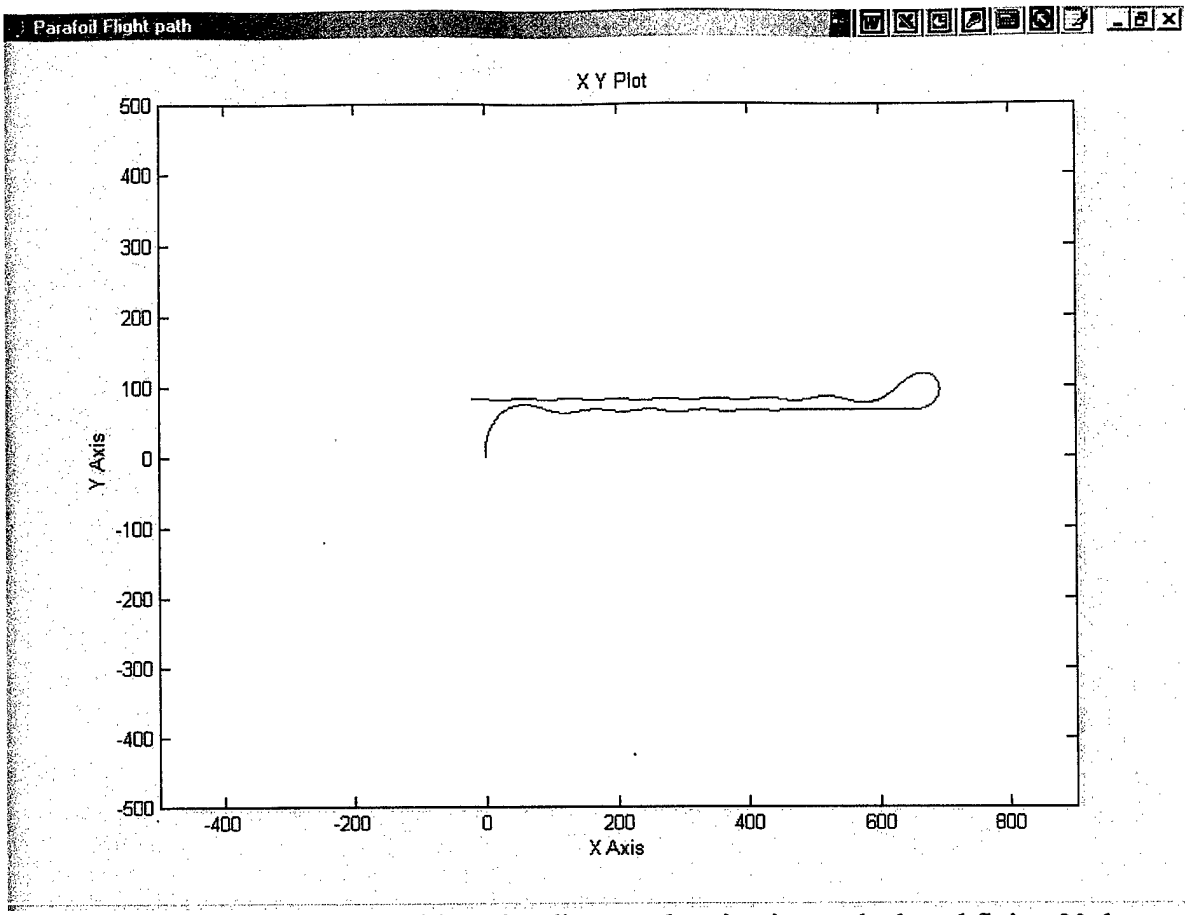


Figure A12 1 180 degree turn <u>without</u> heading acceleration input, deployed flying 90 degrees from desired heading, with wind gusts amp. .1 and freq .5 rad/sec.

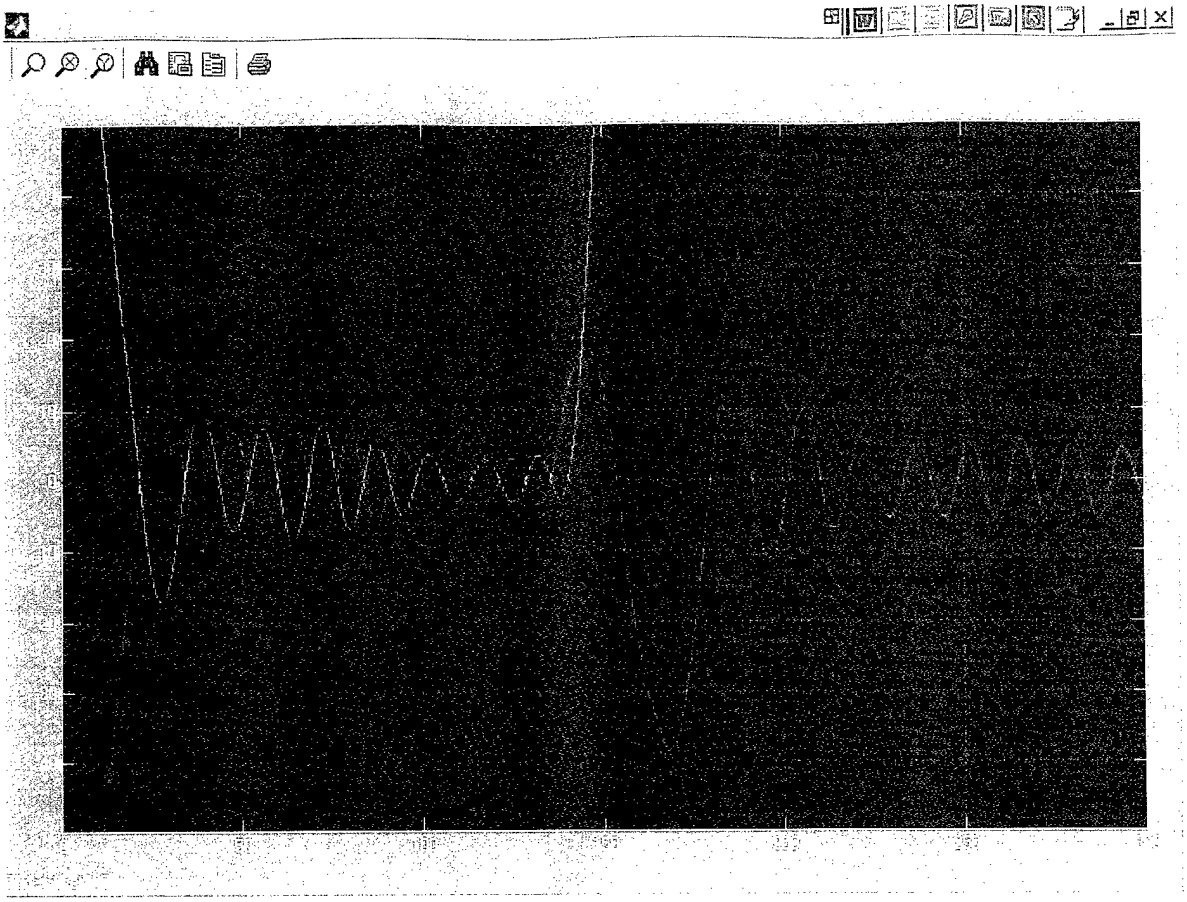


Figure A13 1 180 degree turn without heading acceleration input, deployed flying 90 degrees from desired heading, with wind gusts amp. .1 and freq .5 rad/sec.

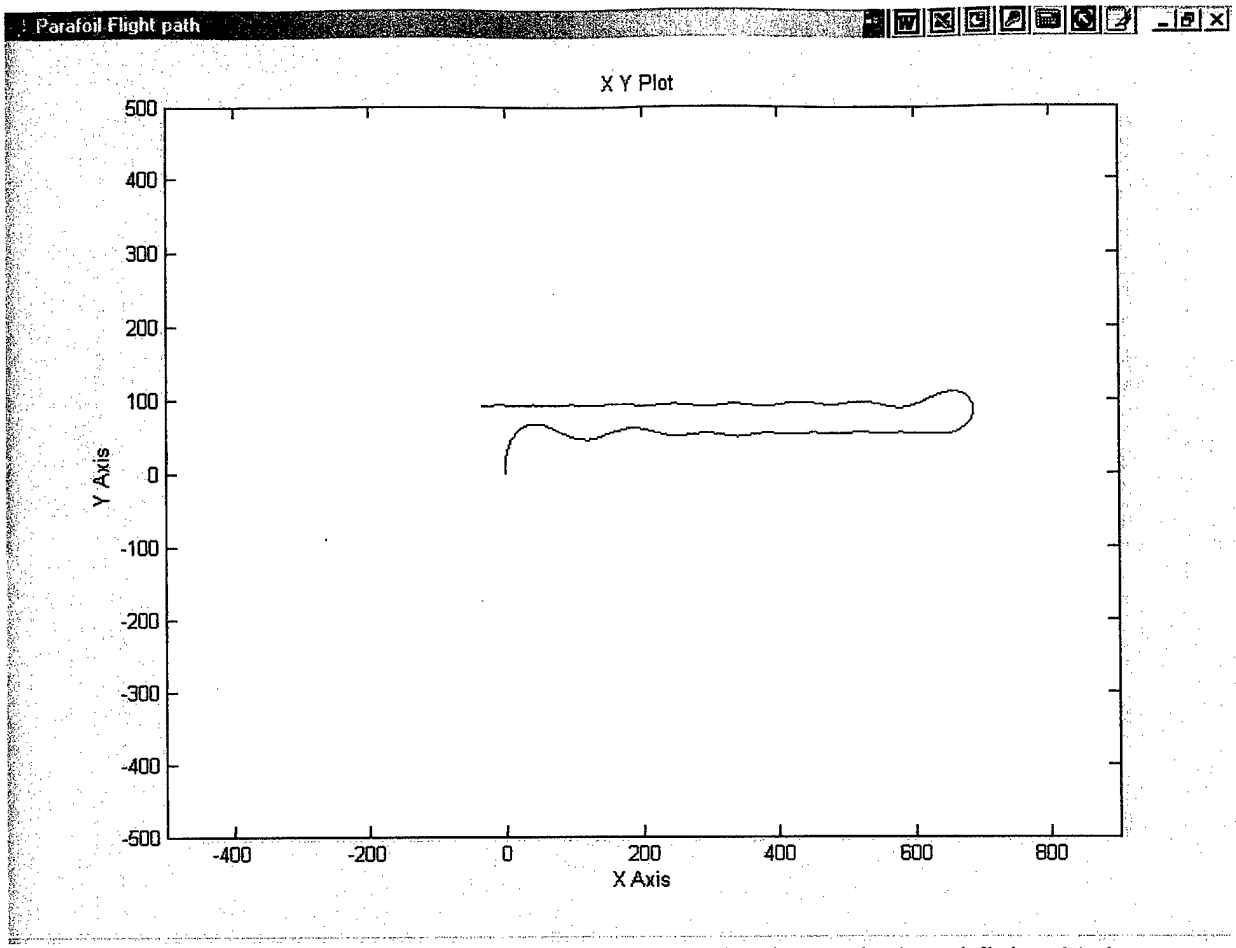


Figure A14 1 180 degree turn without heading acceleration input, deployed flying 90 degrees from desired heading, with noise variance 2.0 degrees

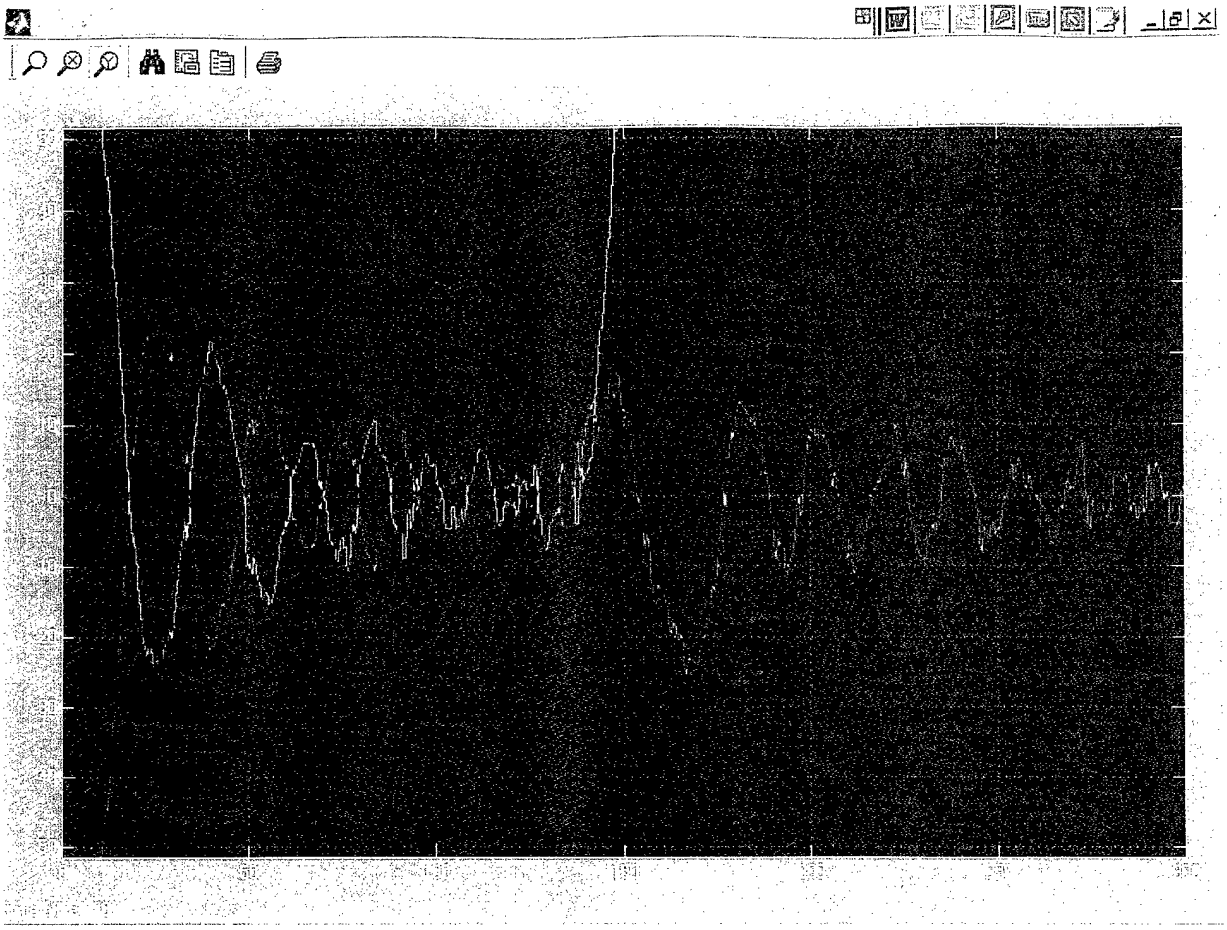


Figure A15 1 180 degree turn without heading acceleration input, deployed flying 90 degrees from desired heading, with noise variance 2.0 degrees

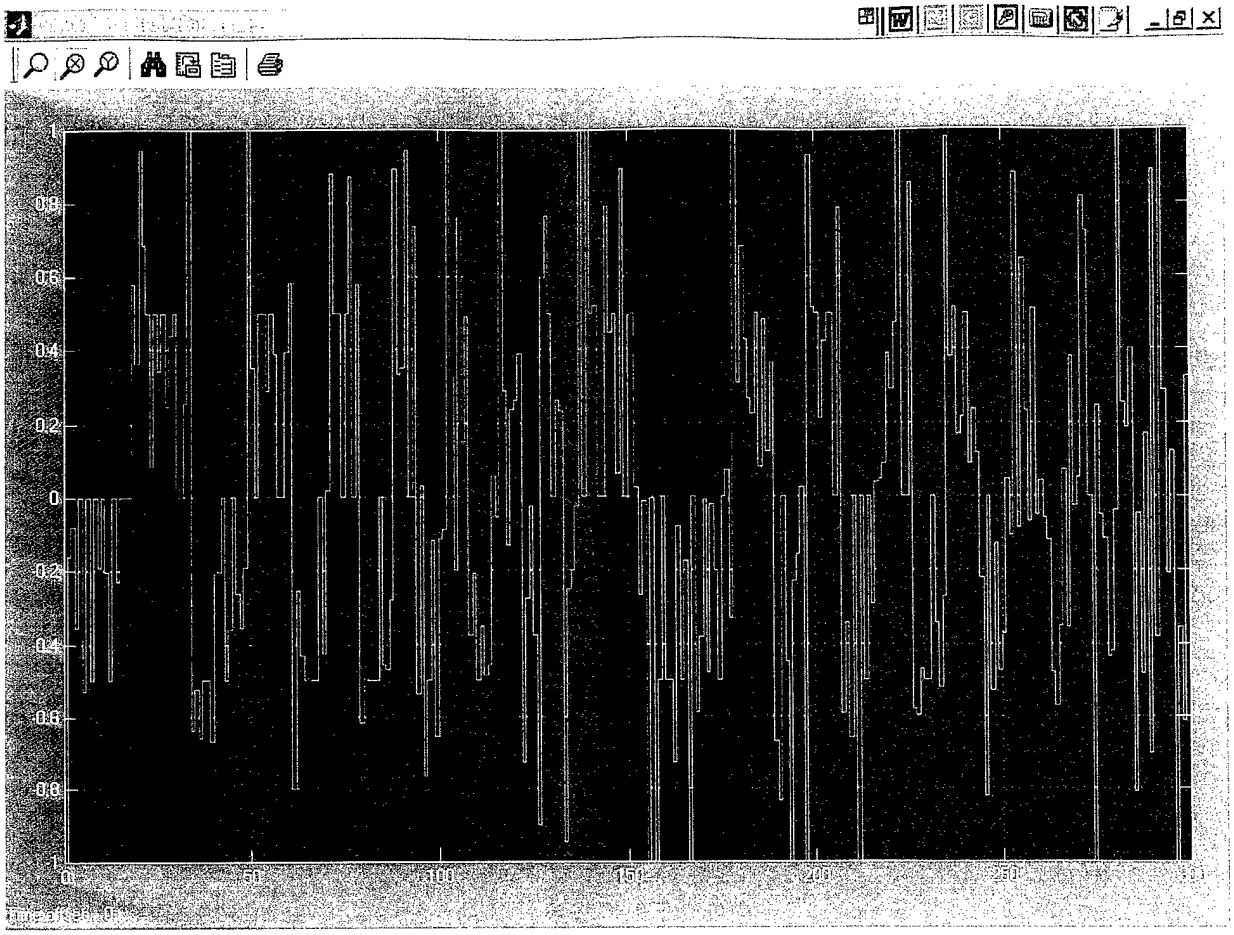


Figure A16 1 180 degree turn without heading acceleration input, deployed flying 90 degrees from desired heading, with noise variance 2.0 degrees

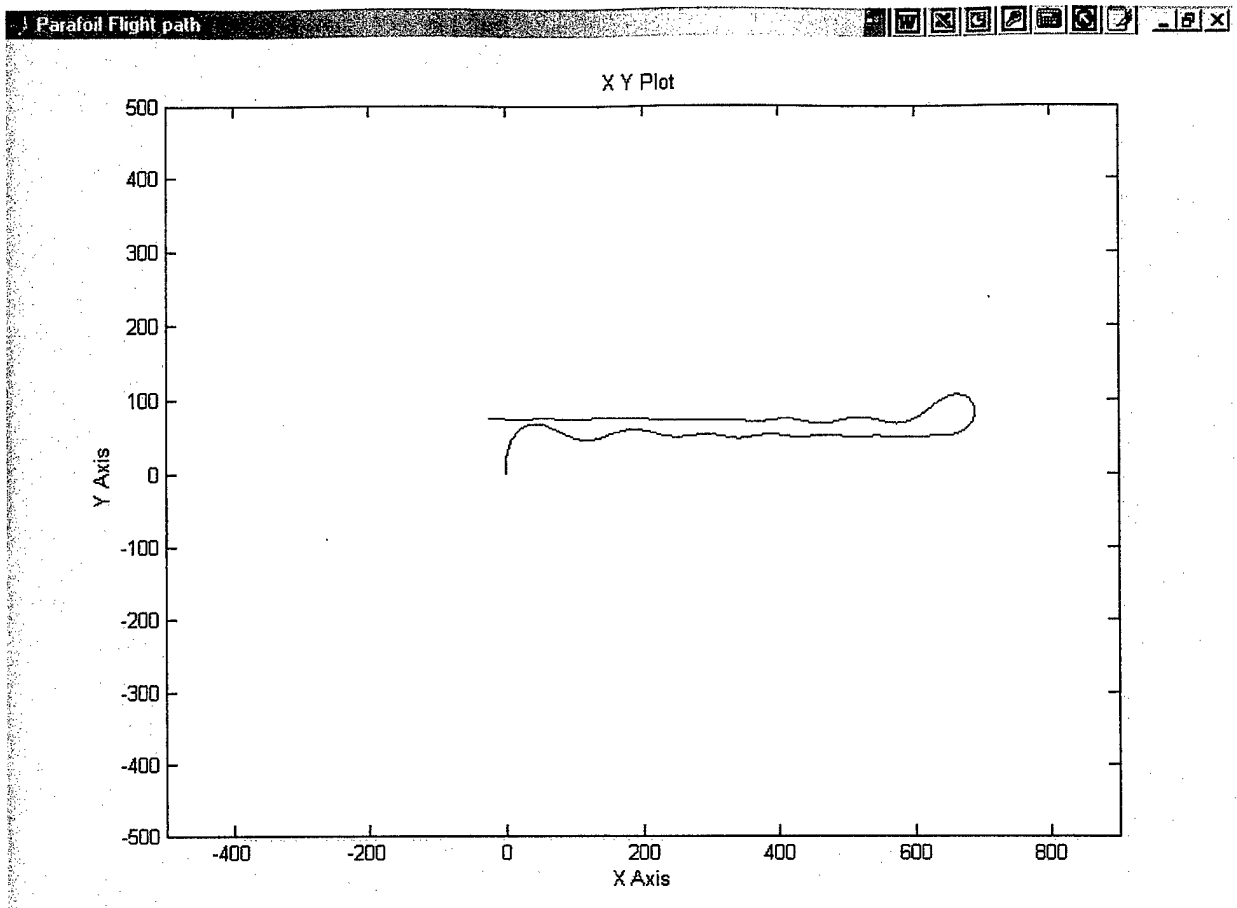


Figure A17 1 180 degree turn with heading acceleration input, deployed flying 90 degrees from desired heading, with noise variance 2.0 degrees

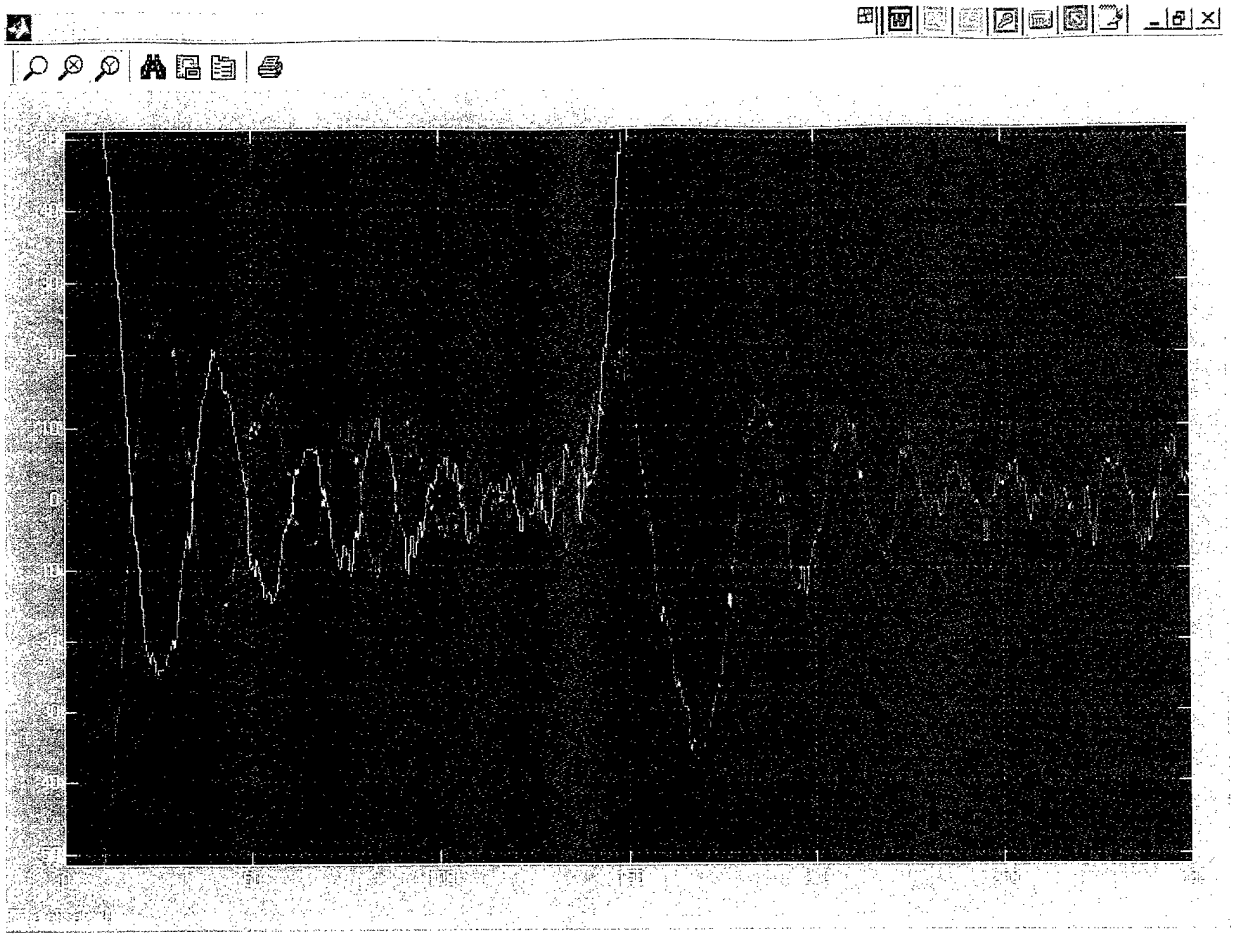


Figure A18 1 180 degree turn with heading acceleration input, deployed flying 90 degrees from desired heading, with noise variance 2.0 degrees

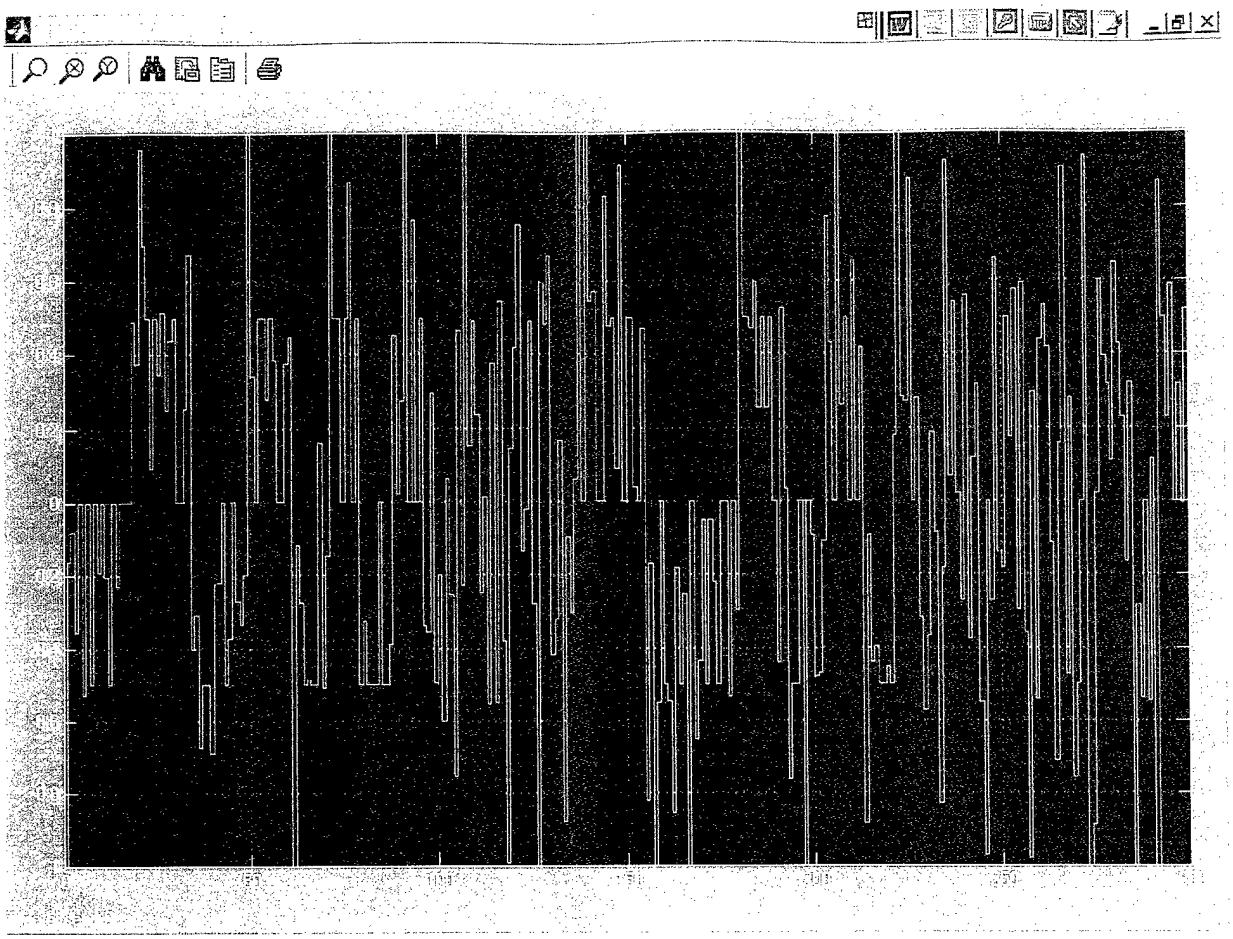


Figure A19 1 180 degree turn with heading acceleration input, deployed flying 90 degrees from desired heading, with noise variance 2.0 degrees

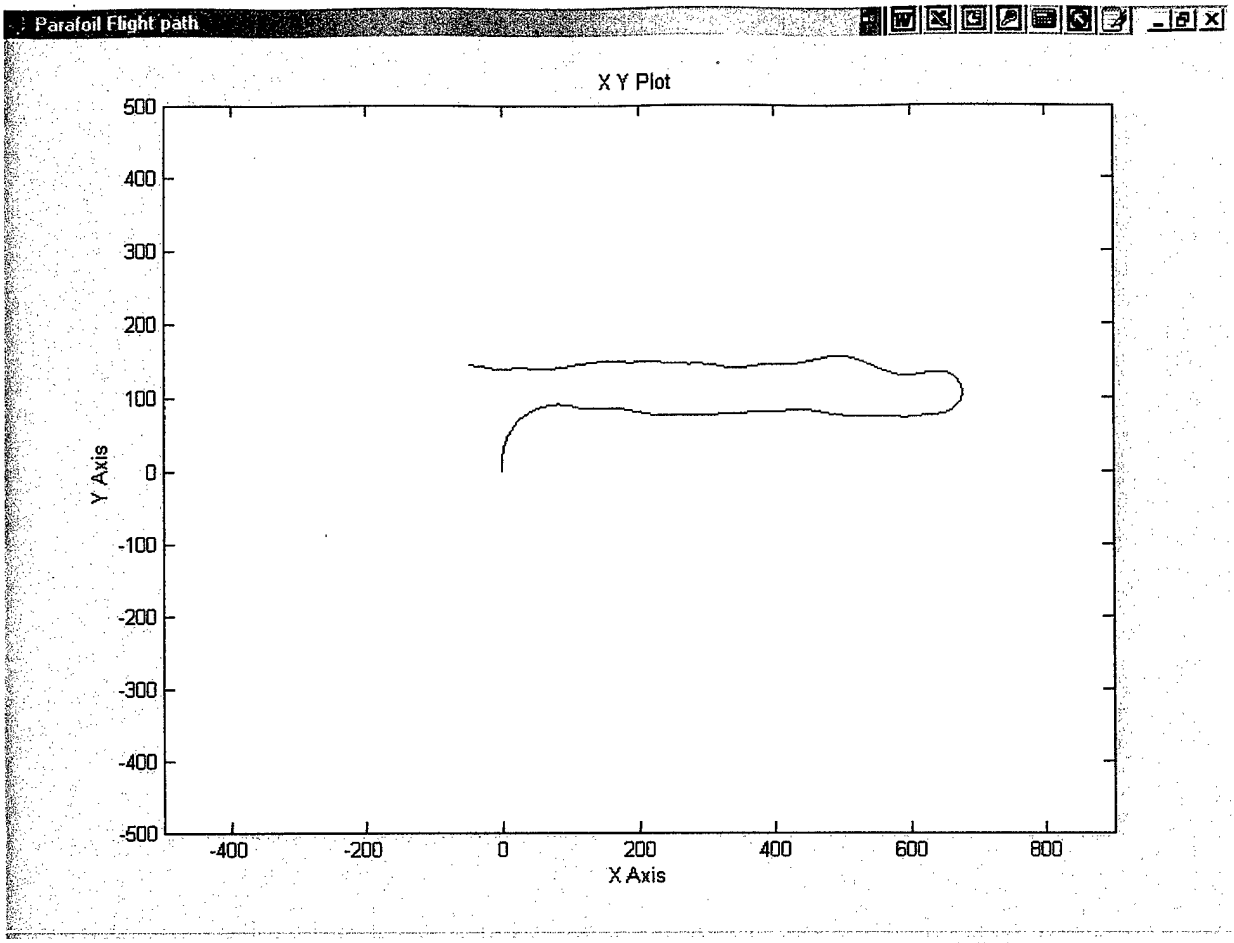


Figure A20 1 180 degree turn without heading acceleration input, deployed flying 90 degrees from desired heading, with wind gusts amp. .1 and freq .5 rad/sec. And noise variance 2.0 degrees

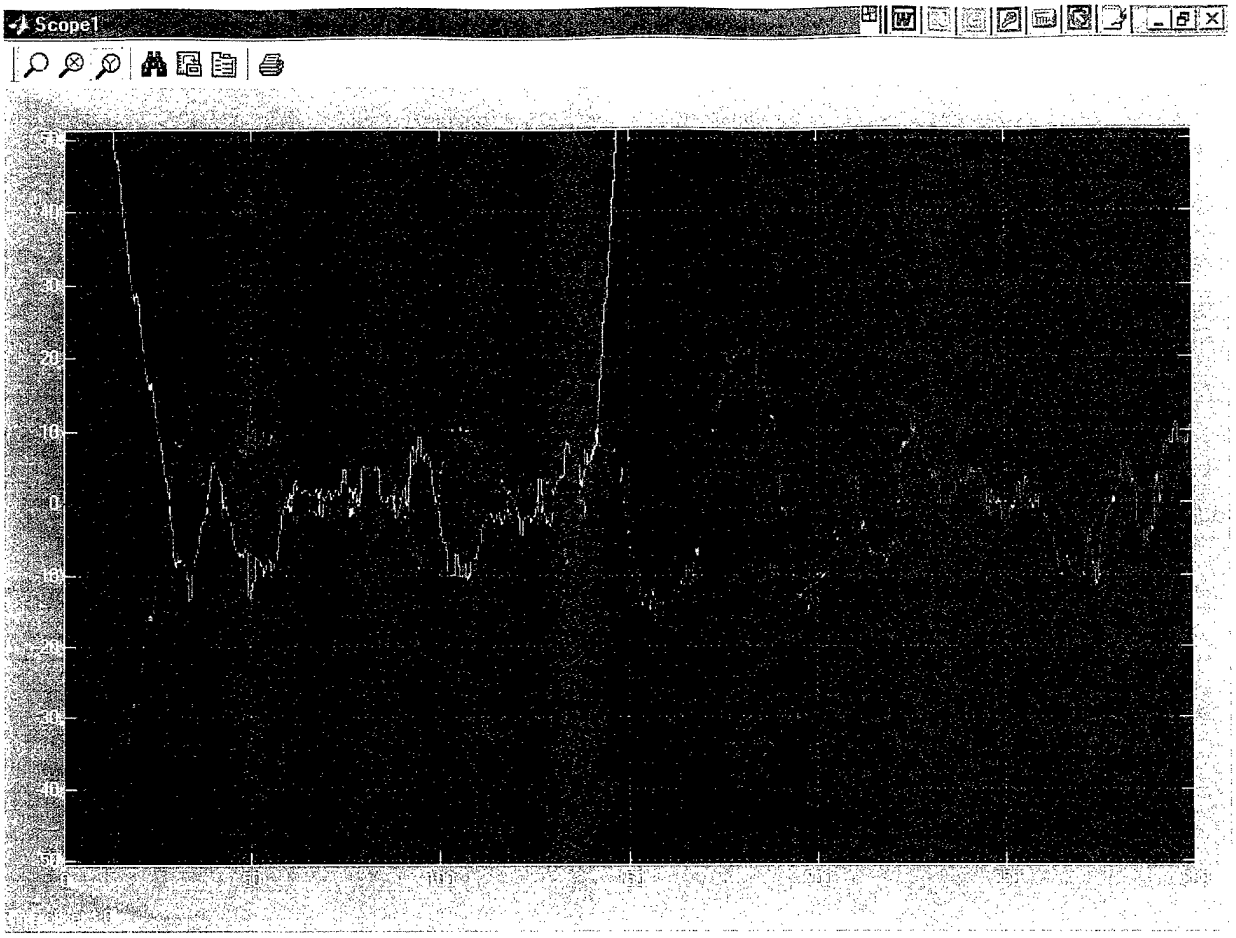


Figure A21 1 180 degree turn without heading acceleration input, deployed flying 90 degrees from desired heading, with wind gusts amp. .1 and freq .5 rad/sec. And noise variance 2.0 degrees

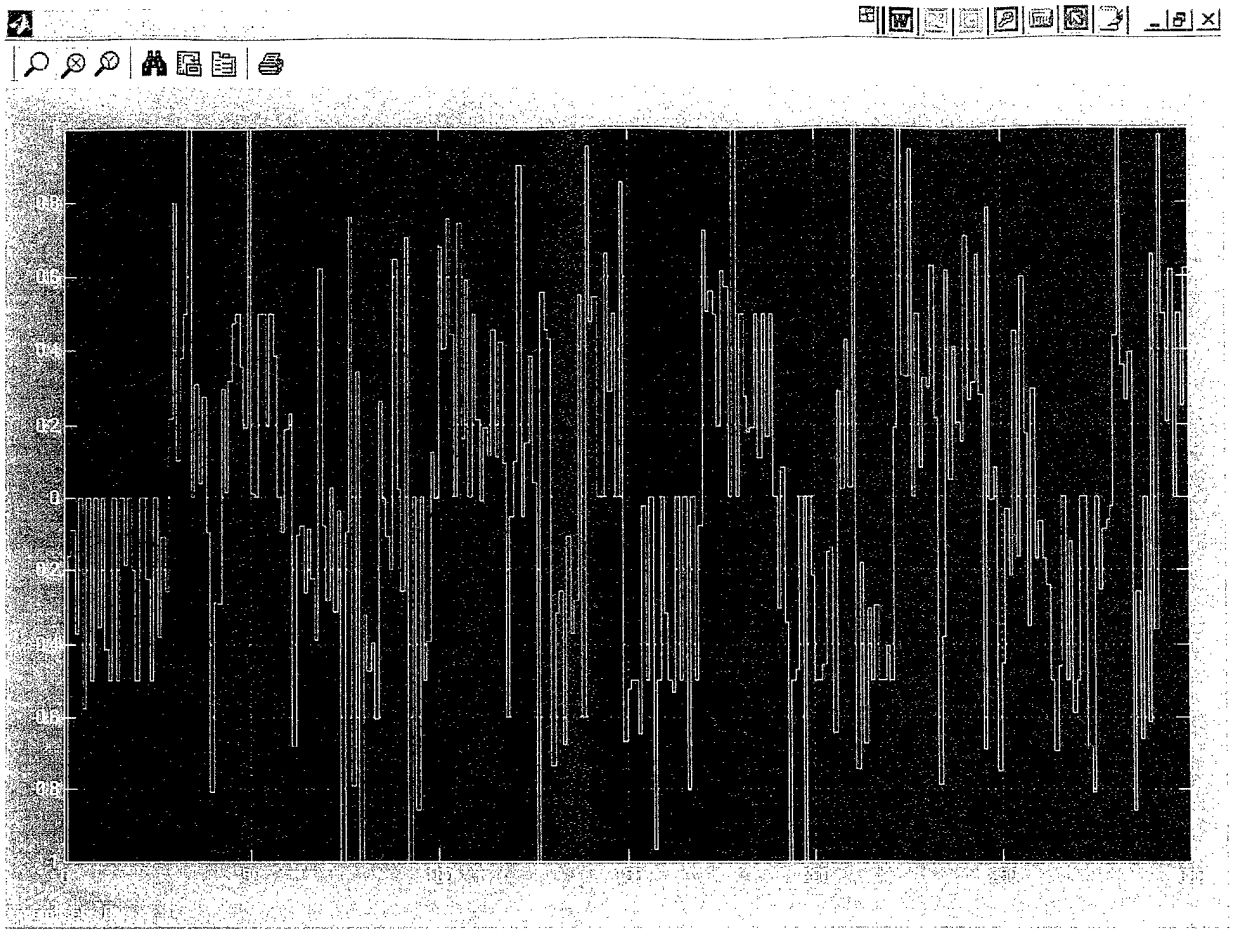


Figure A22 1 180 degree turn without heading acceleration input, deployed flying 90 degrees from desired heading, with wind gusts amp. .1 and freq .5 rad/sec. And noise variance 2.0 degrees

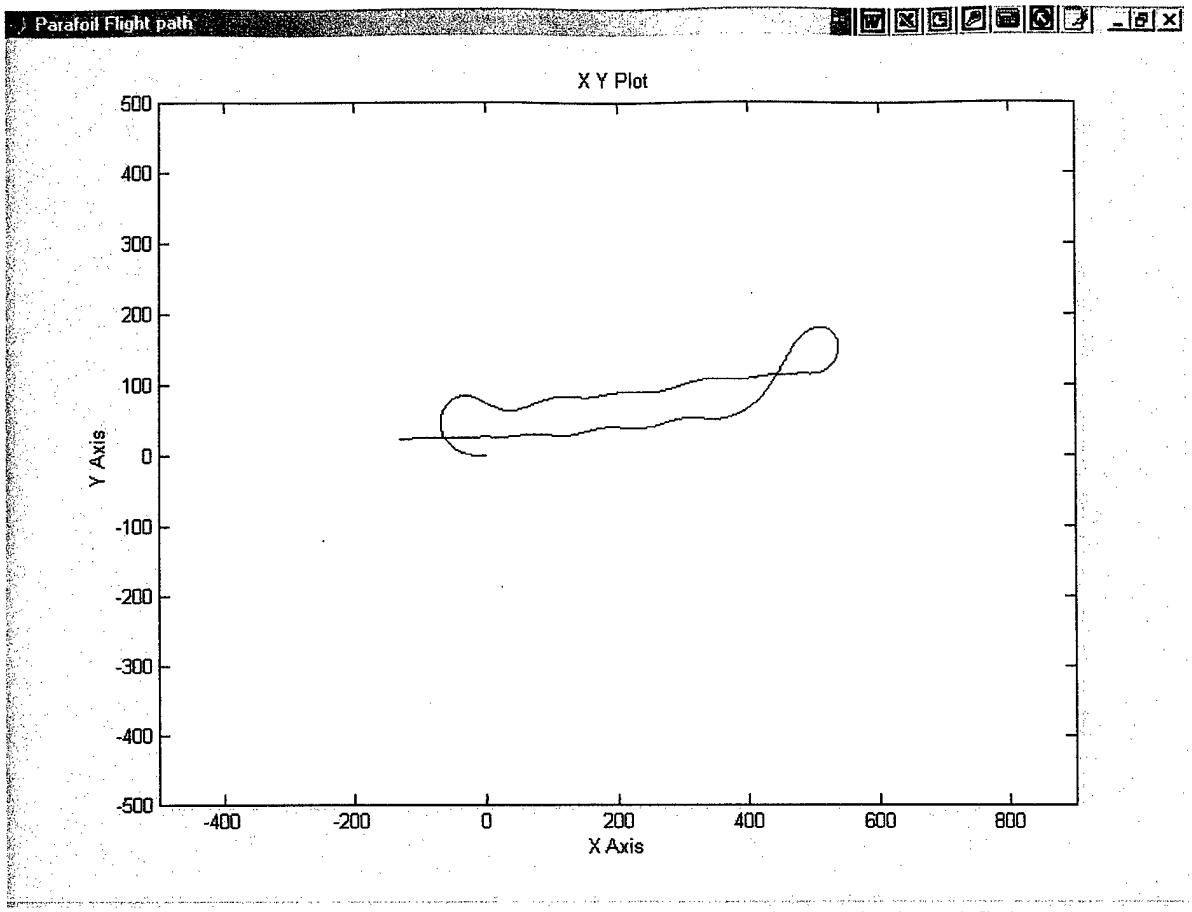


Figure A23 1 180 degree turn without heading acceleration input, deployed flying 180 degrees from desired heading, noise variance 2.0 degrees, with wind gusts amp. .5 and freq .01 rad/sec.

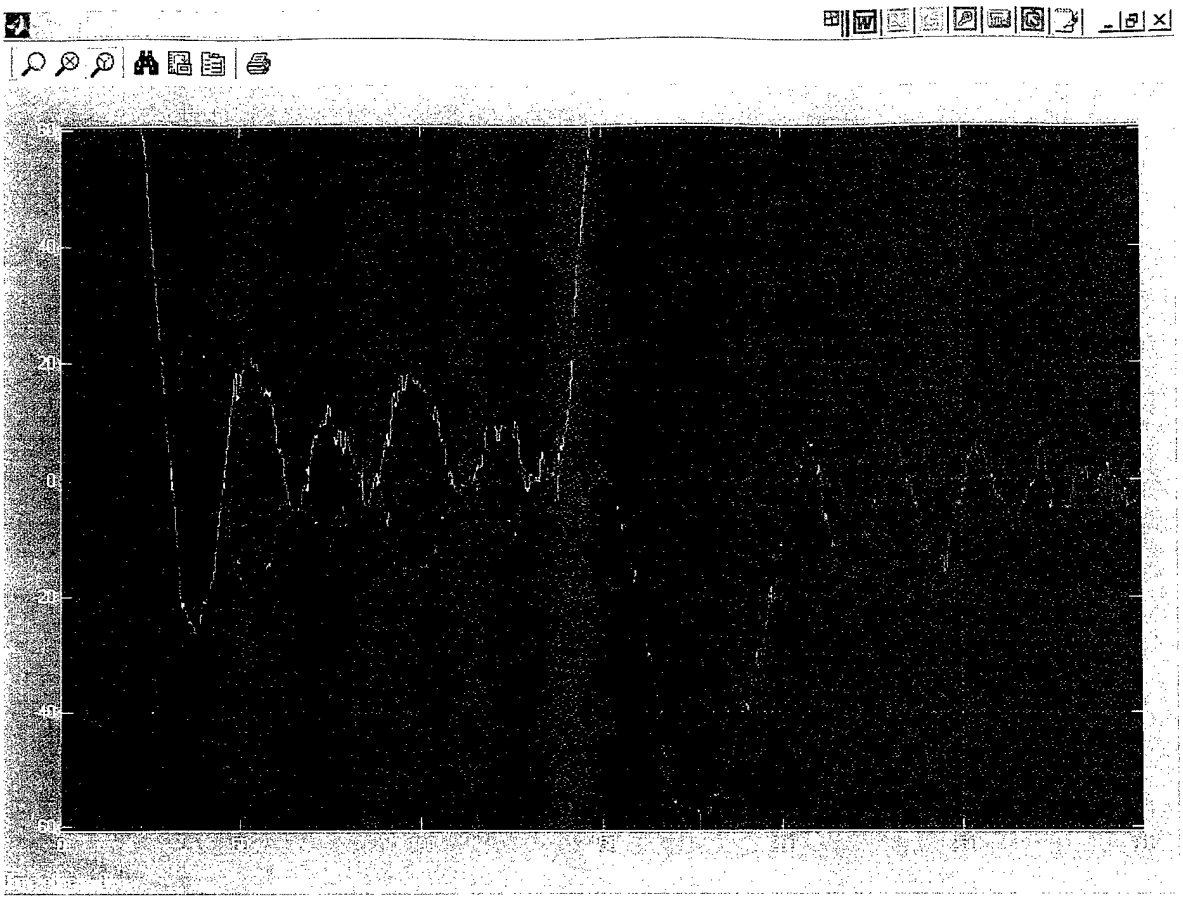


Figure A24 1 180 degree turn without heading acceleration input, deployed flying 180 degrees from desired heading, noise variance 2.0 degrees, with wind gusts amp. .5 and freq .01 rad/sec.

Development of an Aerodynamic Table Lookup System and Landing Gear Model for the Cal Poly Flight Stimulator

Project Investigators:

Daniel J. Biezad, Ph.D.
Professor and Department Chair
Aerospace Engineering Department

Chris Atkinson
Aerospace Engineering Department

Abstract

An aerodynamic table lookup system and a landing gear model have been developed to improve the capabilities of the Cal Poly Flight Simulator. The previous flight simulator model made use of a linearized aerodynamic model for control system analysis and did not model the effects of landing gear on the aircraft. The aerodynamic table lookup system allows a realistic flight model to be created using non-linear aerodynamic data from multiple sources. The landing gear is a configurable component based system that allows any landing gear design to be modeled. With the addition of these two components, the Cal Poly Flight simulator can now simulate pilot tasks such as taxi, takeoff, and landing, and can be used as a design tool for testing aircraft performance under a wide range of flight conditions. The improvements to the flight simulator also serve as a foundation for future projects. Appendices are included to document the flight simulator, allow it to be used as a design tool, and to facilitate future improvements.

Table of Contents

List of Tables and Figures ---- 2
Nomenclature ---- 3
Background ---- 5
Six-Degrees-of-Freedom Model ---- 6
Aerodynamic Table Lookup System ---- 7
Landing Gear Model ---- 8
Results ---- 10
Conclusions ---- 12
Acknowledgments ---- 12
References ---- 12
Appendix Nomenclature ---- 13
Appendix A – Equations of Motion ---- 17
Appendix B – Aerodynamic Table Lookup System ---- 25
Appendix C – Landing Gear ---- 31
Appendix D – Crash Detection System ---- 42
Appendix E – Simulink Block ---- 43
Appendix F – Input File Formats ---- 45
Appendix G – *Vendetta*, Aerodynamic Modeling Example ---- 50

List of Figures and Tables

Figures

Figure 1 – Cal Poly Flight Simulator ---- 5
Figure 2 – Flight Cab and Instruments ---- 5
Figure 3 – Graphics Environment ---- 5
Figure 4 – Heads-up Display ---- 5
Figure 5 – Altitude of Aircraft During Simulated Drop Test ---- 10
Figure 6 – Velocity of Aircraft During Simulated Drop Test ---- 11
Figure 7 – Landing Gear Forces on Aircraft During Simulated Drop Test ---- 11
Figure 8 – Aircraft Euler Angles During Simulated Drop Test ---- 11
Figure B.1 – Orthographic Grid of Function Values and a Desired Value (×) ---- 27
Figure B.2 – Relationship Between Corner Values and Relative Area of Opposite Quadrant ---- 27
Figure B.3 – Three Dimensional Grid Cube with Opposite Volume Relationship ---- 28
Figure B.4 – Extrapolation or Nearest Value Options for Off-Table Values ---- 29
Figure B.5 – Example Table ---- 30
Figure C.1 – Landing Gear Component ---- 31

Figure C.2 – Typical Oleo-Pneumatic Shock Absorber ---- 32
Figure C.3 – Pneumatic Compression Stroke for F-4 Main Gear and Nose Gear ---- 34
Figure C.4 – Total Shock Absorber Force from Simulated Drop Test of F-4 Main Gear ---- 35
Figure C.5 – Coordinate Systems Used in Landing Gear Model ---- 37
Figure G.1 – Aerodynamic Component Breakdown of *Vendetta* ---- 50
Figure G.2 – Wing Lift Curve Slope as a Function of Mach Number ---- 51
Figure G.3 – Spanwise Lift Distribution for Stall Angle-of-Attack Determination ---- 51
Figure G.4 – Subsonic Wing Lift Curve (Mach 0.2) ---- 52
Figure G.5 – Transonic Area Distribution ---- 56
Figure G.6 – Supersonic Area Distribution (Mach 1.6) ---- 56
Figure G.7 – Variation in Parasite and Wave Drag Coefficients with Mach Number (50,000 ft) ---- 57

Tables

Table B.I – Internal Aerodynamic States for Table Lookup ---- 25
Table B.II – Table Lookup Function Methods ---- 26
Table E.I – Simulink Block Parameters ---- 43
Table E.II – Simulink Block Input Ports ---- 43
Table E.III – Simulink Block Output Ports ---- 44
Table F.I – Aerodynamics File Size Limits ---- 45
Table F.II – Aerodynamics File Format ---- 45
Table F.III – Landing Gear File Format ---- 49
Table F.IV – Crash Detection file Format ---- 49
Table G.I – *Vendetta* Model Controls ---- 58
Table G.II – *Vendetta* Model Parameters ---- 58

Nomenclature

AIAA	American Institute of Aeronautics and Astronautics
6DOF	Six-degrees-of-freedom model
C_L	Lift coefficient
C_D	Drag coefficient
C_y	Side force coefficient
C_l	Roll moment coefficient
C_m	Pitch moment coefficient
C_n	Yaw moment coefficient
CFD	Computational Fluid Dynamics
DATCOM	Air Force Data Compendium
F	Force vector, lb
I_{xy}	Moment of inertia in *xy*-plane, slug ft^2
I_{yz}	Moment of inertia in *yz*- plane, slug ft^2
L	Roll moment, ft lb
M	Pitch moment, ft lb
N	Yaw moment, ft lb
NASA	National Aeronautics and Space Administration
S	Wing planform area, ft^2
V	Velocity vector
V_T	True airspeed, ft/s
b	Wing span, ft
$\overline{c}$	Mean aerodynamic chord, ft
h	Altitude
p	Roll rate, rad/s

$\hat{p}$	Non-dimensional roll rate $\hat{p} = \frac{pb}{2V_T}$
q	Pitch rate, rad/s
$\hat{q}$	Non-dimensional pitch rate $\hat{q} = \frac{q\bar{c}}{2V_T}$
r	Yaw rate, rad/s
$\hat{r}$	Non-dimensional yaw rate $\hat{r} = \frac{rb}{2V_T}$
α	Angle-of-attack, rad
$\dot{\alpha}$	Rate of change of angle-of-attack, rad/s
$\hat{\dot{\alpha}}$	Non-dimensional rate of change of angle-of-attack $\hat{\dot{\alpha}} = \frac{\dot{\alpha}\bar{c}}{2V_T}$
β	Sideslip angle, rad
$\dot{\beta}$	Rate of change of sideslip angle, rad/s
$\hat{\dot{\beta}}$	Non-dimensional rate of change of sideslip angle $\hat{\dot{\beta}} = \frac{\dot{\beta}b}{2V_T}$
δ_a	Aileron deflection, rad
δ_e	Elevator deflection, rad
δ_r	Rudder deflection, rad
ρ	Air density, slug/ft^3
ϕ	Roll angle, rad, deg
θ	Pitch angle, rad, deg
ψ	Yaw angle, rad, deg

Subscripts

0	Constant value
a	Air-path axes
$\hat{p}$	Variation with non-dimensional roll rate
$\hat{q}$	Variation with non-dimensional pitch rate
$\hat{r}$	Variation with non-dimensional yaw rate
x	x direction
y	y direction
z	z direction
α	Variation with angle-of-attack,
β	Variation with sideslip angle, rad
δ_a	Variation with aileron deflection
δ_e	Variation with elevator deflection
δ_r	Variation with rudder deflection

Background

The Cal Poly Flight Simulator program began when NASA Dryden donated the flight simulation cab, shown in Figure 1 and Figure 2 to the Cal Poly Aeronautical Engineering Department. The flight cab models the cockpit of an F-4 Phantom and is equipped with flight instrumentation and force feedback flight controls. The force feedback is applied to the elevator, aileron, and rudder controls by three custom built computer controlled electromagnets. Up to fifty pounds of force can be applied to each control. The flight controls and instrumentation are powered by two multi-channel analog computers that amplify and adjust analog signals. The analog computer allows the flight controls and instrumentation to be controlled by a single Windows based PC equipped with analog-to-digital and digital-to-analog cards.

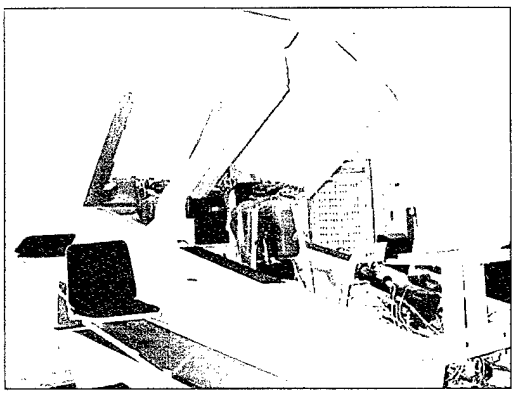


Figure 1 – Cal Poly Flight Simulator

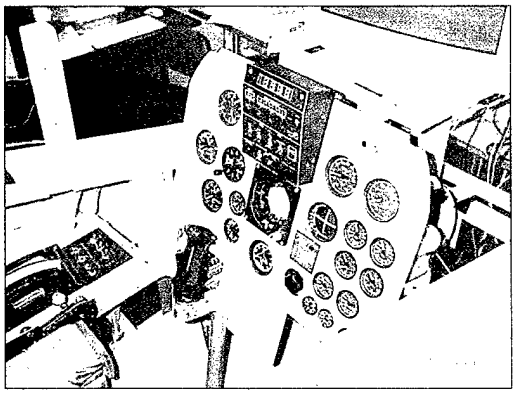


Figure 2 – Flight Cab and Instruments

The simulation graphics are run on separate computers with off-the-shelf OpenGL graphics cards. The position and orientation of the aircraft are sent over a local network for use by the graphics application. The 3DLinX graphics software package by Global Majic is used to display terrain, external views of the aircraft, and an onscreen heads-up display as shown in Figure 3 and Figure 4. The terrain modeled in the graphics is the Mojave airport and surrounding area.

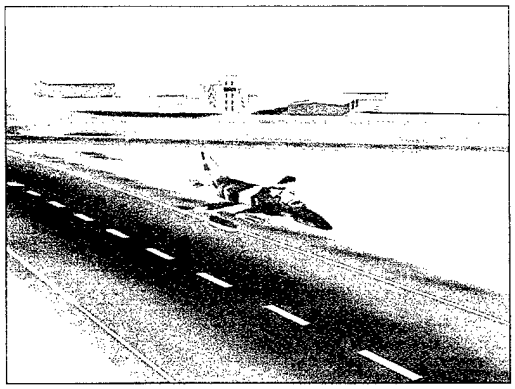


Figure 3 – Graphics Environment

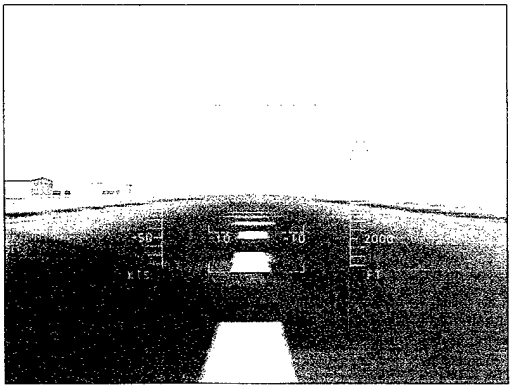


Figure 4 – Heads-up Display

Initially, the graphics was only used for control system design by giving pilots visual feedback for rating aircraft handling qualities. It soon became apparent that the potential existed to develop pilot tasks, such as taxi, takeoff, approach, and landing, based on the airport in the graphics environment. To make use of this potential it was necessary to make a series of upgrades to the existing simulation model.

Background

The Cal Poly Flight Simulator program began when NASA Dryden donated the flight simulation cab, shown in Figure 1 and Figure 2 to the Cal Poly Aeronautical Engineering Department. The flight cab models the cockpit of an F-4 Phantom and is equipped with flight instrumentation and force feedback flight controls. The force feedback is applied to the elevator, aileron, and rudder controls by three custom built computer controlled electromagnets. Up to fifty pounds of force can be applied to each control. The flight controls and instrumentation are powered by two multi-channel analog computers that amplify and adjust analog signals. The analog computer allows the flight controls and instrumentation to be controlled by a single Windows based PC equipped with analog-to-digital and digital-to-analog cards.

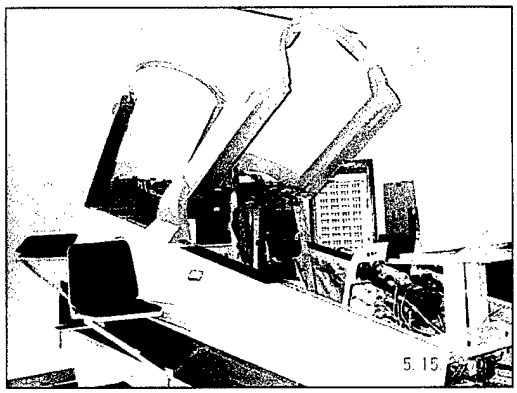


Figure 1 – Cal Poly Flight Simulator

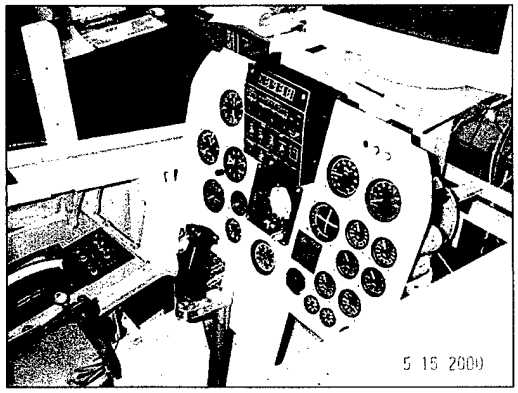


Figure 2 – Flight Cab and Instruments

The simulation graphics are run on separate computers with off-the-shelf OpenGL graphics cards. The position and orientation of the aircraft are sent over a local network for use by the graphics application. The 3DLinX graphics software package by Global Majic is used to display terrain, external views of the aircraft, and an onscreen heads-up display as shown in Figure 3 and Figure 4. The terrain modeled in the graphics is the Mojave airport and surrounding area.

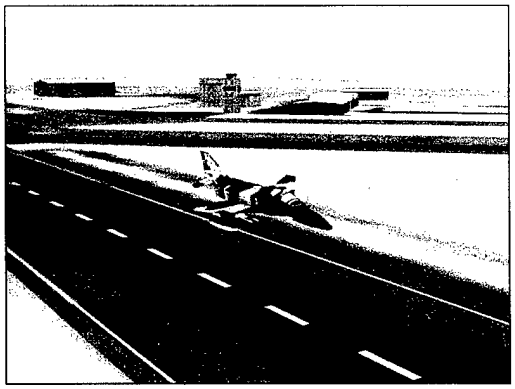


Figure 3 – Graphics Environment

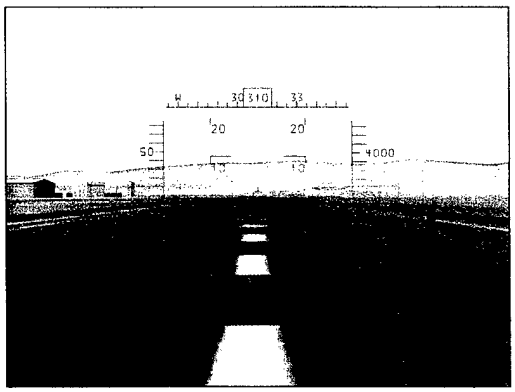


Figure 4 – Heads-up Display

Initially, the graphics was only used for control system design by giving pilots visual feedback for rating aircraft handling qualities. It soon became apparent that the potential existed to develop pilot tasks, such as taxi, takeoff, approach, and landing, based on the airport in the graphics environment. To make use of this potential it was necessary to make a series of upgrades to the existing simulation model.

First, the flight simulator did not include a landing gear or ground model. Because the six-degrees-of-freedom model (6DOF) operates independently of the graphics environment, the aircraft would continue to fly as if it were still in the air even though ground level had been reached in the graphics environment. It was necessary to create a mathematical landing gear model within the 6DOF that would coincide with the graphical terrain and allow the aircraft to taxi, takeoff, and land like a real aircraft.

Second, the flight simulator used a linearized aerodynamic model. This type of aerodynamic model makes use of stability derivatives based on linear Taylor series expansions to define the aerodynamics of the aircraft as shown in Eqs. (1-6).

Longitudinal

$$C_L = C_{L_0} + C_{L_\alpha}\alpha + C_{L_{\dot\alpha}}\hat{\dot\alpha} + C_{L_q}\hat{q} + C_{L_{\delta_e}}\delta_e \tag{1}$$

$$C_D = C_{D_0} + C_{D_\alpha}|\alpha| + C_{D_{\delta_e}}|\delta_e| \tag{2}$$

$$C_m = C_{m_0} + C_{m_\alpha}\alpha + C_{m_{\dot\alpha}}\hat{\dot\alpha} + C_{m_q}\hat{q} + C_{m_{\delta_e}}\delta_e \tag{3}$$

Lateral

$$C_y = C_{y_\beta}\beta + C_{y_p}\hat{p} + C_{y_r}\hat{r} + C_{y_{\delta_a}}\delta_a + C_{y_{\delta_r}}\delta_r \tag{4}$$

$$C_l = C_{l_\beta}\beta + C_{l_p}\hat{p} + C_{l_r}\hat{r} + C_{l_{\delta_a}}\delta_a + C_{l_{\delta_r}}\delta_r \tag{5}$$

$$C_n = C_{n_\beta}\beta + C_{n_p}\hat{p} + C_{n_r}\hat{r} + C_{n_{\delta_a}}\delta_a + C_{n_{\delta_r}}\delta_r \tag{6}$$

The linear aerodynamic model is based on small perturbation control theory around a single flight condition. Although a linear aerodynamics model simplifies control system analysis, it does not accurately simulate the flight characteristics of an aircraft, especially under non-straight and level flight conditions such as takeoff and landing. Since the lift and drag coefficients of an aircraft varies linearly with angle-of-attack, stall characteristics and induced drag are not modeled accurately. To simulate pilot tasks in a wide range of flight conditions, it was necessary to develop a system for defining nonlinear aerodynamics.

Six-Degrees-of-Freedom Model

The first step in updating the flight model in the simulator was to derive, verify, modify, and document the existing equations of motion. The major modifications to the equations of motion from earlier versions were the application of aerodynamic forces in air-path axes, the inclusion of the rate of change of the sideslip angle, the distinction between ground velocity and true airspeed by incorporating atmospheric wind inputs, and the use of radii of gyration to allow the aircraft mass to change in flight.

The equations of motion are derived based on Newton's equations of motion, however they are applied in aircraft body coordinates to simplify application of aerodynamic and external forces. The equations of motion are divided into translational equations, which define the motion of the center of gravity of the aircraft relative to earth, and rotational equations, which define the rotation of the aircraft around it center of gravity. In the simulator, some assumptions are made to simplify the equations of motion.

1. The aircraft is a rigid body with six degrees-of-freedom.
2. The body coordinate system is aligned with the aircraft with its origin at the center of gravity.
3. Earth is a flat surface and it is an inertial reference.
4. The mass of the aircraft is constant.
5. The aircraft is symmetrical about the *xz* plane $\left(I_{xy} = I_{yz} = 0\right)$.
6. Aerodynamic forces and moments operate in the air-path axis system.

Because the equations of motion are applied in the body coordinate system, and Newton's equations are defined relative to an inertial reference frame, the equations of motion make extensive used of the Coriolis theorem to transform derivatives between rotating coordinate systems. Because the earth is assumed to be flat, the latitude and longitude system is not used. Instead, positions on earth are defined by a distance north and a distance east from an arbitrary origin. Although the mass of the aircraft is assumed to be constant, it can be changed in flight. The effects of momentum change due to fuel burned during flight are simulated by an engine model rather than in the equations of motion. Any change of the mass of the aircraft is assumed to be a change in the overall density of the aircraft. This means that the moments of inertia of the aircraft vary proportionally with the mass of the aircraft. To simplify any changes in mass, the moments of inertia of the aircraft are expressed in terms of radii of gyration. Appendix A describes in detailed the derivation of the equations of motion and their implementation in the flight simulator.

Aerodynamic Table Lookup System

The aerodynamic table lookup system designed to allow a user to create an aerodynamic model of an aircraft by supplying aerodynamic data tables from multiple sources. The concept behind the table lookup system is that effect of aerodynamics on an aircraft can be described in terms of three force coefficients and three moment coefficients. These force and moment coefficients define the net aerodynamic force and moment acting on the center of gravity of the aircraft in the air-path axis system. The coefficients are non-dimensionalized using air density, true airspeed, wing planform area, wing span, and wing mean aerodynamic chord. The definitions of the coefficients are given in Eqs. (7-12).

$$C_L = \frac{-F_{z,a}}{\frac{1}{2}\rho V_T^{\,2} S} \quad (7)$$

$$C_D = \frac{-F_{x,a}}{\frac{1}{2}\rho V_T^{\,2} S} \quad (8)$$

$$C_y = \frac{F_{y,a}}{\frac{1}{2}\rho V_T^{\,2} S} \quad (9)$$

$$C_l = \frac{L}{\frac{1}{2}\rho V_T^{\,2} S b} \quad (10)$$

$$C_m = \frac{M}{\frac{1}{2}\rho V_T^{\,2} S \bar{c}} \quad (11)$$

$$C_n = \frac{N}{\frac{1}{2}\rho V_T^{\,2} S b} \quad (12)$$

The table lookup system defines these force and moment coefficients as functions of the aerodynamic state of the aircraft including Mach number, angle-of-attack, sideslip angle, rate of change of angle-of-attack and sideslip angle, aircraft body rates, and geometric characteristics of the aircraft. The geometric characteristics of the aircraft include control surface deflections, high lift devices, landing gear, stores, and center of gravity location. These geometric characteristics are defined by a set of control inputs. The number and type of controls are arbitrary allowing the system to model any aircraft configuration. The table lookup system also allows the user to create a set of user-defined parameters to aid in the table lookup process. The user-defined parameters are defined using tables the same way force and moment coefficients are defined. The parameters can then be used to lookup other values.

The aerodynamics file used by the table lookup system is composed of a header and a series of functions, each defining a single value. The header is the first component of the file and defines the constant properties of the aircraft including planform area, span, mean aerodynamic chord, and radii of gyration. Each function uses one of five methods to define the value of the function. The five methods include a constant value, linear variation with a single parameter, linear variation with the absolute value of a single parameter, the product of multiple parameters and a constant gain, and a tabulated function of multiple parameters. Multiple

functions can be used to define a single value. In this case, the resulting value is the sum of the results of all of the functions that define that value.

The table is the most common method used to define a value. A table can use up to three parameters to define its value. The number of values and a list of the values is given for each parameter along with the function value at every combination of parameter values. The table lookup system interpolates between the given parameter values to create a smooth variation of the function value as the parameters change. If an actual parameter value falls outside of the given range of values, the system can be set to either extrapolate the function value or use the nearest given value. The extrapolation preference can be set independently for each parameter in each table.

The four other function methods are used to describe simpler variation of the function value or to combine multiple parameters into a single value. The constant, linear, and absolute value variations are used to define a simple linear variation of a value around a single condition similar to the linearized system described above. The product method is used to combine multiple values that have been previously defined. For example, the moment coefficient contribution of a component could be defined by the force coefficient contribution and a lever arm, which is defined in terms of Mach number and center of gravity location. The product method can also be used to enable and disable certain component contributions by multiplying by a control input of either 1 or 0. Appendix G gives an example of how the aerodynamics of an aircraft can be defined using a component buildup method.

The table lookup system is designed to allow multiple sources of aerodynamic data to be used to create an aerodynamic model. Possible source of data include classical aerodynamic theory, approximate numerical methods such as those used by DATCOM, CFD results, wind tunnel data, or flight test data. Because none of these methods can completely define an aerodynamic model under all flight conditions, data from multiple methods can be combined to build a complete aerodynamic model.

Landing Gear Model

To simulate the effects of landing gear on an aircraft during taxi, takeoff, and landing, it was necessary to develop a system that would allow the flight simulator to calculate the forces and moments applied to the aircraft by the landing gear based on the position and velocity of the aircraft. The model would require the effects friction, braking, and steering to be modeled so that a pilot could control the aircraft while on the ground. In addition, the landing gear model needed to provide information such as energy absorption characteristics and landing gear loads for uses in preliminary landing gear design. Because the Cal Poly Flight Simulator is used primarily as an aircraft design tool, it was also necessary for the landing gear model to be easily adaptable to different aircraft designs and landing gear configurations.

To meet these requirements, a component based configurable landing gear system was developed. An aircrafts landing gear is modeled as a series of landing gear components. Each landing gear component is modeled as a three axis damped spring system. The axis system used by each component is aligned with the direction of rotation of the wheel. Simple linear damped spring systems are used to model the flexibility of the landing gear component in the direction of rotation of the wheel and perpendicular to the direction of rotation. The vertical axis of the component is modeled as an oleo-pneumatic shock absorber. This allows the energy absorption and nonlinear load characteristics of the shock absorber to be simulated. Frictional forces and braking forces are applied to each landing gear component perpendicular to the ground.

The main concept behind the landing gear model is that the forces produced by a landing gear component can be deflection and rate of deflection of the component. The deflection of a landing gear component is the difference between the non-deflected position and actual position of the wheel. Initially, the

actual position of the wheel is assumed to be at the same location on the ground that it was at during the previous time step. This allows the wheel to remain stationary on the ground if the maximum frictional forces are not exceeded. The deflection must be converted into a coordinate system aligned with the landing gear component to get the deflection of the three damped spring systems. The velocity of the wheel relative to the ground is calculated using the translational and rotational velocity of the aircraft and the position of the landing gear relative to the center of gravity of the aircraft. This velocity is also converted into landing gear coordinates to define the rate of deflection of each of the damped spring systems. The forces produced by the component are calculated by applying simple linear damped spring equations or more complex equations for the oleo-pneumatic shock absorber.

The initial forces produced by the landing gear component are converted into a coordinate system parallel and perpendicular to the ground to get the normal and frictional forces on the wheel. The frictional forces can then be compared to maximum frictional forces to determine whether the wheel will slide across the ground. The maximum frictional force is equal to the force normal to the ground times the coefficient of friction between the tire and the ground. The static or kinetic coefficient of friction is used depending on whether the tire was sliding across the ground during the previous time step. If this maximum frictional force is exceeded, the force applied to the landing gear component is corrected to the kinetic frictional force. In the direction of rotation of the wheel, the force on the landing gear is limited to the braking force applied to the wheel. If the braking force is exceeded, then the wheel will roll across the ground. The resulting forces on the landing gear are converted into aircraft body coordinates to be applied to the 6DOF. The moment contributions from the forces are also calculated based on the position of the wheel relative to the center of gravity of the aircraft. Appendix C gives a detailed description of the entire landing gear calculation process.

Steering the aircraft can be accomplished in two ways, by using differential braking or by steering one or more wheels. Consider a typical tricycle landing gear configuration. Because the braking force on each landing gear component can be controlled independently, a braking force can be applied to only one of the rear landing gear legs. The moment created by the asymmetrical forces on the aircraft will tend to yaw the aircraft in the direction of the braking force. For differential braking to work effectively, it is important for the nose wheel leg to be able to rotate around its vertical axis. This is because when a yaw moment is applied to the aircraft, a side force will be created in the nose wheel. The yaw moment from this side force will cancel out the original yaw moment from the braking force and prevent the aircraft from turning. However, if the wheel is positioned slightly behind the vertical axis of the landing gear, the side force will create a yaw moment around the nose wheel's vertical axis. Since the leg is allowed to rotate, the moment will steer the wheel in the direction of the original yaw moment and allow the aircraft to turn. To control the rotation of the landing gear, the vertical axis of each landing gear component is modeled as a torsional damped mass spring system. The offset of the wheel relative to the vertical axis of the each landing gear component is also defined. The frictional moment created by the tire rotating on the ground also contributes to the moment on the landing gear component. The angle of the wheel can be controlled directly to simulate nose wheel or tail wheel steering.

For each landing gear component, a set of maximum forces are defined relative to the landing gear component. These forces represent the maximum design load of the landing gear component. During each time step, the forces on the landing gear are compared to this maximum load. If the maximum force is exceeded, the landing gear has failed. A final addition to the landing gear model is a crash detector. The crash detector helps the landing gear model to avoid the impossible situation where the aircraft is underground. The crash detector works by defining a series of points at critical locations around the aircraft. The position of these crash points are converted into earth coordinates and compared with the ground elevation. If any point goes below ground, then the aircraft has crashed. This indicates that the results obtained from the landing gear model are not accurate because interactions between the ground and other parts of the aircraft are not accounted for. If the

landing gear fail or a crash is detected, the simulation can be automatically stopped or reset to indicate that the aircraft has been damaged.

Results

The results of the table lookup system are difficult to quantify because they depend on the data supplied to the system. By adding a typical lift curve, pitch moment characteristics based on a given static margin, and realistic induced drag, great improvement over the linear model can be seen. The aircraft will have typical stall characteristics and will tend to pitch upward once the stall angle-of-attack is reached. Under landing conditions, the aircraft exhibits a realistic power curve requiring more power to be added at lower speeds due to induced drag. The same drag increase can be seen in high load factor turns. If a more complicated component buildup is used, loss of control due to tail surface stalling can be modeled, and static margin characteristics can be predicted.

The results obtained from the landing gear model can be best illustrated with data from a simulated drop test. In this test, the F-4 Phantom model was dropped from a height of 2 ft above ground resulting in an approximate 8 ft/s sink rate on impact. The aircraft was given an initial 20 ft/s forward velocity, an initial sideways velocity of 1 ft/s to the right, and an initial pitch angle of 2 degrees. An 8,000 lb braking force was slowly applied to the two rear landing gear components 3 seconds into the test. Figures 5-8 show the resulting landing gear forces, aircraft velocity, Euler angles, and height as a function of time.

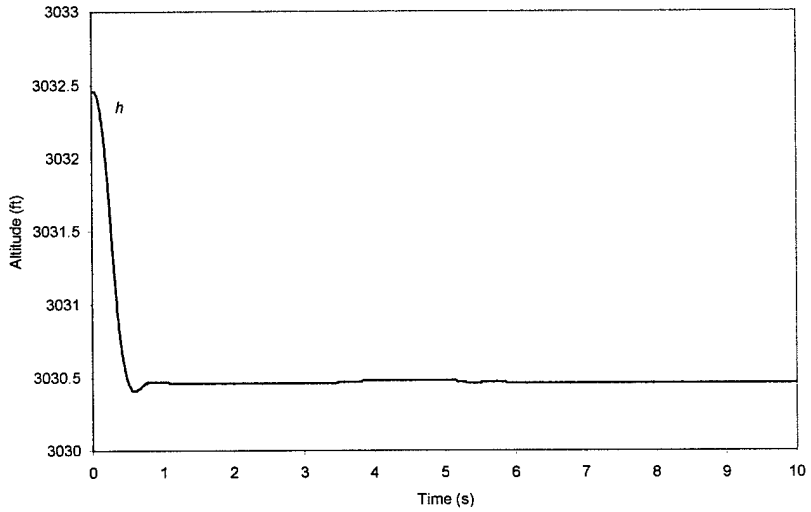


Figure 5 – Altitude of Aircraft During Simulated Drop Test

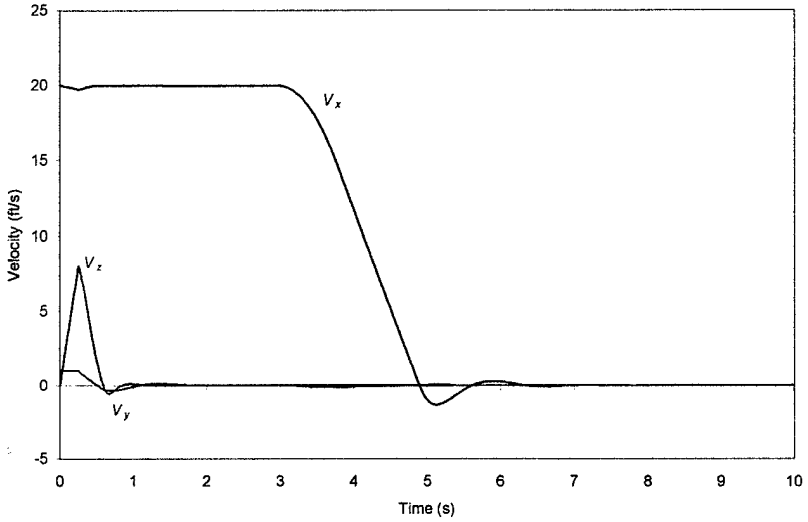


Figure 6 – Velocity of Aircraft During Simulated Drop Test

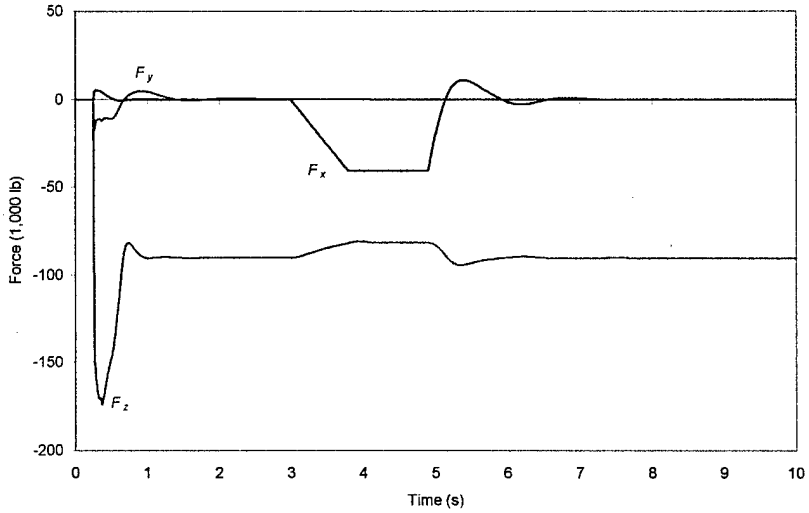


Figure 7 – Landing Gear Forces on Aircraft During Simulated Drop Test

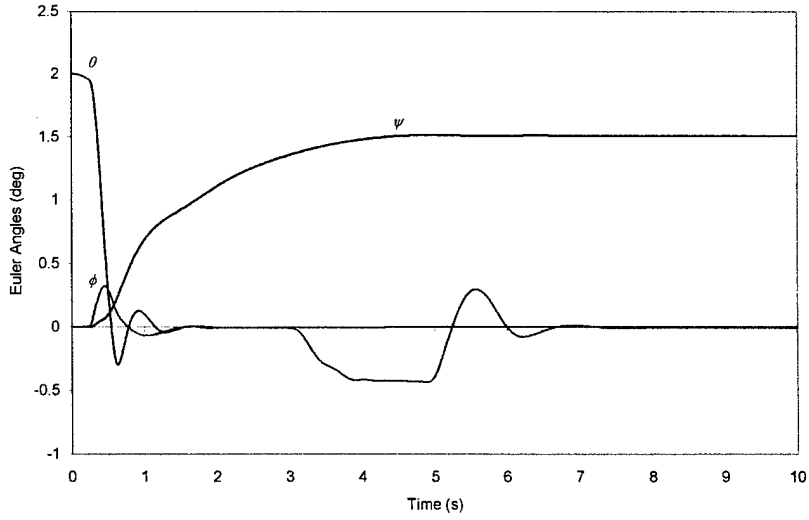


Figure 8 – Aircraft Euler Angles During Simulated Drop Test

Initially, the aircraft accelerates downward under the influence of gravity. Approximately a quarter of a second into the test, the wheels contact the ground. A side force stops the sideways velocity of the aircraft and causes the aircraft to roll slightly to the right. The vertical compression stroke and damping slow the vertical velocity to a static condition, but the aircraft continues to roll forward at 20 ft/s. After three seconds, the brakes are slowly applied and the aircraft pitches forward due to the moment created by braking. At approximately 5 seconds, the forward velocity of the aircraft has reached zero, the brakes lock, and the aircraft comes to a rest as a typical damped mass spring system.

Flying the F-4 with the landing gear model shows realistic interaction between the aircraft and the ground. The aircraft can be taxied on the ground using either differential braking or nose wheel steering, although nose wheel steering makes the aircraft unstable at speeds above 20 knots. The aircraft can rotate and takeoff under its own power, and it can be landed realistically on the runways in the graphics environment. In general, the simulator is now capable of accurately simulating takeoff and landing tasks.

Conclusions

The improvements made to the Cal Poly Flight Stimulator allow it to be used to simulate a wider range of pilot tasks and increase its capabilities as a design tool. The aerodynamic table lookup system allows the flight characteristic of any aircraft to be simulated given aerodynamic data. The landing gear model allows the effects of the landing gear of any aircraft to be modeled. This has the potential to act as a foundation for future projects and design work. A project is already underway to model runway surface shape and roughness in the landing gear model. Other possible additions to the simulator include: an atmospheric wind and turbulence model, modeling of control feedback from aerodynamic forces on control surfaces, and implementation of landing approach control systems. The appendices included are intended to aid future work on the flight simulator by providing detailed information about how the systems work and how they are used.

Acknowledgments

I gratefully acknowledge funded support from the Office of Naval Research through Dr. Susan Opava, Dean of Research and Graduate Programs at Cal Poly. I also recognize the assistance of Mr. Ed Burnett at Lockheed Martin Skunk Works in Palmdale CA. I would also like to thank Peter Siebold at Scaled Composites, Dan Salluce, and the rest of the team for their contributions to the flight simulator, which made this project possible.

References

Currey, N. S., *Aircraft Landing Gear Design: Principles and Practices*, AIAA Educational Series, AIAA, 1988.

Brandt, S.A., Stiles, R.J., Bertin, J.J., and Whitford, R., *Introduction to Aeronautics: A Design Perspective*, AIAA Educational Series, AIAA, 1997.

Jones, R.T., *Theory of Wing-Body Drag at Supersonic Speeds*, NACA 1284.

Milwitzky, B., and Cook, F.E., *Analysis of Landing-Gear Behavior*, NACA 1154.

Davis, P.A., *Quasi-Static and Dynamic Response Characteristics of F-4 Bias-Ply and Radial-Belted Main Gear Tires*, NASA Technical Paper 3586, 1997.

Keiser, K., Droney, C., Schnaible, N., Atkinson, C., Maglio, C., and Salluce, D., *Vendetta, 2001/2002 AIAA Foundation Undergraduate Team Aircraft Design Competition*, 2002.

Appendix Nomenclature

A	Fuselage cross-sectional area, ft^2
A_{max}	Maximum fuselage cross-sectional area, ft^2
$A_{contact}$	Wheel contact area, ft^2
AR	Wing aspect ratio
C	Transformation matrix
C_L	Lift coefficient
$C_{L_{LE\ flap}}$	Lift coefficient with leading edge flaps
$C_{L_{no\ LE\ flap}}$	Lift coefficient without leading edge flaps
C_D	Drag coefficient
C_{D_i}	Induced drag coefficient
$C_{D_{LE\ flap}}$	Drag coefficient with leading edge flaps
$C_{D_{no\ LE\ flap}}$	Drag coefficient without leading edge flaps
$C_{D_{wave}}$	Wave drag coefficient
$C_{D_{wave}}'$	Wave drag coefficient of a perfect Sears-Haack body
C_y	Side force coefficient
C_l	Roll moment coefficient
C_m	Pitch moment coefficient
C_n	Yaw moment coefficient
C_N	Normal force coefficient
E_{WD}	Wave drag efficiency factor
$\vec{F}$	Force vector $\vec{F} = \begin{bmatrix} F_x & F_y & F_z \end{bmatrix}^T$, lb
F_{brake}	Braking force, lb
F_{max}	Maximum friction force, lb
I	Moment of inertia matrix
I_x	Moment of inertia around x-axis, slug ft^2
I_y	Moment of inertia around y-axis, slug ft^2
I_z	Moment of inertia around z-axis, slug ft^2
I_{xy}	Moment of inertia in xy-plane, slug ft^2
I_{xz}	Moment of inertia in xz-plane, slug ft^2
I_{yz}	Moment of inertia in yz- plane, slug ft^2
$I_{z\,t}$	Gear leg moment of inertia, slug ft^2
K_{SM}	Static margin
L	Roll moment, ft lb
M	Pitch moment, ft lb
M_∞	Mach number
$M_{C_{D_0}max}$	Mach number for maximum wave drag
N	Yaw moment, ft lb
N_t	Gear leg torsional yaw moment, ft lb
N_{max}	Maximum frictional gear leg torsional yaw moment, ft lb
P	Pressure, lb/ft^2
P_{int}	Shock absorber internal pressure, lb/ft^2
$P_{int\,u}$	Shock absorber uncompressed internal pressure, lb/ft^2
P_{ext}	Shock absorber external pressure, lb/ft^2
S	Wing plan form area, ft^2
$\vec{T}$	Moment vector $\vec{T}_b = \begin{bmatrix} L & M & N \end{bmatrix}^T$
$\vec{V}$	Velocity vector, ft/s
$\vec{V}_k$	Ground velocity vector $\vec{V}_{k,b} = \begin{bmatrix} u & v & w \end{bmatrix}^T$, ft/s

$\vec{V}_T$	True airspeed velocity vector $\vec{V}_{T,b} = \begin{bmatrix} u_T & v_T & w_T \end{bmatrix}^T$, ft/s
$\vec{V}_{atm}$	Atmospheric velocity vector $\vec{V}_{atm,b} = \begin{bmatrix} u_{atm} & v_{atm} & w_{atm} \end{bmatrix}^T$, ft/s
$\dot{\vec{V}}$	Acceleration vector, ft/s^2
$\dot{\vec{V}}_k$	Ground acceleration vector $\dot{\vec{V}}_{k,b} = \begin{bmatrix} \dot{u} & \dot{v} & \dot{w} \end{bmatrix}^T$, ft/s^2
$\dot{\vec{V}}_T$	True airspeed acceleration vector $\dot{\vec{V}}_{T,b} = \begin{bmatrix} \dot{u}_T & \dot{v}_T & \dot{w}_T \end{bmatrix}^T$, ft/s^2
$\dot{\vec{V}}_{atm}$	Atmospheric acceleration vector $\dot{\vec{V}}_{atm,b} = \begin{bmatrix} \dot{u}_{atm} & \dot{v}_{atm} & \dot{w}_{atm} \end{bmatrix}^T$, ft/s^2
W	Aircraft Weight, lb
a	Constant
b	Wing span, ft
b	Damping constant, lb s/ft
b_t	Torsional damping constant, ft lb s/rad
$\bar{c}$	Mean aerodynamic chord, ft
d	Shock absorber piston diameter, ft
dt	Time step, s
f	Function
g	Acceleration due to gravity, ft/s^2
h	Altitude, ft
h_{ground}	Ground elevation, ft
k	Spring constant, lb/ft
k_t	Torsional spring constant ft lb/rad
k_{tire}	Tire vertical spring constant, lb/ft
$k_{t\,tire}$	Tire torsional spring constant, ft lb/rad
k_l	Induced drag term
l	Aircraft length, ft
m	Aircraft mass, slug
n	Load factor (g's)
p	Parameter
p	Roll rate, rad/s
p_0	Initial roll rate, rad/s
$\hat{p}$	Non-dimensional roll rate $\hat{p} = \frac{pb}{2V_T}$
$\dot{p}$	Roll rate acceleration, rad/s
$\vec{p}$	Position vector $\vec{p} = \begin{bmatrix} x & y & z \end{bmatrix}^T$
$\Delta\vec{p}$	Displacement vector $\Delta\vec{p} = \begin{bmatrix} \Delta x & \Delta y & \Delta z \end{bmatrix}^T$
q	Pitch rate, rad/s
q_∞	Dynamic pressure, lb/ft^2
q_0	Initial pitch rate, rad/s
$\hat{q}$	Non-dimensional pitch rate $\hat{q} = \frac{q\bar{c}}{2V_T}$
$\dot{q}$	Pitch rate acceleration, rad/s
r	Yaw rate, rad/s
r_0	Initial yaw rate, rad/s
$\hat{r}$	Non-dimensional yaw rate $\hat{r} = \frac{rb}{2V_T}$
$\dot{r}$	Roll rate acceleration, rad/s
r_x	Radius of gyration about x-axis, ft
r_y	Radius of gyration about y-axis, ft
r_z	Radius of gyration about z-axis, ft

r_{xz}	Radius of gyration in xz-plane, ft
u	Forward velocity, ft/s
u_0	Initial forward velocity, ft/s
$\dot{u}$	Forward acceleration, ft/s^2
v	Sideways velocity, ft/s
v_0	Initial sideways velocity, ft/s
$\dot{v}$	Sideways acceleration, ft/s^2
w	Vertical velocity, ft/s
w_0	Initial vertical velocity, ft/s
$\dot{w}$	Vertical acceleration, ft/s^2
x	x earth position, ft
x_0	Initial x earth position, ft
x_1	Longitudinal position on aircraft, ft
x_2	Longitudinal position on aircraft, ft
Δx	x displacement, ft
x'	Previous x earth position, ft
$\dot{x}$	x earth velocity, ft/s
x_{ac}	x position of aerodynamic center of component, ft
y	y earth position, ft
y_0	Initial y earth position, ft
Δy	y displacement, ft
y'	Previous y earth position, ft
$\dot{y}$	y earth velocity, ft/s
y_{ac}	y position of aerodynamic center of component, ft
z	z earth position, ft
z_0	Initial z earth position, ft
Δz	z displacement, ft
$\dot{z}$	z earth velocity, ft/s
z_{ac}	z position of aerodynamic center of component, ft
z_{stroke}	Shock absorber stroke, ft
z_{disp}	Shock absorber displacement, ft
z_{tire}	Tire displacement, ft
Λ_{LE}	Wing leading edge sweep, deg
α	Angle-of-attack, rad
α_w	Wheel angular acceleration, rad/s^2
α_H	Effective angle of attack of horizontal stabilizer, rad
α_V	Effective angle of attack of vertical stabilizer, rad
$\dot{\alpha}$	Rate of change of angle-of-attack, rad/s
$\hat{\dot{\alpha}}$	Non-dimensional rate of change of angle-of-attack $\hat{\dot{\alpha}} = \frac{\dot{\alpha}\bar{c}}{2V_T}$
β	Sideslip angle, rad
$\dot{\beta}$	Rate of change of sideslip angle, rad/s
$\hat{\dot{\beta}}$	Non-dimensional rate of change of sideslip angle $\hat{\dot{\beta}} = \frac{\dot{\beta}b}{2V_T}$
γ	Polytropic Compression Exponent
δ_x	Ground slope angle around earth x-axis $\delta_x = \tan^{-1}\frac{\partial h_{ground}}{\partial y_e}$, rad
δ_y	Ground slope angle around earth y-axis $\delta_y = \tan^{-1}\frac{\partial h_{ground}}{\partial x_e}$, rad
δ_e	Elevator deflection, rad
δ_a	Aileron deflection, rad

δ_r	Rudder deflection, rad
$\delta_{f_{LE}}$	Leading edge flap deflected
$\delta_{f_{LE}}'$	Leading edge flap not deflected
ϕ	Roll angle, rad
ϕ_0	Initial roll angle position, rad
$\dot{\phi}$	Roll angle rate, rad/s
θ	Pitch angle, rad
θ_0	Initial pitch angle position, rad
$\dot{\theta}$	Pitch angle rate, rad/s
ψ	Yaw angle, rad
ψ_0	Initial yaw angle position, rad
ψ_w	Wheel yaw angle, rad
ψ'_w	Previous wheel earth yaw angle, rad
$\dot{\psi}$	Yaw angle rate, rad/s
μ_s	Static coefficient of friction
μ_k	Kinetic coefficient of friction
$\vec{\omega}$	Angular velocity $\vec{\omega}_{b/e,b} = \begin{bmatrix} p & q & r \end{bmatrix}^T$, rad/s
$\dot{\vec{\omega}}$	Angular velocity $\dot{\vec{\omega}}_{b/e,b} = \begin{bmatrix} \dot{p} & \dot{q} & \dot{r} \end{bmatrix}^T$, rad/s
ω_w	Wheel angular velocity, rad/s

Subscripts

a	Air-path coordinates
aero	Aerodynamic force or moment
atm	Atmospheric velocity or acceleration
b	Body coordinates
e	Earth coordinates
ext	External force or moment
g	Gear coordinates
gear	Gear force or moment
i	Inertial reference
k	Ground velocity
max	Maximum force
offset	Offset distance
point	Crash point
T	True airspeed
u	Uncompressed
w	Wheel coordinates
x	*x* direction
y	*y* direction
z	*z* direction
$\hat{\varepsilon}$	Euler coordinates

Appendix A – Equations of Motion

Derivation

Translational

The translation of the aircraft is calculated using Newton's second law of motion, which states that the force on an object is equal to the rate of change of the linear momentum of the object relative to an inertial reference frame. If $\vec{F}_b$ is the force vector on the aircraft in body coordinates, m is the mass of the aircraft, and $\vec{V}_{i,b}$ is the velocity of the aircraft relative to an inertial reference frame expressed in body coordinates,

$$\vec{F}_b = \frac{d_i}{dt}\left(m\vec{V}_{i,b}\right) \tag{A-1}$$

If we assume that the earth is an inertial reference frame, then $\vec{V}_{i,b} = \vec{V}_{k,b}$, where $\vec{V}_{k,b}$ is the aircraft's velocity relative to the earth (ground velocity) in body coordinates. We also assume that the mass of the aircraft is constant. Eq. (1) can be rewritten as,

$$\vec{F}_b = m\frac{d_e}{dt}\vec{V}_{k,b} \tag{A-2}$$

To find the acceleration of the aircraft relative to the body frame, we must apply the Coriolis Theorem. The Coriolis Theorem states that the derivative of a vector relative to one reference frame is equal to the derivative of the vector relative to a second frame plus the cross product of the rotation vector of the second frame relative to the first frame and the vector. If $\vec{\omega}_{b/e,b}$ is the rotation vector of the body frame with respect to the earth frame expressed in body coordinates, then,

$$\vec{F}_b = m\left(\dot{\vec{V}}_{k,b} + \vec{\omega}_{b/e,b} \times \vec{V}_{k,b}\right) \tag{A-3}$$

If u, v, and w are the components of $\vec{V}_{k,b}$, and p, q, and r are the components of $\vec{\omega}_{b/e,b}$, then we can separate Eq. (3) into its components.

$$F_{x,b} = m\left(\dot{u} + qw - rv\right) \tag{A-4a}$$

$$F_{y,b} = m\left(\dot{v} + ru - pw\right) \tag{A-4b}$$

$$F_{z,b} = m\left(\dot{w} + pv - qu\right) \tag{A-4c}$$

Solving Eqs. (4a – 4c) for $\dot{u}$, $\dot{v}$, and $\dot{w}$, we get,

$$\dot{u} = \frac{F_{x,b}}{m} + rv - qw \tag{A-5a}$$

$$\dot{v} = \frac{F_{y,b}}{m} + pw - ru \tag{A-5b}$$

$$\dot{w} = \frac{F_{z,b}}{m} + qu - pv \tag{A-5c}$$

Rotational

Newton's second law also states that the moment on an object is equal to the rate of change of the angular momentum of the object relative to an inertial reference frame. If $\vec{T}_b$ is the moment vector on the aircraft in body coordinates, I, is the moment of inertia matrix of the aircraft, and $\vec{\omega}_{b/i,b}$, is the rotation of the body frame (or aircraft) with respect to an inertial reference point expressed in body coordinates,

$$\vec{T}_b = \frac{d_i}{dt}\left(I\,\vec{\omega}_{b/i,b}\right) \tag{A-6}$$

where,

$$I = \begin{bmatrix} I_x & -I_{xy} & -I_{xz} \\ -I_{xy} & I_y & -I_{yz} \\ -I_{xz} & -I_{yz} & I_z \end{bmatrix} \tag{A-7}$$

Again, we assume that earth is an inertial reference frame, so $\vec{\omega}_{b/i,b} = \vec{\omega}_{b/e,b}$, and Eq. (6) can be written as,

$$\vec{T}_b = \frac{d_e}{dt}\left(I\,\vec{\omega}_{b/e,b}\right) \tag{A-8}$$

Using the Coriolis Theorem to find the moments expressed in the body coordinate system,

$$\vec{T}_b = I\,\dot{\vec{\omega}}_{b/e,b} + \vec{\omega}_{b/e,b} \times I\,\vec{\omega}_{b/e,b} \tag{A-9}$$

Multiplying out the moment of inertia matrix in Eq. (9), we get,

$$\vec{T}_b = \begin{bmatrix} \dot{p}I_x - \dot{q}I_{xy} - \dot{r}I_{xz} \\ -\dot{p}I_{xy} + \dot{q}I_y - \dot{r}I_{yz} \\ -\dot{p}I_{xz} - \dot{q}I_{yz} + \dot{r}I_z \end{bmatrix} + \vec{\omega}_{b/e,b} \times \begin{bmatrix} pI_x - qI_{xy} - rI_{xz} \\ -pI_{xy} + qI_y - rI_{yz} \\ -pI_{xz} - qI_{yz} + rI_z \end{bmatrix} \tag{A-10}$$

Expanding the cross product in Eq. (10) and separating the components of the moment vector, L, M, and N,

$$L = \dot{p}I_x - \dot{q}I_{xy} - \dot{r}I_{xz} + qrI_z - pqI_{xz} + \left(r^2 - q^2\right)I_{yz} - qrI_y + prI_{xy} \tag{A-11a}$$

$$M = \dot{q}I_y - \dot{p}I_{xy} - \dot{r}I_{yz} - prI_z + \left(p^2 - r^2\right)I_{xz} + pqI_{yz} + prI_x - qrI_{xy} \tag{A-11b}$$

$$N = \dot{r}I_z - \dot{p}I_{xz} - \dot{q}I_{yz} + pqI_y + \left(q^2 - p^2\right)I_{xy} - prI_{yz} - pqI_x + qrI_{xz} \tag{A-11c}$$

For an aircraft of standard configuration, $I_{xy} \approx 0$ and $I_{yz} \approx 0$, so Eqs. (11a – 11c) can be simplified to,

$$L = \dot{p}I_x - \dot{r}I_{xz} + qrI_z - pqI_{xz} - qrI_y \tag{A-12a}$$

$$M = \dot{q}I_y - prI_z + \left(p^2 - r^2\right)I_{xz} + prI_x \tag{A-12b}$$

$$N = \dot{r}I_z - \dot{p}I_{xz} + pqI_y - pqI_x + qrI_{xz} \tag{A-12c}$$

Finally, solving Eqs. (12a – 12c) for $\dot{p}$, $\dot{q}$, and $\dot{r}$,

$$\dot{p} = \frac{L I_z + N I_{xz} + pq\left(I_x I_{xz} - I_y I_{xz} + I_z I_{xz}\right) + qr\left(I_y I_z - I_{xz}^{\ 2} - I_z^{\ 2}\right)}{I_x I_z - I_{xz}^{\ 2}} \tag{A-13a}$$

$$\dot{q} = \frac{M + pr\left(I_z - I_x\right) + \left(r^2 - p^2\right) I_{xz}}{I_y} \tag{A-13b}$$

$$\dot{r} = \frac{N + pq\left(I_x - I_y\right) + \left(\dot{p} - qr\right) I_{xz}}{I_z} \tag{A-13c}$$

The three translational Eqs. (5a – 5c) and three rotational Eqs. (13a – 13c) are the general equations of motion used in the flight simulator.

$$\dot{u} = \frac{F_x}{m} + rv - qw \qquad \dot{p} = \frac{L I_z + N I_{xz} + pq\left(I_x I_{xz} - I_y I_{xz} + I_z I_{xz}\right) + qr\left(I_y I_z - I_{xz}^{\ 2} - I_z^{\ 2}\right)}{I_x I_z - I_{xz}^{\ 2}}$$

$$\dot{v} = \frac{F_y}{m} + pw - ru \qquad \dot{q} = \frac{M + pr\left(I_z - I_x\right) + \left(r^2 - p^2\right) I_{xz}}{I_y}$$

$$\dot{w} = \frac{F_z}{m} + qu - pv \qquad \dot{r} = \frac{N + pq\left(I_x - I_y\right) + \left(\dot{p} - qr\right) I_{xz}}{I_z}$$

Implementation

Translational

In the flight simulator, the forces on the aircraft are the sum of the external forces, aerodynamic forces, and gravitational forces.

$$F_x = F_{x\,ext,b} + F_{x\,gear,b} + F_{x\,aero,b} + m g_{x,b} \tag{A-14a}$$

$$F_y = F_{y\,ext,b} + F_{y\,gear,b} + F_{y\,aero,b} + m g_{y,b} \tag{A-14b}$$

$$F_z = F_{z\,ext,b} + F_{z\,gear,b} + F_{z\,aero,b} + m g_{z,b} \tag{A-14c}$$

Using Eqs. (4a – 4b) derived above, we can solve for the acceleration of the aircraft.

$$\dot{u} = \frac{F_x}{m} + rv - qw$$

$$\dot{v} = \frac{F_y}{m} + pw - ru$$

$$\dot{w} = \frac{F_z}{m} + qu - pv$$

The acceleration of the aircraft is integrated over time with an initial velocity to get the velocity of the aircraft.

$$u = \int \dot{u}\, dt + u_0 \tag{A-15a}$$

$$v = \int \dot{v}\, dt + v_0 \tag{A-15b}$$

$$w = \int \dot{w}\, dt + w_0 \tag{A-15c}$$

The velocity of the aircraft is then converted into earth coordinates.

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = C_b^e \begin{bmatrix} u \\ v \\ w \end{bmatrix} \tag{A-16}$$

where,

$$C_e^b = \begin{bmatrix} \cos\theta\cos\psi & \cos\theta\sin\psi & -\sin\theta \\ \sin\phi\sin\theta\cos\psi - \cos\phi\sin\psi & \sin\phi\sin\theta\sin\psi + \cos\phi\cos\psi & \sin\phi\cos\theta \\ \cos\phi\sin\theta\cos\psi + \sin\phi\sin\psi & \cos\phi\sin\theta\sin\psi - \sin\phi\cos\psi & \cos\phi\cos\theta \end{bmatrix} \qquad C_b^e = C_e^{b^T} \tag{A-17}$$

The earth coordinates are then integrated with an initial position to get the position of the aircraft in earth coordinates.

$$x = \int \dot{x}\, dt + x_0 \tag{A-18a}$$

$$y = \int \dot{y}\, dt + y_0 \tag{A-18b}$$

$$z = \int \dot{z}\, dt + z_0 \tag{A-18c}$$

Rotational

The moments on the aircraft in the flight simulator are the sum of the external moments and aerodynamic moments.

$$L = L_{ext} + L_{gear} + L_{aero} \tag{A-19a}$$

$$M = M_{ext} + M_{gear} + M_{aero} \tag{A-19b}$$

$$N = N_{ext} + N_{gear} + N_{aero} \tag{A-19c}$$

The moments of inertia of the aircraft are calculated from the radii of gyration and the mass of the aircraft. The sign of the moment of inertia in the *xz*-plane can be negative, so the sign of the radius of gyration must be preserved.

$$m = \frac{W}{g} \tag{A-20}$$

$$I_x = m r_x^2 \tag{A-21a}$$

$$I_y = m r_y^2 \tag{A-21b}$$

$$I_z = m r_z^2 \tag{A-21c}$$

$$I_{xz} = m \operatorname{sgn}(r_{xz}) r_{xz}^2 \tag{A-21d}$$

Eqs. (13a – 13c) derived above are used to calculate the angular acceleration of the aircraft.

$$\dot{p} = \frac{L I_z + N I_{xz} + pq\left(I_x I_{xz} - I_y I_{xz} + I_z I_{xz}\right) + qr\left(I_y I_z - I_{xz}^2 - I_z^2\right)}{I_x I_z - I_{xz}^2}$$

$$\dot{q} = \frac{M + pr\left(I_z - I_x\right) + \left(r^2 - p^2\right) I_{xz}}{I_y}$$

$$\dot{r} = \frac{N + pq\left(I_x - I_y\right) + \left(\dot{p} - qr\right) I_{xz}}{I_z}$$

The angular acceleration is integrated with an initial angular velocity to get the angular velocity of the aircraft.

$$p = \int \dot{p}\, dt + p_0 \tag{A-22a}$$

$$q = \int \dot{q}\, dt + q_0 \tag{A-22b}$$

$$r = \int \dot{r}\, dt + r_0 \tag{A-22c}$$

The angular velocity is then converted into Euler coordinates to get the rate of change of the Euler angles.

$$\begin{bmatrix} \dot{\psi} \\ \dot{\theta} \\ \dot{\phi} \end{bmatrix} = C_b^{\hat{e}} \begin{bmatrix} p \\ q \\ r \end{bmatrix} \tag{A-23}$$

where,

$$C_b^{\hat{e}} = \begin{bmatrix} 0 & \sin\phi\sec\theta & \cos\phi\sec\theta \\ 0 & \cos\phi & -\sin\phi \\ 1 & \sin\phi\tan\theta & \cos\phi\tan\theta \end{bmatrix} \tag{A-24}$$

The Euler rates are integrated with initial Euler angles to get the Euler angles of the aircraft.

$$\psi = \int \dot{\psi}\, dt + \psi_0 \tag{A-25a}$$

$$\theta = \int \dot{\theta}\, dt + \theta_0 \tag{A-25b}$$

$$\phi = \int \dot{\phi}\, dt + \phi_0 \tag{A-25c}$$

Euler Angle Correction

The pitch angle, θ, ranges from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$. If these limits are exceeded then the aircraft has pitched past vertical and the pitch angle must be measured from the opposite horizon. The pitch angle is corrected with the given formula and the yaw and roll angles are reversed.

$$\begin{aligned} &\text{if } |\theta| > \tfrac{\pi}{2} \\ &\theta = \begin{cases} \pi - \theta & \theta > 0 \\ -\pi - \theta & \theta < 0 \end{cases} \\ &\psi = \psi + \pi \\ &\phi = \phi + \pi \end{aligned} \tag{A-26}$$

The aircraft's yaw and roll angles, ψ and ϕ, range from $-\pi$ to π. If these limits are exceeded, then the angles are corrected using the modulus function. The modulus function is defined here so that it will never return a negative value.

$$\begin{aligned} \psi &= \text{mod}(\psi + \pi, 2\pi) - \pi \\ \phi &= \text{mod}(\phi + \pi, 2\pi) - \pi \\ 0 &\le \text{mod}(a,b) < b \end{aligned} \tag{A-27}$$

Angle-of-Attack and Sideslip Angle

To calculate the angle-of-attack and sideslip angle of the aircraft, we must first calculate the air velocity of the aircraft in body coordinates. The air velocity is equal to the ground velocity of the aircraft minus the wind velocity in body coordinates.

$$\vec{V}_{T,b} = \vec{V}_{k,b} - \vec{V}_{atm,b} \tag{A-28}$$

$$u_T = u - u_{atm} \tag{A-29a}$$
$$v_T = v - v_{atm} \tag{A-29b}$$
$$w_T = w - w_{atm} \tag{A-29c}$$

The acceleration of the air relative to the aircraft is needed to find $\dot{\alpha}$ and $\dot{\beta}$, so the derivative of the Eq. (26) taken with respect to body coordinates. Since the acceleration of the wind is given relative to earth coordinates, the Coriolis Theorem must be applied.

$$\dot{\vec{V}}_{T,b} = \dot{\vec{V}}_{k,b} - \frac{d_e}{dt}\vec{V}_{atm,b} - \vec{\omega}_{e/b,b} \times \vec{V}_{atm,b} \tag{A-30}$$

$$\dot{u}_T = \dot{u} - \dot{u}_{atm} + q\,w_{atm} - r\,v_{atm} \tag{A-31a}$$
$$\dot{v}_T = \dot{v} - \dot{v}_{atm} + r\,u_{atm} - p\,w_{atm} \tag{A-31b}$$
$$\dot{w}_T = \dot{w} - \dot{w}_{atm} + p\,v_{atm} - q\,u_{atm} \tag{A-31c}$$

The angle-of-attack and sideslip angle can then be calculated from the components of the air velocity. The rates of change of the angle-of-attack and sideslip angle are calculated by taking the derivatives of the angles.

$$\alpha = \tan^{-1}\frac{w_T}{u_T} \tag{A-32}$$

$$\dot{\alpha} = \frac{u_T\,\dot{w}_T - \dot{u}_T\,w_T}{u_T^{\,2} + w_T^{\,2}} \tag{A-33}$$

$$\beta = \sin^{-1}\frac{v_T}{\sqrt{u_T^{\,2} + v_T^{\,2} + w_T^{\,2}}} \tag{A-34}$$

$$\dot{\beta} = \frac{1}{\cos\beta}\left(\frac{\dot{v}_T}{\sqrt{u_T^{\,2} + v_T^{\,2} + w_T^{\,2}}} - v_T\frac{u_T\,\dot{u}_T + v_T\dot{v}_T + w_T\,\dot{w}_T}{\left(u_T^{\,2} + v_T^{\,2} + w_T^{\,2}\right)^{\frac{3}{2}}}\right) \tag{A-35}$$

Aerodynamic Forces

The aerodynamic forces and moments on the aircraft are defined using the table lookup system described in section 4. The forces and moments are defined in air-path axes and then converted to body coordinates. The derivatives defined by the tables are non-dimensional and must be dimensionalized using the airspeed, wing planform area, span, and mean aerodynamic chord.

$$V_T = \sqrt{u_T^{\,2} + v_T^{\,2} + w_T^{\,2}} \tag{A-36}$$

$$F_{x\,aero,a} = -\tfrac{1}{2}\rho V_T^{\,2} S \sum_{Tables} C_D \tag{A-37a}$$

$$F_{y\,aero,a} = \tfrac{1}{2}\rho V_T^{\,2} S \sum_{Tables} C_y \tag{A-37b}$$

$$F_{z\,aero,a} = -\tfrac{1}{2}\rho V_T^{\,2} S \sum_{Tables} C_L \tag{A-37c}$$

$$L_{aero,a} = \tfrac{1}{2}\rho V_T^{\,2} S\, b \sum_{Tables} C_l \tag{A-38a}$$

$$M_{aero,a} = \tfrac{1}{2}\rho V_T^{\,2} S\, \bar{c} \sum_{Tables} C_m \tag{A-38b}$$

$$N_{aero,a} = \tfrac{1}{2}\rho V_T^{\,2} S\, b \sum_{Tables} C_n \tag{A-38c}$$

$$F_{aero,b} = C_a^b\, F_{aero,a} \tag{A-39}$$

$$LMN_{aero,b} = C_a^b\, LMN_{aero,a} \tag{A-40}$$

where,

$$C_a^b = \begin{bmatrix} \cos\alpha\cos\beta & -\cos\alpha\sin\beta & -\sin\alpha \\ \sin\beta & \cos\beta & 0 \\ \sin\alpha\cos\beta & -\sin\alpha\sin\beta & \cos\alpha \end{bmatrix} \tag{A-41}$$

Load Factor (g's)

The load factor is calculated by dividing the sum of the non-gravitational forces in the z direction by the weight of the aircraft.

$$n = -\frac{F_{z\,aero,b} + F_{z\,ext,b} + F_{z\,gear,b}}{m\,g} \tag{A-42}$$

Integration

All numerical integration for the equations of motion is performed using the Adams-Bashforth-Moulton method.

$$x = \int \dot{x}\,dt \quad \rightarrow \quad x_n = x_{n-1} + \tfrac{1}{2}\left(3\dot{x}_n - \dot{x}_{n-1}\right) dt \tag{A-43}$$

Appendix B – Aerodynamic Table Lookup System

Overview

All of the aerodynamic forces acting on an aircraft can be expressed in terms of three force coefficients and three moment coefficients. The force and moment coefficients are defined in the air-path coordinate system so that the forces they describe are always parallel or perpendicular to the freestream. The definitions of the six coefficients are given in Eqs (1-6).

$$C_L = \frac{-F_{z,a}}{\frac{1}{2}\rho V_T{}^2 S} \quad \text{(B-1)} \qquad C_l = \frac{L}{\frac{1}{2}\rho V_T{}^2 S b} \quad \text{(B-4)}$$

$$C_D = \frac{-F_{x,a}}{\frac{1}{2}\rho V_T{}^2 S} \quad \text{(B-2)} \qquad C_m = \frac{M}{\frac{1}{2}\rho V_T{}^2 S \bar{c}} \quad \text{(B-5)}$$

$$C_y = \frac{F_{y,a}}{\frac{1}{2}\rho V_T{}^2 S} \quad \text{(B-3)} \qquad C_n = \frac{N}{\frac{1}{2}\rho V_T{}^2 S b} \quad \text{(B-6)}$$

The table lookup system defines these coefficients as functions of the aerodynamic state of the aircraft and control surface deflections. The internal states listed in Table B.I are accessible to the table lookup system. Because the force and moment coefficients are non-dimensionalized using airspeed and density, true airspeed and altitude are usually not used to define force and moment coefficients, however the may be useful for describing Reynold's number or atmospheric effects on the aircraft. Predefined force and moment coefficients can be also be used to define other coefficients.

Table B.I – Internal Aerodynamic States for Table Lookup

M	Mach number
V_T	True airspeed, ft/s
α	Angle-of-attack, rad
β	Sideslip angle, rad
$\hat{\dot{\alpha}}$	Non-dimensional rate of angle-of-attack
$\hat{\dot{\beta}}$	Non-dimensional rate of sideslip angle
$\hat{p}$	Non-dimensional roll rate
$\hat{q}$	Non-dimensional pitch rate
$\hat{r}$	Non-dimensional yaw rate
h	Altitude, ft

In addition to the internal aerodynamic states of the aircraft, aircraft geometry can affect the aerodynamic force and moment coefficients. The most obvious geometric parameters of an aircraft are control surface deflections. For this reason, all geometric parameters are considered controls. Other geometric parameters could include high lift devices, air brakes, landing gear, stores, or the aircraft center of gravity. The center of gravity can be considered a geometric parameter since the geometry of the aircraft is defined relative to it. The table lookup system accounts for these geometric parameters by reading in a set of control inputs. The control inputs can then be used to define the aerodynamic force and moment coefficients.

Finally, a set of user-defined parameters can be defined to simplify the lookup of the force and moment coefficients. For example, the downwash angle at the tail can be defined using a table and then used to define the lift and pitching moment contributions of the horizontal stabilizer. These user-defined parameters are important, because they can be used to minimize the number of dimension in a table.

There are five methods for defining force and moment coefficients and user-defined parameters. These methods are described below. Typically, the table method is used most often, but the other methods can be used to define simple relationships or to combine user defined parameters.

Table B.II – Table Lookup Function Methods

Constant	Constant value without variation	$f = c$
Linear	Linear variation with one parameter	$f = a\,p$
Abs	Linear variation with the absolute value of one parameter	$f = a\,\lvert p \rvert$
Product	Product of multiple parameters and a constant gain	$f = a\prod p_n$
Table	Tabulated function of multiple parameters	$f = f(p_1,\ldots p_n)$

Multiple functions can be used to define a single force or moment coefficient or user defined parameters. In this case the resulting value is the sum of all of the individual functions.

To define the aerodynamics of a specific aircraft, the variation of the force and moment coefficients must be known. Some possible sources of data are listed below. Typically multiple sources of data are combined. Appendix G illustrates a component build up process based on the first three sources.

- Theory
- Linear stability derivatives
- DATCOM
- CFD
- Wind Tunnel
- Flight Test
- All of the above

Interpolation

The table lookup system used to define the aerodynamics of the aircraft is based on a method to lookup and interpolate between values in the table. The interpolation method is base on a weighted average of surrounding values in the table. Consider a two-dimensional table in which some function, f, is defined by parameters x and y. If the value of f is defined at specific values of x and y, and it is defined at every combination of the given x and y values, then the table can be represented as a orthographic grid similar to the grid shown in Figure B.1. Any desired value of x and y within the boundaries of the table will always fall within a rectangle defined by four known values.

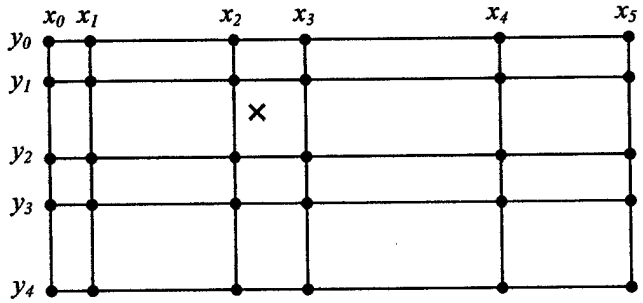


Figure B.1 – Orthographic Grid of Function Values and a Desired Value (×)

The value at the desired location is a combination of the four surrounding values. As the desired point approaches a known point, the value at the desired point should approach that known value. This relationship is illustrated in Figure B.2. To describe this relationship quantitatively, the influence of each known point is proportional to the area of the quadrant of the rectangle opposite to that point.

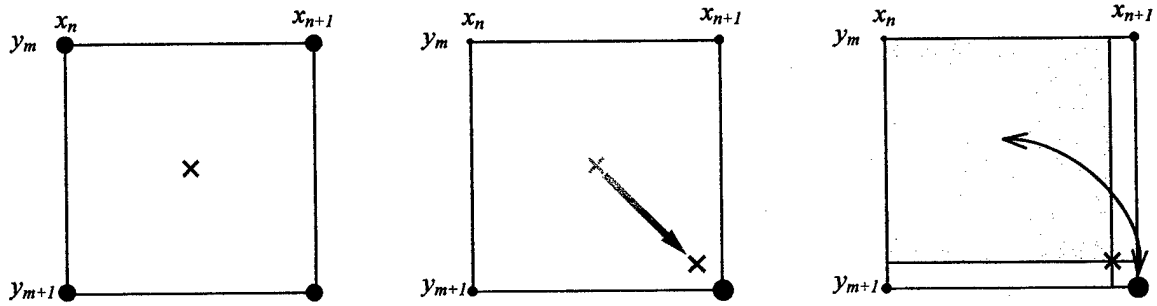


Figure B.2 – Relationship Between Corner Values and Relative Area of Opposite Quadrant

By summing the contribution of all four known values, we can get the approximate value of the function at the desired point.

$$f(x,y) = \sum_{\substack{i=n\to n+1 \\ j=m\to m+1}} f(x_i,y_j)\left|\frac{(x-x_{i'})(y-y_{j'})}{(x_i-x_{i'})(y_j-y_{j'})}\right| \tag{B-7}$$

$$i=n\to i'=n+1,\ \ i=n+1\to i'=n$$
$$j=m\to j'=m+1,\ \ j=m+1\to j'=m$$

Expanding the concept into three dimensions, a cubic grid can be defined by eight known values. Here the weight of each corner value is proportional to the relative volume of the opposite quadrant as shown in Figure B.3.

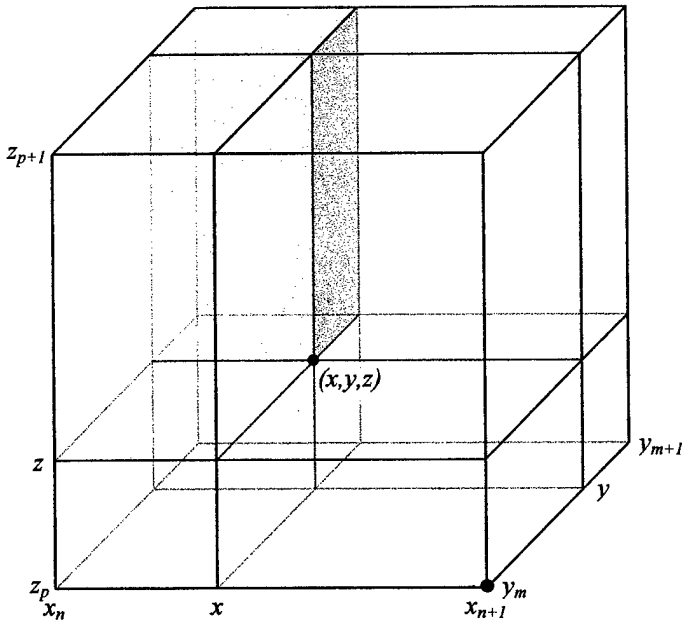


Figure B.3 – Three Dimensional Grid Cube with Opposite Volume Relationship

The resulting three-dimensional equation is,

$$f(x,y,z) = \sum_{\substack{i=n\to n+1 \\ j=m\to m+1 \\ k=p\to p+1}} f(x_i,y_j,z_k)\left|\frac{(x-x_{i'})(y-y_{j'})(z-z_{k'})}{(x_i-x_{i'})(y_j-y_{j'})(z_k-z_{k'})}\right| \tag{B-8}$$

$$i=n\to i'=n+1,\ \ i=n+1\to i'=n$$
$$j=m\to j'=m+1,\ \ j=m+1\to j'=m$$
$$k=p\to k'=p+1,\ \ k=p+1\to k'=p$$

To apply these equations, the index values n, m, and p must be known. These values are defined as the nearest known value in the list that is smaller than the desired. To find these index values, they are iterated until the criteria are met. Since the parameters used for the lookup typically do not change much from one time step to the next, the index iteration process can be made more efficient by using the previous value as a starting point. The iteration process is detailed below.

$$\begin{array}{ll} while\left(n>0\ and\ x<x_n\right) & while\left(n<n_{max}-2\ and\ x\geq x_n+1\right) \\ \quad n=n-1 & \quad n=n+1 \end{array} \tag{B-9}$$

If the desired value of x or y fall off of the table then the two options are given. The data in the table can be extrapolated in that dimension, or the closest known value can be used. These two options are illustrated in Figure B.4. To extrapolate the data in the table, the same equations can be used without modification. To use the nearest value in the table, that value is used in place of the desired value. The table format allows the user to specify whether each dimension of the table should be extrapolated.

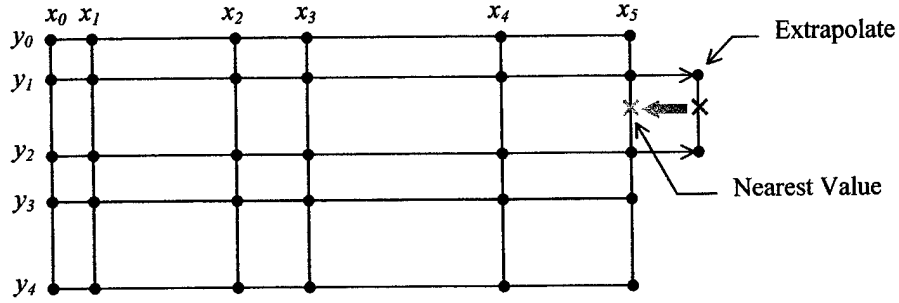


Figure B.4 – Extrapolation or Nearest Value Options for Off-Table Values

Table Formatting

To apply this interpolation system to a table lookup system, each table must meet with the following criteria:

- A list of parameter values must be given for each table dimension.
- Each list must be monotonically increasing.
- The table value must be defined at every combination of parameter values.

The format used to define lookup tables is based on the criteria above. The first line specifies the variable to be defined by the table, second line contains the word "Table", and the third line contains the number of dimensions in the table. For each dimension, two lines are given. The first specifies the parameter for that dimension, the number of values given, and a Boolean value (1 or 0) that defines whether that dimension will be extrapolated. The second line contains the list of parameter values in ascending order. After all dimensions have been defined, the table values are given. The values for the first dimension are listed across the file to the right. The values for the second dimension are listed down the file. The values for the third dimension are expressed a multiple tables listed down the file. Any additional characters in a line are ignored, so comments can be added after each line. Blank lines are ignored, but comments cannot be place in these lines. An example table is shown on the next page in Figure B.5.

```
CL                    // Lift Coefficient
Table
3                     // Three Dimensions

Alpha 7 0             // Angle-of-Attack (7 values, not extrapolated)
-0.6 -0.4 -0.2 0.0 0.2 0.4 0.6

Mach 5 0              // Mach Number (5 values, not extrapolated)
0.0 0.5 1.0 1.5 2.0

Control 0 3 1// Elevator Deflection (3 values, extrapolated)
-1.0 0.0 1.0
```

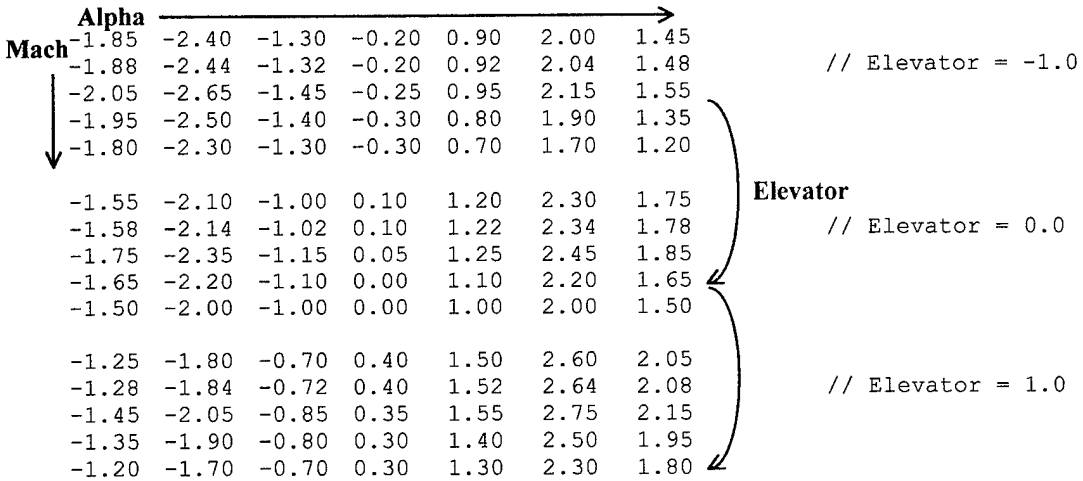


Figure B.5 – Example Table

Appendix C – Landing Gear

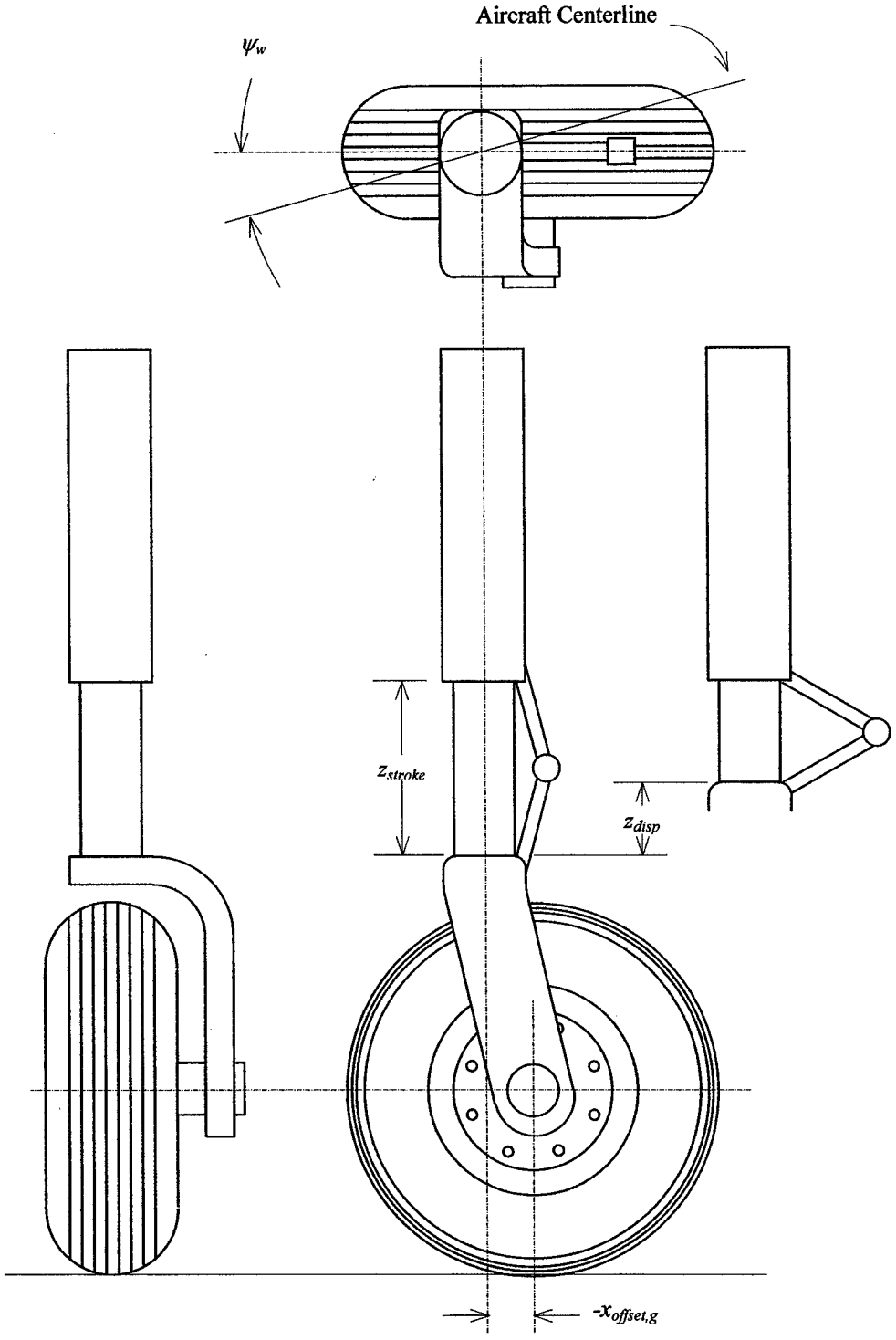


Figure C.1 – Landing Gear Component

Shock Absorber Compression Stroke

The landing gear model simulates an oleo-pneumatic shock absorber typically used in aircraft landing gear. The shock absorber compresses chamber of air to produce a spring force. Another fluid, usually oil, is also forced through an orifice to crate a damping force. The total stroke of the shock absorber is the difference between the uncompressed and fully compressed lengths of the absorber. A simplified schematic of a typical oleo-pneumatic shock absorber is shown in Figure C.2.

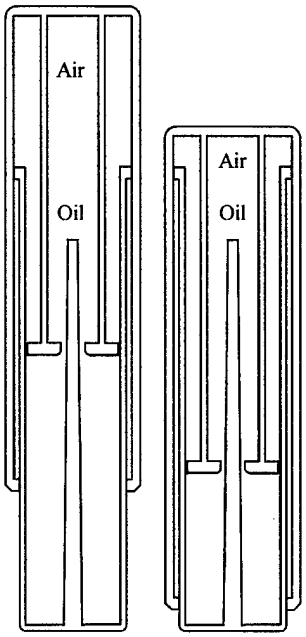


Figure C.2 – Typical Oleo-Pneumatic Shock Absorber

The equation for the compression stroke of the oleo-pneumatic shock absorber is derived starting with the equation for a polytropic compression. Where the exponent γ is determined by the heat transfer that occurs during the compression. A γ value of 1 represents an isothermal compression, and a value of 1.4 represents an adiabatic compression for air. The actual compression stroke during landing should be somewhere in-between.

$$PV^{\gamma} = constant \tag{C-1}$$

The volume of air in the shock absorber is equal to the cross-sectional area of the piston times the length of the air chamber. The air chamber length is equal to the difference between total stroke of the shock absorber and the displacement.

$$V = \tfrac{\pi}{4} d^2 \left(z_{stroke} - z_{disp} \right) \tag{C-2}$$

By substituting Eq. (2) into Eq. (1), we get,

$$P_{int} \left(\tfrac{\pi}{4} d^2 \left(z_{stroke} - z_{disp} \right) \right)^{\gamma} = constant \tag{C-3}$$

The total uncompressed internal pressure is given for a displacement of zero. This pressure and volume can be used to define the constant in Eq. (3)

$$P_{int,u}\left(\tfrac{\pi}{4}d^2 z_{stroke}\right)^{\gamma} = P_{int}\left(\tfrac{\pi}{4}d^2\left(z_{stroke}-z_{disp}\right)\right)^{\gamma} \tag{C-4}$$

Solving Eq. (4) for P_{int},

$$P_{int} = \frac{P_{int,u}\, z_{stroke}^{\gamma}}{\left(z_{stroke}-z_{disp}\right)^{\gamma}} = \frac{P_{int,u}}{1-\left(\dfrac{z_{disp}}{z_{stroke}}\right)^{\gamma}} \tag{C-5}$$

The total force produced by the shock absorber is equal to the piston area times the difference between the internal and external pressures. Since the internal pressure is greater than the external pressure, the result is negative meaning that the force is upward. Negative displacements indicate that the wheel is off of the ground so the strut produces no force.

$$F_{z\,gear,g} = \begin{cases} \tfrac{\pi}{4}d^2\left(P_{ext} - \dfrac{P_{int,u}}{1-\left(\dfrac{z_{disp}}{z_{stroke}}\right)^{\gamma}}\right) & z_{disp} \geq 0 \\ 0 & z_{disp} < 0 \end{cases} \tag{C-6}$$

The difference between the uncompressed internal pressure of the shock absorber and the external pressure creates a discontinuity in the force at a zero displacement. This makes it difficult to solve for the force on the shock absorber from a given displacement. To remove this discontinuity, the elastic effects of the rubber tires on the aircraft must be modeled. Since the spring rate of the tire is known, the displacement of the tire when the shock absorber begins to compress can be calculated. This length is added to the beginning of the shock absorber compression to create a linear slope up to the initial shock absorber force. The resulting compression stroke is shown for the F-4 in Figure C.3.

$$F_{z\,gear,g} = \begin{cases} \tfrac{\pi}{4}d^2\left(P_{ext} - \dfrac{P_{int,u}}{1-\left(\dfrac{z_{disp}}{z_{stroke}}\right)^{\gamma}}\right) & z_{disp} > 0 \\ -k_{tire}\left(z_{tire}+z_{disp}\right) & -z_{tire} < z_{disp} \leq 0 \\ 0 & z_{disp} \leq -z_{tire} \end{cases} \tag{C-7}$$

$$z_{tire} = \frac{\tfrac{\pi}{4}d^2\left(P_{int,u}-P_{ext}\right)}{k_{tire}} \tag{C-8}$$

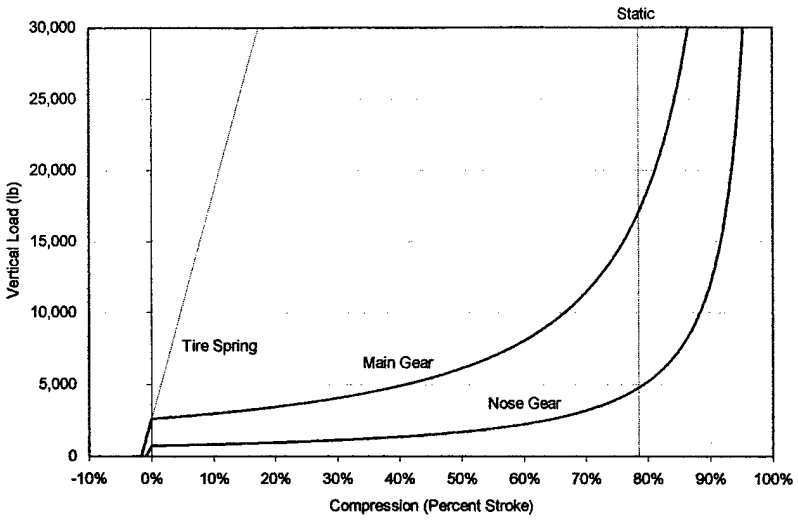


Figure C.3 – Pneumatic Compression Stroke for F-4 Main Gear and Nose Gear

Another consideration for the shock absorber is its energy absorption characteristics. When an aircraft lands, the shock absorbers must absorb both the vertical kinetic energy of the aircraft and some portion of the potential energy. The remainder of the potential energy is taken up by lift. The total energy absorbed by the shock absorber is the integral of the vertical force over the compression stroke. An ideal shock absorber would exert a constant force over its compression stroke, because this would maximize the energy absorption for a given maximum load. In an oleo-pneumatic shock absorber, hydraulic damping is used to improve the energy absorption characteristics. Hydraulic damping operates by forcing oil or some other fluid through an orifice, creating a force that is a function of the rate of compression. A metering pin is used to change the size of the orifice depending on the compression of the strut. The effect of damping in the shock absorber is modeled using a damping constant that varies linearly with shock absorber compression. An initial uncompressed damping constant and a final compressed damping constant are used to define this variation. The damping force is given by the equation below.

$$F_{z\,gear,g} = \left(b_{z\,initial}\left(1 - \frac{z_{disp}}{z_{stroke}}\right) + b_{z\,final}\frac{z_{disp}}{z_{stroke}} \right) V_{z\,wheel,g} \qquad \text{(C-9)}$$

Because the damping force is a function of vertical velocity, the total shock absorber compression stroke depends on the sink rate of the aircraft on touchdown. Usually a worst-case sink rate is defined and used to optimize the compression stroke. The total compression stroke from simulated drop tests of the F-4 main gear is shown in Figure C.4 with different initial sink rates.

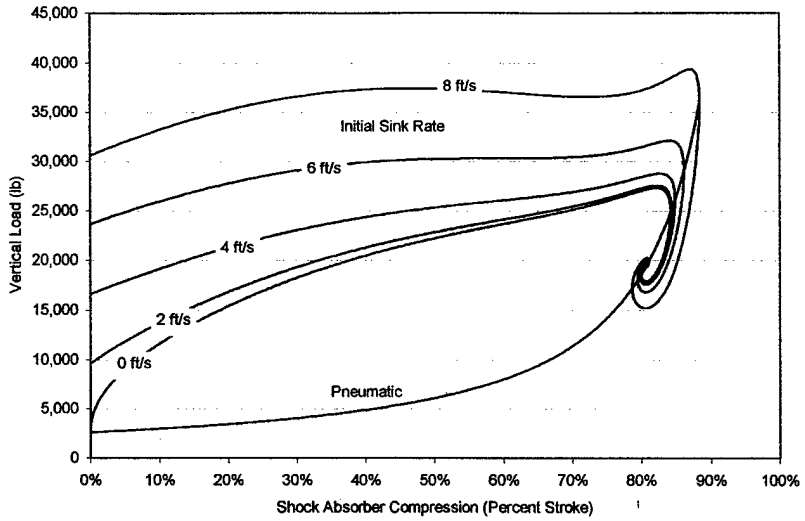


Figure C.4 – Total Shock Absorber Force from Simulated Drop Test of F-4 Main Gear

Landing Gear Calculation Process

Coordinate Systems & Transformations

The first step in calculating the forces on the aircraft due to the landing gear is to define the coordinate systems and coordinate transformation matrices used in the model. The four coordinate systems used in the landing gear model are earth coordinates, body coordinates, gear coordinates, and wheel coordinates. Earth coordinates and body coordinate systems are the same as those used in the six-degrees-of-freedom model. The coordinate transformation matrices from earth to body coordinates and vice-versa are given below.

$$C_e^b = \begin{bmatrix} \cos\theta\cos\psi & \cos\theta\sin\psi & -\sin\theta \\ \sin\phi\sin\theta\cos\psi - \cos\phi\sin\psi & \sin\phi\sin\theta\sin\psi + \cos\phi\cos\psi & \sin\phi\cos\theta \\ \cos\phi\sin\theta\cos\psi + \sin\phi\sin\psi & \cos\phi\sin\theta\sin\psi - \sin\phi\cos\psi & \cos\phi\cos\theta \end{bmatrix} \qquad C_b^e = C_e^{bT} \qquad \text{(C-10)}$$

The gear coordinate system is the same as the body coordinate system except that the x and y axes follow the wheel when it rotates relative to the aircraft for steering. The coordinate transformation from body to gear coordinates thus only has a yaw component equal to the wheel angle.

$$C_b^g = \begin{bmatrix} \cos\psi_w & \sin\psi_w & 0 \\ -\sin\psi_w & \cos\psi_w & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad C_g^b = C_b^{gT} \qquad \text{(C-11)}$$

The wheel coordinate system is parallel and normal to the ground with the x-axis aligned in the direction of the wheel. To allow for the future inclusion of a runway surface model, the gradient angles of the ground δ_x and δ_y are taken into account. The elevation of the earth is defined as a function of the x and y earth position. A runway model would be needed to define this value of this function. The earth gradient angles, $\delta_{x,e}$ and $\delta_{y,e}$, are defined as the angle of the ground around the earth x and y axes. The transformation matrix from wheel to gear coordinates applies these angles converted into gear coordinates as pitch and roll angles.

$$C_w^g = \begin{bmatrix} \cos(\theta-\delta_{y,g}) & 0 & -\sin(\theta-\delta_{y,g}) \\ \sin(\phi-\delta_{x,g})\sin(\theta-\delta_{y,g}) & \cos(\phi-\delta_{x,g}) & \sin(\phi-\delta_{x,g})\cos(\theta-\delta_{y,g}) \\ \cos(\phi-\delta_{x,g})\sin(\theta-\delta_{y,g}) & -\sin(\phi-\delta_{x,g}) & \cos(\phi-\delta_{x,g})\cos(\theta-\delta_{y,g}) \end{bmatrix} \quad C_g^w = C_w^{g^T} \tag{C-12}$$

$$h_{ground} = f(x_e, y_e) \tag{C-15}$$

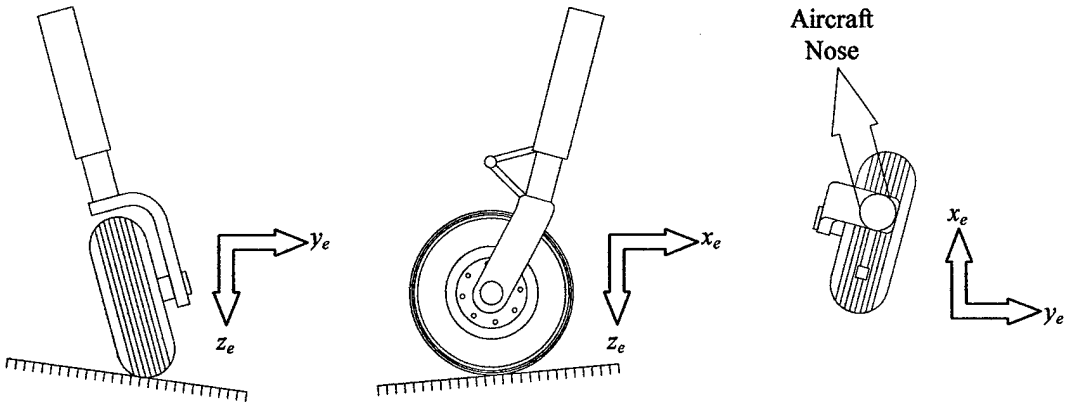
$$\delta_{x,g} = \delta_{x,e}\cos(\psi+\psi_w) - \delta_{y,e}\sin(\psi+\psi_w) \tag{C-13}$$

$$\delta_{x,e} = \tan^{-1}\frac{\partial h_{ground}}{\partial y_e} \tag{C-16}$$

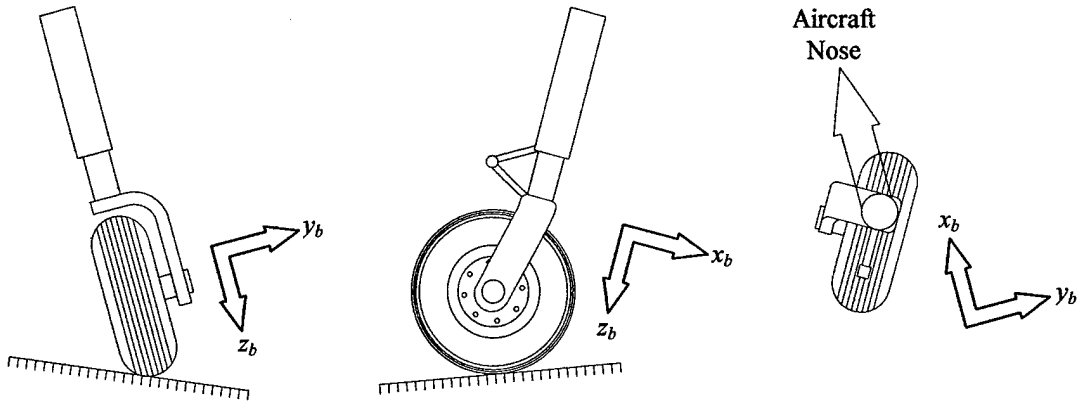
$$\delta_{y,g} = \delta_{x,e}\sin(\psi+\psi_w) + \delta_{y,e}\cos(\psi+\psi_w) \tag{C-14}$$

$$\delta_{y,e} = \tan^{-1}\frac{\partial h_{ground}}{\partial x_e} \tag{C-17}$$

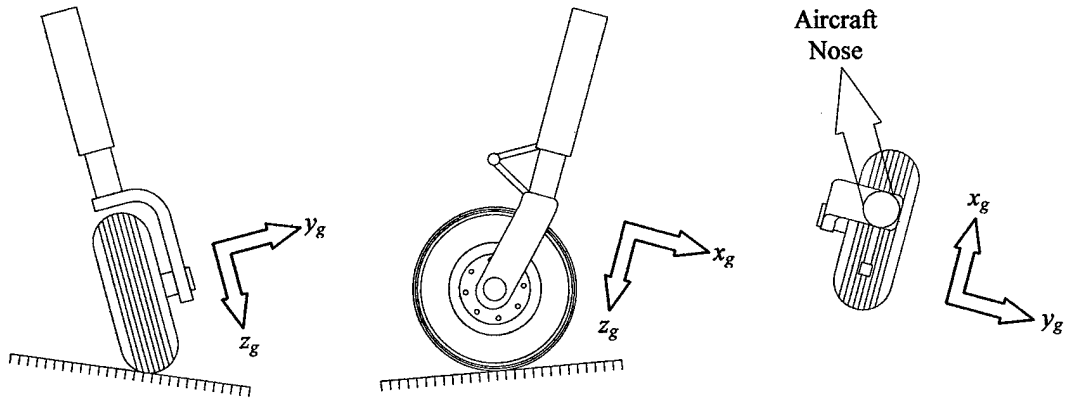
The four coordinate systems are illustrated in Figure C.5 in relation to a typical landing gear component.



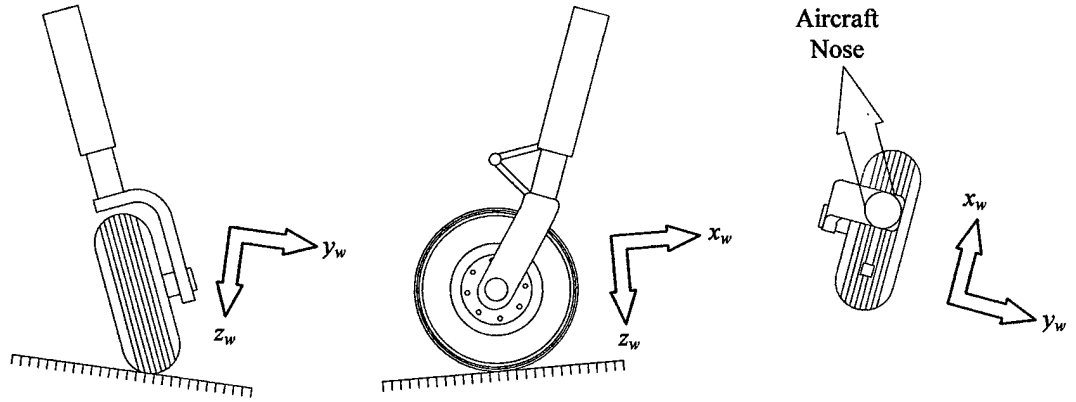
Earth Coordinates



Body Coordinates



Gear Coordinates



Wheel Coordinates

Figure C.5 – Coordinate Systems Used in Landing Gear Model

To calculate the forces on each landing gear component, we first must know the position of the wheel when the shock absorber is uncompressed. The position of the wheel in earth coordinates is equal to the position of the aircraft plus the position of the landing gear relative to the aircraft converted into earth coordinates plus the offset of the wheel in gear coordinates converted into earth coordinates.

$$\vec{p}_{u\,wheel/earth,e} = \vec{p}_{aircraft/earth,e} + \vec{p}_{gear/aircraft,e} + \vec{p}_{wheel/gear,e} \tag{C-18}$$

$$\vec{p}_{aircraft/earth,e} = \begin{bmatrix} x_{aircraft,e} \\ y_{aircraft,e} \\ z_{aircraft,e} \end{bmatrix} \tag{C-19}$$

$$\vec{p}_{gear/aircraft,e} = C_b^e \begin{bmatrix} x_{gear,b} \\ y_{gear,b} \\ z_{gear,b} \end{bmatrix} \tag{C-20}$$

$$\vec{p}_{wheel/gear,e} = C_b^e C_g^b \begin{bmatrix} x_{offset,g} \\ 0 \\ 0 \end{bmatrix} \tag{C-21}$$

The actual position of the wheel is the position of the wheel on the ground. Initially, the wheel is assumed to remain at the same position as the previous time step. This position may be corrected if the wheel experiences enough forces to roll or slide across the ground.

$$\vec{p}_{wheel/earth,e} = \begin{bmatrix} x'_{wheel,e} \\ y'_{wheel,e} \\ -h_{ground} \end{bmatrix} \tag{C-22}$$

The displacement of the landing gear can be then calculated as the difference in the actual position of the wheel and the uncompressed position of the wheel. The displacement is converted into body coordinates and gear coordinates.

$$\Delta\vec{p}_{wheel,e} = \vec{p}_{wheel/earth,e} - \vec{p}_{u\,wheel/earth,e} \tag{C-23}$$

$$\Delta\vec{p}_{wheel,b} = C_e^b \, \Delta\vec{p}_{wheel,e} \tag{C-24}$$

$$\Delta\vec{p}_{wheel,g} = C_b^g \, \Delta\vec{p}_{wheel,b} \tag{C-25}$$

The velocity of the wheel can be calculated in body coordinates using the velocity of the aircraft in body coordinates minus the cross product of the body rates of the aircraft and the position of the wheel. The velocity is converted into both gear coordinates and wheel coordinates.

$$\vec{V}_{wheel/earth,b} = \vec{V}_{aircraft/earth,b} - \vec{\omega}_{b/e,b} \times \vec{p}_{wheel/aircraft,b} \tag{C-26}$$

$$\vec{V}_{wheel/earth,b} = \begin{bmatrix} u \\ v \\ w \end{bmatrix} + \begin{bmatrix} q\,z_{wheel/aircraft,b} - r\,y_{wheel/aircraft,b} \\ r\,x_{wheel/aircraft,b} - p\,z_{wheel/aircraft,b} \\ p\,y_{wheel/aircraft,b} - q\,x_{wheel/aircraft,b} \end{bmatrix} \tag{C-27}$$

$$\vec{V}_{wheel/earth,g} = C_b^g \, \vec{V}_{wheel/earth,b} \tag{C-28}$$

$$\vec{V}_{wheel/earth,w} = C_g^w \, \vec{V}_{wheel/earth,w} \tag{C-29}$$

The displacement of the shock absorber is the z component of the displacement of the landing gear. To avoid problems if the shock absorber is compressed past the stroke distance, the displacement is limited to the 99.9% of the stroke.

$$\begin{aligned} z_{disp} &= -\Delta z_{wheel,g} \\ z_{disp} > 0.999\, z_{stroke} &\rightarrow z_{disp} = 0.999\, z_{stroke} \end{aligned} \tag{C-30}$$

Now the force from the shock absorber can be calculated from Eqs. (7-8) derived above.

$$z_{tire} = \frac{\frac{\pi}{4} d^2 \left(P_{int,u} - P_{ext} \right)}{k_{tire}}$$

$$F_{z\,gear,g} = \begin{cases} \frac{\pi}{4} d^2 \left(P_{ext} - \dfrac{P_{int,u}}{1 - \left(\dfrac{z_{disp}}{z_{stroke}} \right)^{\gamma}} \right) & z_{disp} > 0 \\ -k_{tire} \left(z_{tire} + z_{disp} \right) & -z_{tire} < z_{disp} \le 0 \\ 0 & z_{disp} \le -z_{tire} \end{cases}$$

The forces from the linear spring and damping forces can be added to the shock absorber force using Eq. (9) to define the damping force.

$$F_{z\,gear,g} = F_{z\,gear,g} + k_z\, \Delta z_{wheel,g} - \left(b_{z\,initial} \left(1 - \frac{z_{disp}}{z_{stroke}} \right) + b_{z\,final} \frac{z_{disp}}{z_{stroke}} \right) V_{z\,wheel,g} \tag{C-31}$$

The x and y axes are modeled as simple linear damped springs. This allows for-and-aft forces and side forces to be absorbed by the landing gear.

$$F_{x\,gear,g} = k_x\, \Delta x_{wheel,g} - b_x\, V_{x\,wheel,g} \tag{C-32}$$

$$F_{y\,gear,g} = k_y\, \Delta y_{wheel,g} - b_y\, V_{y\,wheel,g} \tag{C-33}$$

The maximum frictional forces are calculated using the coefficients of friction and the normal force on the wheel. If the wheel is already sliding, the kinetic coefficient of friction is used, otherwise the static coefficient is used.

$$F_{max} = \begin{cases} \mu_s\, F_{z\,gear,w} & \text{Not Sliding} \\ \mu_k\, F_{z\,gear,w} & \text{Sliding} \end{cases} \tag{C-34}$$

The forces on the landing gear can now be corrected for the maximum friction and braking forces.

$$\vec{F}_{gear,w} = C_g^w\, \vec{F}_{gear,g} \tag{C-35}$$

$$\vec{V}_{wheel,w} = C_g^w\, \vec{V}_{wheel,g} \tag{C-36}$$

Both braking and friction forces are applied in wheel coordinates, and the direction of the force is always opposite of the direction of motion of the wheel.

$$\left|F_{x\,gear,w}\right| > F_{brake} \rightarrow F_{x\,gear,w} = -F_{brake}\,\mathrm{sgn}\left(V_{x\,wheel,w}\right) \tag{C-37}$$

$$\left|F_{x\,gear,w}\right| > F_{max} \rightarrow F_{x\,gear,w} = -F_{max}\,\mathrm{sgn}\left(V_{x\,wheel,w}\right) \tag{C-38}$$

$$\left|F_{y\,gear,w}\right| > F_{max} \rightarrow F_{y\,gear,w} = -F_{max}\,\mathrm{sgn}\left(V_{y\,wheel,w}\right) \tag{C-39}$$

The forces are then converted back into gear and body coordinates. The body coordinate forces are the gear forces used in the equations of motion.

$$\vec{F}_{gear,g} = C_w^g\,\vec{F}_{gear,w} \tag{C-40}$$

$$\vec{F}_{gear,b} = C_g^b\,\vec{F}_{gear,g} \tag{C-41}$$

If the braking force or friction force is exceeded, then the forces on the wheel are no longer in equilibrium. In this case, the wheel will begin to roll or slide across the ground, and its displacements must be re-calculated based on the new force on the landing gear neglecting the damping force. The new displacement is converted into body and earth coordinates and the new position of the wheel relative to earth is calculated.

$$\Delta x_{wheel,g} = F_{x\,gear,g} / k_x \tag{C-42}$$

$$\Delta y_{wheel,g} = F_{y\,gear,g} / k_y \tag{C-43}$$

$$\Delta\vec{p}_{wheel,b} = C_g^b\,\Delta\vec{p}_{wheel,g} \tag{C-44}$$

$$\Delta\vec{p}_{wheel,e} = C_g^e\,\Delta\vec{p}_{wheel,g} \tag{C-45}$$

$$x'_{wheel,e} = x_{wheel/earth,e} + \Delta x_{wheel,e} \tag{C-46}$$

$$y'_{wheel,e} = y_{wheel/earth,e} + \Delta y_{wheel,e} \tag{C-47}$$

To allow for both nose-wheel steering and differential braking, the rotation of the landing gear component around its own z-axis is modeled in two ways. In steering mode, the angle of the wheel is controlled directly by an input to the model. In differential braking mode, the wheel is allowed to rotate as a torsional damped spring system. If the nose wheel were not allowed to rotate, a side force from the nose wheel would counteract the yaw moment created by differential braking. The side force and the offset of the wheel create a moment that rotates the wheel in the direction of the braking moment and allows the aircraft to rotate on the ground. The moment caused by the tire twisting on the ground is calculated as the torsional spring constant of the tire times the angle that the tire has twisted on the ground. This moment is calculated in wheel coordinates so that it is normal to the ground.

$$N_{t,w} = -k_{t\,tire}\left(\psi + \psi_w - \psi'_w\right) \tag{C-48}$$

The maximum frictional moment from the wheel twisting on the ground is calculated with the following formula adapted from *Aircraft Landing* Gear *Design.* This moment is only applied when the wheel is not rolling or sliding in the x or y-axis. The frictional moment when the wheel is rolling, or tire scrubbing, is not modeled.

$$N_{max} = \begin{cases} \mu_s \left(0.002428\sqrt{A_{contact}} - 0.15\right) F_{z\,gear,w} & Not\ Rotating \\ \mu_k \left(0.002428\sqrt{A_{contact}} - 0.15\right) F_{z\,gear,w} & Rotating \end{cases} \tag{C-49}$$

If the tire moment exceeds the maximum frictional moment, then the moment is corrected and the angle of the tire on the ground is re-calculated.

$$\left|N_{t,w}\right| > N_{max} \rightarrow N_{t,w} = -N_{max}\,\mathrm{sgn}\left(\omega_w\right) \tag{C-50}$$

$$\psi_w' = \psi + \psi_w + \frac{N_{t,w}}{k_{t\,tire}} \tag{C-51}$$

The moment is then converted into gear and body coordinates for use in the torsional damped spring system.

$$\vec{T}_{t,g} = C_w^g \begin{bmatrix} 0 \\ 0 \\ N_{t,w} \end{bmatrix} \tag{C-52}$$

$$\vec{T}_{t,b} = C_g^b \vec{T}_{t,g} \tag{C-53}$$

The angular acceleration of the landing gear is calculated from the total moment and the moment of inertia of the landing gear component. The angular acceleration is then integrated to get the angular rate, which is integrated to get the angle.

$$\alpha_w = \frac{F_{y\,gear,g}\,x_{offset,g} - k_t\,\psi_w - b_t\,\omega_w + T_{zt,g}}{I_{zt}} \tag{C-54}$$

$$\omega_w = \omega_w + \alpha_w\,dt \tag{C-55}$$

$$\psi_w = \psi_w + \omega_w\,dt \tag{C-56}$$

The moment contribution of the landing gear component in body coordinates is the cross product of the force on the wheel and the position of the wheel plus the frictional moment from the tire.

$$\vec{T}_{gear,b} = \vec{F}_{gear,b} \times \vec{p}_{wheel/aircraft,b} + \vec{T}_{t,b} \tag{C-57}$$

$$L_{gear} = F_{z\,gear,b}\,y_{wheel/aircraft,b} - F_{y\,gear,b}\,z_{wheel/aircraft,b} + T_{xt,b} \tag{C-58}$$

$$M_{gear} = F_{x\,gear,b}\,z_{wheel/aircraft,b} - F_{z\,gear,b}\,x_{wheel/aircraft,b} + T_{yt,b} \tag{C-59}$$

$$N_{gear} = F_{y\,gear,b}\,x_{wheel/aircraft,b} - F_{x\,gear,b}\,y_{wheel/aircraft,b} + T_{zt,b} \tag{C-60}$$

Appendix D – Crash Detection System

A crash detection system is an important addition to the landing gear model. The crash detector prevents the landing gear model from encountering impossible situations where the aircraft is underground. The crash detection system works by defining a series of points around the aircraft at critical locations. The positions of these points are converted into earth coordinates and compared with the ground elevation. If any point goes below ground, the aircraft is assumed to have crashed, and the crash port on the block will output 1.

First, the body coordinates to earth coordinate transformation matrix is calculated.

$$C_e^b = \begin{bmatrix} \cos\theta\cos\psi & \cos\theta\sin\psi & -\sin\theta \\ \sin\phi\sin\theta\cos\psi - \cos\phi\sin\psi & \sin\phi\sin\theta\sin\psi + \cos\phi\cos\psi & \sin\phi\cos\theta \\ \cos\phi\sin\theta\cos\psi + \sin\phi\sin\psi & \cos\phi\sin\theta\sin\psi - \sin\phi\cos\psi & \cos\phi\cos\theta \end{bmatrix} \qquad C_b^e = C_e^{bT} \qquad \text{(D-1)}$$

For each crash point, the position of the point in body coordinates is converted into earth coordinates and added to the position of the aircraft to get the position of the point relative to earth.

$$\vec{p}_{point/earth,e} = \vec{p}_{aircraft/earth,e} + \vec{p}_{point/aircraft,e} \qquad \text{(D-2)}$$

$$\vec{p}_{aircraft/earth,e} = \begin{bmatrix} x_{aircraft,e} \\ y_{aircraft,e} \\ z_{aircraft,e} \end{bmatrix} \qquad \text{(D-3)}$$

$$\vec{p}_{point/aircraft,e} = C_b^e \begin{bmatrix} x_{point,b} \\ y_{point,b} \\ z_{point,b} \end{bmatrix} \qquad \text{(D-4)}$$

The elevation of the ground is then calculated a function of the x and y earth coordinates of the crash point using a runway or terrain model.

$$h_{ground} = f\left(x_{point,e}, y_{point,e}\right) \qquad \text{(D-5)}$$

If the altitude of the crash point is below ground, then the airplane has crashed.

$$\begin{gathered} if\left(-z_{point,e} < h_{ground}\right) \\ crash = true \end{gathered} \qquad \text{(D-6)}$$

Appendix E – Simulink Block

The 6DOF, landing gear model, crash detector, and future ground model are implemented as a Simulink S-function block. The block takes five parameters as listed in Table E.1. The parameters are input in the parameters text box in the S-function dialog window.

Table E.I – Simulink Block Parameters

1. Aerodynamics File Name (String)
2. Landing Gear File Name (String)
3. Crash Detection File Name (String)
4. Runway Model File (String)
5. Inner Loop Multiplier (Integer)

Example: 'F-4Table.txt','F-4Gear.txt','F-4Crash.txt','Runway.txt',100

The input and output ports are listed in Table E.2 and Table E.3 along with the width of the ports. The width of a port is the number of signals that that port carries. A "mux" block can be used to combine individual signals into one input, and a "demux" block can be used to separate an output into its individual signals.

Table E.II – Simulink Block Input Ports

1.	Aircraft Weight mg, lb	1
2.	Controls	$n_{controls}$
3.	Air Density ρ, slugs/ft^3	1
4.	External Body Forces F_x, F_y, F_z, lb	3
5.	External Body Moments L, M, N, ft lb	3
6.	Atmospheric Earth Velocity $V_{x\,atm}$, $V_{y\,atm}$, $V_{z\,atm}$, ft/s	3
7.	Atmospheric Earth Acceleration $\dot{V}_{x\,atm}$, $\dot{V}_{y\,atm}$, $\dot{V}_{z\,atm}$, ft/s	3
8.	Initial Earth Position x, y, z, ft	3
9.	Initial Euler Angles, ψ, θ, ϕ, rad	3
10.	Initial Body Velocity u, v, w, ft/s	3
11.	Initial Body Rates p, q, r, rad/s	3
12.	Speed of Sound, ft/s	1
13.	Coefficients of Friction, μ_s, μ_k	2
14.	Brake Force, lb	n_{gear}
15.	Wheel Angle, rad	n_{gear}
16.	Steering Engaged (1 – yes, 0 – no)	n_{gear}
17.	Gear Down (1 – yes, 0 – no)	1
18.	Reset (1 – yes, 0 – no)	1

Table E.III – Simulink Block Output Ports

1.	Earth Position x, y, z, ft	3
2.	Euler Angles, ψ, θ, ϕ, rad	3
3.	Body Velocity u, v, w, ft/s	3
4.	Body Airspeed Velocity u_T, v_T, w_T, ft/s	3
5.	Body Rates p, q, r, rad/s	3
6.	Angle-of-Attack & Sideslip Angle, α, β, rad	2
7.	Load Factor (g's)	1
8.	Force & Moment Derivatives, C_L, C_D, C_y, C_l, C_m, C_n	6
9.	User-Defined Parameters	n_{param}
10.	Landing Gear Body Forces, $F_{x\,gear}$, $F_{y\,gear}$, $F_{z\,gear}$, lb	3
11.	Landing Gear Body Moments, L_{gear}, M_{gear}, N_{gear}, ft lb	3
12.	Maximum x Gear Force, lb	n_{gear}
13.	Maximum y Gear Force, lb	n_{gear}
14.	Maximum z Gear Force, lb	n_{gear}
15.	Wheel Angle, rad	n_{gear}
16.	On Ground (1 – yes, 0 – no)	1
17	Gear Fail (1 – yes, 0 – no)	1
18.	Crash (1 – yes, 0 – no)	1

Appendix F – Input File Formats

Aerodynamics File

The aerodynamics file is constructed as a header, which defines the constant parameters of the aircraft, and a series of functions that define the aerodynamic force and moment coefficients or user-defined parameters. Each function defines a single value using one of five function methods: Constant, Linear, Abs, Product, or Table. Multiple functions can be used to define a single value. The resulting value is the sum of all of the functions that define that value. The values that can be defined by functions are the aerodynamic force and moment coefficients and user-defined parameters. Each function uses parameters to define the function value. These parameters include: internal aerodynamic states, pre-defined force or moment coefficients, pre-defined custom parameters, and controls. The functions are evaluated in order from the beginning of the file to the end of the file, so if a parameter defined by other functions is used; the value defined only by previous functions is used.

The size limitations shown in Table F.1 are defined in the program. These values can be changed in the source code; however, they are limited by memory allocation. If it were necessary to increase these values (or add additional dimensions), the arrays used to store the data could be declared dynamically as the file is being read in to make more efficient use of memory.

Table F.I – Aerodynamics File Size Limits

Maximum number of functions	120
Maximum number of dimensions per table	3
Maximum number of values in each parameter list	85
Maximum number of controls	100
Maximum number of user-defined parameters	100

Table F.II – Aerodynamics File Format

Header

Header	
Description: File header that defines the constant parameters of the aircraft	
Format:	Sample: F-4

Wing Planform Area (ft^2)
Wing Span (ft)
Mean Aerodynamic Chord (ft)
x-Axis Radius of Gyration (ft) $I_x = mr_x^2$
y-Axis Radius of Gyration (ft) $I_y = mr_y^2$
z-Axis Radius of Gyration (ft) $I_z = mr_z^2$
xz-Plane Radius of Gyration (ft) $I_{xz} = m\,\mathrm{sgn}(r_{xz})r_{xz}^2$
Number of Controls
Number of Parameters
	`530.0 // Area (ft^2)`
	`38.7 // Span (ft)`
	`16.0 // Chord (ft)`
	`4.5411 // rx (ft)`
	`10.036 // ry (ft)`
	`10.751 // rz (ft)`
	`1.348 // rxz (ft)`
	`3 // Number of Controls`
	`0 // Number of Parameters`

Table F.II – Aerodynamics File Format Cont.

Values

Function Values
Description: Value defined by a function
Legal Values (case sensitive): Force or Moment Coefficients: `CL` – Lift Coefficient `CD` – Drag Coefficient `Cy` – Side Force Coefficient `Cl` – Roll Moment Coefficient `Cm` – Pitch Moment Coefficient `Cn` – Yaw Moment Coefficient Parameters: `Parameter` *n* – Parameter Number *n*

Parameter
Description: Used to define function value
Legal Values (case sensitive): Internal Values: `Alpha` – Angle-of-Attack (rad) `Beta` – Sideslip Angle (rad) `Mach` – Mach Number `Airspeed` – True Airspeed (ft/s) `Altitude` – Altitude Above Sea-Level (ft) `p` – Non-Dimensional Roll Rate $\left(\hat{p}=\frac{pb}{2V_T}\right)$ `q` – Non-Dimensional Pitch Rate $\left(\hat{q}=\frac{q\bar{c}}{2V_T}\right)$ `r` – Non-Dimensional Yaw Rate $\left(\hat{r}=\frac{rb}{2V_T}\right)$ `AlphaDot` – Non-Dimensional Rate of Change of Angle-of-Attack $\left(\hat{\dot{\alpha}}=\frac{\dot{\alpha}\bar{c}}{2V_T}\right)$ `BetaDot` – Non-Dimensional Rate of Change of Sideslip Angle $\left(\hat{\dot{\beta}}=\frac{\dot{\beta}b}{2V_T}\right)$ Force or Moment Coefficients (values defined by tables above are used): `CL` – Lift Coefficient `CD` – Drag Coefficient `Cy` – Side Force Coefficient `Cl` – Roll Moment Coefficient `Cm` – Pitch Moment Coefficient `Cn` – Yaw Moment Coefficient Parameters/Controls: `Parameter` *n* – Parameter Number *n* `Control` *n* – Control Number *n*

Table F.II – Aerodynamics File Format Cont.

Function Methods

Constant	
Description: Sets the function value (*f*) equal to a constant value (*a*)	
Equation: $f = a$	

Format (case sensitive):
 	
Function
`Constant`
a
	Sample: $C_D = 0.015$
	
	`CD  // Drag Coefficient`
	`Constant`
	`0.015  // CD0`

Linear	
Description: Sets the function value (*f*) to vary linearly with parameter value (*p*)	
Equation: $f = a\,p$	

Format (case sensitive):
 	
Function
`Linear`
p
a
	Sample: $C_L = 5.6\alpha$
	
	`CL  // Lift Coefficient`
	`Linear`
	`Alpha`
	`5.6  // CLalpha`

Absolute Value	
Description: Sets the function value (*f*) to vary with the absolute value of parameter value (*p*)	
Equation: $f = a\lvert p\rvert$	

Format (case sensitive):
 	
Function
`Abs`
p
a
	Sample: $C_D = 0.01\lvert\delta_e\rvert$
	
	`CD  // Drag Coefficient`
	`Abs`
	`Control 0  // Elevator (rad)`
	`0.01  // CDde`

Product	
Description: Sets the function value (*f*) equal to the product of *n* parameters (p_n) and a constant.	
Equation: $f = a\prod p_n$	

Format (case sensitive):
 	
Function
`Product`
n [$n \le 3$]
p_1
...
p_n
a
	Sample: $C_m = \dfrac{(x_n - x_{cg})C_L}{\bar{c}} = K_{SM}\,C_L$
	
	`Cm  // Pitch Moment Coefficient`
	`Product`
	`2  // 2 Variables`
	`CL  // Lift Coefficient`
	`Parameter 1  // Wing Arm (ft)`
	`0.1  // 1/Chord`

Table F.II – Aerodynamics File Format Cont.

Table (See also Appendix B)	
Description: Defines function value (f) using an n dimensional table with n parameters (p_n.)	
Equation: $f = f(p_1,...p_n)$	

Format (case sensitive):
 	
Function
Table
n $[n \le 3]$
$p_1\ m_1\ e_1$
$a_1 \ldots a_{m_1}$
$p_2\ m_2\ e_2$
$b_1 \ldots b_{m_2}$
$p_3\ m_3\ e_3$
$c_1 \ldots c_{m_3}$
 	
$f(a_1,b_1,c_1)\ f(a_2,b_1,c_1) \ldots f(a_{m_1},b_1,c_1)$
$f(a_1,b_2,c_1)\ f(a_2,b_2,c_1) \ldots f(a_{m_1},b_2,c_1)$
...
$f(a_1,b_{m_2},c_1)\ f(a_2,b_{m_2},c_1) \ldots f(a_{m_1},b_{m_2},c_1)$
...
...
$f(a_1,b_1,c_{m_3})\ f(a_2,b_1,c_{m_3}) \ldots f(a_{m_1},b_1,c_{m_3})$
$f(a_1,b_2,c_{m_3})\ f(a_2,b_2,c_{m_3}) \ldots f(a_{m_1},b_2,c_{m_3})$
...
$f(a_1,b_{m_2},c_{m_3})\ f(a_2,b_{m_2},c_{m_3}) \ldots f(a_{m_1},b_{m_2},c_{m_3})$
	Sample: $C_L = f(M,\alpha,\delta_e)$
	
	`CL  // Lift Coefficient`
	`Table`
	`3  // 3 Dimensions`
	
	`Alpha 5 0  // 5 Angles of Attack`
	`-0.10 -0.05 0.00 0.05 0.10`
	`Mach 3 0  // 3 Mach Numbers`
	`0 1 2`
	`Control 4 3 1  // 3 Elevator Deflections`
	`-1.0 0.0 1.0`
	
	`-0.61 -0.33 -0.05 0.23 0.51`
	`-0.80 -0.45 -0.10 0.25 0.60`
	`-0.55 -0.30 -0.05 0.20 0.45`
	
	`-0.56 -0.28 0.00 0.28 0.56`
	`-0.70 -0.35 0.00 0.35 0.70`
	`-0.50 -0.25 0.00 0.25 0.50`
	
	`-0.51 -0.23 0.05 0.33 0.61`
	`-0.60 -0.25 0.10 0.45 0.80`
	`-0.45 -0.20 0.05 0.30 0.55`

Landing Gear File

The landing gear file defines the characteristics of multiple landing gear components. The first line in the file contains the number of landing gear components defined in the file. For each component, the first line is left as a label and the following parameters are defined in the given order.

Table F.III – Landing Gear File Format

$x_{gear,b}$	x position of wheel in body coordinates from c.g., ft
$y_{gear,b}$	y position of wheel in body coordinates from c.g., ft
$z_{gear,b}$	z position of wheel in body coordinates from c.g., ft
k_x	x axis spring constant, lb/ft
k_y	y axis spring constant, lb/ft
k_z	z axis spring constant, lb/ft
b_x	x axis damping constant, lb s/ft
b_y	y axis damping constant, lb s/ft
$b_{z\,initial}$	z axis uncompressed damping constant, lb s/ft
$b_{z\,final}$	z axis compressed damping constant, lb s/ft
$x_{offset,g}$	x direction wheel offset in gear coordinates, ft
k_t	Torsional spring constant, ft lb/rad
b_t	Torsional damping constant, ft lb s/rad
$I_{z\,t}$	Moment of inertia of gear leg around z axis, slug ft^2
absorb	Boolean value whether to use shock absorber (1 – yes, 0 – no)
z_{stroke}	Shock absorber stroke, ft
d	Shock absorber piston diameter, ft
P_{ext}	External pressure, lb/ft^2
$P_{int,u}$	Uncompressed internal pressure, lb/ft^2
γ	Polytropic compression exponent
k_{tire}	Tire spring constant, lb/ft
$k_{t\,tire}$	Tire torsional spring constant, ft lb/rad
$A_{contact}$	Tire contact area, ft^2
$F_{x\,max}$	Maximum load in x direction, lb
$F_{y\,max}$	Maximum load in y direction, lb
$F_{z\,max}$	Maximum load in z direction, lb

Crash Detection File

The crash detection file defines the locations of multiple crash detection points around the aircraft. The first line in the file contains the number of crash detection points. For each point, the first line is left as a label and the x, y, and z body coordinates of the point are listed.

Table F.IV – Crash Detection file Format

$x_{point,b}$	x position of crash point in body coordinates from c.g., ft
$y_{point,b}$	y position of wheel in body coordinates from c.g., ft
$z_{point,b}$	z position of wheel in body coordinates from c.g., ft

Appendix G – *Vendetta*, Aerodynamic Modeling Example

Vendetta is a preliminary design for a Mach 1.6 supercruising advanced deep interdictor aircraft. It was submitted as to the AIAA as one of two Cal Poly, San Luis Obispo entries in the 2001-2002 undergraduate team aircraft design competition. As part of the design process, a simulation model of *Vendetta* was created and flown in the Cal Poly Flight Simulator to verify the performance and stability and control characteristics of the design. The aerodynamic modeling method used to create this model is outlined here, because the resulting model is the most sophisticated model currently used in the simulator. *Vendetta* was modeled using a aerodynamic component build-up method. The seven components defined in the model are shown in Figure G.1. The methods used to model each component are described below. The actual function definitions from the aerodynamic file are included where applicable, but tables were to large to be printed.

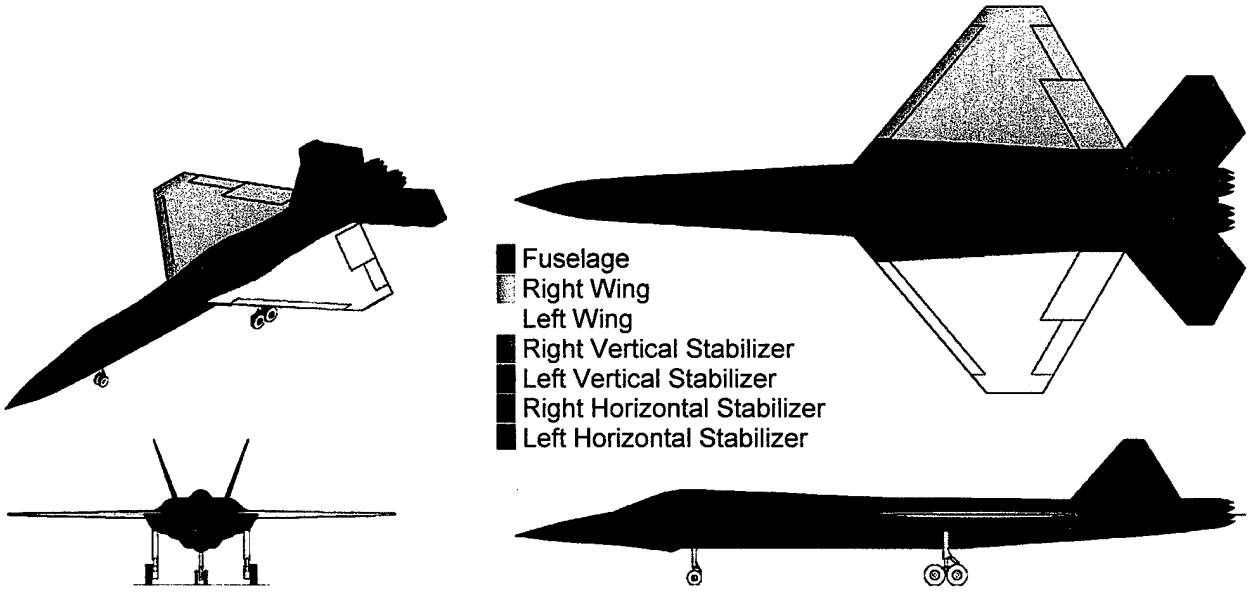


Figure G.1 – Aerodynamic Component Breakdown of *Vendetta*

Wings

The lift curve and induced drag were modeled for the aerodynamic contribution of the wings. The lift curve was estimated using the theoretical lift curve slope based on standard compressible subsonic theory and linear supersonic theory. The resulting lift curve slope is shown as a function of Mach number in Figure G.2. The stall angle-of-attack was then estimated using the spanwise lift distribution and maximum section lift coefficient as shown in Figure G.3. Lift coefficient and stall angle-of-attack increments were then added for leading edge and trailing edge flaps. The resulting lift curve is shown in Figure G.4.

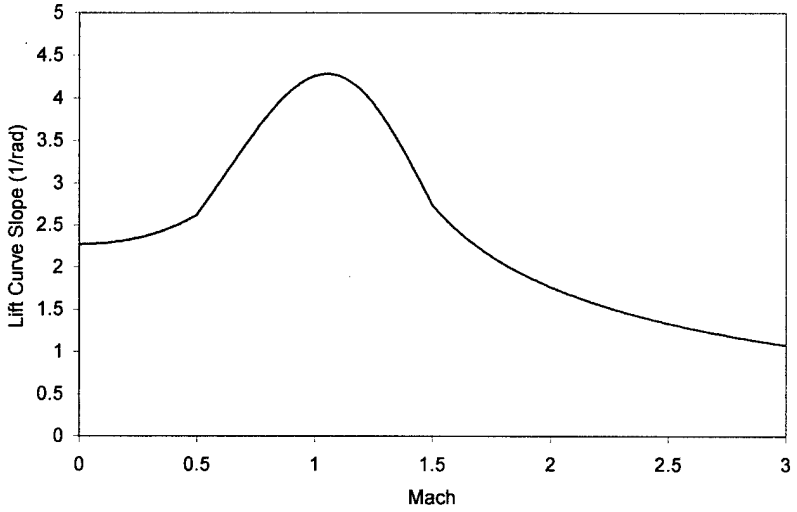


Figure G.2 – Wing Lift Curve Slope as a Function of Mach Number

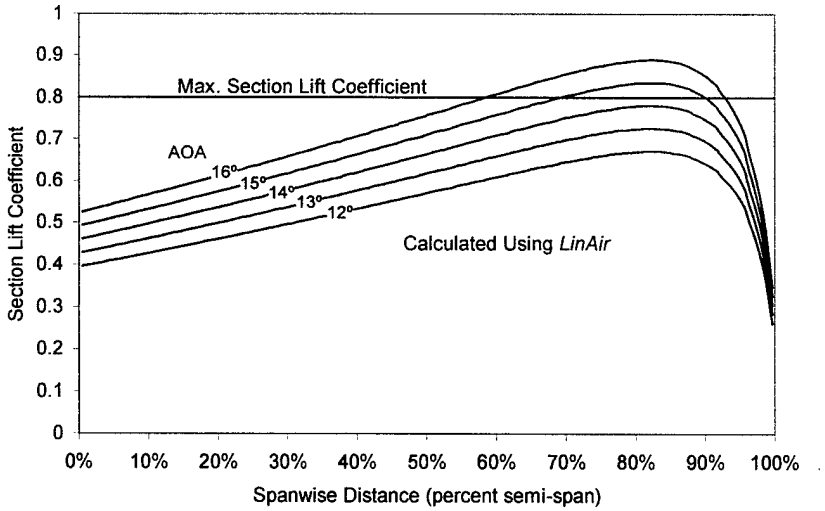


Figure G.3 – Spanwise Lift Distribution for Stall Angle-of-Attack Determination

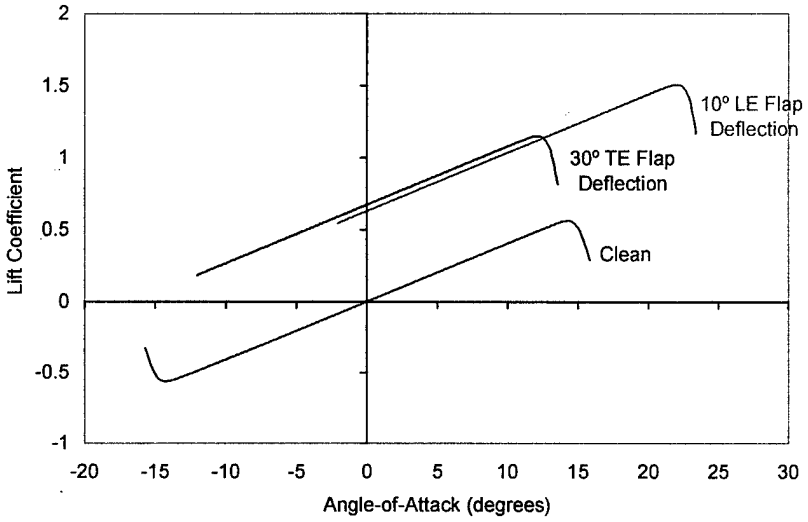


Figure G.4 – Subsonic Wing Lift Curve (Mach 0.2)

The induced drag coefficient for each wing was calculated using Eq. (1) from standard aerodynamic theory. Since induced drag continues to increase with angle-of-attack even after a wing has stalled, the linearized form of the lift coefficient was used, resulting in Eq. (2). Here C_{L_0} and C_{L_α} are functions of Mach number and flap deflections.

$$C_{D_i} = \frac{C_L^{\,2}}{\pi e AR} \tag{G-1}$$

$$C_{D_i} = \frac{\left(C_{L_0} + C_{L_\alpha}\alpha\right)^2}{\pi e AR} \tag{G-2}$$

Since the table lookup system limits tables to three dimensions, two sets of tables were created; one with leading edge flaps extended and one with leading edge flaps retracted. Each set of tables defined lift coefficient and drag coefficient in terms of Mach number, angle-of-attack, and trailing edge flap deflection. The left and right wings were modeled separately, so half of the lift and drag coefficients were used for each side. The results were stored as user-defined parameters so the correct coefficient could be chosen based on the leading edge flap input. The leading edge flap deflection was defined by a control input that ranged from zero to one. A value of zero indicated that the flap was retracted and a value of one indicated that it was extended. Another parameter was defined as the opposite of the leading edge flap deflection. The two lift coefficients and two drag coefficients were then combined using these two parameters as shown in Eqs. (4-5). An additional lift coefficient increment for the aileron contribution was added as a function of Mach number and aileron deflection. Again the results were stored as parameters so that the coefficients could be used for moment calculation.

$$\delta'_{f_{LE}} = 1 - \delta_{f_{LE}} \tag{G-3}$$

$$C_L = C_{L_{LE\,flap}} \delta_{f_{LE}} + C_{L_{no\,LE\,flap}} \delta'_{f_{LE}} \tag{G-4}$$

$$C_D = C_{D_{LE\,flap}} \delta_{f_{LE}} + C_{D_{no\,LE\,flap}} \delta'_{f_{LE}} \tag{G-5}$$

```
Parameter 32    // No Right Slats
Constant
1.0

Parameter 32    // No Right Slats
Linear
Control 8       // Right Slat
-1.0
```

```
Parameter1      // Right Wing Lift
Product
2
Parameter 3     // Right Wing Lift (No Slats)
Parameter 32    // No Right Slats
1.0

Parameter 1     // Right Wing Lift
Product
2
Parameter 5     // Right Wing Lift (Slats)
Control 8       // Right Slat
1.0
```

```
Parameter 2     // Right Wing Drag
Product
2
Parameter 4     // Right Wing Drag (No Slats)
Parameter 32    // No Right Slats
1.0

Parameter 2     // Right Wing Drag
Product
2
Parameter 6     // Right Wing Drag (Slats)
Control 8       // Right Slat
1.0
```

Once the lift and drag coefficients were defined for each wing, the moment contributions of each force could be calculated. The forces were assumed to act at the aerodynamic center of the mean aerodynamic chord of the wing. Because the center of gravity was variable, and because the aerodynamic center of a wing shifts with Mach number, the arm from the c.g. to the aerodynamic center (x_{mac}) was looked up as a function of c.g. location and Mach number. Once the position of the aerodynamic center was known, the moment coefficients could be calculated as shown in Eq. (6-8).

$$C_m = \frac{C_L x_{ac}}{\bar{c}} - \frac{C_D z_{ac}}{\bar{c}} \tag{G-6}$$

$$C_l = -\frac{C_L y_{ac}}{b} \tag{G-7}$$

$$C_n = \frac{C_D y_{ac}}{b} \tag{G-8}$$

```
Cm              // Pitch
Product
2
Parameter 1     // Right Wing Lift
Parameter 0     // Wing Arm
0.03125         // 1/chord

Cm              // Pitch
Linear
Parameter 2     // Right Wing Drag
0.00781         // -zmac/chord
```

```
Cl              // Roll
Linear
Parameter 1     // Right Wing Lift
-0.19161        // -ymac/span

Cn              // Yaw
Linear
Parameter 2     // Right Wing Drag
0.19161         // ymac/span
```

Horizontal Stabilizers

The horizontal stabilizer on *Vendetta* is a full flying surface, which means that the angle-of-attack of the stabilizer is controlled directly with the elevator deflection. Because the horizontal stabilizer is located behind the wing, the downwash from the wing affects the effective angle-of-attack of the stabilizer. The effective angle-of-attack of the horizontal stabilizer is given by Eq. (9). The downwash angle at the horizontal tail, ε_H, was calculated as a function of angle of attack and Mach number using DATCOM and defined with a table.

$$\alpha_H = \alpha + \delta_e - \varepsilon_H \tag{G-9}$$

```
Parameter 15    // Right Horizontal Angle
Linear
Alpha
1.0

Parameter 15    // Right Horizontal Angle
Linear
Control 0       // Right Elevator
1.0

Parameter 15    // Right Horizontal Angle
Linear
Parameter 14    // Horizontal Downwash
-1.0
```

The lift and induced (trim) drag coefficients for each horizontal stabilizer were then looked up as functions on Mach number and effective angle-of-attack similar to the wings. The coefficients are non-dimensionalized using the wing planform area. The lever arm of the horizontal tail was looked up as a function of center of gravity location and Mach number. The moment coefficient contributions were then calculated using the same equations (6-8) as the wings, except using the position of the aerodynamic center of the horizontal stabilizer.

Vertical Stabilizers

The vertical stabilizers on *Vendetta* are canted 20° off of vertical. This means that the effective angle of attack of the stabilizers are made up of components of both angle-of-attack and sideslip angle. The normal (lift) forces produced by the stabilizers have components in the lift and side force directions. The effective angle-of-attack of the vertical stabilizer was calculated using the components angle-of-attack and sideslip angle as shown in Eq. (10). The sign depends on which stabilizer is being defined.

$$\alpha_V = \pm\alpha \sin 20° - \beta \cos 20° \tag{G-10}$$

```
Parameter 22    // Right Vertical Angle
Linear
Beta
-0.93969        // -cos 20°

Parameter 22    // Right Vertical Angle
Linear
Alpha
-0.34202        // -sin 20°
```

The normal force and drag coefficients were looked up as functions of Mach number, effective angle-of-attack, and rudder deflection. The coefficients are non-dimensionalized using the wing planform area. The lift and side force coefficients were then calculated using the components of the normal force as shown in Eqs. (11-12). Again the sign depends on which stabilizer is being defined.

$$C_L = \pm C_N \sin 20° \tag{G-11}$$

$$C_y = C_N \cos 20° \tag{G-12}$$

```
Parameter 25   // Right Vertical Lift
Linear
Parameter 23   // Right Vertical Normal
-0.34202       // -sin 20°

Parameter 26   // Right Vertical Y
Linear
Parameter 23   // Right Vertical Normal
0.93969        // cos 20°
```

Again the lever arm of the vertical stabilizer was looked up as a function of center of gravity location and Mach number, and the moment coefficient contributions were calculated using the position of the aerodynamic center of the vertical stabilizer.

$$C_m = \frac{C_L x_{ac}}{\bar{c}} - \frac{C_D z_{ac}}{\bar{c}} \tag{G-13}$$

$$C_l = -\frac{C_L y_{ac}}{b} - \frac{C_y z_{ac}}{b} \tag{G-14}$$

$$C_n = \frac{C_y x_{ac}}{b} + \frac{C_D y_{ac}}{b} \tag{G-15}$$

```
Cm             // Pitch
Product
2
Parameter 25   // Right Vertical Lift
Parameter 21   // Vertical Arm
0.03125        // 1/chord

Cm             // Pitch
Linear
Parameter 24   // Right Vertical Drag
0.1527         // -zmac/chord

Cl             // Roll
Linear
Parameter 25   // Right Vertical Lift
-0.03245       // -ymac/span

Cl             // Roll
Linear
Parameter 26   // Right Vertical Y
0.08917        // -zmac/span

Cn             // Yaw
Product
2
Parameter 26   // Right Vertical Y
Parameter 21   // Vertical Arm
0.01825        // 1/span

Cn             // Yaw
Linear
Parameter 24   // Right Vertical Drag
0.03245        // ymac/span
```

Fuselage

The aerodynamic contributions of the fuselage molded were, drag, lift, pitch moment, side force, and yaw moment. The parasite and wave drag generated by the entire aircraft were assumed to be contributed by the fuselage. The parasite drag was calculated using a typical component drag build up with form factors and interference factors. To account for Reynold's number effects, the drag coefficient was defined at different altitudes using the standard atmosphere to define viscosity. The wave drag was approximated using the area ruling of the aircraft. The cross-sectional area

distribution of the aircraft, shown in Figure G.5 and Figure G.6, was integrated using the de Kármán integral in Eq. (16) to approximate the wave drag efficiency factor, E_{WD}, by comparing with the wave drag of a perfect Sears-Haack body as shown in Eqs. (17-18). The resulting efficiency factor was then used to approximate the variation with Mach number using Eq. (19). The resulting drag variation is shown in Figure G.7.

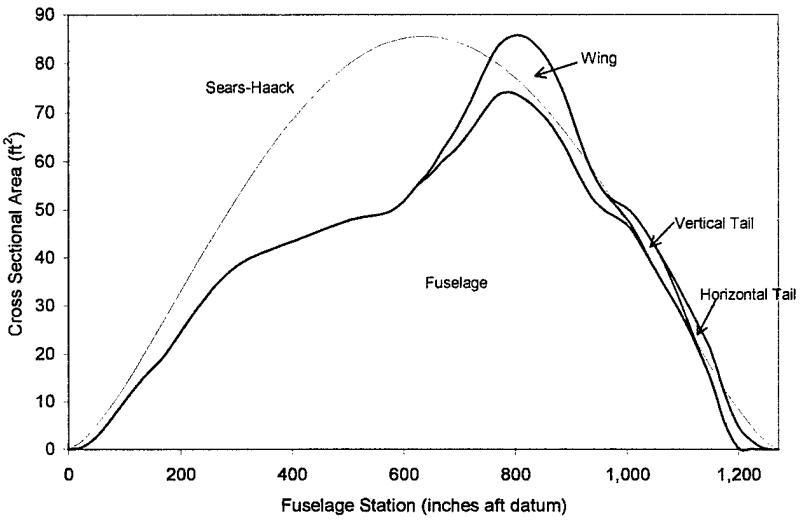


Figure G.5 – Transonic Area Distribution

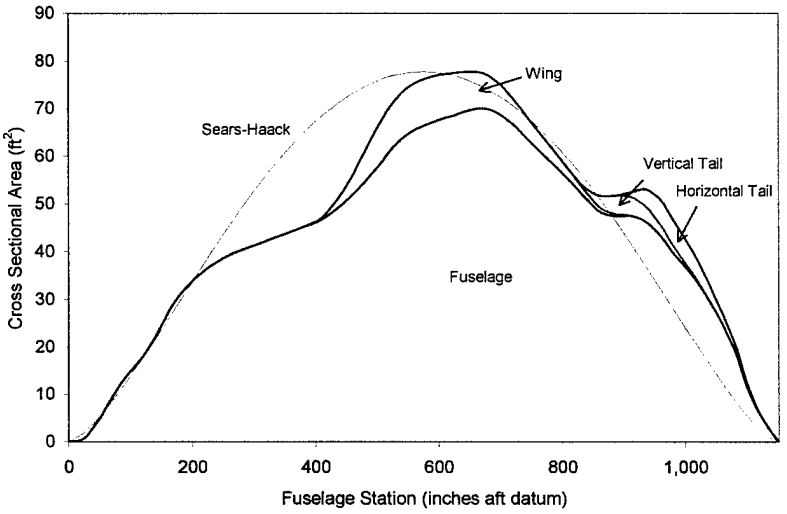


Figure G.6 – Supersonic Area Distribution (Mach 1.6)

$$C_{D\,wave} = -\frac{1}{2\pi S}\int_0^l\int_0^l \frac{\partial^2 A}{\partial x_1^{\,2}}\frac{\partial^2 A}{\partial x_2^{\,2}}\ln\left|x_1 - x_2\right| dx_1\, dx_2 \qquad \text{(G-16)}$$

$$C_{D\,wave}{}' = \frac{4.5\pi}{S}\left(\frac{A_{max}}{l}\right)^2 \qquad \text{(G-17)}$$

$$E_{WD} = \frac{C_{D\,wave}}{C_{D\,wave}{}'} \qquad \text{(G-18)}$$

$$C_{D\,wave} = \frac{4.5\pi}{S}\left(\frac{A_{max}}{l}\right)^2 E_{WD}\left(0.74 + 0.37\cos\Lambda_{LE}\right)\left(1 - 0.3\sqrt{M - M_{C_{D_0}\,max}}\right)$$

$$M_{C_{D_0}\,max} = \frac{1}{\cos^{0.2}\Lambda_{LE}} \qquad \text{(G-19)}$$

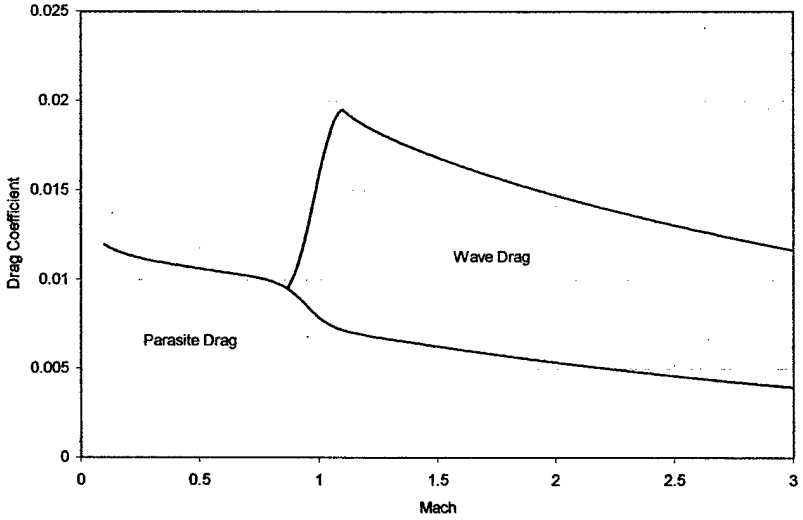


Figure G.7 – Variation in Parasite and Wave Drag Coefficients with Mach Number (50,000 ft)

The lift and pitch moment coefficient contributions of the fuselage were calculated using DATCOM and tabulated as functions of angle-of-attack and Mach number. Because DATCOM can not calculate fuselage side force and yaw moments, it was assumed that the fuselage was symmetrical about the *x*-axis. This means that the same data for lift and pitch moment in terms of angle-of-attack could be used for side force and yaw moment in terms of sideslip angle. The moment coefficients need to be corrected for the difference in non-dimensionalization as shown in Eq. (20).

$$C_n(\beta) = C_m(\alpha)\frac{\bar{c}}{b} \qquad \text{(G-20)}$$

In addition to the aerodynamic component buildup described above, linear stability derivatives for body rates and rate of change of angle-of-attack and sideslip angle from DATCOM were also included. The resulting aerodynamic model for *Vendetta* is made up of a total of 107 functions, 12 control inputs, and 34 user-defined parameters. The control input and user-defined parameter numbering are listed in Table G.I and G.II.

Table G.I – *Vendetta* Model Controls

0 Right Elevator Deflection, rad
1 Left Elevator Deflection, rad
2 Right Aileron Deflection, rad
3 Left Aileron Deflection, rad
4 Right Rudder Deflection, rad
5 Left Rudder Deflection, rad
6 Right Flap Deflection, rad
7 Left Flap Deflection, rad
8 Right Slat Extended (0 or 1)
9 Left Slat Deflection (0 or 1)
10 Center of Gravity Location From Nose, ft
11 Landing Gear Extended (0 or 1)

Table G.II – *Vendetta* Model Parameters

0 Wing Arm, ft
1 Right Wing Lift Coefficient
2 Right Wing Drag Coefficient
3 Right Wing Lift Coefficient (No Slat)
4 Right Wing Drag Coefficient (No Slat)
5 Right Wing Lift Coefficient (Slat)
6 Right Wing Drag Coefficient (Slat)
7 Left Wing Lift Coefficient
8 Left Wing Drag Coefficient
9 Left Wing Lift Coefficient (No Slat)
10 Left Wing Drag Coefficient (No Slat)
11 Left Wing Lift Coefficient (Slat)
12 Left Wing Drag Coefficient (Slat)
13 Horizontal Arm, ft
14 Horizontal Downwash Angle, rad
15 Right Horizontal Angle, rad
16 Right Horizontal Lift Coefficient
17 Right Horizontal Drag Coefficient
18 Left Horizontal Angle, rad
19 Left Horizontal Lift Coefficient
20 Left Horizontal Drag Coefficient
21 Vertical Arm, ft
22 Right Vertical Angle, rad
23 Right Vertical Normal Force Coefficient
24 Right Vertical Drag Coefficient
25 Right Vertical Lift Coefficient
26 Right Vertical Y Force Coefficient
27 Left Vertical Angle, rad
28 Left Vertical Normal Force Coefficient
29 Left Vertical Drag Coefficient
30 Left Vertical Lift Coefficient
31 Left Vertical Y Force
32 No Right Slat (0 or 1)
33 No Left Slat (0 or 1)

Development of Field Rechargeable Gas Mask Filters

Project Investigators:

Katherine C. Chen, Ph.D.
Associate Professor
Materials Engineering Department

Kevin B. Kingsbury, Ph.D.
Associate Professor
Chemistry and Biochemistry Department

Ronald F. Brown, Ph.D.
Professor
Physics Department

Christopher L. Kitts, Ph.D.
Associate Professor
Biology Department

C3RP FINAL REPORT

Development of Field Rechargeable Gas Mask Filters

K. Chen
Materials Engineering, Cal Poly
December 18, 2002

A collaborative project between Cal Poly and Sun Microsystems Advanced Development demonstrated a proof of concept for a field rechargeable gas mask filter. Based on patented adsorption technology licensed to Sun, an innovative system was developed to model the collection and purging of specific contaminants, rendering replacement filters and cartridges unnecessary. Quick regeneration of active adsorption sites (i.e., "desorption") of the filter prevents plugging and allows continuous use of gas masks in industrial and military settings.

The feasibility study involved setting up testing capabilities and examining adsorption/desorption mechanisms. A test chamber was built and then remodified several times to improve the testing capabilities. Characterization of the process and the materials involved (i.e., adsorbent and adsorbate) were executed through experiments and interpreted with adsorption theories. The work was carried out by Cal Poly students and faculty. Sun Microsystems contributed their intellectual property and in-kind support to investigate the adsorption/desorption technology on a potential new application.

Project Summary

The objectives proposed in the grant were met (Table I), and the completed work has nicely set up capabilities for future efforts. Results are discussed in more detail later in this report. Most importantly, a sound, working relationship has blossomed between the Cal Poly research team and Dennis Pfister at Sun Microsystems. Numerous discussions and meetings occurred at Cal Poly and at Sun, Advanced Development.

Table I. Objectives of the project met

√	establish working relationships between Cal Poly faculty and Sun Microsystems
√	test the feasibility of a regenerated gas mask filter using adsorption technology
√	explore the possibility of decontamination through intense localized heating
√	provide technical expertise to Sun Microsystems Advanced Development, SLO
√	create research projects and entrepreneurial experiences for Cal Poly students
√	encourage interdisciplinary research projects across Cal Poly colleges
√	increase local employment opportunities for Cal Poly spouses and graduates
√	investigate additional funding from other agencies based on results from project

Because the project is of immense interest to Sun, they showed great commitment to our efforts. A list of donations (or long-term loans) from Sun is given in Table II. Due to the proprietary nature of the technology, there are no publications dealing with this particular adsorption/desorption technology.

Table II. Donations to the gas mask project by Sun Microsystems

Equipment and supplies donated or loaned to Cal Poly research group	
HP Power supply (~ $5000)	Teflon tubing
LabView software and manuals (~$1000)	copper mesh
data acquisition card & signal conditioning box (~$1k)	activated carbon cloth
aluminum sample chamber	check valves
thermocouples	Swagelok fittings

The project enabled the hiring of a Cal Poly Materials Engineering (MATE) graduate and MATE part-time lecturer. These individuals were able to stay in the SLO area due to the supplemental salaries from the grant, and they proved to be vital members to the research team. Students were also hired on the project, and as a nice surprise, the project has unintentionally piqued the interest of several others. An overview of the project and some preliminary results were presented at the MATE poster session this past spring. In addition, other Cal Poly faculty from other departments and colleges have been involved with the project.

Some discussions with Mr. Pfister have taken place to strategize how to approach for outside funding for this project. We agree that we first need to first build a prototype that will demonstrate the capability to regenerate or recharge the gas mask filters. Contact with an Air Force program manager was made in reference to novel forms and applications of carbon. Mr. Pfister is also in the process of spinning off into his own company that will be backed by Sun Microsystems and General Electric (GE).

Significance of the Technology

Presently, most gas masks use adsorption to filter out contaminants, but an innovative concept is to *renew or recharge the active filter sites in a controlled and precise manner.* The technology to accomplish such a task is available, and the project was a logical extension of a patented technology using adsorption principles. While Sun is currently utilizing the technology for cooling applications, the same scientific theories and technology can be applied to a design of rechargeable gas mask filters.

The unique feature of this design is the regeneration of active adsorption sites after becoming saturated with contaminants. Thus, the fundamental challenge involves *desorption.*

Instead of molecules adhering to a surface, the reaction is reversed, and molecules *debond* and are removed from the active sites. Desorption can be achieved by supplying the correct amount of energy (i.e., heat) to the system by using electrical current.

The ability to recharge gas mask filters enables several improvements and provides significant impacts. The use of gas masks as safety protection devices can be extended to field uses where time and location would prohibit simple exchange of cartridges. Threats of biological and chemical warfare underscore the need for application of advanced technology to ensure national security. In addition, the proposed technology has potential to be developed to *decontaminate and destroy harmful species*. Toxic substances or microorganisms that are captured by the gas mask could then be subjected to intense heat at the bound sites. Instant remediation and neutralization of harmful substances are possible.

From an environmental standpoint, used canisters and cartridges would no longer need to be disposed. While the military applications are quite evident, the developed technology could also be transferred to industrial safety uses.

Testing Capabilities

The testing apparatus went through two different rounds of design and construction. The team learned immeasurable amounts through the experiences, and had the opportunity to work closely with Dennis Pfister through these efforts. Each side has been able to offer different solutions to problems and bring new ideas to the project. A third and final test cell was built that avoids major leaks and allows several different data acquisition capabilities.

Figure 1 displays the initial test set-up that allowed some preliminary data in order to create a better sample chamber and to design better testing capabilities. Control of the gas inlet and quantitative measurement of the amount of gas adsorped/desorped was added. In addition, **real-time data acquisition** on temperature, pressure, power, and weight gain/loss were designed into the system. Computer controls and data collection via LabView software were installed, and are ready for use in future studies. The entire endeavor was a challenging engineering achievement in itself, and Figure 2 depicts the redesigned fixtures and test set-up.

The simultaneous acquisition of temperature and weight gain/loss as a function of time enables us to understand the *science* behind the technology. During *adsorption*, gas molecules essentially drop down onto the activated carbon (or the *adsorbent*) as a solid, and thus heat is released. As a result, the temperature within the sample chamber increases. In addition, a weight gain is noted. Furthermore, the desorption process can be monitored in a similar manner. This feature of attempting to understand and monitor the physics behind the application offers

capability not currently being pursued at Sun. The testing capabilities will enable meaningful data and assist in future product development.

Results

The ability to adsorp and desorp gases has been successfully demonstrated. Two different adsorbates (propane and freon) have been tested thus far. Eight different thermocouples were placed in various positions within the sample chamber to monitor the temperature increase associated with adsorption. Dynamic weight gain was also detected. Controlled desorption was also confirmed, yet not efficiently. Improvements to the process have been discussed and will the focus for the future work.

Initiation of the desorption process to regenerate the active adsorption sites has been demonstrated, and fairly high temperatures are achievable through the applied power. However, more directed work is needed to explore the high temperatures capability for decontamination efforts.

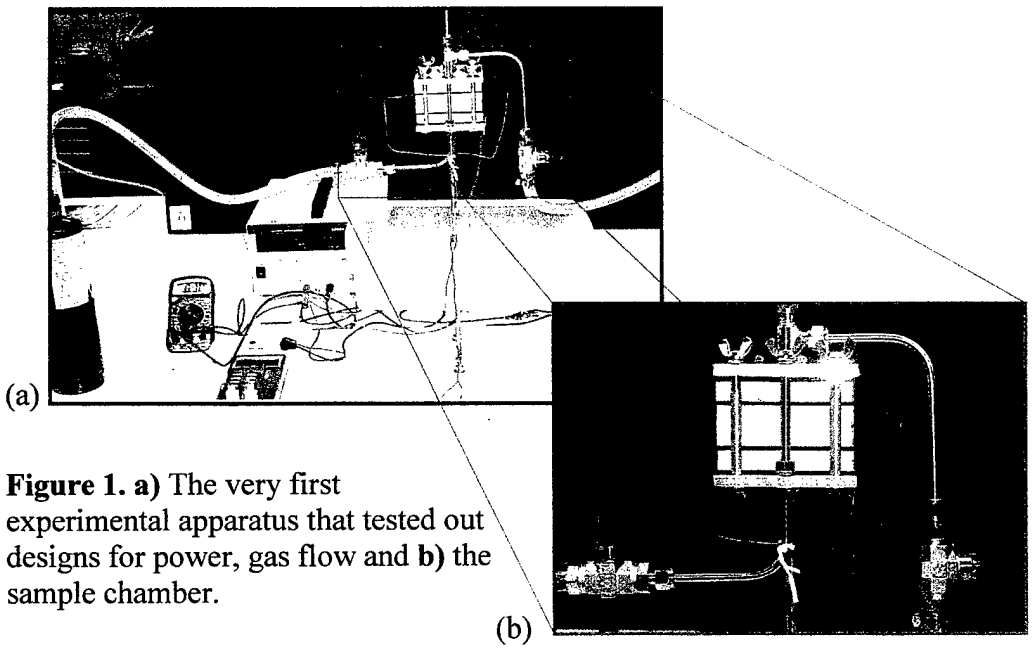
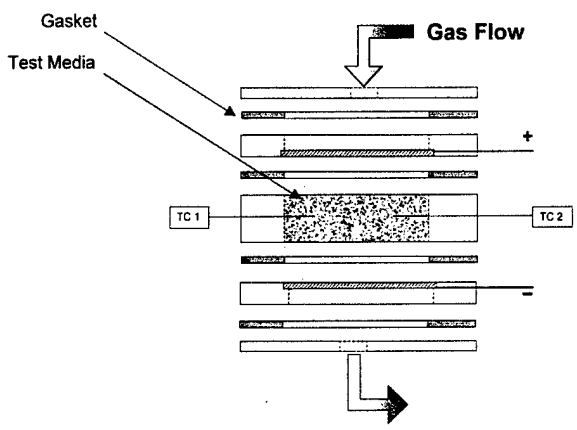
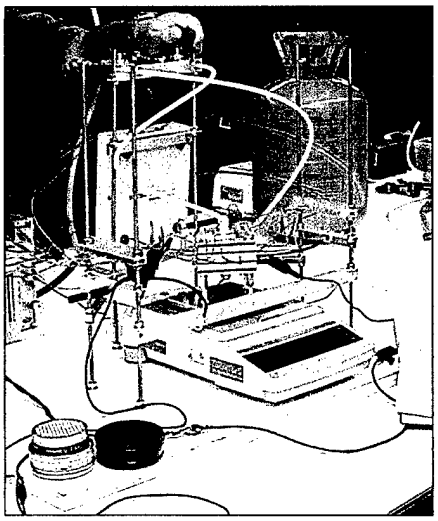


Figure 1. a) The very first experimental apparatus that tested out designs for power, gas flow and **b)** the sample chamber.



(a)



(b)

Figure 2. a) Schematic of the sample chamber and **b)** the second redesign of the testing apparatus that includes many modifications that enable simultaneous data acquisition of temperature, power, and weight as a function of time.

Exploitation of Network Bandwidth and the Ethernet/IP Application Layer Standard for Automation Networks

Project Investigator:

Kurt Colvin, Ph.D.
Professor
Industrial Engineering Department

Exploitation of Network Bandwidth and the Ethernet/IP Application Layer Standard for Automation Networks

Project Final Report, September, 2002

Kurt Colvin
IME Department, Cal Poly, San Luis Obispo

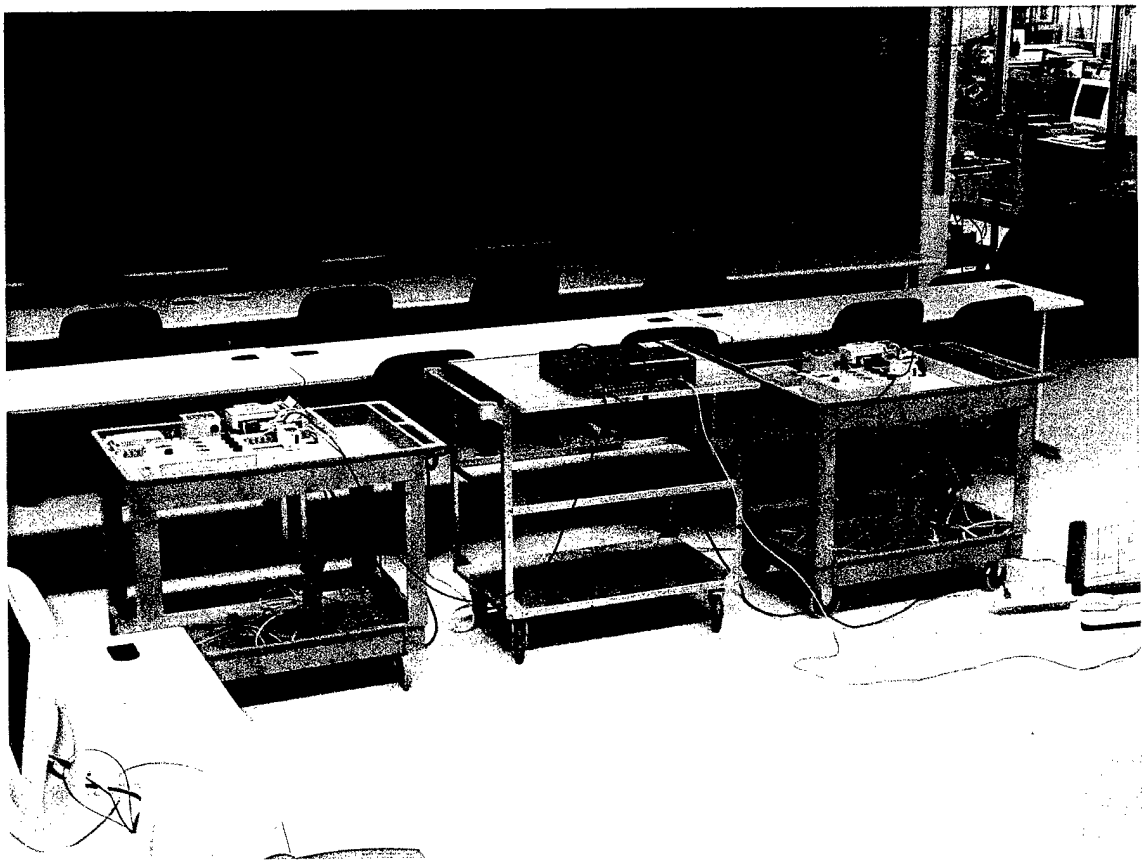


Figure 1. Ethernet/Industrial Protocol Prototype System. The device on the left is an automation cart with an Allen-Bradley PLC configured as a TCP/IP node on an Ethernet network. The Cisco Ethernet switch in the center isolates communications links between devices. The cart on the right is a second PLC configured as a TCP/IP node. Messages were successfully sent between the devices allowing for remote sensing and control of the devices anywhere on the Internet. A full-scale implementation of this system in the IME automation lab is planned for the 2002-03 academic year.

Summary

The primary goal of this project was to experiment with the exploitation of high-speed Ethernet network bandwidth for use in industrial automation applications. A prototype system has been developed and tested successfully in our lab (see Figure 1). This project has resulted in the development of an "Automation Networking" module to be included

in the IME 356 (Manufacturing Automation). This module includes 3 hours of lecture material, several reading assignments and two 3-hour labs. The labs will fit nicely into the overall course, as they are capstone in nature, integrating many of the concepts covered throughout the course. Several other unexpected benefits emerged as we worked through this project and are explained in the body of this report. It is anticipated that future projects will follow as there is much interest in the Ethernet/IP topic from potential industrial partners (specifically, Rockwell Automation and their partners).

Tasks Completed

1. Literature review. An extensive literature search was performed to learn more about the technologies and their use in research and the commercial domain. We found more than 100 relevant articles with abstracts. We reviewed the majority of these and developed these generalizations:
    - Industrial automation over the Ethernet is very much in its infancy.
    - Ethernet networks are still largely misunderstood by traditional automation engineers and managers/sales people. (They can't get past the fact that Ethernet is NOT a deterministic network.)
    - Information systems people have a good understanding of Ethernet, but little knowledge of how automation devices share information on a network.
    - A few are starting to see the benefits of using Ethernet for not only their information systems, but their automated device-level systems as well.
    - Ethernet/IP
        - Interest, enthusiasm and activities are building around use of the Ethernet/IP specification and devices/support are on the way.
        - Control and Information Protocol (CIP) is the common upper layer protocol for the major automation networks:
            - DeviceNet (device-level networks)
            - ControlNet (deterministic control-level networks)
            - Ethernet/IP (information-level networks capable of both device-level and control-level functionality. The best of all worlds)

    From this literature review, several excellent reading assignments have been identified that will be included in the IME 356 course.

2. Prototype equipment. Our prototyping equipment was ordered and setup in our lab.

    Ethernet system is composed of the following (See Figure 1):
    - Cisco Ethernet Switch (donated by Cal Poly ITS, Jerry Handley)
    - Two Allen-Bradley MicroLogix 1500 Programmable Logic Controllers (PLCs) (Already owned by the IME Dept.)
    - Two Allen-Bradley Ethernet/IP Interface devices (NET-ENI). These enable the PLCs to communicate over an Ethernet network (See Figure 2).

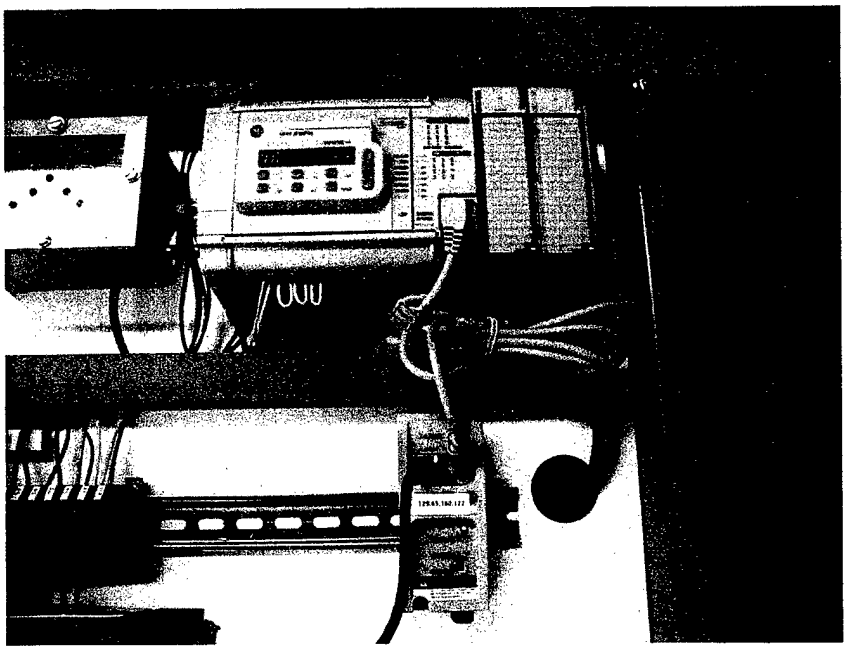


Figure 2. Ethernet/IP Interface (ENI). Configured with serial connection to the PLC and a patch cable to the Cisco Ethernet switch. The ENI is assigned an Internet Protocol (IP) address making it accessible by any device on the Internet. This allows for remote sensing, monitoring and control of this PLC across any network capable of running the TCP/IP protocol.

DeviceNet system (See figure 3):

- As a comparison of competing automation network technologies, we setup a DeviceNet network (we already had most of the hardware in place to do this). We will use this system as a comparison of a deterministic network compared to the ***nearly-deterministic*** characteristics of the Ethernet system. This system is not yet fully functional due to a failure of a DeviceNet scanner device. Replacement hardware is on back order, but it is anticipated that the system will be running by Mid-December.
    - DeviceNet scanner.
    - Allen Bradley DeviceNet Interface (DNI). Already owned by the IME Dept.
    - Allen Bradley KFD. Already owned by the IME Dept.
    - NetWorx Software. DeviceNet configuration software.
    - Rightsight Photosensor (a DeviceNet input device)
    - Festo Pneumatic Manifold (A DeviceNet output device)

3. Functionality of systems
    - Task 1: Enable all PLC's on an Ethernet network. In this task, we brought up 2 PLCs on the network and verified communications. This will enable any PLC to be program/monitored from and PC on the Ethernet network.
    - Task 2: Enable Peer to Peer messaging between PLC's. I this task, we programmed ladder logic to allow one PLC to modify the state of another PLC.

- Task 3: Enable DeviceNet network. This task involved configuring 6 components into a small DeviceNet network. Limited functionality has currently been achieved, but we are still waiting on backordered equipment.

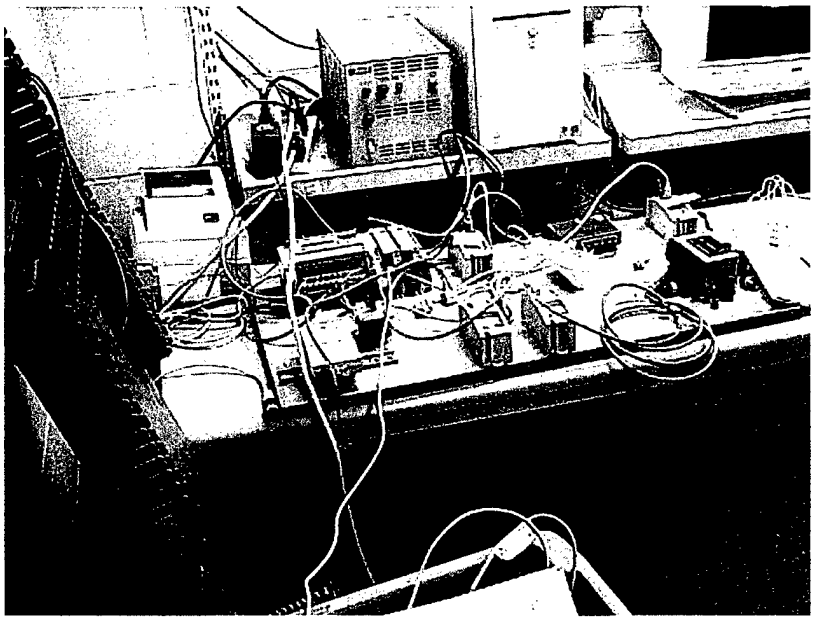


Figure 3. DeviceNet prototype system. A PLC, Photoelectric sensor and pneumatic manifold make up a DeviceNet network. DeviceNet is a truly deterministic network with limited bandwidth (250 Kbits/second). DeviceNet is characteristic of conventional automation networks.

Deliverables of This Project

Lab equipment (See Figures 1 and 2):
The prototype lab equipment was purchased and developed in our lab. This equipment can be directly utilized in our IME 356 labs. The additional equipment needed to outfit our entire lab has been submitted to the Principle Investigator of the SME grant. This requires an equipment purchase of about $5000. We hope to implement this during Fall term, 2002.

Lab assignments and solutions:
Two labs have been developed for the IME 356 lab.

Automation Networking Lab 1 (see Appendix I):
In this lab, the benefits of automation networks are presented. There is a significant lecture component needed to introduce networking concepts. The practical application includes the ability to program any PLC connected to the Ethernet network and implementation of basic digital I/O between PLCs.

Automation Networking Lab 2 (see Appendix II):
This lab builds not only on the preceding networking lab, but also requires extensive use of previous labs including: digital and analog I/O, sequencers, timers, counters and use of the Human Machine Interface (HMI) software RSView. This is an excellent capstone project, designed to implement most of the topics covered in the course.

Revelations and Benefits of This Project

Is it really Ethernet/IP?

We discovered that several vendors sell products as Ethernet/IP-compatible, but in reality they are not. For example, Allen-Bradley says that their ENI is "Ethernet/IP" compatible (in fact, they have Ethernet/IP printed on the device). However, in reality, when we investigated, it turns out that their application-layer protocol of our Micrologix PLCs is a proprietary protocol (PCCC) that we do not have access to. In fact, this device does not respond to Ethernet/IP messages. THIS IS NOT AN OPEN SYSTEM! Other manufacturers (Omron, Siemens, Mitsubishi) are similar. Conversations with Rockwell Automation (the parent company of Allen-Bradley) revealed that they are currently working on new products that are truly Ethernet/IP-compatible. Rockwell Automation is dedicated to the success of Ethernet/IP and is interested in future collaboration. We we are attempting to develop a future proposal together.

CIP is the key

The control and information protocol (CIP) is the key to the Ethernet/IP system. This is really an open protocol and is common in most of the major automation device manufacturers. Allen-Bradley's new product will all implement this application-layer protocol (as will Siemens, Omron, GE, Mitsubishi and other manufacturers).

We need an Ethernet/IP Interface Device!

Due to the infancy of the Ethernet/IP system, there does not yet exist a true Ethernet/IP interface device. As stated above, that is what we thought we were getting from Allen Bradley, but it turns out that this device is not yet available commercially. This includes both hardware and software.

A low cost Ethernet/IP Interface Device?

Industry (and the research arena) is in need of a generic, low cost Ethernet/IP interface device. This would be a microcontroller capable of Ethernet communications, enough memory and functionality to implement the Ethernet/IP specification and standard digital input/output and serial communications. This device would be the "translator" between an automated device (sensor, bar code reader, servo motor amplifier, etc.) and the Ethernet network. This device would have the functionality to communicate with a central server running a SCADA (Supervisory Control and Data Acquisition) system. The proposal to develop this device has been submitted to C3RP for funding. Additionally, a proposal to Rockwell Automation for funding is in progress.

An Ethernet/IP ActiveX component

While the C and C++ languages are currently being used in the development of Ethernet/IP devices, there does not yet exist similar functionality in the Visual Basic language. Due to Visual Basic's wide acceptance and interoperability with many applications though Visual Basic APIs (application programmers interface), there is a need for Ethernet/IP code in Visual Basic in the form of an ActiveX object. This aspect is included in the proposal mentioned above.

Conclusions

This project has been an excellent opportunity to experiment with the user of Ethernet networks in an industrial automation setting. Through the project, and extensive literature review was performed and we have become familiar with the current state of the automation industry with regard to Ethernet networking. Further, we were able to develop appropriate course material and implement both lecture content and practical, hands-on labs to convey some of this knowledge to our students. Finally, we have identified a true need in industry (Ethernet/IP hardware and software) and we are currently working to continue work in this area.

Appendix I

Introduction to Automation Networking

The objective of this lab is to learn practical skills in working with an automation network (or automation bus). An automation bus has at least the following advantages:

- Allows an engineer to program the PLC remotely (no need for a dedicated serial connection). You can program many PLCs from a single location.
- Allows for peer-to-peer messaging (passing data from one PLC to another).
- Allows a device to share its data with many devices, such as multiple HMIs.

Do not roll the carts to the lab computers for this lab. Instead, plug in an Ethernet patch cable between the ENI box on your cart and the Cisco Ethernet switch. Make sure your cart is powered on.

Step 1: Configure a communications driver in RSLinx

The first step is to configure a driver so that the PC can communicate with the PLC over the Ethernet network. This is again accomplished with RSLinx. This is similar to the other labs, however this time you will configure a Ethernet connection instead of a serial connection.

1. Start RSLinx from the start menu (Start→Programs→Rockwell Software→RSLinx→RSLinx)
2. In RSLinx, go to the Communications menu and select Configure Drivers.
3. If there is already an existing Ethernet driver (probably named "AB_ETH-1"), first remove it from the list of drivers before continuing.
4. In the top drop-down box, select the 'Ethernet Devices' driver and select the Add New button. Click OK and take the default name "AB_ETH-1".
5. In the configure driver dialog box, enter station **1** and the ***IP address of the PLC that you will program***. You must configure this correctly for you to communicate with you PLC. See Figure 1. (If you want to program other PLCs, you would need to enter those stations and IP addresses here.)
6. Select the OK button.
7. You can close the "Configure Drivers" dialog box and minimize the RSLinx application.

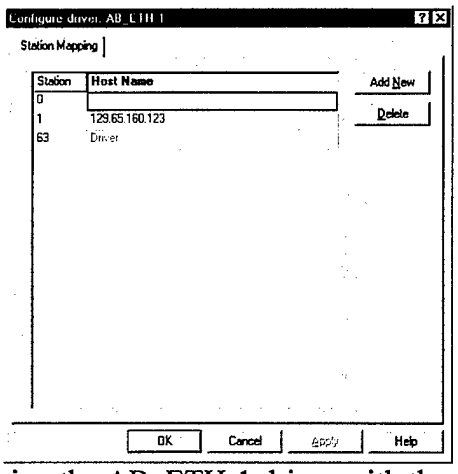


Figure 1. Configuring the AB_ETH-1 driver with the correct IP address

Step 2: Write a very simple ladder program to verify communications settings are correct.

Using RSLogix, create a very simple ladder and download to your PLC. It can be as simple as examining the first button and turning on a light if the button is pressed.

The whole trick to this step is to correctly identify your PLC in the communications setup. Double click on the Controller Properties (on the left window of RSLogix), and select the Controller Communications tab (see Figure 2). Make sure you have the AB_ETH-1 driver selected and the correct Processor Node.

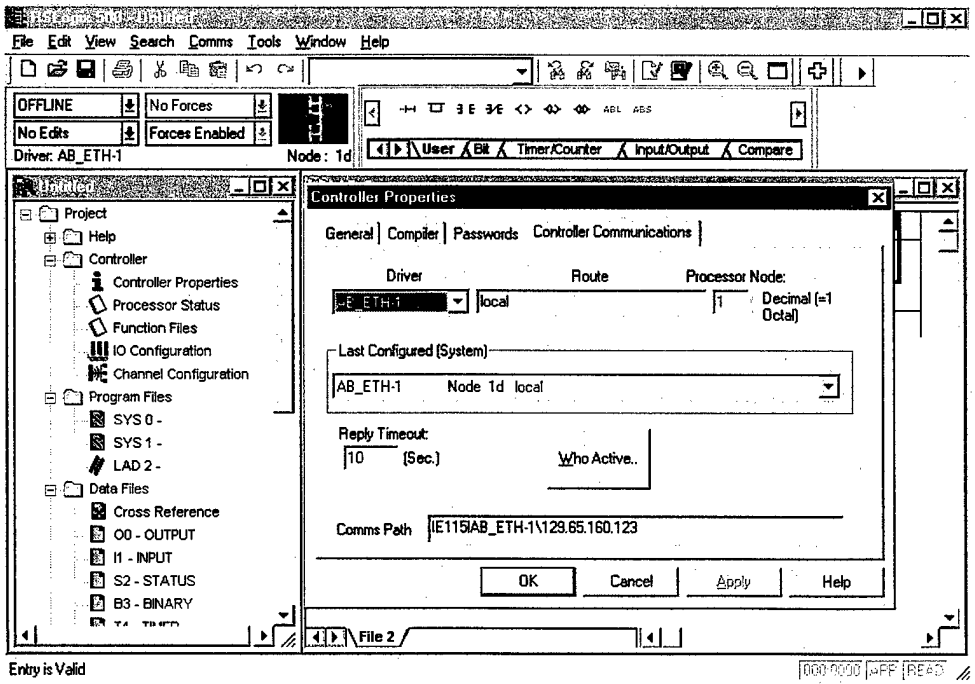


Figure 2. Controller Communications. Select the Ethernet driver.

You must make this work before continuing this lab.

Step 3: Program another PLC on the network.

Find another group and get their PLC's IP address (and give them your IP address). Configure RSLinx to communicate with this additional PLC by entering another station in the Ethernet driver configuration. Refer to Step 1 of this lab for help. Now download your program to their PLC (and vise versa). This is one advantage to an Automation Network (can program multiple PLCs from anywhere on the network).

Once this is completed, go back to programming your own PLC.

Step 4: Write a ladder program that controls a light on another PLC (peer-to-peer messaging).

In this step, you will write a program that changes the state of a binary bit (B3:1/0) on a remote PLC (and they will set a bit (B3:2/0) on your PLC). For this step, you will again need another group's IP address of their PLC.

1. The first rung of your program is very important. It will configure the Ethernet Interface (ENI) box on your cart. To do this, you will use a new instruction "MSG." This instruction is available on the Input/Output tab of the instruction toolbar in RSLogix. Once you have the rung built, click on the "Setup Screen" of the MSG instruction. It will bring up the dialog box. All of the needed details of how to configure this are included below. Your ladder should look exactly like Figure 3.

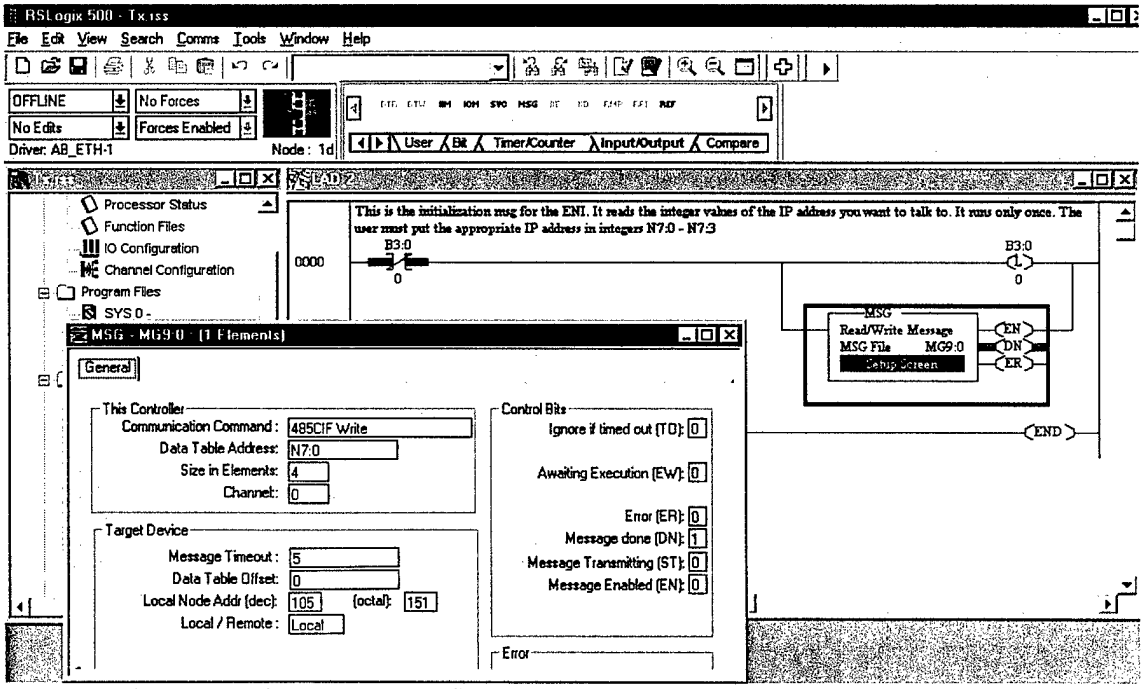


Figure 3. The rung to configure the ENI box. Yours should be identical.

Notice that we use B3:0/0 to make sure that this rung only executes one time, on the initial startup. It will never execute again until the program is restarted.

2. Now we must somehow input the IP address of the remote PLC that we will communicate with. This is done by using the first 4 integers in the integer file (N7:0 – N7:3). It should look similar to Figure 4, but with the appropriate address.

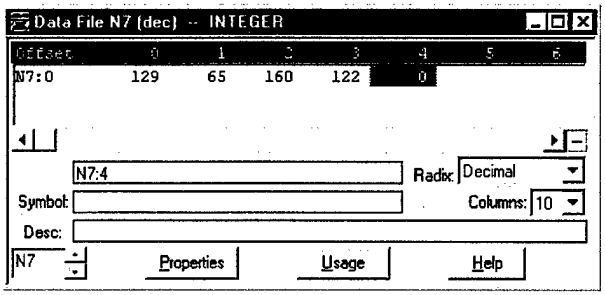


Figure 4. Entering the IP address of the remote PLC into the data file (N7:0-N7:3)

3. Now write a rung that does the following: If the first button is pushed, latch a binary bit AND send that binary word to the remote PLC using another MSG instruction. Once you get the rung setup, then this is what the MSG Setup Screen should look like:

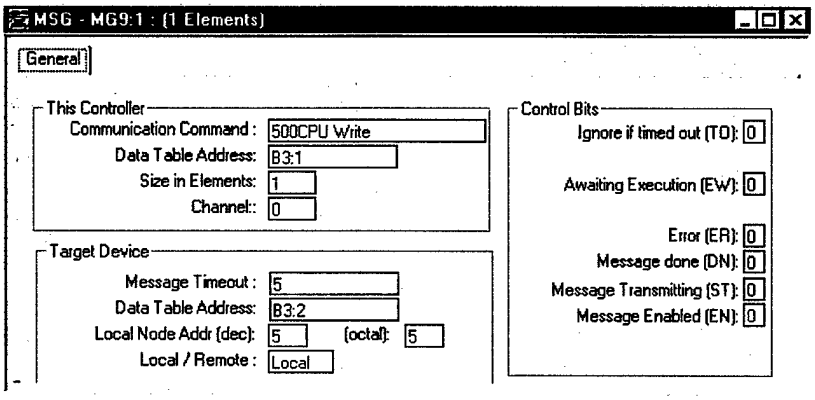


Figure 5. This MSG Setup Screen says, "Write the binary word (B3:1) from this PLC to the binary word (B3:2) on the remote PLC." Whenever this rung executes, the data will be sent across the network to the remote PLC.

4. In the next rung, write this functionality: If the second button is pressed, unlatch the binary bit that you set in the previous rung AND send that binary word to the remote PLC. This is very similar to the previous rung. You figure this one out.

5. Write the ladder logic that illuminates light 1 if B3:1/0 is true (the bit that you set when you press button 1). Also, create a run that illuminates when B3:2/0 is true (this is the bit that the remote PLC will set on your PLC).

Step 5: Write the Report

Each of you will write an ***individual*** report.

The report should be in the form of documentation the work performed in this lab. Document your system so that another engineer could come in, read the documentation and modify your program.

Include all pertinent information, including such things as: how the system operates, instructions used, I/O addresses and similar details.

What to turn in:

Demo your programs to the instructor by the end of the lab. The report is due at the beginning of the next lab period.

Appendix II

Automation Networking – 2

(Analog values and HMI Applications)

The objective of this lab is to build on the previous lab and perform more advanced networking communications, control and HMI applications.

Do not roll the carts to the lab computers for this lab. Instead, plug in an Ethernet patch cable between the ENI box on your cart and the Cisco Ethernet switch. Make sure your cart is powered on.

Step 1: Write a very simple ladder program to verify communications settings are correct.

Similar to the last lab, you must configure the communications driver so you can download and communicate with your PLC over the network. Refer back to the last lab documentation for help on configuring the communication driver in RSLinx.

Using RSLogix, create a very simple ladder and download to your PLC. It can be as simple as examining the first button and turning on a light if the button is pressed.

The whole trick to this step is to correctly identify your PLC in the communications setup. Double click on the Controller Properties (on the left window of RSLogix), and select the Controller Communications tab (see Figure 2). Make sure you have the AB_ETH-1 driver selected and the correct Processor Node.

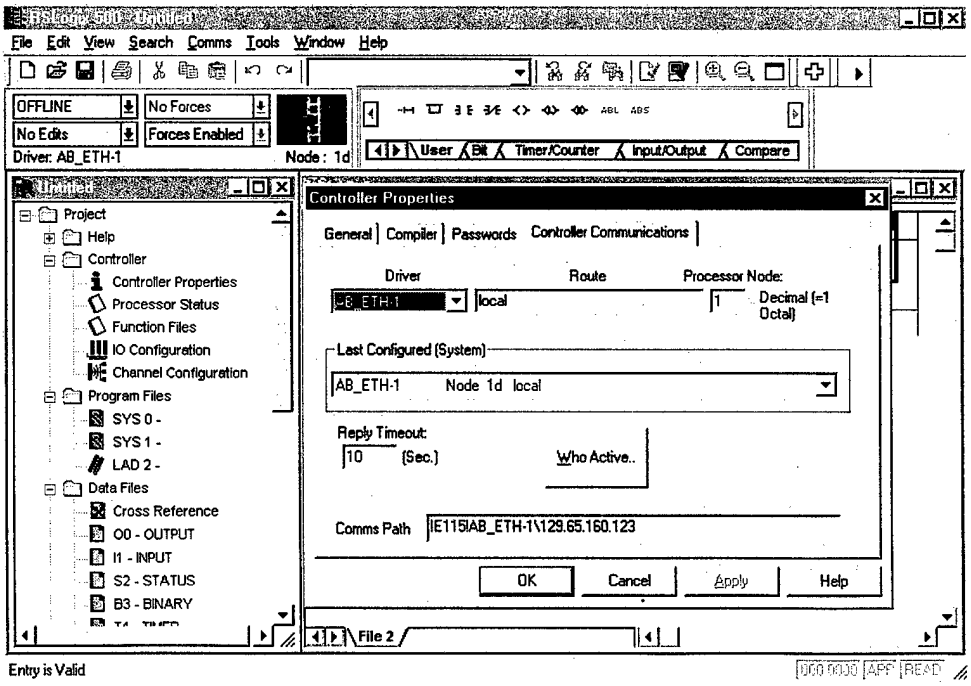


Figure 2. Controller Communications. Select the Ethernet driver.

You must make this work before continuing this lab.

Step 2: Send the analog input value to a remote PLC

In this step, you will send the value from your potentiometer (I:1.0) across the network to a remote PLC. That remote PLC will display the value on their voltmeter (O:2.0) and vice versa (you will display their analog value on your voltmeter).

- You must identify a partner PLC and obtain their IP address.
- Create a new ladder file. Make sure you configure and enable the analog input and output modules.
- The first rung of your program should be identical to the last lab. This is the configuration message that is send to the ENI box.
- Make sure you put the IP address of your partner PLC into integers N7:0-N7:3.
- Next, create a Timer ON that completes every 250 milliseconds. Wire this timer to run continually. You will use this timer to trigger the sending of the analog value over the network.
- One limitation of the message instruction (MSG) is that it cannot send input or output values directly. Therefore, you must store the analog input value into an integer location, then send the integer value using the MSG instruction. So, in the 3rd rung (which is triggered by the DN bit of the timer), perform the following tasks:
    - Move the value from the analog input into an open integer location (N7:4).
    - Using a MSG instruction, send the value in N7:4 to the remote PLC's N7:5.
    - Reset the timer so the whole process can continue.
- Finally, in the last rung, display the value that is put into your N7:5 location (by the remote PLC) on your voltmeter (O:2.0).
- Download and test the functionality. Troubleshoot if needed.

If everything goes right, when you turn your POT on your panel, the voltmeter on the remote PLC should move and vice versa. Notice the behavior of the voltmeter. Is it smooth or jerky? Why? How might you tweak its behavior? What is the advantage or disadvantage to tweaking this parameter?

Step 3: Add digital input/output over the network and build an HMI.

Now, add digital input/output functionality into your existing ladder. This should be similar to the last lab. You are free to do as simple or sophisticated digital control as you wish (perhaps you could use a timer, counter, sequencer or some other instruction in some capacity).

Once your digital functionality is working, using RSView, create an HMI to monitor and control your system. Not much in the way of direction in this step, you are free to experiment with the ladder logic and HMI as you wish. The result should exhibit your understanding of the concepts covered in your application.

Step 5: Write the Report

Each of you will write an ***individual*** report.

The report should be in the form of documentation the work performed in this lab. Document your system so that another engineer could come in, read the documentation and modify your program.

Include all pertinent information, including such things as: how the system operates, instructions used, I/O addresses and similar details.

What to turn in:

Demo your programs to the instructor by the end of the lab. The report is due at the beginning of the next lab period.

Range Sensing and Real-Time Registration

Project Investigator:

Fred W. DePiero, Ph.D.
Associate Professor
Electrical Engineering Department

Final Project Report
Range Sensing and Real-Time Registration (C3RP - 54360)

Fred DePiero, fdepiero@calpoly.edu, (805) 756-2917
Electrical Engineering Department, CalPoly State University

We would like to express our sincere thanks to the Office of Naval Research and the California Central Coast Research Partnership for sponsoring this project.

Research Goals and Project Definition

The goal of this research is to develop new algorithms that rapidly align surface data. Range sensors acquire this type of data, and depending on an application, sensors will typically need to be located at different vantage points in order to observe all the surfaces of an entire scene. Rapid and deterministic alignment of the data acquired from different vantage points is a key technical hurdle that will enable many new applications.

Registration, or alignment, of sensor data is necessary because range sensors are line-of-sight devices. Hence only a portion of a scene can be observed from a given location. For example if data from the surface of a planet is to be collected by a satellite-based sensor, then views of each hemisphere would require alignment. The alignment is needed so that the entire data set (planet surface) is expressed with respect to the same coordinate frame – located at the planet center, for example. Such alignment is non-trivial if the sensor motion is not know precisely, which is common for sensors on mobile platforms. Once the sensor location for each data set is known, the acquired data can be combined directly into a digital map.

Other application areas for range sensing and registration include control and remote viewing. For example, a location under surveillance by an airborne sensor could be compared against previous observations stored in a digital map. If the current view is registered to a stored view, then changes could be identified. Rapid and deterministic processing would allow such information to be used for targeting purposes, for example. Remote viewing applications could also be greatly enhanced with a real-time registration capability, particularly if both range and color data is acquired. A digital elevation map augmented with natural color data would permit views to be rendered from an arbitrary observation point (distinct from a sensor location). Rendering pairs of such views would benefit virtual reality applications.

The goal of this project has been to investigate a new method for registration that is based on landmarks. We seek methods of alignment that are based solely on the examination of scene content, without knowledge of sensor movement. This is the more general case. We are also interested in combining both range and color data for remote viewing. To facilitate robotic control and remote viewing applications, a processing rate of 10Hz is targeted as an eventual goal.

Accomplishments

In this past year the basic registration algorithm was first implemented and tested. Results were compared against a traditional algorithm and are quite promising. Tests showed that the new landmark-based approach could achieve results faster and more accurately than traditional methods, with an accuracy of 0.6 degrees and a

deterministic processing rate above 4 Hz rate. (See Appendix 2 and 3). Efforts in the first year focused on the alignment of aerial range data, such as would be associated with a sensor on an airplane or satellite.

In addition to the algorithmic development, an inexpensive range sensor was also built. See Appendix 1. The sensor is quite useful for testing and algorithm development because it permits scene content and sensing conditions (sensor position, for example) to be controlled. A relief map was scanned to generate 'aerial' range data in our lab. The map was positioned with a micrometer drive to provide known test conditions.

We are grateful that funding permitted a graduate student to be employed on this project. Our team is expanding to 3 students in the 2002-2003 year.

Future Efforts

Planned efforts for a 2nd year (2002-2003) include improvements to our registration algorithms, sensing capabilities, and computing environment.

We are improving the in-house sensor to acquire both range and color data. This will support our investigation of remote viewing applications. We are also switching our focus to indoor scenes. Algorithmic improvements will target better accuracy of the location and description of landmarks in the range data.

A distributed computing platform is being setup for the registration system. Four Linux machines are being configured in a ring topology to run the registration algorithms and support the rendering calculations for viewing. The goal for 'the ring' is process data at a 10Hz rate. Although the improved sensor will not acquire data at this rate, the system will support rapid playback of stored sensor images. Our intention is to demonstrate the viability of real-time processing with the registration algorithms, if married with a state-of-the-art range sensor (that cost $100k).

Technical Publication

Appendix 2 contains a publication that appeared in the Proceedings of the 1st Int. Symposium on 3-D Data Processing, Visualization and Transmission (3DPVT) and was presented in a poster session at the conference in Padova, Italy June 19-21, 2002. Appendix 3 has the slides used in the poster presentation.

Fast Landmark-Based Registration via Deterministic and Efficient Processing, Some Preliminary Results[Ω]

Fred DePiero
CalPoly State University, San Luis Obispo, CA 93407, fdepiero@calpoly.edu

Abstract

Preliminary results of a new method for range view registration are presented. The method incorporates the LeRP Algorithm, which is a deterministic means to approximate subgraph isomorphisms. Graphs are formed that describe salient scene features. Graph matching then provides the scene-to-scene correspondence necessary for registration. A graphical representation is invariant with respect to sensor standoff. Test results from real and synthetic images indicate that a reasonable tradeoff between speed and accuracy is achievable. A mean rotational error of ~1 degree was found for a variety of test cases. Mean compute times were found to be better than 2 Hz, with image sizes varying from 128x200 to 240x320. These tests were run on a 900 MHz PC. The greatest challenge to this approach is the stable localization and invariant characterization of image features via fast, deterministic techniques.

1. Approach Based on Subgraph Matching

The goal of this research is to pursue a technique that can perform view registration at rates approaching 10 Hz, without any user input for initial estimates. This performance goal is set to match the data rates of range cameras. It is also desired to have a method that computes the rigid transformation (translation plus rotation) in the presence of possible scale changes. Registration at these rates could permit sensor motion to be tracked in real-time. This would permit unconstrained movement of a sensor across a large scene.

To achieve fast and deterministic processing, iterative [1] [8] [13] [17] [18], compute intensive [10], or random [6] approaches were avoided. Established methods do not typically separate the steps of determining corresponding points and determining the transform. This limits compute speed. In the new approach these steps have been kept separate, and are implemented in a non-iterative fashion. This is an important difference for the new approach. Another difference is that correspondence between the data sets is determined only for select feature points. This improves processing speed, but it does limit accuracy.

Correspondence between the data sets is determined via graph matching. Graphs are formed using salient features in each range image. Graph matching is accomplished using the LeRP Algorithm [2][4]. LeRP approximates a subgraph isomorphism via a deterministic procedure, based on the comparison of length-r paths. The LeRP algorithm yields a set of corresponding locations in the two input range images, from which the absolute orientation may be found via closed-form solution [9].

A typical graph appears in Figure 1. The white segments designate an accurately matched subgraph for the scene.

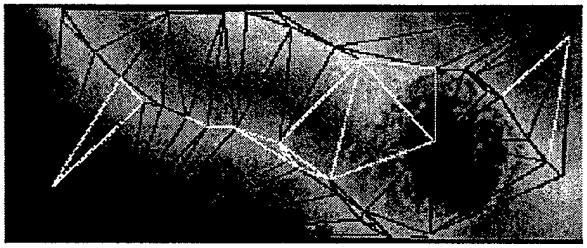


Figure 1. Graph associated with Mars Odyssey data. Nodes are associated with local maxima. Edges are determined via a Delaunay triangulation. White edges indicate a matched subgraph that contains nodes with an accurate mapping into the other scene.

1.1. Approach – Salient Features

Graphs describe the salient features in each range image. For rapid detection, local, well isolated, peaks in the range data were used to form the image features. These are desirable because they are likely to remain in view with small shifts in scene.

The significant percentage of black segments shown in Figure 1 demonstrates a lack of stability in feature detection. Stable and invariant feature detection – via fast, efficient, deterministic means – remains the greatest challenge in this new approach. While the lack of stability shown in the figure is undesirable, it was deemed an appropriate tradeoff in terms of processing speed. Other techniques employ more robust local features [10] but require more processing time. The approach taken here

[Ω] Project sponsored, in part, by the Department of the Navy, Office of Naval Research.

was to rely on the LeRP matching algorithm to determine appropriate correspondences, despite noisy features.

Some reported techniques use invariant features that involve curvature, moments, or spherical harmonics [18]. These types of features react to jump discontinuities that may occur at a limb [19]. Such feature points were avoided in this new approach. Consider the occurrence of a limb where the line-of-sight of a sensor becomes tangent to a hillside. Slight movement of the sensor would alter the tangential viewing conditions. Hence jump discontinuities may be unstable in some situations. Ridge curves [20][21] or simply the use of isolated peaks are believed to be more stable.

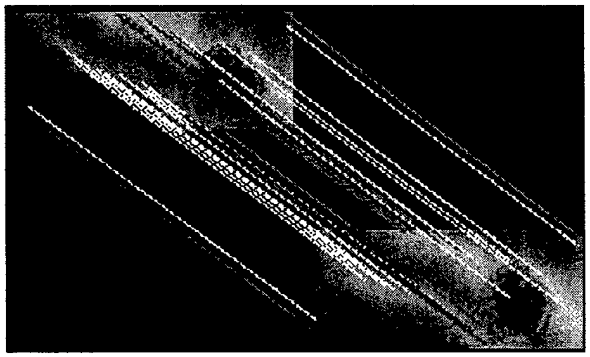


Figure 2. The connecting line segments indicate correspondences found via graph matching. White lines indicate acceptable matches and gray lines show correspondences that failed a test of the residual fit error.

To improve speed and to reduce sensor noise, the range images were subsampled by a factor of 8, making images 1/64 the original area. A 3x3-averaging kernel was used repeatedly, to subsample. During the subsampling process, occurrences of pixels with missing range data (due to occlusions, for example) were reduced in number.

When finding local peaks in the range data, a 3x3 window was used in the smallest subsampled image. This corresponded to a 24x24 window in the original image. A peak was defined as being a pixel with at most 1 other higher (3x3) neighbor.

To help recover the accuracy lost by peak detection in the subsampled images, feature locations were adjusted along the local gradient. A fixed number of steps, S-2, were used for an S-factor of subsampling (S=8, here).

1.2. Approach – Graphical Representation

Nodes were associated with each feature (local peak) in an image and were colored by the sharpness of the peak. Sharpness was computed by finding the volume enclosed under a fitted 3x3 surface patch, centered at the peak.

The existence of an edge was established via a Delaunay triangulation [5] [16]. Distances between nodes were used to color edges. (Distances were computed in world coordinates, not just a pixel distance, making the edge coloring tolerant to standoff changes.) This formed an object-centered representation that could be compared without first aligning the range images.

A Delaunay triangulation was chosen because it is efficient, requiring on the order $O(\mathbf{F}^2)$ effort, for **F** features [5]. This is the case for a 2-D mesh, associated with the 2 ½-D range image. A Delaunay triangulation is also invariant with respect to translation and rotation, and with respect to node ordering. Also, the triangulation yields a graph (as opposed to a tree) and this works better with the LeRP algorithm.

1.3. Approach – Finding Subgraph Isomorphisms

Noisy sensor data introduces some fundamental limitations to the repeatability and stability of extracted features. This limits the similarity of scene graphs. Varying occlusion with different sensor viewpoints also limits the similarity of scene graphs. For these reasons, graphs made from real sensor data tend to be imperfect representations. Since the graphs are imperfect, an exact method of graph matching has limited use – and consumes inappropriate compute time. Hence using an approximate method of graph matching is a reasonable approach.

The LeRP algorithm [4] [11] approximates subgraph isomorphisms by comparing the number of length-r paths in each graph. These are found via $\mathbf{A}^{\mathbf{R}}$, where **A** is the adjacency matrix [7]. LeRP identifies the node-to-node mapping between the graphs by finding matching path counts, favoring assignments with higher values of **R**. As nodes are added to the mapping, the structural consistency with previously mapped nodes is enforced.

Node and edge colors are also compared during the matching process. As these colors are continuous quantities, some threshold on color differences was needed to verify similarity. A threshold of 2% was used for node colors and ½% for edge colors, in all the tests reported herein.

Processing effort for the graph matching is on the order of $\mathbf{O(F^3D^2R)}$, where **F** is the number of nodes, **D** is the average degree. The parameter **R** is actually a weak function of **F** (see [4]) but was set to a constant in all tests reported herein. If the number of features increased significantly, then additional subsampling would be appropriate to maintain reasonable processing times. This would permit **R** to remain constant. In the reported trials,

F ranged from 25-80, approximately. Of these nodes, 20%-40% were typically matched.

1.4. Algorithm – Absolute Orientation

Horn's method [9] was used to determine absolute orientation. This reveals the rigid transformation as well as any scale changes. It operated on the corresponding features identified by the graph matching routine.

The residual fitting error of each corresponding feature was checked. The process of finding the transform and checking consistency was repeated (fewer than) **F** times, for **F** local peaks. Hence the effort in this stage of the processing is bounded by $O(\mathbf{F}^2)$.

2. Summary of New Algorithm for Registration

1) Find salient features in range images.
2) Refine feature locations.
3) Form graph. Use Delaunay triangulation to establish edges and compute feature descriptions.
4) Match graphs with LeRP algorithm to find corresponding features.
5) Find absolute orientation and check residual error.

Summary of Processing Effort

Step	Sub-Step	Effort
Find Range Features	Subsample	O(N M)
	Find peaks	O(N M)
Refine Features	Refine peak locations	O(F)
Form Graph	Delaunay triangulation	$O(F^2)$
	Find feature descriptions	O(F)
Match Graphs	LeRP algorithm	$O(F^3D^2)$
Find Absolute Orientation	Horn's technique	$O(F^2)$

Table 1. Summary of processing effort required for each step in the algorithm. Each step has a polynomial bound on worst-case effort. Images were NxM, containing F features. D is the mean degree of feature graphs.

Each of the above steps requires worst-case effort that has polynomial bound. The effort indicated assumes: **NxM** range images and **F** range features. Note that there are additional parameters that effect the processing time, such as the degree of subsampling for the original range image, and the window size used when looking for range features. These were omitted from the Table 1, for clarity.

3. Testing

Referring to Table 2, three types of range data were used. The 'NASA' data was from the Odyssey probe, acquired Nov. 13, 2001. This is actually an intensity image in the visible - thermal spectrum, not a range image. The 'Sensor' data was acquired by scanning a relief map of the Great Smoky Mountains National Park. The 'Synthetic' data was generated randomly. The size of the various images is given in the table.

The 'Sensor' images appearing herein were acquired using a structured light range sensor, similar to [3]. The sensor used a laser line generator that casts a plane of light from a diode source. An inexpensive web camera then observed the intersection of the laser illumination with an object in the scene. One profile of an object was revealed in each camera image. A linear positioner advanced the laser and camera across the scene. Sensor calibration is also described in [3].

This sensor is relatively slow; relative to available 3-D range cameras. Furthermore, it is these range cameras that are driving the goals for the new registration technique. Hence range images from the sensor were acquired and then stored for use in test trials. Also note the time necessary to load the range images was not included in the measures of execution time that are presented below.

3.1. Testing - Results

Image data was resampled at a random rotation in each test trial. The measured and true rotation angles were compared to estimate the accuracy of the registration measurement. Accuracy is described in terms of the mean absolute error of the rotation angles.

The ICP algorithm was used for comparison purposes [1][18]. The simplex optimization routine [15] was used to implement ICP. To reduce effort for ICP, only lateral displacements and a rotation in the image plane were optimized (3 DOF, not 6). This was done for simplicity. Also, test cases were constructed that made it easy for the ICP algorithm to converge in each case. In this way an apples-to-apples comparison with the new algorithm could be more easily made.

The nominal rotational difference was 5 degrees. This small rotational difference helped ensure the convergence of the ICP algorithm. Exactly 20 iterations of ICP were run, i.e. the termination was deterministic - resulting in a varying accuracy, rather than a varying processing time.

See Table 2. Twenty-five trials were run for each data set. Tests were run on a 900 MHz PC.

Results show that the new algorithm can achieve processing rates of 2 Hz or better. This is considered good. However, the new algorithm did not always yield 100% successful results (as in the 75 trials reported below). This was more common in test cases with higher rotational offsets. Improved feature stability should help improve reliability and overall accuracy.

Data Set	Size of Data Set	Abs. Rotational Error [New]	Abs. Rotational Error [ICP-3DOF]	Duration [New]	Duration [ICP-3DOF] (20 Iterations, Fixed)
NASA	128x320	0.78 Deg	0.46 Deg	0.26 Sec	1.6 Sec
Sensor	128x200	1.1 Deg	0.56 Deg	0.28 Sec	1.6 Sec
Synthetic	240x320	0.59 Deg	0.45 Deg	0.45 Sec	1.6 Sec

Table 2. Accuracy and execution speed for image registration test trials. Results indicate that a reasonable tradeoff between speed and accuracy is achievable. Twenty-five trials were run for each data set. For simplicity, only 3 DOF were optimized in the ICP routine and iterations were limited to 20.

4. Conclusion and Future Studies

The graph-based approach appears to provide a desirable tradeoff between speed and accuracy. It is possible that this type of approach could eventually be used to determine sensor motion on-line.

This new approach has its share of challenges. Most significant is the stable computation of invariant features, when computed via deterministic techniques that permit high-speed registration. The method employed herein is simplistic and further study and refinement is appropriate.

Future extensions include marrying the new technique with ICP in a post-processing step. ICP could be run with a fixed number of iterations to help improve accuracy (as all the image data would then be employed). Processing video streams could be enhanced by using results from previous images to predict subsequent scene conditions. Also, some experiments have been performed using features that are based on ridge curves [20]. These appear to be more stable than the simple method using isolated peaks described herein.

5. Acknowledgement

The Mars images from the NASA Odyssey probe were downloaded from the JPL website.

6. References

[1] P.J. Besl, N.D. McKay, A method for registration of 3-D shapes, IEEE Trans. Pattern Analysis and Machine Intelligence, 14 (2) (1992) 239-256.

[2] F. W. DePiero, M. M. Trivedi and S. Serbin, Graph Matching Using a Direct Classification of Node Attendance, Pattern Recognition Journal 29 (6) (1996) 1031-1048.

[3] F. W. DePiero and M. M. Trivedi, "3-D Computer Vision Using Structured Light: Design, Calibration and Implementation Issues," AIC, vol.43 (1996).

[4] F. W. DePiero and D.W. Krout, LeRP: An algorithm using length-r paths to approximate subgraph isomorphism, Accepted by Pat Rec J. See www.ee.calpoly.edu/~fdepiero.

[5] O. Faugeras, 3-D Computer Vision – A Geometric Viewpoint, MIT Press, Cambridge, MA, 1993.

[6] M. Fischler, R. Bolles, Random Consensus: a paradigm for model fitting with applications in image analysis and automated cartography, Communications of the ACM, 24 (1981) 381-395.

[7] L.R. Foulds, Graph Theory Applications, Springer-Verlag, New York, 1992.

[8] B. Luo and E.R. Hancock, Structural graph matching using the EM algorithm and singular value decomposition, IEEE Trans. Pattern Analysis and Machine Intelligence, 23 (10) (2001) 1106-1119.

[9] B.K.P. Horn, Closed-form solution of absolute orientation using unit quaternions, J. Optical Society of America A, 4 (4) (1987) 629-642.

[10] A. E. Johnson and S. B. Kang. "Registration and integration of textured 3-D data." Image and Vision Comput., 17, 135-147, '99.

[11] D.W. Krout, LeRP: An Algorithm for Finding Subgraph Isomorphisms with Applications to VLSI, Master's Thesis, CalPoly State University, San Luis Obispo, CA ('01).

[12] N. Nikolaidis, I. Pitas, 3-D Image Processing Algorithms, John Wiley & Sons, NY (2001).

[13] C. Kapoutsis, C. Vavoulidis, and I. Pitas, Morphological iterative closest point algorithm, IEEE Trans. On Image Processing, v. 8, no. 11, pp 1644-1646, Nov. (99).

[14] Tefas, C. Kotropoulos, I. Pitas, Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication, IEEE Trans. Pattern Analysis and Machine Intelligence, 23 (7) (2001) 735-746

[15] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C, Cambridge University Press, NY, 1988.

[16] J.R. Shewchuk, Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator, Proc. First Workshop on Applied Computational Geometry, Philadelphia, Pennsylvania, pages 124-133, ACM (May 1996).

[17] S. Hsu, Multiple-view constrained video registration and its applications, Workshop on Video Registration, IEEE Computer Society, Vancouver Canada, (July, 2001).

[18] G. C. Sharp, S. W. Lee, D. K. Wehe, ICP registration using invariant features, IEEE PAMI, 24 (1) (2002) 90-102.

[19] L.G. Shapiro, G.C. Stockman, Computer Vision, Prentice-Hall, NJ, 2001.

[20] J. T. Kent, K. V. Mardia, J. M. West, Ridge curves and shape analysis, British Machine Vision Conference, University of Edinburgh (1996).

[21] X. Pennec, N. Ayache, J-P. Thirion, Landmark-based registration using features identified through differential geometry, Handbook of Medical Imaging, I. N. Bankman ed., 499-513, Academic Press (Sept 2000).

[22] A. E. Johnson, Surface landmark selection and matching in natural terrain, IEEE Computer Vision and Pattern Recognition, v. 2, 413-420 (2000).

Appendix I

Appendix – LeRP Algorithm for Approximating Subgraph Isomorphism

Main Routine

Input: Graph **G** with nodes g_i, $0<=i<N_G$ and Graph **H** with nodes h_k, $0<=k<N_H$

Output: Mapping **m()**, that gives $h_k = m(g_i)$.

Steps:

1. Compute powers of adjacency matrices A^R and B^R for graphs **G** and **H**
1. $beta_{peak}[][] =$ ***find_best_beta***(G,H,A^r,B^r)
2. Clear node-to-node mappings
3. For each **L**, $0<=L<minimum(N_G,N_H)$
    a. Let **peak = 0**
    b. For each unmapped node g_i
    c. For each unmapped node h_k
        i. Verify consistency of mapping g_i to h_k given current **m()**
        ii. **rho = 0**
        iii. For each mapped edge e_{ij}
            1. lookup associated edge e_{kl} where **l=m(j)**
            2. **beta =** ***compare*****(i,j,k,l)**
            3. **gamma =** ***compare*****(j,j,l,l)**
            4. **rho = 1 – (1-rho)(1-beta)(1-gamma)**
        iv. Next **j**
        v. **alpha =** ***compare*****(i,i,k,k)**
        vi. **rho = 1 – (1-rho)(1- alpha)(1-** $beta_{peak}$**[i][k])**
        vii. If **rho>peak** Then
            1. g_{peak}**=i**
            2. h_{peak} **=k**
            3. **peak=rho**
        viii. End If
    d. Next **k**
    e. Next **i**
    f. If **peak=0** Then GoTo **END**
    g. Let $m(g_{peak})=h_{peak}$
4. Next **L**
5. If **(L=**N_G**)** and **(L=**N_H**)** Then **G** *is ISOMORPHIC to* **H**, refer to mapping **m()**.
6. Else a subgraph isomorphism exists between **G** and **H**, refer to mapping **m()**.
7. END

Function: find_best_beta(G,H,A^r,B^r)

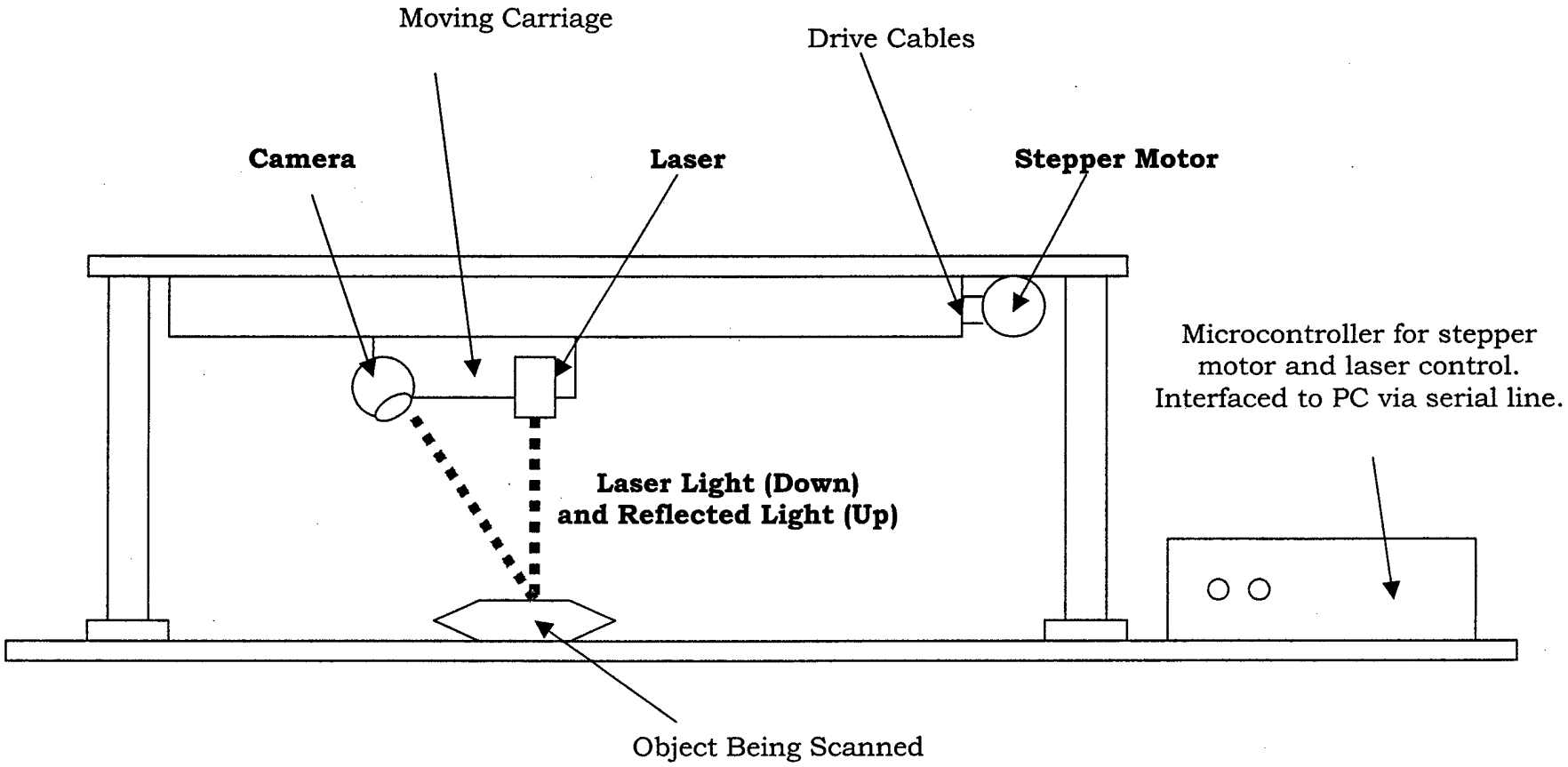
a. For each node g_i
b. For each node h_k
    i. For each edge e_{ij}
    ii. For each edge e_{kl}
        1. **beta =** ***compare*** **(i,j,k,l)**
        2. Save $beta_{peak}$**[i][k]=beta** if maximal for nodes i,k
    iii. Next **l**
    iv. Next **j**
c. Next **k**
d. Next **i**
e. Return $beta_{peak}[][]$

Function: compare(i,j,k,l)

1. For **1<=r<=R**
    a. If $a_{ij}^{(r)}$!= $b_{kl}^{(r)}$ Then Break
2. Next **r**
3. Return $(r/N)^2$

Appendix II

Structured Light Range Sensor



Fast Landmark-Based Registration via Deterministic and Efficient Processing, Some Preliminary Results

Fred DePiero

Electrical Engineering
CalPoly State University, San Luis Obispo, CA, USA

Goals for Registration Target 3-D Data Streams

- Desire registration techniques that are suitable for use with range + color sensors.
- *Goal* for registration rate: 10 Hz, or better.
- Challenge: Provide stable and accurate registration for 3-D data streams, at sensor data rates.

Fred DePiero, CalPoly State University, San Luis Obispo, CA, USA

Real-Time Range Acquisition and Registration Enables Many Applications

- Military: Improved Surveillance

  Complete views of people, for viewing, tracking or targeting within a room.
- Military: Improved Aerial Reconnaissance (RST-V)

  Sense new buildings and equipment, immediately register against existing maps for targeting or reporting.
- 3D-TV with Arbitrary View Direction
  - Viewing direction and standoff selected by user, not director.
  - Requires range and color data across all surfaces, registered (aligned) before transmission.
  - 'TV' replaced with graphics workstation for rendering. Use one for each eye, with appropriate interocular spacing.
- Telemedicine
  - Actuated scalpel that limits penetration depth, for example during removal of lung tissue from a breathing patient. Range sensing and registration needed for volumetric modeling of lungs.
- More (non-real-time) applications:

  CAD modeling, object recognition, planet mapping, robotic surgery

Fred DePiero, CalPoly State University, San Luis Obispo, CA, USA

Non-Iterative Approach

1. Reduce resolution of images by 8x in each direction
   Using successive applications of 3x3 averaging kernel.
2. Find salient features in each, compute description.
   Find local peaks or ridge points (3x3 window).
   Find sharpness of peaks (~volume under 5x5 window).
3. Form graph connecting neighboring features.
   Delaunay triangulation to define edges.
   Peak sharpness colors nodes, inter-peak distance colors edges.
4. Match graphs to find corresponding locations.
   LeRP Algorithm to approximate subgraph isomorphism, by DePiero.
   Apply tolerance check when comparing graph coloring.
5. Use Horn's method to find alignment transformation.
   Prune point-pairs based on residual error, 3-passes only.

Fred DePiero, CalPoly State University, San Luis Obispo, CA, USA

LeRP Algorithm Requires Deterministic Calculations

- Approximates maximal subgraph
- Fast

  25 nodes in 0.1 sec

  50 nodes in 0.5 sec

  (Comparing 100 node vs. 50 node graphs w/ 50 node matching subgraph).
- Handles low overlap

Fred DePiero, CalPoly State University, San Luis Obispo, CA, USA

Preliminary Results Indicate Approach Has Merit

- Non-iterative.
- Faster than ICP, for a given accuracy.
- Aerial data, resampled at varying orientations.
- Potential to tolerate scale changes.
- *Future Challenge*:

  Improve feature stability

	New	**ICP** *60 iterations*
Rate	4.75 Hz	0.25 Hz
Error *Degrees*	0.14 Median 0.61 Mean 1.90 S.Dev.	1.60 Median 2.27 Mean 2.83 S.Dev.

- True Rotation varied 2-20 degrees.
- Using absolute error, mean rate.
- 320x125 Range Images.
- All 55 trials reported, all successful.
- Benchmarked on a 900MHz PC

Fred DePiero, CalPoly State University, San Luis Obispo, CA, USA

SIPTool Development Platform: Versatile, Easy to Use & Free

- Multimedia demos and programming environment.
- Designed for student projects
- Uses Visual C++, Hides details of Windows programming
- User Configuration via a text file

  Number of tab sets, type of tab children, user menu
- Many Data Window Types:

  Sound I/O, WAV I/O, Camera, Image, Range, 2-D plot, 3-D plot, Slider Controls, Color Images
- No barriers to integration:

  Serial, Sockets, Cameras, Microphone, Acquisition boards.
- Math support routines:

  Simultaneous equations, Eigenvectors/values, FFT.

Fred DePiero, CalPoly State University, San Luis Obispo, CA, USA

When Johnny's Dad Can't Read: Using Technology to Close the Adult Literacy Gap

Project Investigator:

Roberta J. Herter, Ph.D.
Associate Professor
University Center for Teacher Education

Table of Contents

Table of Contents 1
Language eXperience Environment (LXE) 2
Phase I Final Report 2
Executive Summary 2
About the Prototype 3
System Overview Chart 4
Internet Data Miner 4
Internet Data Miner 5
Word Master Maintenance 6
Word List 6
Definitions 6
Relationships 7
Questions 7
Word Classes 7
Phonetic Rules 7
Students 7
Voice File Recorder 8
Exam Module 9
Question types 11
Summary 13
Credits 13

Language eXperience Environment (LXE)

Phase I Final Report

July 29, 2002

Executive Summary

In Phase I the team finished a functional proof of concept prototype, a vision for Phase II, and identified customers with specific needs for the extended version of this learning tool. The LXE project goal is to provide technology support to the process of language literacy instruction for adult students. While initially proposed as a 1st language (English) literacy system, the team now believes that it also has application to 2nd language learners. This possible 2nd language use is currently a critical area to the DoD and other government agencies and we are pursuing specific application to these needs.

In our initial Phase I proposal we discussed development of a broad based literacy training system but as our effort evolved we realized that much of the presentation systems discussed could be purchased from various vendors. What we discovered was that the existing literacy assessment tools lack the ability to intelligently determine what it should present as the next activity. This has evolved as the focus of our research and development.

Language literacy instruction requires the management of large volumes of data consisting of potentially millions of facts. LXE assists the instructor by maintaining a database of the student vocabulary and the training experience. This data will be used in our next generation of software to make intelligent decisions for the next activity based on the knowledge of the specific student. Adult learners are goal orientated readers and differ from general education in that they are highly focused on the goal. The Internet represents the world's largest collection of reading materials, and LXE leverages that by allowing the instructor to construct a vocabulary from any Internet document or set of documents.

LXE supports eleven different types of questions including flash cards, true/false, multiple choice, matching, and word sequencing. Each question type is designed to test and record experiences in reading, writing, listening, speaking, and comprehension. LXE makes extensive use of audio, graphic, and text output as well as voice, keyboard, and mouse inputs. This allows questions that target specific instructional goals to be built, delivered, and monitored by the system.

The future of LXE is in what we are calling Next-Activity logic. While consensus about the best instructional form is tenuous, most agree that a one-on-one human tutor is most effective. However, the challenge with human tutors is the cost and the inability to scale the solution. Our future development will strive to incorporate the success secret of the human tutor model, which is the ability to intelligently select the next activity.

While Phase I is not fully scalable technology, it lays the groundwork for Phase II. The database design supports multiple students, thus this completed product would be usable for trainees throughout an entire organization and the training data available to management. Our core element is the definition, not the word, and we hope this will provide us with language neutrality. Of course, this is as yet an unproven goal.

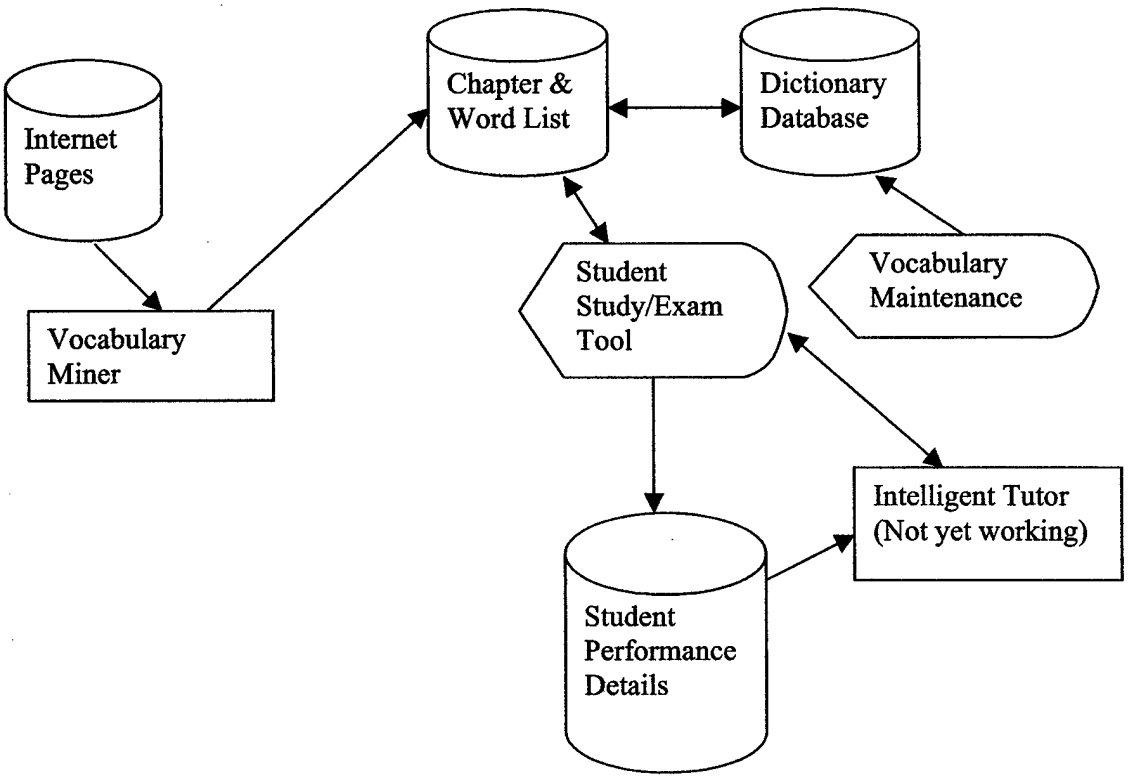
LXE is an integrated tool to assist the student in language skill acquisition using an assessment driven training model. In the following section we will present the results of the Phase I funding in the form of our prototype.

In conjunction with this project, Dr. Herter involved her Education 530 students with the practical function of building materials for literacy assessment and instruction. This resulted in practical work for the students and data for our system tests. The data has been gathered, although not yet tested through the system. The student work brought a few new items for consideration and will likely influence some of the Phase II effort.

About the Prototype

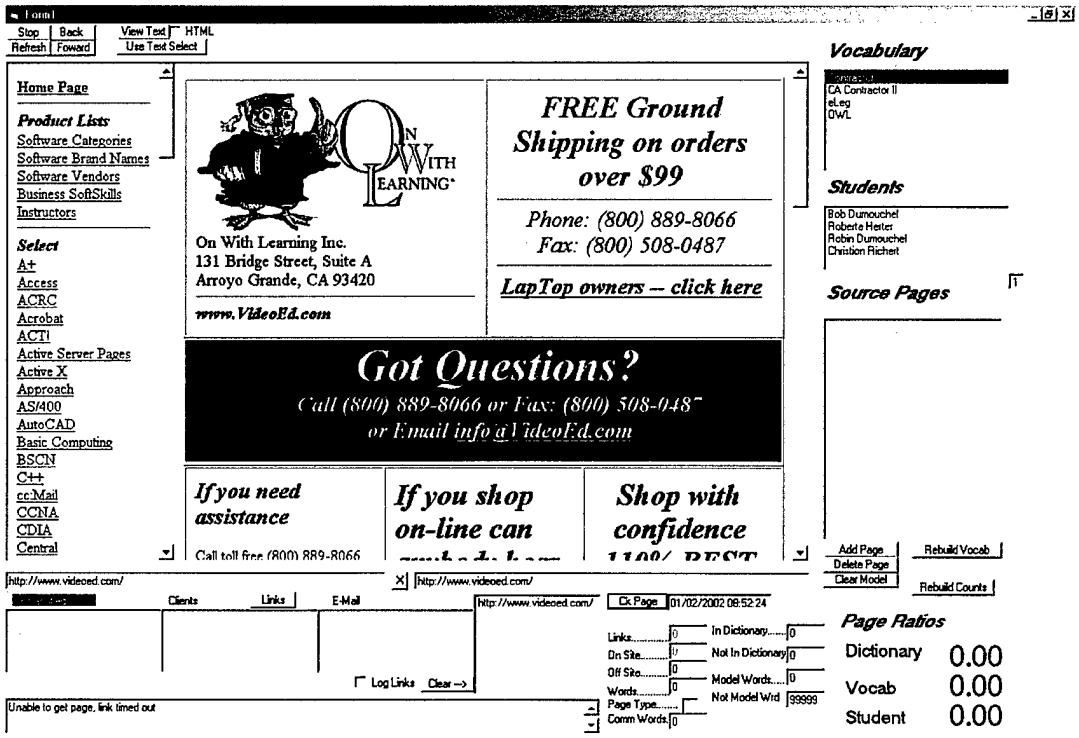
The prototype applications were created as proof-of-concept software. In Phase II we intend to take what we have learned and extend that into a beta version (working product tested in real world environment) level that could be place in limited field situations in order to get reactions to our training approach. In the following section, we discuss the functions and relationships of the proof-of-concept applications.

System Overview Chart



In the Vocabulary Miner application the instructor finds Internet documents that represent the objective vocabulary. This assigns the document and its associated words to the instructional chapter, which drives the study/exam tool. Once a word is in the system the Vocabulary Maintenance allows the attachment of supporting voice files, definitions, word relationships, and other extended information. The output of the student study/exam tool is the student performance details. This is data that we intend to use in Phase II to support the Intelligent Tutor's next activity logic.

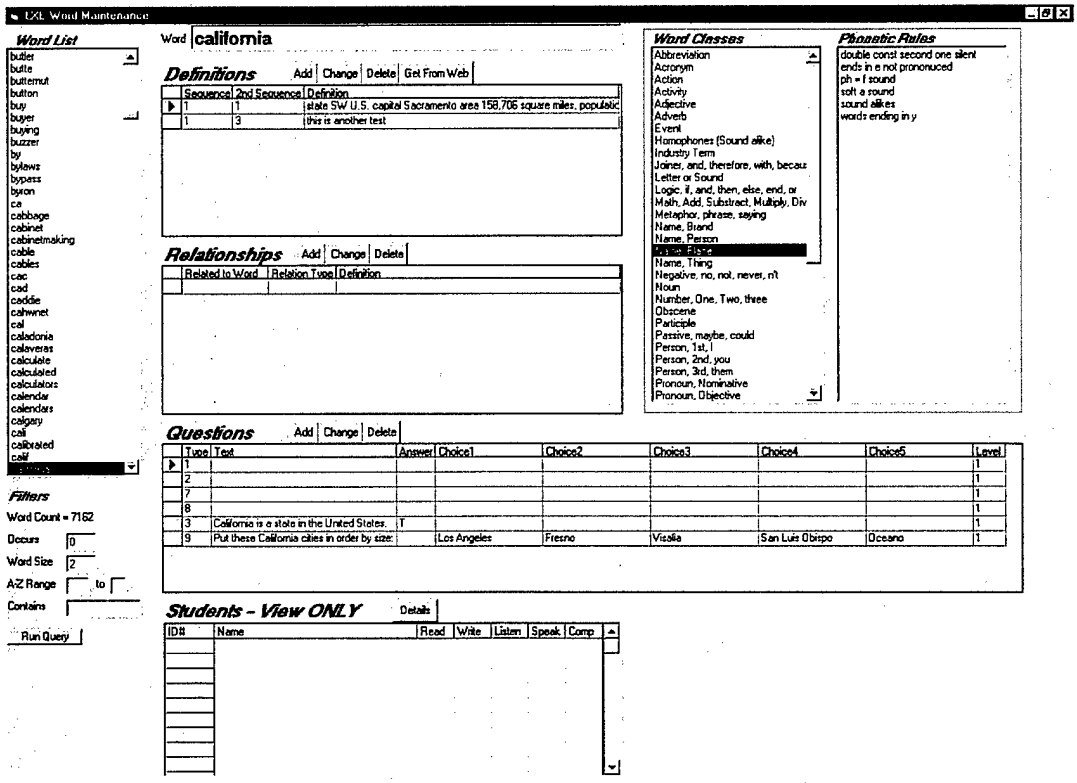
Internet Data Miner



The Vocabulary Miner application provides the instructional designer with the ability to link Internet resources to specific vocabularies. Adult literacy students are goal specific readers and goal can be to read a collection of document. The example selected by Dr. Herter for her graduate class was the goal of obtaining contractor's license,which is a significant need in the local community. The Vocabulary Miner created vocabularies from different sources including a construction trade site and the CA Contractor's Regulations on the State of California site. This application is largely based on the extraction tools in OWL's NetListening product.

This is a full function browser allowing navigation to any Internet resource and the user can add the page being viewed. The "Ck Page" command button performs the analysis of the page resulting in data such as links, on-site, off-site, Words, Words already in the dictionary, words not in the dictionary and ratios. The Ratios relate to the dictionary, vocabulary, and specific student (if selected) and they indicate the percentage of words that are already in a specific resource. For example a 9.0 dictionary rating would indicate that 90% of the words on this page are already in the resource. The student ratio indicates the percentage of words that the student has some experience with. In Phase II we will spread the word count based on the skill level of the word within the student history.

Word Master Maintenance



This application provides maintenance to the word master and its associated support tables. Each area of the screen is discussed below.

Word List

The word list allows the user to navigate to different words in the dictionary. The filters at the bottom of the list allow the user to create various word groups. The filters include: occurs, word size, A to Z range, and contains. Occurs indicates how many references exist for a word. Using this the instructor can concentrate on the most important words first. Word size will allow selection based on the length of the word allowing the user to view long or short words. The A to Z range selects words based on the starting and ending range. Entering a letter such as A in the first box will result in all words starting with "A" because the to range default is "AZZZ". Contains is used for isolating words with the same set of letters such as words with "ph"

Definitions

Definitions provide an area for input of the meaning of the word. When words are first added to the system a default entry is made in this table. We intend to hook this up to an external resource for the base data but we have not yet negotiated that relationship. We have had initial conversations with Random House in New York to get access to the

Webster dictionary and Roget's Thesaurus. We believe that this licensing will ultimately cost the project about $5,000 plus a per copy fee when the product is distributed.

Relationships
Relationships are links between different words. These relationships include things such as synonyms, anonyms, languages, plurals, and many others.

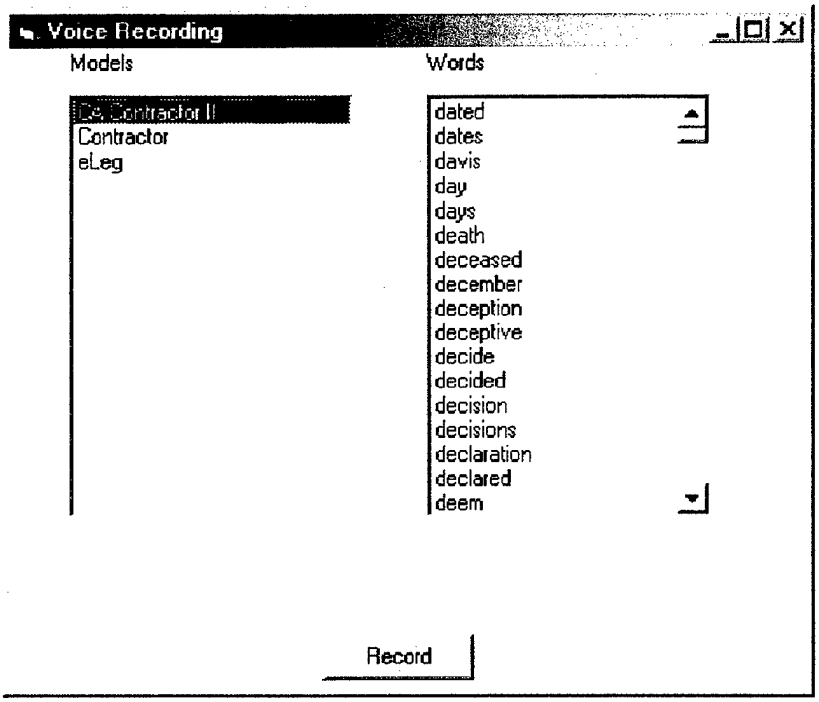
Questions
The questions table provides the supporting information to ask the various types of questions. This is where the exam module looks when it wants to extract a work file for the student. These are not simply copied to another file since the student configuration table contains instructions to the exam module on what type and how many of each question type the student needs.

Word Classes
Word classes provide ways categorizing words into large groups. Word classes and relationships might seem like duplications but they are not. Word classes group large sets of words such as words that relate to objects while relationships point from one specific definition to another. Word classes are intended to provide a linkage between instructional methodologies and the specific words. For example, teaching objects by using a visual makes sense but trying to teach action words with a visual is much more difficult and the system knows what question types use visual support.

Phonetic Rules
This table links the phonetic rules to the word. In the future this will allow the system to chase a phonetic rule in the presentation. For example, once a student has mastered "Phone" the system will be able to chase other words with the ph = f rule.

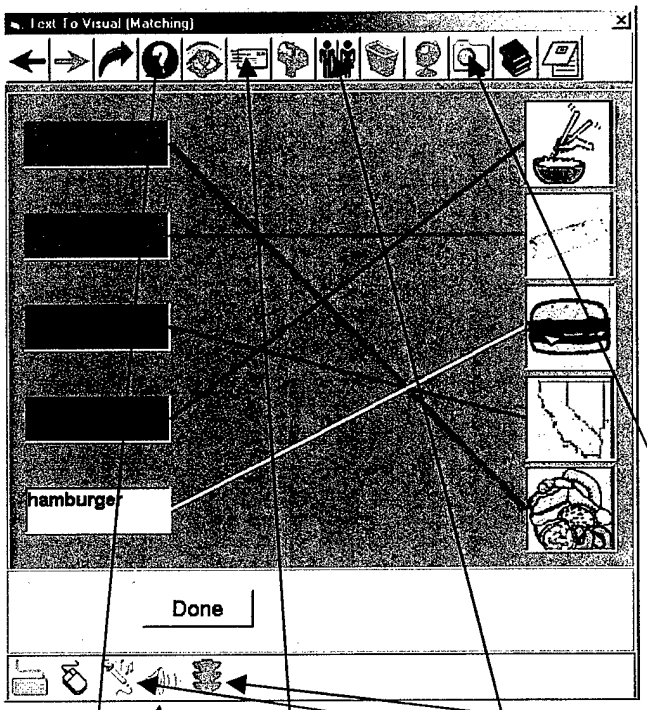
Students
This table provides a view into the specific word and students with skills or experience with that word. Populating this table is a key objective of the Phase II effort. We know at this point that we want a 1-5 skill rating based on the DLPT (Defense Language Proficiency Test) scale standards but we have yet to work out the details that would support this.

Voice File Recorder



This simple application allows the user to select a vocabulary and it will produce a list of the words within the system that do not have recorded voice files. The instructor simply selects the word and clicks on record to link the word to its new voice file. This application knows what words have been recorded for other vocabularies so new word sets require much less maintenance.

Exam Module



Question Mark	Mail Icon	Peer Group Mail	Input Icons	Traffic Light
Speaker Icon			Explorer Icon	

This screen sample is the text to visual mapping. On the toolbar we have the standard navigation arrows for forward, back, and skip. The question mark provides clues based on the type of question being processed. In the case of the text to visual matching, the clue will attach one text element to its related picture for the student. The other icons do not work in this prototype but we can discuss their functionality. The mail icon allows the student to send email to the assigned instructor or mentor. Since we might be dealing with language literacy this email will support attachment of voice file and automatic transmission of the current student and program information.

The group icon is an email/chat subsystem (male & female) for peer study groups. The explorer icon will provide Internet access to the source document of the word they are working on.

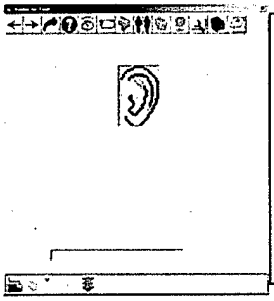
The toolbar on the bottom of the screen indicates the input that the system is waiting to receive from the keyboard, mouse, and microphone. The speaker icon is active when the system is playing audio to the student for clues or questions such as audio to text. The traffic light icon is used for the flashcard mode of the application. When in this mode, it

presents a word as in a flashcard. Then the light goes from green to yellow to red so the student will know how much time is remaining.

The exam module outputs the student's question and answer table. This includes the question type, question, answer, elapsed time, and number of clues given. The analysis of this data is the core of our Phase II efforts to create a proof of concept intelligent software tutor capable of picking the next activity. Speed of response is an important element in the analysis of the student's knowledge and we will provide an icon to stop the clock, so the student can do other things.

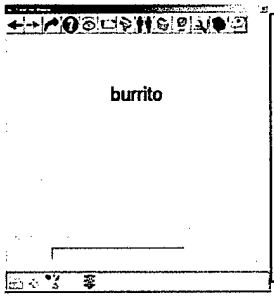
Matching questions presented a few issues regarding time recording. What the time on these records indicates is the length of time since the prior activity. Our database design handles the fact that the student may answer the same question multiple times. It keeps record of each Matching questions are actually five questions presented at the same time and each time a match is made, it records the elapsed time. If an answer is changed, the Q & A table will have multiple answers with each amount of time recorded separately. There are some flaws in this logic since the student may be analyzing multiple questions simultaneously. When considered as an overall block, all the time is accounted for.

Question types



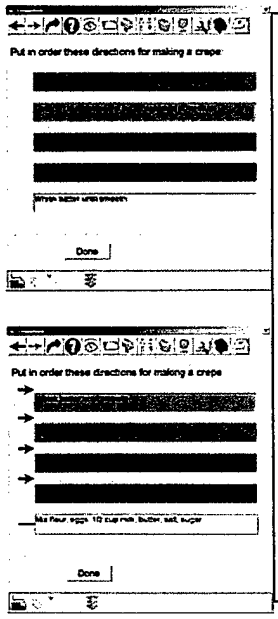
Audio to Text (Flash card)

Student listens to the word and types in the response. This is a flash card question with a speed drill available. When equipped with ViaVoice, this format can be used for audio to voice. When in speed drill mode, the traffic light icon on the bottom is active and goes through a green, yellow, red sequence.



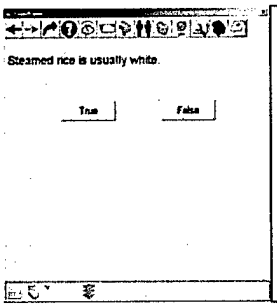
Text to Voice (Flash card)

Displays text and the students respond verbally. This is another form of flash card with speed drill features. Via Voice is required for this question type.



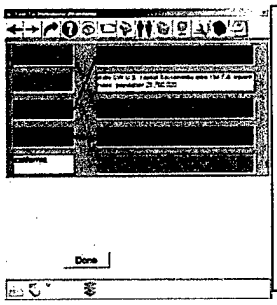
Sequencing

In this question type, the student is presented with a series of items and must move them into the proper sequence. The system randomizes the list so the student cannot learn a pattern. Drag and drop skills are at best difficult for the literacy student. This question changes to the sample below when an item is selected. The student then clicks on the arrow where they want the item placed and the system moves it. We believe that this is a more intuitive and simpler method to instruct students.



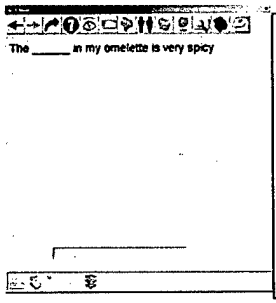
True / False

A simple true-false question where the student clicks on the response.



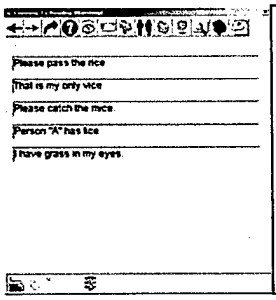
Text to Definition Mapping

System selects 5 words and definitions then randomizes the order. The student selects



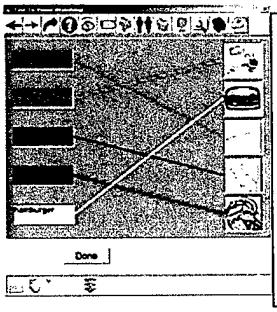
Cloze/Fill in the blank

Fill in the blank is a common question type and the response can be either verbal or keyboard. Clues on this question type display possible answers effectively converting it to a multiple choice question.



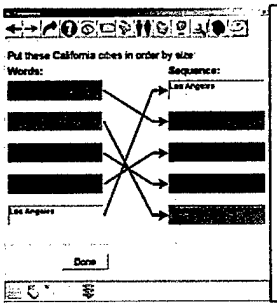
Listening to Reading

In this question type, the system reads the sentence to the student who then clicks on the correct matching sentence. Again the answers are randomized by the system so the student cannot learn the pattern of the answer.



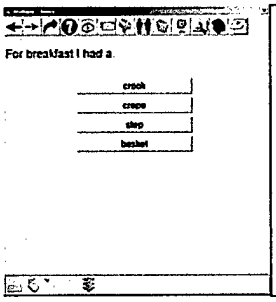
Text to Visual Mapping

In this question type the student matches text to pictures. This is an effective question type when dealing with word classes like objects while trying to build the student's comprehension.



Text Sequencing

In this matching, the student is to place the list into a predetermined sequence. Once a box is selected from the left side, clicking on a position on the right will select that and copy the text to that box.



Multiple Choice

The clues function reduces the number of options starting with the dumb choices. This question type can be filled by the system or by the instructor. Letting the system select the wrong answers results in a more randomized question, but the question designer can put specific answers creating more difficult questions.

Summary

The current prototype is a proof of concept application with technical challenges left to be resolved but it provides a framework for the project in Phase II. While it is not as smooth as a production system, it will allow attachment of documents, and it demonstrates the flow to individuals outside the project team. This also provides for the extensive data recording needed to test training concepts.

Credits

This project includes work performed by

Person	Role within Project
Dr. Roberta Herter	Education Research
Dr. Joseph Grimes	Computer Systems Research
Robert Dumouchel	Commercial Research, Software Development Manager
Christion Richert	Programmer
Yvonne	Educational Research Assistants

Correlation of Milk Composition and Fouling with Biofilm Formation and Microbial Spore Production in Heat Exchangers

Project Investigators:

Rafael Jimenez-Flores, Ph.D.
Professor
Dairy Products Technology Center

Yarrow Nelson, Ph.D.
Associate Professor
Civil and Environmental Engineering Department

Daniel Walsh, Ph.D.
Associate Dean
College of Engineering

California Central Coast Research Partnership

Correlation of Milk Composition and Fouling with Biofilm Formation and Microbial Spore Production in Heat Exchangers

Final Report September 16, 2002

Dr. Rafael Jimenez-Flores, Professor, Dairy Products Technology Center
Dr. Yarrow Nelson, Asst. Professor, Dept. of Civil and Environmental Engineering,
Dr. Daniel Walsh, Professor and Assoc. Dean, College of Engineering
Mr. Kamran Ghashghaei, Graduate Student, Biochemical Engineering
Cal Poly State University, San Luis Obispo, CA

Executive summary

Biofouling in dairy processing was investigated using pilot-scale equipment to determine the effect of milk composition on biofouling rates, to analyze the protein composition of materials deposited during biofouling, to determine metalurgical influences on biofouling, and to examine spore deposition during biofouling. During this first year of research, significant progress was made in all of these areas. A pilot heat exchanger system was completed and used for measuring biofouling rates for different types of milk products. This apparatus was used to determine biofouling rates by monitoring inlet and outlet milk temperatures using thermocouples connected to a data logger, and also by determining reductions in milk flow rates. Results indicate that fouling is accelerated at low fluid velocities, and that fouling rates were 25% lower for reconstituted skim milk compared to fresh whole milk, indicating that milk fat increases biofouling slightly. Work is in progress to measure biofouling rates for milk from genetically modified cows to determine of these genetic variants produce less biofouling. Proteins deposited during biofouling were analyzed quantitatively using Kjeldahl analysis of nitrogen and qualitatively using electrophoresis with SDS-PAGE. Materials testing methods were initiated to investigate the effect of alloy composition on biofouling and the formation of biofilms. A method was also developed for enumerating spores in milk biofilms using epifluorescence microscopy. Graduate students Kamran Ghashghaei (Biochemical Engineering) and Stephen Nelson (Biomaterials Engineering) have received training and performed the experimental research described here.

Biofouling rate measurements

Biofouling heat exchanger apparatus

Biofouling was measured directly in a stainless steel plate heat exchanger by monitoring the increase in the temperature differential as indicated by lowering milk outlet temperatures. The plate heat exchanger (PHE) used was a "Junior" stainless steel plate heat exchanger obtained from APV Crepaco , Inc. It has 17 plates and 4 passes and both counter-current and co-current flow. Four thermocouple probes(NPT series type K) were installed onto the inlets and outlets. A data logger (OM-3001 from Omega Engineering) was used to record all four temperatures (two inlet plus two outlet).

The product (milk) at about 40 F is supplied from a balance tank and pumped to the heating section of the plate heat exchanger as shown in Figure 1 . Hot water at about 206 F is pumped from the heating medium tank through the plates and is then recirculated to the heating medium tank where it is maintained at constant temperature using steam.

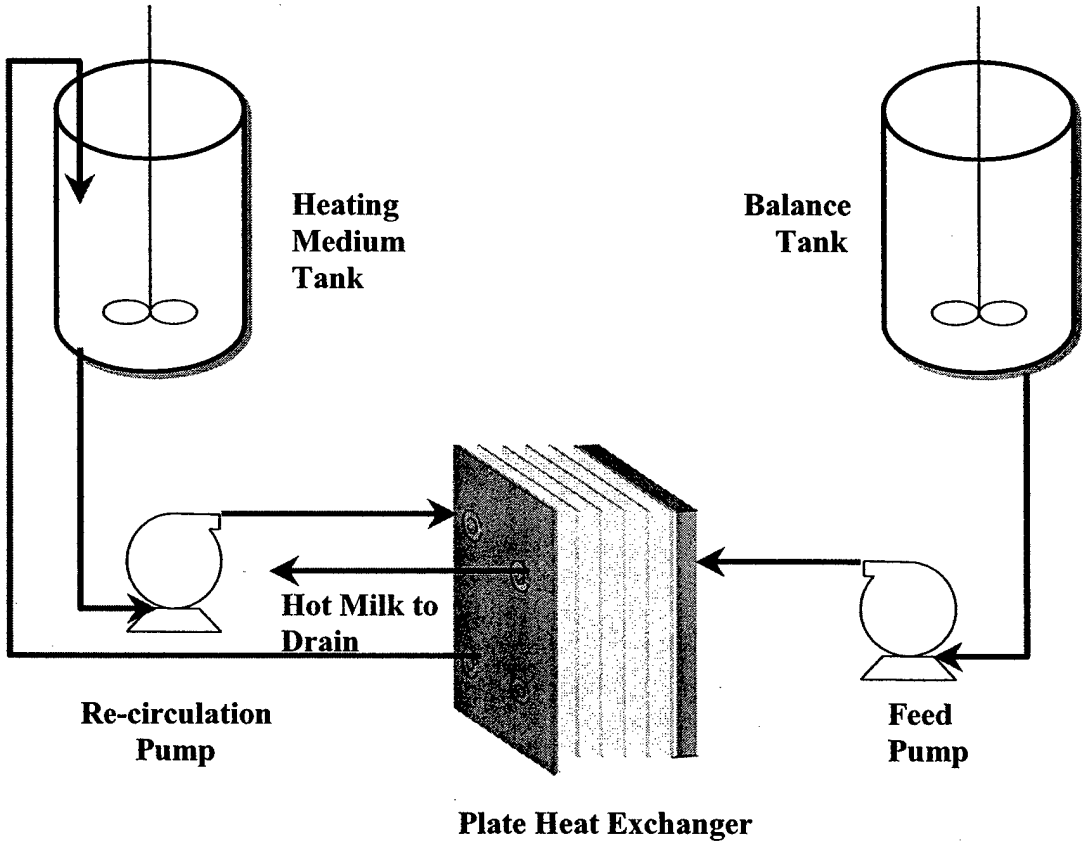


Figure 1. Process flow chart for indirect heating in plate heat exchanger

Biofouling Experiment 1
Whole milk; high flow rate

This experiment tested biofouling with whole milk at a flow rate of 1 L/min. The milk entered the heat exchanger at 40 F. The heating medium is hot water at flow rate of 10 L/min and inlet temperature of 206 F. Run time was 210 minutes. The results of this experiment are shown in Table 1 and Figure 2. The temperature drop caused by the biofouling in this experiment was 0.0028 °F/min.

Table 1- Conditions and Results for Fouling Experiment 1: Whole milk; high flow rate

	Milk IN	Milk OUT	Water IN	Water OUT
Minimum Temp (F)	38.7	203.5	205.5	190.6
Maximum Temp(F)	40.7	205.3	207	194.9
Start value	40.4	203.9	205.5	190.6
End value	40.7	204.4	206.7	192.6
Run Time (Min)	210	210	210	210
Regulated flow rate ?	Yes	Yes	Yes	Yes
Fouling rate (F/min)		0.0028		
Overall fouling		0.59		

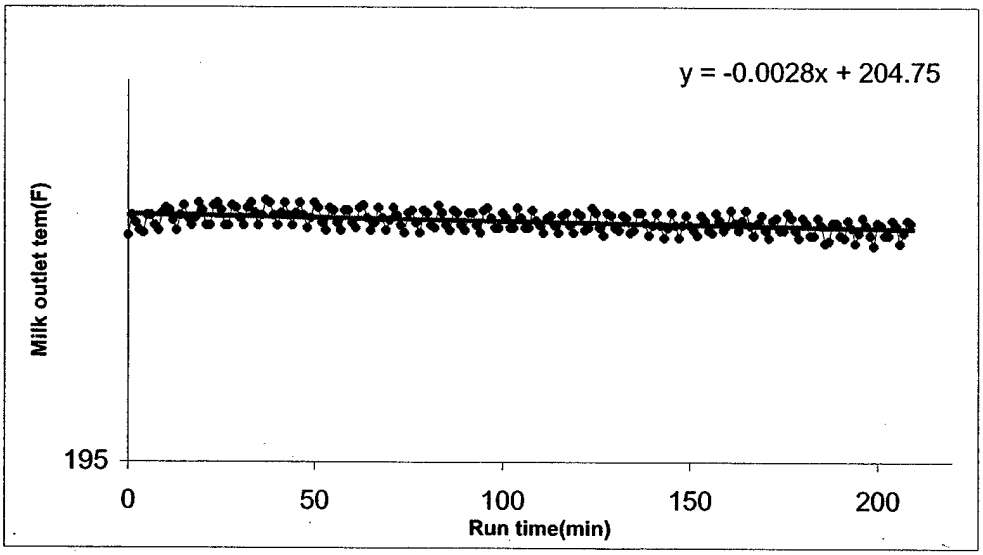


Figure 2 – Milk outlet temperature profile for Experiment 1

Biofouling Experiment 2
Whole milk; low flow rate

Experiment 2 was run again with whole milk, but in this case the flow rate is lower than last experiment (Table 2). Thus, the purpose of this experiment was to find out the relation between flow rate and fouling. While the milk flow rate in Experiment 1 was 1000 mL/min, the flow rate in this experiment was maintained at 600 mL/min. The run time was also longer for this experiment (total run time increased to 325 min). Summary results for this experiment are shown below in Table 2 and Figure 3. Relative to Experiment 1 (with a milk flow rate of 1 L/min), the biofouling rate increased by only about 10%, to 0.0031 °F/min.

Table 2 Conditions and Results for Fouling Experiment 2: Whole milk; low flow rate

	Milk In	Milk Out	Water In	Water Out
Minimum Temp(F)	39.9	202.2	205.5	197.3
Maximum temp(F)	45.2	204.1	206.8	199.4
Start value	44.5	203.4	206.1	198.1
End Value	45.2	202.3	205.9	198.6
Run time(Min)	325	325	325	325
Regulated Flow rate ?	Yes	Yes	Yes	Yes
Fouling rate (F/min)		.0031		
Overall fouling		1.01		

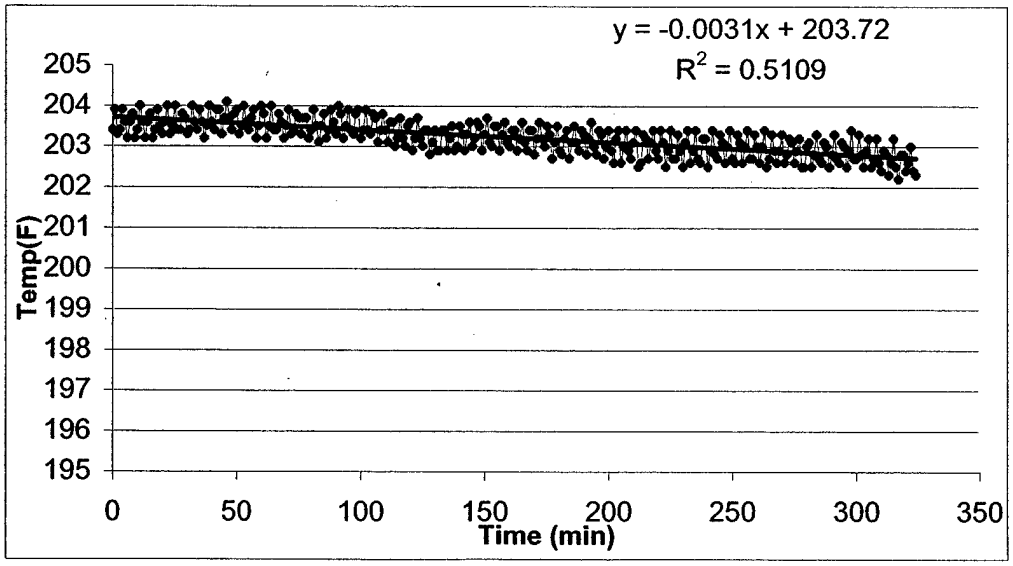


Figure 3 – Milk outlet temperature profile for Experiment 2.

Biofouling Experiment 3
Reconstituted milk powder, flow controlled at 600 mL/min.

Milk powder was reconstituted with 10.5% w/w solids enters the PHE at 40°F. The heating medium is hot water at flow rate of 10 L/min and inlet temperature at 206°F . Milk flow rate was 600 ml/min, and the total run time was 325 min. Conditions and results are summarized in Table 3, and the milk outlet temperature profile is shown in Table 3 and Figure 4. The fouling rate for the milk powder under these conditions was 0.0023 °F/min. (Table 3 and Figure 4). For comparison, the fouling rate for whole milk at this same flow rate was 0.0031°F/min (Table 2 and Figure 3). Thus, fouling was about 25% lower for reconstituted non-fat milk powder compared to whole milk.

Table 3 Conditions and Results for Fouling Experiment 3: Reconstituted milk powder, flow controlled at 600 mL/min.

	Milk IN	Milk OUT	WATER IN	WATER OUT
Minimum Temp(F)	39.9	202.2	205.5	197.3
Maximum Temp(F)	45.2	204.1	206.8	199.4
Start value	44.5	203.4	206.1	198.1
End value	45.2	202.3	205.9	198.6
Run time (Min)	325	325	325	325
Regulated flow rate ?	Yes	Yes	Yes	Yes
Fouling rate (F/min)		.0023		
Overall fouling		0.75		

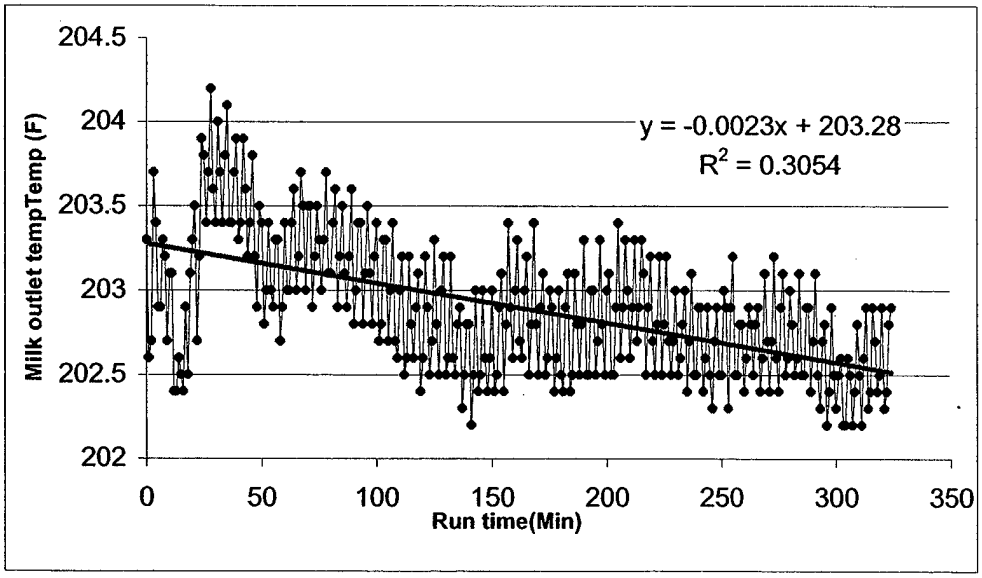


Figure 4 – Milk outlet temperature profile for Experiment 3

Biofouling Experiment 4
Reconstituted milk powder; Initial flow rate = 600 mL/min – not regulated

Reconstituted milk powder as described above entered the PHE at 43°F and was heated with hot water at a flow rate of 10 L/min and an inlet temperature at 206°F. Milk flow rate was initially 600 mL/min. Run time was only 180 min. because the high fouling rate led to a rapid decrease in milk flow rate under these conditions. With unregulated flow, the milk flow rate decreased to a final flow rate of 110 mL/min and led to a fouling rate of 0.028°F/min. (Table 4 and Figure 5), which is about one order of magnitude greater than that observed with regulated flow at 600 mL/min.

Table 4. Results for Experiment 4.
Reconstituted milk powder; Initial flow rate = 600 mL/min – not regulated

	Milk In	Milk Out	Water In	Water Out
Minimum Temp(F)	43.1	198	205.7	197.4
Maximum Temp(F)	48.9	203.6	207	204.4
Start value	43.2	203	206.1	197.4
End value	46.7	198	206.8	204.4
Run time (Min)	181	181	181	181
Regulated flow rate ?	No	No	Yes	Yes
Fouling Rate (F/min)		0.0283		
Overall fouling		5.12		

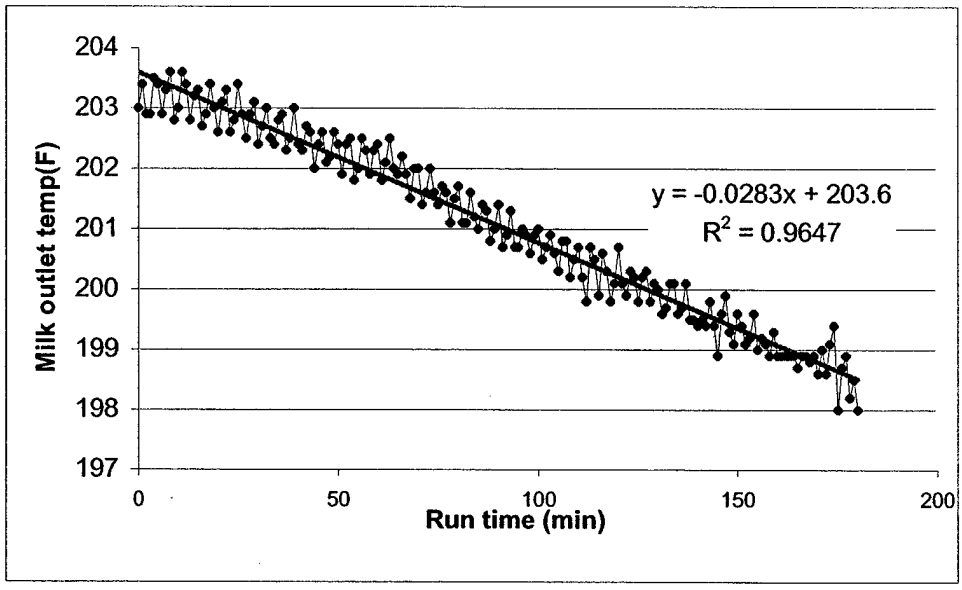


Figure 5 – Skim milk outlet temperature profile for Experiment 4 during fouling with unregulated flow rate

Conclusions from biofouling rate measurements:

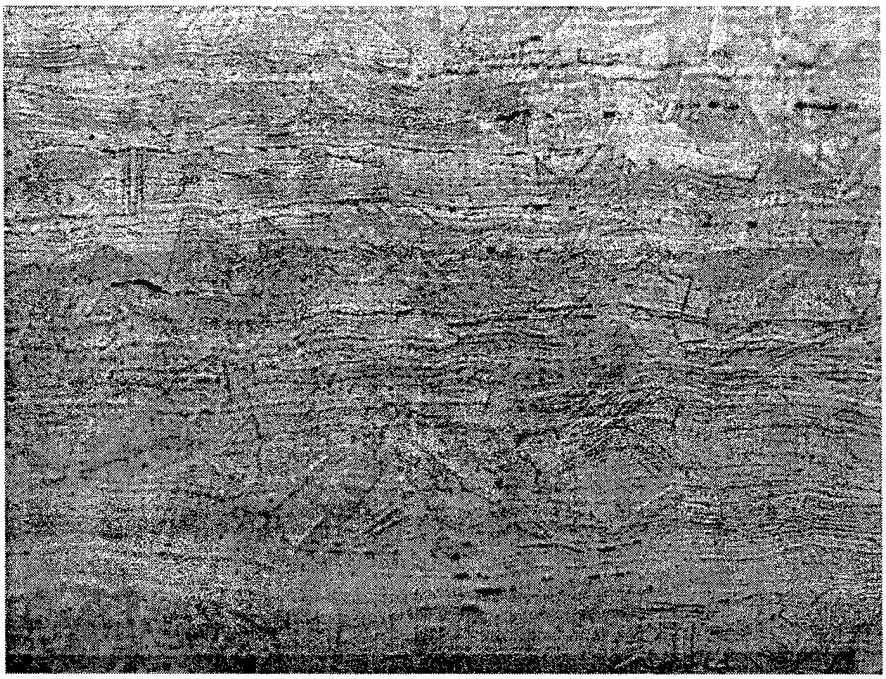
- The heat exchanger apparatus has proven to be a useful tool for measuring biofouling rates.
- Decreasing milk flow rate from 1 L/min to 600 mL/min caused about a 10% increase in biofouling rate.
- Biofouling was 25% lower for reconstituted dry non-fat milk compared to whole milk.
- Very high biofouling rates were observed when the milk flow rate was permitted to decrease during fouling.

Metallurgy of biofouling

To get a better understanding of how biofouling occurs when milk is in contact with metals we are examining the metal surface structure using microscopy and investigating the effect of metal composition on fouling rates. In particular, microscopic examination of the metal surfaces is being used to determine if there are any preferential locations on the metal where proteins can bind and where biofilm-forming bacteria can thrive and form a colony. This is accomplished by placing metal samples into heat exchangers in contact with the heated milk. Ultimately, we hope establish if is a direct correlation between alloy and film formation.

Initial experiments were conducted with a Type 304 Stainless Steel with both high and low sulfur content, and titanium alloy 6Al-4V. Metallography was done on the materials mounted so that the cross sectional microstructure of the metal can be seen (Figure 6). The cross sections include directions perpendicular and parallel to the rolling direction of the plate. Image analysis will be done on the samples to determine grain size, inclusion characteristics, inclusion distribution per area, aspect ratio of the inclusions, for both directions. Metal coupons were also cut for surface film characterization; they will be dipped into a controlled temperature bath of milk. The biofilm formed will then be characterized for each alloy system with surface profiles from an atomic resolution microscope.

Figure 6. Photomicrographs of 304 stainless steel coupons with high and low sulfur content (magnification = 200x).



A. Electrolytic: 10% Oxalic Acid SS 304L Low Sulfur 200x



B. Electrolytic: 10% Oxalic Acid SS 304L High Sulfur 200x

Protein Analyses

In order to develop a better understanding of the mechanisms of formation of milk biofilms, proteins deposited during biofouling in the pilot-scale system were analyzed. Proteins deposited during biofouling were analyzed quantitatively using Kjeldhal analysis of nitrogen and qualitatively using electrophoresis with SDS-PAGE. For these analyses, dried attached deposits were collected by scraping the surface of the plate heat exchanger which had direct contact with raw milk while processing.

Quantitative Protein Analyses: The Kjeldhal method was used to determine the nitrogen content of the milk deposits, and percent protein was calculated from the nitrogen content. The milk deposits were collected as described above and digested in H_2SO_4, using $CuSO_4 5\ H_2O$ as catalyst with K_2SO_4 as a boiling point elevator. This digestion releases nitrogen from the proteins and produces ammonium salt. NaOH is then added to hydrolyze ammonium and release NH_3, which is distilled, collected in H_3BO_3 solution and titrated with 0.1 N HCL to a pink endpoint. Protein content was calculated by assuming that milk protein is 15.7 % N, which is standard for milk proteins. Protein content was calculated as follows:

Nitrogen % = [1.4007 * Vs * N] / W

where Vs = ml HCL titrant used,
N = Normality of HCL solution, and
W = sample weight , gr

Percent Protein = Nitrogen % * 6.38

For the milk biofilms collected from the heat exchanger, the following results were obtained:

Vs = 15.2 ml HCL
W = 0.5 gr
Nitrogen % = 4.26
Protein % = 27

Thus, the milk biofilm contained 27% protein by weight. In future experiments, we hope to determine the protein content of milk biofilms as a function of deposition time.

Qualitative Protein Analyses (SDS-PAGE): Milk proteins deposited during biofouling were analyzed quantitatively using electrophoresis with SDS-PAGE. This method separates proteins on a gel based on their molecular size. SDS (Sodium Dodecyl Sulfate) is a surfactant that can dissolve hydrophobic molecules but also has a polar group (sulfate) that renders it water soluble. The proteins are separated in a polyacrylamide gel, thus the name Poly Acrylamide Gel Electrophoresis (PAGE). Essentially, proteins of smaller size will move faster toward in the gel, allowing separation and comparison to model proteins.

Gels were prepared with 12 % acrylamide as follows:

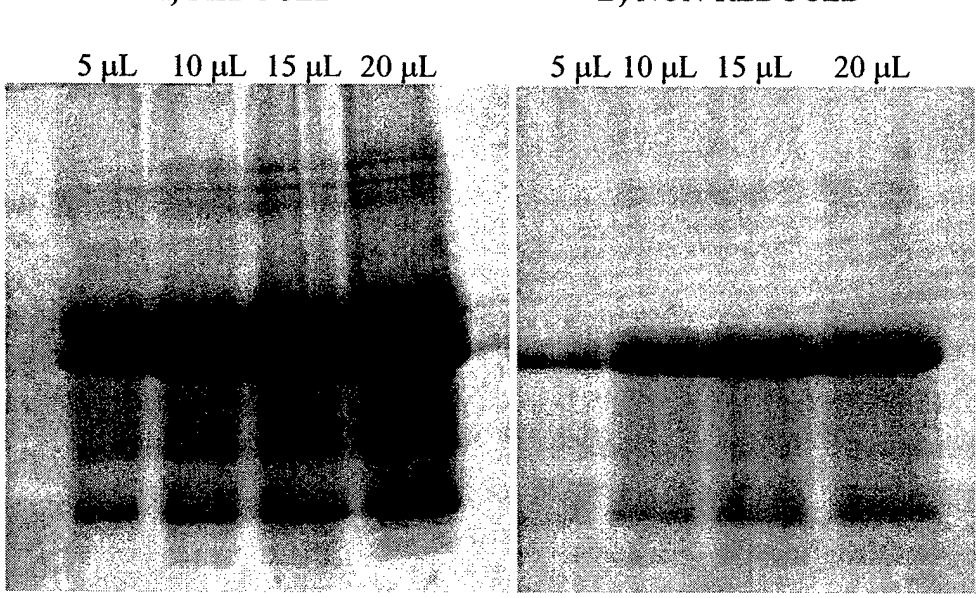
H_20	8.8 ml
Resolving Buffer	5 ml
10 % SDS	0.2 ml
40 % Acrylamide	6 ml
10 % APS	100 ul ,
Temed	10 ul

A biofouling sample was collected by scraping 0.1 g of biofilm from the heat exchanger. Grinding of the dry deposits was conducted with an Ultra-Turrax T8 to dissolve sample. Gels were prepared as described above. A minimal amount of butanol was appplied on the top of the gel to smooth out the gel. After dissolving samples and processing , the gels were loaded as follows:

Well #	*Sample ID*	*Reducing Gel*
1	Biofilm	5 ul
2	Biofilm	10 ul
3	Biofilm	15 ul
4	Biofilm	20 ul
Well #	*Sample ID*	*Non-reducing Gel*
1	Biofilm	5 ul
2	Biofilm	10 ul
3	Biofilm	15 ul
4	Biofilm	20 ul

Electrophoresis was run at 80 Volt for 30 min. and then increased to 100 Volt for an additional 30 min. After elution, the gel was stained with Comassie Blue overnight. Excess stain was removed by pouring off the Comassie Blue and adding de-stain to visualize bands. Photos were taken and gels were viewed for protein content (Figure 7).

Figure 7. SDS-PAGE gels of milk proteins collected from biofouled heat exchanger.



Endospore Analyses

Some progress was made toward developing a method for enumerating spores in milk biofilms using epifluorescence microscopy. Spores are biotinulated and then stained using a Texas Red fluorescent stain. A new lens set was purchased for the epifluorescence microscope for observation of the stained spores. Preliminary results show promise for enumeration of spores in biofilms using this method.

Microbiological Assessment of Biofilm.

Assesment of spore-former organisms and endospore detection using TRF patterns was developed in the pilot plant, using our evaporator. Biofilm study is not yet considered at this point because final characterization of the heat exchange system needs to be completed. However, we accomplished a full set of tests for detection and spore determination in a milk processing line at the DPTC pilot plant.

Endospore survival and detection was modeled as a low heat skim milk powder processing run. The experiments consisted on sixty gallons (226 liters) of skim milk was pasteurized, condensed, and spray dried using the DPTC pilot plant. Thirty gallons (113 liters) of the raw skim milk was inoculated with approximately 10^7 endospores per liter using an endospore mix of the detrimental endospore strains. Samples were collected from the raw, pasteurized, condensed and powder stages during the spiked and non-spiked processing runs. The processing runs were performed in triplicate. DNA extractions, using an optimized procedure, were performed on the samples to ensure that DNA from endospores and vegetative cells was obtained. The Polymerase Chain Reaction (PCR) was used to amplify the 16s rDNA regions to produce a product of approximately 500 base pairs (Figure 1). These fragments were digested with three restriction enzymes (*HaeIII, HhaI, DpnII*) to produce TRFPs as an electropherogram (Figure 2). Standard peaks are used to determine base pair fragment length.

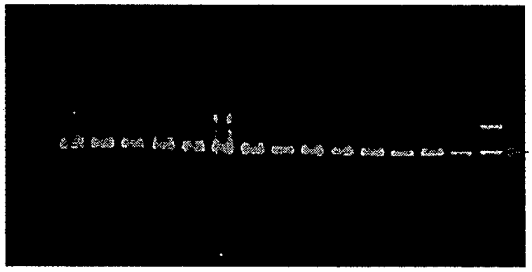


Figure 1. 16s rDNA products

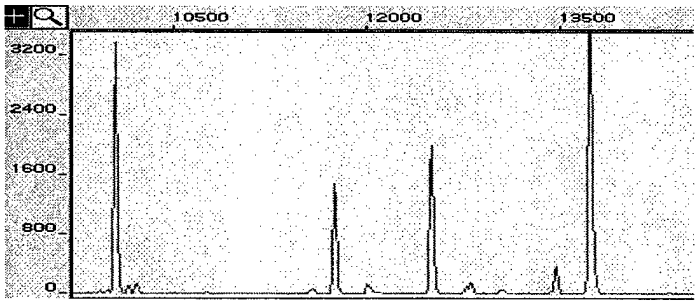


Figure 2. Sample TRFP electropherogram

Using previously designed GerC3 primers, a PCR was optimized using temperature profiles in order to obtain a single germination gene amplification band of approximately 480 base pairs. GerC3 PCR was performed on the low heat skim milk powder processing run.

Results

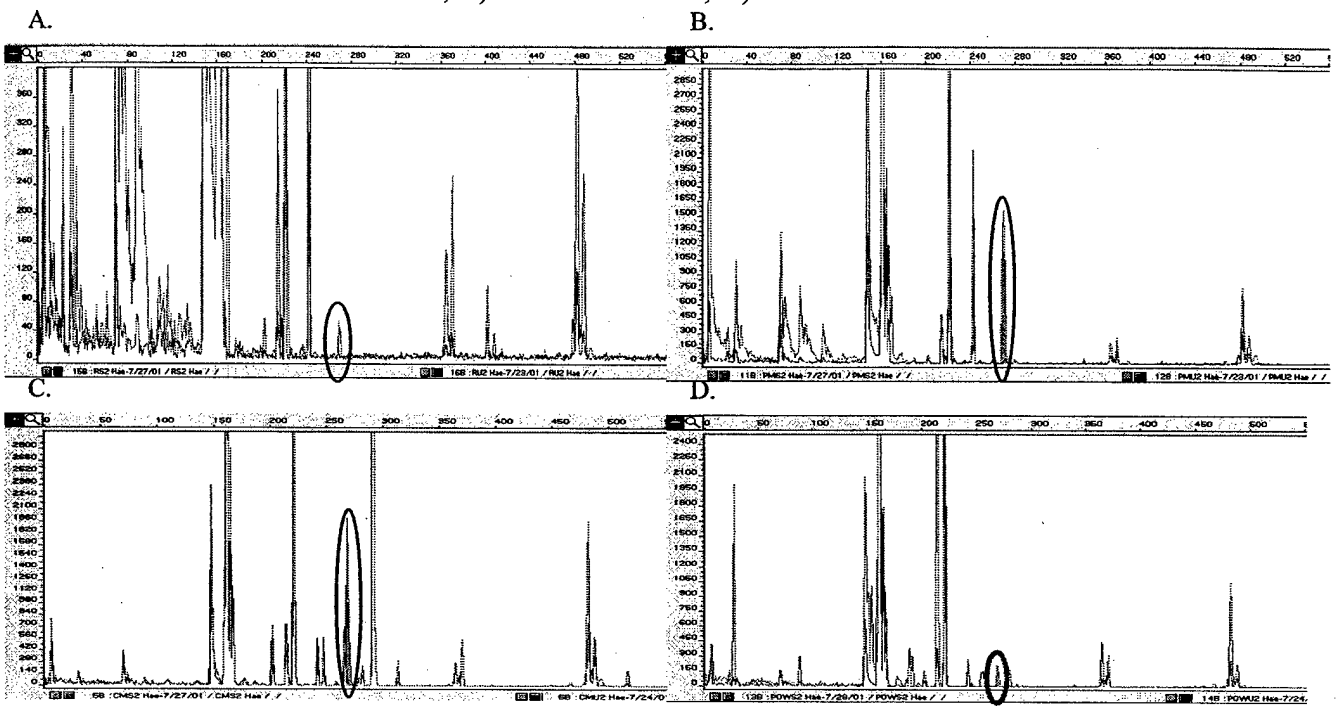
Using Terminal Restriction Fragment Patterns, we are able to positively identify endospore-formers during the low heat skim milk powder processing run. This detection method was based on the predicted base pair fragment sizes of the model *Bacillus* strains that were added to the raw skim milk. Table 1 shows the predicted base pair fragment size for the model strains according to the enzyme used.

Table 1. Predicted base pair fragment size of model strains.

Enzyme	**Fragment size (in bp)**
HhaI	199 to 202
DpnII	262 to 265
HaeIII	268 to 270

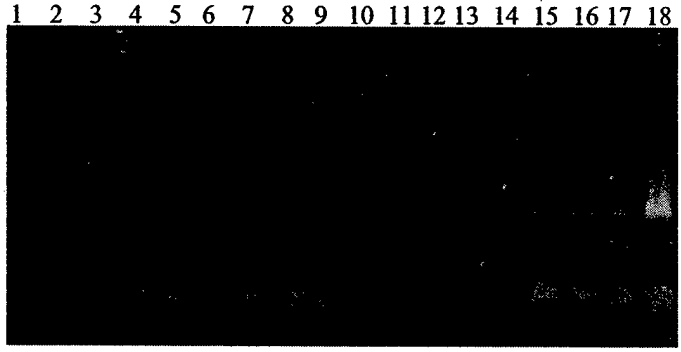
TRF patterns for the *HaeIII* enzyme are shown for raw, pasteurized, condensed and powdered samples (Figure 3). Results are indicative for all three enzymes. The red peaks are representative of the non-spiked skim milk samples and the blue peaks are representative of the spiked samples. Results show, indicated by a circled peak, that we are able detect the presence of the endospores in our samples.

Figure 3. TRFP of low heat skim milk powder processing run. A) Raw milk, B) Pasteurized milk, C) Condensed milk, D) Powdered milk.



Using GerC3 PCR on the low heat skim milk powder processing samples, we are able to specifically detect endospore formers (Figure 4).

Figure 4. GerC PCR on low heat skim milk powder processing run. 1.100bp ladder, 2. Raw unspiked, 3. Pasteurized unspiked beginning, 4. Pasteurized unspiked middle, 5. Pasteurized unspiked end, 6. Condensed unspiked beginning, 7. Condensed unspiked middle, 8. Condensed unspiked end, 9. Powder unspiked, 10. Raw spiked, 11. Pasteurized spiked beginning, 12. Pasteurized spiked middle, 13. Pasteurized spiked end, 14. Condensed spiked beginning, 15. Condensed spiked middle, 16. Condensed spiked end, 17. Powder spiked, 18. Positive control



Our results showed that we are able to detect a high concentration of endospores during low heat milk powder processing using Terminal Restriction Fragment Patterns. This has proven to be a sensitive detection method and further experimentation will determine detection limits associated with this technique. TRFPs also prove to be effective in determining microorganisms present throughout each processing step. This can lead to microbial ecology studies dealing with contamination parameters during powder production. The electropherograms (Figure 3) show the differences between the spiked and non-spiked samples within each processing step as well as the ecological changes that occur from raw to powder milk. Using TRFPs in conjunction with the RDP, we can observe the community transformations that occur during normal milk powder production in a pilot plant or in an industrial setting. We are able to observe how adding a high concentration of endospores will alter the microbial community and the interactions that are taking place within each sample.

In summary, the results of the microbiological detection of spores is feasible. Using a fouling film, or biofilm material scraped from the surface of the heat exchanger should be an appropriate sample to detect the presence of spore formers, and determine some information regarding the microbial ecology of that sample.

References

1. Pitesky, M., *The Microbial Ecology of Milk Powder Production Using Terminal Restriction Fragment Patterns and the Development of a Rapid PCR Assay for the Detection of Mesophilic Bacillus Endospores in Milk Powder*, in *Agriculture*. 2000, California Polytechnic State University: San Luis Obispo.
2. Lin S., S.H., Odumeru J.A., Griffeths M. W., *Identification of contamination sources of Bacillus cereus in pasteurized milk.* International Journal of Food Microbiology, 1998. **43**(3): p. 159-171.
3. DP, C., *Redefining relativity: quartitative PCR at low template concentrations for industrial and environmental microbiology.* Journal of Industrial Microbiology and Biotechnology, 1998. **21**: p. 128-140.
4. Clement B., L.K., C. Kitts, *Terminal Restriction Patterns (TRFP's), a rapid, PCR-based method for the comparison of complex bacterial communities.* Journal of Microbiological Methods, 1998. **31**: p. 135-142.

The Role of Discovery in Context-Building Decision-Support Systems

Project Investigators:

Jens G. Pohl, Ph.D.
Executive Director
CAD Research Center

Steven J. Gollery
Software Engineer
CAD Research Center

The Role of Discovery in Context-Building Decision-Support Systems

Steven J. Gollery
Collaborative Agent Design Research Center (CADRC)
California Polytechnic State University (Cal Poly)
San Luis Obispo, CA

Introduction

There are many sources of on-line information available to those responsible for making decisions in complex situations. However, most of that information is either intended for human use or is available only in custom or proprietary formats. Both of these conditions reduce the ability of computer software to perform automatic reasoning about this information. As a result, the usefulness of on-line information in intelligent decision-support systems is currently limited to the few sources that are implemented specifically for those systems. The vast mass of sources that do not fit that description are virtually invisible.

Current practices for integrating multiple information sources have proven too costly to implement and have produced inflexible systems. These practices, which are discussed further below, cannot cope with the magnitude or the diversity of the available information.

This paper describes the TEGRID project, which demonstrates the use of a *service oriented architecture* to enable a more flexible, loosely coupled system of interoperable information sources and consumers.

Disaster Management Requires Information from Diverse Sources

To determine how best to manage disaster or emergency response, the decision-maker requires access to information from many sources. These sources might include law enforcement agencies, hospitals, ambulance services, the weather service, city traffic control, the National Guard, and so on. Each of these sources is controlled by different organizations, some non-governmental, some local, some state, and some federal; some organizations are civilian, some military. A decision-support system that assists users in planning responses to ongoing emergencies must incorporate information from as many of these sources as possible.

One often-used approach to the problem of inter-system communication is to create an interface agreement (i.e., an exact definition of the format that will be used to communicate data and information). Each system then creates a translation between its own internal data or information model and the format defined by the interface agreement. While this approach is conceptually simple, it poses problems at both the technical and the cultural levels.

It can be very difficult for multiple organizations to reach agreement on an interface format. Such agreements take time and a great deal of focused effort. At the same time, it can be difficult to arrive at common understandings of the information represented by each system. A single model representing all the kinds of information available in all systems may be so extensive that it becomes impractical to implement.

By far the most problematic aspect of the interface agreement approach is its lack of flexibility and extensibility. When any of the systems changes its internal model, this change may make the interface agreement obsolete. In that case, the agreement would have to be renegotiated and modifications would need to be made to all systems, with associated funding requirements. The level of effort required to change an interface once it has been implemented across multiple systems tends to create distributed systems that are brittle, static, and resistant to evolving quickly to meet the changing needs of their users.

Current approaches to inter-system communication too often result in tightly coupled systems. In extreme cases, the coupling becomes so tight that nothing can be changed in any individual system without requiring equivalent changes in other systems. Over time, the coupling of such systems tends to become tighter as more software is written based on assumptions about the exact format of communicated data and information.

Adding more information sources to such a system can also be difficult. As each new system is added, the communication among systems tends to become more complex. Each new revision of the interface agreement becomes more difficult to define, since any changes to the interface require changes to larger sections of the software. The result is a further slowing of growth and change.

This gradual reduction in the amount of change possible in a given period of time is especially injurious to decision-support systems. Decision-support systems, especially in the area of emergency management, should be implemented in a timely manner in order to provide assistance to their users as soon as possible. Furthermore, when users identify a requirement for change to the system, it must be feasible for developers to implement that change quickly so that new functionality is available when the decision-maker needs it.

These requirements, and the problems of the interface agreement approach, led us to consider a different type of architecture for our demonstration system. That architecture must provide (at least) two benefits over architectures that require tightly-coupled communications: first, the architecture must support the rapid addition of new sources of information; and second, the architecture must allow individual information sources to change their communication format while requiring little or no reworking of the systems consuming this information.

The proposed solution involves the loose coupling provided by *web services*, combined with the self-describing information model of the *semantic web*. Both of these concepts are discussed further in the following section.

Service-Oriented Architectures and Web Services

For our demonstration system, we explored the feasibility of replacing the tight coupling of earlier distributed systems with a more loosely-coupled architecture. We implemented a system based on web service standards that allowed us to use a discovery process to construct the system at run-time based on the information needs of the clients. The use of discovery enables each participant in the system to build an awareness of the context in which it is operating. This section defines these concepts, starting with the most basic: the service-oriented architecture.

In a Service-Oriented Architecture (SOA), each information source is considered to be a separate service, providing information to remote clients. Each service is developed and deployed by the organization that provides the information, eliminating the need for complex interface

agreements. Also, each service is remotely accessible, usually over the Internet. Services are generally not constrained to work with a single distributed system. This helps to keep the degree of coupling low, and allows the same service to be used in multiple contexts.

Web Services are a specialization of the more general Service-Oriented Architecture. The definition of Web Service has been hard to pin down, but the World Wide Web Consortium (W3C) has provided the following definition: "A Web service is a software application identified by a URI, whose interfaces and bindings are capable of being defined, described, and discovered as XML artifacts. A Web service supports direct interactions with other software agents using XML-based messages exchanged via Internet-based protocols." (see web site at: (http://www.w3.org/TR/wsa-reqs#IDAIO2IB)) Current implementations of web service standards specialize this farther: most web services use HTTP (Hyper-Text Transfer Protocol) to exchange messages defined in SOAP (Service Oriented Architecture Protocol). The SOAP standard defines an XML language and a set of rules for serializing and de-serializing objects and data, regardless of programming language, operating system, or hardware platform.

The description of web services is handled by WSDL (the Web Services Description Language), while discovery is provided by UDDI (Universal Description, Discovery, and Integration). UDDI defines an XML language for accessing a repository to register and locate web services according to the attributes of the service. The repository may be one of the public registries operated by Microsoft, IBM, or SAP. However, for most uses it is likely that participants will use a community registry accessible only to authorized partners. This paper discusses some of the limitations of WSDL and UDDI, below.

Why Discovery Matters in a Decision-Support System

As discussed earlier, in order to provide the broadest possible support for human decision-making, a decision-support system needs to provide all the information that is: (a) currently available; and, (b) currently relevant. This generally requires bringing together information from multiple sources.

Existing decision-support systems generally require that all potential sources of information be identified when the system is being designed and implemented. This requirement stems from the need to build concrete knowledge about each information source into the decision-support system. The information may include the location of the source, its access protocol, the data or information format, and the type of security used, among others.

The requirement that all sources of information be known prior to deployment prevents the decision-support system from taking advantage of information whose source is not identified (and may not even exist) while the system is being built. Most significantly, this requirement prevents the system from solving the 'transient need' problem. This problem arises when an unexpected situation requires a kind of information that is not present among the sources known to the system. For example, the information may be useful to a decision maker for a short period of time, but irrelevant after that point. Systems that can only use information sources that are known to the designers of the system tend to lack the flexibility required to deal with changes to the information environment and to the needs of the decision-maker.

A system built around web services solves the 'unknown information source' problem by adding the ability to discover services at run-time. If the service consumer can understand that a particular service provides information that the consumer can use, in a form that the consumer can process, then the consumer can use information from a previously-unknown source.

Some Limitations of Current Web Service Standards for Discovery

There are two important gaps in the ability of standard web services to discover and use unknown services. First, the standard discovery protocol only allows consumers to search for very specific types of services. In many cases, this amounts to a key-word search, which may result in missed opportunities and mistaken connections. Second, the standard method of defining service operations and parameters limits a would-be consumer to those services that implement operations using the names and classes that the consumer expects.

The difficulty is that the current web service standard for defining services and operations (Web Service Definition Language, or WSDL) does not include any information about the intent and meaning of an operation or its parameters. The best that a client can do is to search for operations based on the model that a given service implements, where the models are publicly defined either by the consumer, the providers, or by an industry standards body. The client will miss services that provide the same functionality but use different models.

For example: in the TEGRID demonstration project, we have created a definition of the operations and parameters for a 'publish-and-subscribe' service that allows clients to ask for information to be sent to them on topics of interest, and to send information regarding those topics. This service definition is then placed in the TEGRID service registry as a model of a service that the designers of publish-and-subscribe services may examine and implement. The SubscriptionManager service is one such service. In a real service-oriented system, there might be several services implementing the same model. Information about the SubscriptionManager is then entered into the service registry, including the fact that the SubscriptionManager implements the publish-and-subscribe model.

Participants in the TEGRID system that are interested in either sending or receiving information must implement the consumer side of the publish-and-subscribe model. That is, these participants must be able to discover those services that adhere to this specific model. In other words, they must be able to invoke operations using exact names and constructing parameters using class definitions that must be built-in to the client. Finally, the consumer must be able to receive values of a given class and map those values to objects within the consumer's own information model.

If another service were to provide the same functionality, but used different operation names and different classes for the input and output values, consumers searching the registry for publish-and-subscribe services would miss this equivalent service completely. The problem is that the registry does not describe the *purpose* (or *intent*) of a publish-and-subscribe service. As a result, consumers cannot find services based on a description of their own requirements, but only based on keywords and model definitions. The same problem also prevents consumers from locating new kinds of services (i.e., consumers are unable to describe their functional needs in a way that allows the registry to match those needs to specific service providers).

In short, using current web service standards, the 'discovery' process is limited in practical terms to discovering service locations. Adding entirely new types of services after deployment is only possible in limited circumstances. This in turn limits the ability of the client to draw information from all available relevant sources, and moves the effort of defining usable service types back from post-deployment to the development phase. Web service standards increase the flexibility of distributed systems, but they do not take us as far in that direction as we would like.

These limitations do not mean that web services as they are defined today provide no value. Many useful systems can be, will be, and are being built based on current standards. Service models are being defined by consortiums and standards bodies for specific vertical markets. These models will enable wide-spread interoperability. Service producers and consumers written for those models have the potential to provide unprecedented levels of system-to-system communication throughout an industry, resulting in significant increases in productivity.

Additionally, WSDL (the language used for defining service models) is designed to enable automated generation of client and server interfaces. The effort of incorporating a new service model defined in WSDL into a consumer is therefore generally very small, so that modifying a consumer to access a new type of information server can be done quickly and at little cost. Many (perhaps most) web service consumers will never have a need to dynamically locate sources of entirely unknown kinds of information. For these consumers, the limitations of the current methods of locating web services are simply irrelevant.

In order to provide full support for decision makers in areas where time is critical, on the other hand, these limitations do matter. The goal is to be able to give decision makers access to the information they need, when they need it, even when this involves types of information that were not anticipated by the designers of the system. At the CADRC, we are engaged in ongoing research exploring the use of semantic information to extend web services. We hope that such an approach will eventually allow systems to become progressively more flexible and responsive to the needs of their users. Some background on and explanation of the semantic approach is given in the next section.

Semantic Web Services

Several years ago, Tim Berners-Lee (the originator of the World Wide Web) began to discuss his vision of the future of the web. The web as it currently exists consists mainly of human-readable information. The markup in web pages is dedicated almost entirely to presentation instructions. This means that the only machine-processable content of most web pages is concerned with how the page should look.

Berners-Lee envisioned a web whose contents would include a new kind of markup that would enable software to reason about the *meaning* of the contents of a page. This would enable very intelligent automated processing of information on the web. Essentially, it would turn the contents of the web into an enormous knowledge base. Berners-Lee calls this vision of the future the *Semantic Web*. Semantic and logic languages are being developed to support this vision. The most influential of these languages is called RDF (Resource Definition Framework). RDF is deceptively simple, but has very deep underpinnings in the constructs and theory of formal logic.

Another language related to the Semantic Web is DAML-OIL (DARPA Agent Markup Language – Ontology Inference Layer). DAML-OIL builds on RDF, adding more complex

object-oriented concepts. DAML-OIL was submitted to the World Wide Web Consortium's Semantic Web working group, as the basis for the standardized OWL language, which is currently under development. It is likely that when OWL emerges from committee, it will bear a strong resemblance to DAML-OIL, although there may be significant differences.

Although web services and the semantic web are being defined by two very different communities, there is a growing realization of the potential synergy between the two. The basic idea here is that web services are a powerful means to deliver semantic information, while enhancing web service standards with semantic information will increase their flexibility and their effectiveness. The combination of Web Services with the Semantic Web is referred to (not surprisingly) as *Semantic Web Services*.

How Does Semantic Information Improve Service Discovery?

As discussed above, the current methods of discovering and invoking services and operations do not deal with the meaning of either the operations themselves or the classes of the parameters and return values (the input to and output from the operations). This limitation can be described succinctly as the *lack of semantic information*.

The premise of the research project currently in progress at the CADRC is as follows: If service definitions were to be expanded to include a formal description of the purpose of each operation, as well as an ontology that relates the classes of objects being passed in and out to other classes and concepts in the domain of knowledge, it would become possible for a would-be consumer to make a more intelligent determination of the suitability of a given operation to the consumer's own needs.

Semantic information concerning the service's domain of knowledge may also allow the prospective consumer to map operation parameters and return values to the consumer's own representation. This mapping is critical to the possibility of accessing services with unknown operation models. In other words, the consumer needs to be able to determine what information must be sent each operation, and what to do with the information received from the service. Without the ability to create mappings to and from its own information model at run-time, a consumer is again restricted to exactly those services that implement a service model that is built into the consumer during design and development.

In a previous pilot project, CADRC developers defined several ontologies using DAML (DARPA Agent Markup Language) and demonstrated the ability of a client program to automatically merge ontologies from multiple services under controlled conditions. This demonstration project also showed that users could extend the information model of a program at execution time, and that inference rules written for a specific ontology could also operate on instances of classes that the client had received from service provider, even though the classes were not known to the developers that wrote the rules. A future project will extend this investigation to include DAML-S (DAML Services), a vocabulary with semantics for defining the capabilities of services. We hope to learn whether the use of DAML-S to define the semantics of a service can enable a consumer to discover and utilize services without the need for each service to implement a specific interface.

The goal of adding semantic-level service descriptions is to enable consumers to locate services based on each service's purpose, rather than the names of the service's operations or the types of its parameters. This description of purpose is stated in a formal language that can be interpreted by the client. Automated reasoning can then determine the relationships between each operation and the definition of services needed by the client. Semantic description will allow service discovery to become more flexible, and will eventually lead to systems that can evolve as more services become available and the needs of the users of client program change over time.

The TEGRID Demonstration System

To demonstrate the use of web services and discovery in a decision-support system, the CADRC implemented a system within the context of emergency management of the rolling power outages experienced in many parts of California during the summer of 2001. This system is called TEGRID (i.e., 'Taming the Electric Grid').

Based on knowledge acquisition performed under the auspices of the National Institute for Urban Search and Rescue (NIUSR), and with the cooperation of the Los Angeles Sheriff's Department in the Fall of 2001, we identified several distinct entities that would be involved in planning and executing responses to power outage situations. Among these: the local sheriff stations (LSS); rapid response teams (RRT); the power supply organization (PSO); the traffic control organization (TCO); and the emergency operations bureau (EOB). The remainder of this section describes the responsibilities and actions of each of these major participants in the TEGRID demonstration system.

In the demonstration scenario, the Emergency Operations Bureau (EOB) is responsible for coordinating responses to the announcement of power outages. This coordination potentially includes a wide variety of decisions and communications. For the purposes of the scenario, we implemented only the assignment of Rapid Response Teams to provide support for local sheriff stations at priority locations.

Since the focus of this demonstration was on constructing the system through discovery and loosely-coupled communication through web services, we simulated the existence of several information sources:

1. A database at each local sheriff station that contains current officer assignments, equipment manifests and status, and priority infrastructure and intersections.

2. Lists of Rapid Response Teams (RRTs) and their primary and alternative assignments, maintained by the Emergency Operations Bureau (EOB).

3. Current power supply information, along with alerts about planned and current power outages, maintained and disseminated by a Power Supply Organization (PSO).

4. Traffic information, especially alternative route planning, supplied by the Traffic Control Organization (TCO).

5. Incident reports, fed into the system from 911 emergency lines and other sources.

Most of these information sources do exist, but are not (currently) available as web services. With the possible exception of the alternative route information, creating web services for these systems would be straightforward, given the cooperation of the agencies and organizations involved.

In addition, we implemented two services that are not part of the problem statement but are essential to the operation of the system. The first service is the Web Services Kiosk (WSK). Currently, this is an implementation of the Universal Description, Discovery, and Integration (UDDI) standard, but over time we expect this to evolve into an expanded service that will provide advanced semantic-based discovery services and will be the key to future semantic web services. The second infrastructure service is the Subscription Manager. This is our implementation of a web service that provides the ability for entities in the system to register their interests in information on specific topics, and to publish information that may be of interest to other participants in the system. The information is published as XML documents so that subscribers are not dependent on any static definition of the contents.

The TEGRID Demonstration Scenario: Initial Discovery Phase

The primary visible participants in the demonstration are the Emergency Operations Bureau (EOB) and the two Local Sheriff Stations, Lomita and East Los Angeles. Each of them starts with only one piece of information: the location of the Web Services Kiosk (WSK). The process of discovery is similar regardless of the order in which the three primary participants are started, but for the purposes of this discussion, we will assume that the EOB is started first.

The EOB queries the WSK for the location of a service that can provide publish-and-subscribe functionality. This functionality is defined, as described above, through the use a publicly accessible service model. The WSK finds the only match, which is the Subscription Manager, and returns the Subscription Manager's location (its URI) to the EOB. The EOB also searches the WSK for a service that can monitor power supply levels and send alerts when a power outage is about to begin. The WSK returns the URI of the Power Supply Organization service.

Using the Subscription Manager, the EOB subscribes to notifications about the creation of any Local Sheriff Station (LSS). The EOB also publishes a description of itself, in case there are other elements of the system that have subscribed to the creation of EOB entities. Additionally, the EOB subscribes to notifications of power outages.

Now we start one of the LSS clients – Lomita, for instance. Lomita goes through the same process of using its knowledge of the WSK to discover the Subscription Manager, and registers its interest in receiving messages regarding the creation of EOBs. Lomita also publishes information about itself. Since there is already a subscriber for notifications of LSS creation, the Subscription Manager passes along Lomita's information to the EOB. The EOB adds the Lomita information to its knowledge base, and replies with its own information. Now the EOB and the Lomita station know each other's web location, which means that they can communicate directly with each other, and the EOB has information about Lomita, particularly the RRTs assigned there, and the station resources. When the East Los Angeles station is started, the same process occurs, resulting in East Los Angeles and the EOB learning each other's web location, and the EOB learning all the information that East Los Angeles has provided.

As each participant has entered the system, several agents resident within each one have also subscribed to information on various topics. We will see the effects of these subscriptions in a later phase.

At this point, we have demonstrated the use of two different kinds of discovery. First, there is registry-based discovery, which allowed all the participants to locate the Subscription Manager, and the EOB to find the Power Supply service. Each participant has now created its own awareness of the context in which it is operating, and has exchanged information with other entities in the system. TEGRID has constructed itself in an ad hoc manner, based on available information sources, and has established a loosely-coupled communication system. This is the end of the discovery phase.

The TEGRID Demonstration Scenario: Operational Phase

The operational phase begins when the Power Supply Organization (PSO) determines that a rolling power blackout is imminent (i.e., is planned to begin in fifteen minutes). The PSO publishes that information to all subscribers, which in this case includes only the EOB.

On receiving the imminent power outage alert, the EOB immediately broadcasts the alert to all Local Sheriff Stations. The EOB then uses its Station Monitor Agent to determine which LSSs will experience the outage within their jurisdiction, using the information provided by each station at the time of discovery. These LSSs receive a second alert at a higher priority level. In this scenario, the only station directly affected is East Los Angeles. The EOB also alerts the RRTs assigned to assist the affected stations, so that these RRTs will begin to prepare for deployment. When an LSS receives a power outage alert, it assumes a state of readiness consistent with whether the outage is within its jurisdiction or not.

The second stage in the operational phase occurs when the power outage actually occurs. This is a repeat of the previous step, except that all elements move to a higher state of readiness. The third stage of the operational phase begins with a report of a traffic accident within East Los Angeles' jurisdiction, sent into the system by the Incident Report information source. The East Los Angeles Sheriff Station determines that it does not have sufficient resources to cover the traffic accident due to the power outage. East Los Angeles therefore requests additional resources from the EOB.

The EOB service receives the request and uses its own Scheduling Agent to assign an RRT and other equipment to the traffic accident. In addition, the EOB creates an Incident Agent to monitor further messages relating to the accident.

The fourth (and final) stage of the operational phase begins with the RRT finding that the route to the traffic accident is blocked by traffic due to the fact that many signals are not functioning because there is no power in the area. The RRT requests assistance from the EOB in finding an alternative route. The EOB Incident Agent queries the Web Service Kiosk for a service that implements an alternative traffic route model. The WSK responds with the address of the Traffic Control Service. The Incident Agent sends the request for assistance to the Traffic Control Service, which utilizes its own Routing Agent to determine an alternative route to the traffic accident. The Traffic Control Service sends this route as a reply to the EOB Incident Agent.

At this point, the Incident Agent displays the alternative route to the user of the EOB client system. This allows a human being to examine this route and accept it, alter it, or ask for another

route. Keeping a human in the loop is vital here and at many other places in the system, because the human being will almost always be aware of information that is not available through the system. With the user's acceptance, the alternative route is sent to the RRT, which also accepts it. This concludes the demonstration scenario.

Aspects of the TEGRID Demonstration

The TEGRID demonstration system shows the ability of a system to configure itself based on the available participants and their intents and interests. It also demonstrates the ability of loosely-coupled systems to exchange object-oriented information when the communications protocol is chosen for this purpose. We demonstrated further that participants in a distributed system can extend their awareness of the information available to them, and do not need to rely on pre-defined knowledge of information sources.

On the negative side, we also proved to ourselves the limitations of existing web service standards due to the lack of semantic information. Elements of TEGRID were not required to know exactly what servers they would be using, but we could not eliminate totally the necessity for each element to have knowledge of the interfaces provided by different kinds of services.

Future Directions

As indicated previously, future work will be focused in the area of adding semantic information to the descriptions of services and to the ontology of a service. We will be examining the suitability of the new DAML-Services language for this purpose, by defining representations of the information needs of each participant, and determining whether DAML-S can be used to locate services without an exact match.

We will also be exploring further the use of DAML as an ontology representation language, and creating tools and techniques that will eventually allow systems built around the concept of the semantic web to merge disparate ontologies into a single information model.

Other researchers working in similar areas can be found through articles listed below in the reference section.

References

(It should be noted that there are a large number of resources on these topics, most of which are on the World Wide Web. The following is a short list of starting points, and should not be taken as definitive or complete.)

Special Theme Issue: The Semantic Web, ERCIM News, No. 51, October 2002, http://www.ercim.org/publication/Ercim_News/enw51/

Berners-Lee, Tim, James Hendler and Ora Lassila, The Semantic Web; Scientific American, May 2001

Graham, Steve, et al., Building Web Services with Java: Making Sense of XML, SOAP, WSDL and UDDI; Sams Publishing, Indianapolis, IN, 2001

McIlraith, Sheila A., Tran Cao Son and Honglei Zeng, Semantic Web Services; IEEE Intelligent Systems, March/April 2001 (pp. 46-5)

Oellermann, William L., Architecting Web Services; Apress, Berkeley, CA, 2001

Pallos, Michael S., Service-Oriented Architecture: A Primer; eAI Journal, December 2001, http://www.eaijournal.com/PDF/SOAPallos.pdf

Appendix A: Semantic Web Services and Network-Centric Systems

A recently announced goal of the Defense Information Systems Agency (DISA) is a move to what has been dubbed 'net-centricity' or 'network-centric'. This goal is directly related to the enterprise architecture known as the 'Global Information Grid' (GIG). As John Osterholz, the director of architecture and interoperability in the Department of Defense's CIO office recently stated: "We believe ultimately that the key to managing data overload is making commanders responsible for pulling the data that they need into their decision space rather than having some galactic, on-line genius decide what they need". (Osterholz's statement was reported on the following web site: http://www.gcn.com/21_29/management/20098-1.html.)

This vision has a deep correlation with the goals of the project described in this paper. In order for each commander (or any other human or software agent) to pull the data (information) that suits his, her, or its current needs, most existing systems would require the user to identify an exact set of information sources. But in a distributed system of the size envisioned by the Department of Defense, it is unlikely that any individual would be able to be aware of all the potential information sources within this network-centric system.

Clearly, a form of automated discovery is necessary for the success of net-centricity. There may be a temptation to repeat past practices and create a (very large) number of standard service models, so that service providers can search registries for service providers implementing those models. If this occurs, commanders using specific client software programs will be restricted to accessing information providers whose models are built into the software. This will be problematic if the commander is in a fluid situation that might benefit from the input of unplanned types of information.

A far more flexible solution will include the addition of semantic descriptions of service capabilities to the discovery system. Client software will then be able to describe the information needs of the user in order to locate appropriate information sources. As the user's needs change, he or she will be able to direct the client software to seek out new services to provide new kinds of information. The result will be the right information delivered to the right person at the right time.

Web services in general also address two other problems that continue to obstruct progress in moving toward transformation of the Department of Defense's information technology systems: interoperability; and ,legacy systems. Web services help to solve both problems by putting a web-based interface on legacy systems. This interface is by definition interoperable with web service clients, which may include other systems that could use the information contained in the legacy system. Adding a web service interface to a legacy system is far more cost-effective than

re-engineering or replacing the system, and in the case of client/server systems, constructing a web service interface can generally be performed over a short period of time, contrasting with the major effort involved in developing replacement systems. Finally, in the case where multiple legacy systems are planned to be replaced by a single new system in the future, the web service interface can serve as a portal that allows clients to access information from the legacy system now, while forwarding requests to the new system as it comes on-line, without rewriting the web service client.

Appendix B: Organizations and Standards

Resource Description Framework (RDF)
http://www.w3.org/RDF/

DAML Organization
www.daml.org
The source for information on the DARPA Agent Markup Language

Semantic Web Working Group of the World Wide Web Consortium
http://www.w3.org/2001/sw/

Web Services Working Group of World Wide Web Consortium
http://www.w3.org/2002/ws/

UDDI
http://www.uddi.org/

WSDL
http://www.w3.org/TR/wsdl

The TEGRID Semantic Web Application: A Service-Oriented Distributed Approach to Disaster Management Decision-Support Systems

Project Investigators:

Jens G. Pohl, Ph.D.
Executive Director
CAD Research Center

Steven J. Gollery
Software Engineer
CAD Research Center

DEMONSTRATION
A system with discovery, reasoning, and learning capabilities.
[ONR Workshop (2002): Wednesday, Sep.18 at 1 pm]

The TEGRID Semantic Web Application

Steven Gollery, Senior Software Engineer
Jens Pohl, Ph.D., Executive Director
Collaborative Agent Design Research Center
Cal Poly, San Luis Obispo, California

Introduction

Over the past several years there has been an increasing recognition of the shortcomings of message-passing data-processing systems that compute data without understanding, and the vastly superior potential capabilities of information-centric systems that incorporate an internal information model with sufficient context to support a useful level of automatic reasoning.

The key difference between a data-processing and an information-centric environment is the ability to embed in the information-centric software some understanding of the information being processed. The term ***information-centric*** refers to the representation of information in the computer, not to the way it is actually stored in a digital machine. This notion of ***understanding*** can be achieved in software through the representational medium of an ontological framework of objects with characteristics and interrelationships (i.e., an internal information model). How these objects, characteristics and relationships are actually stored at the lowest level of bits in the computer is immaterial to the ability of the computer to undertake reasoning tasks. The conversion of these bits into data and the transformation of data into information, knowledge and context takes place at higher levels, and is ultimately made possible by the skillful construction of a network of richly described objects and their relationships that represent those physical and conceptual aspects of the real world that the computer is required to reason about.

In a distributed environment such information-centric systems interoperate by exchanging ontology-based information instead of data expressed in standardized formats. The use of ontologies is designed to provide a context that enhances the ability of the software to reason about information received from outside sources. In the past, approaches to inter-system communication have relied on agreements to use pre-defined formats for data representation. Each participant in the communication then implemented translation from the communication format to its own internal data or information model. While relatively simple to construct, this approach led to distributed systems that are brittle, static, and resistant to change.

It is the premise of the TEGRID (Taming the Electric Grid) proof-of-concept demonstration that, for large scale ontology-based systems to be practical, we must allow for dynamic ontology definitions instead of static, pre-defined standards. The need for ontology models that can change after deployment can be most clearly seen when we consider providing information on the World Wide Web as a set of web services augmented with ontologies. In that case, we need to allow

client programs to discover the ontologies of services at run-time, enabling opportunistic access to remote information. As clients incorporate new ontologies into their own internal information models, the clients build context that enables them to reason on the information they receive from other systems. The flexible information model of such systems allows them to evolve over time as new information needs and new information sources are found.

The TEGRID Demonstration Context

Since mid-2001 the Emergency Operations Bureau of the Los Angeles Sheriff's Department has been assigned the additional task of coordinating the response to expected rolling electric power blackouts, as California's demand for electric power came perilously close to exceeding availability. While both the power outage areas and individual blackout periods are predefined in terms of a large number of power grid units that are distributed throughout the Los Angeles County, the emergency events that are likely to be triggered by blackout conditions (e.g., multi-vehicle accidents, carbon monoxide poisoning in enclosed parking garages, fires, criminal activities, and other disturbances) are less determinate.

The TEGRID proof-of-concept system has been designed to assist the Los Angeles Sheriff's Department by addressing this potentially chaotic situation in an autonomously evolving, just-in-time manner. TEGRID does not exist as a pre-configured system of tightly bound components that know about the existence of each other, have predefined connections, and predetermined capabilities. In fact at the beginning of the demonstration TEGRID, as a system, does not really exist at all. What does exist is a set of cooperating Semantic Web Services, based on standard Web Service specifications (e.g., SOAP, UDDI, WSDL, and XML) enhanced by the ability of sharing semantic-level descriptions of their own internal information models.

In essence TEGRID involves sharing information among a number of separate organizations, including local police stations, the Emergency Operations Bureau, a power supply management and monitoring organization, and a traffic control system. The proof-of-concept relies on a set of assumptions about the existing resources available from each of the organizations involved.

1. That each local sheriff's station has a database that includes (at least): current officer assignments; equipment manifests and status; and, priority infrastructure and intersections.
2. That the Emergency Operations Bureau has a list of Rapid Response Teams and their primary and alternative assignments.
3. That there exists some kind of Power Supply Organization that has a database of recent history of power consumption, plus the ability to provide a real-time feed of current power levels.
4. That there exists some kind of Traffic Control Organization that has some method of determining acceptable alternative routes for reaching a particular destination from a given starting location.

Another underlying assumption is that all of these organizations have Internet connections and either have an existing web site or are willing to establish one. TEGRID builds on these existing information and data sources to construct a web service infrastructure that allows information-sharing and automated decision-support.

Since the proof-of-concept system does not have access to live databases, it simulates them, using sample data to implement the demonstration scenario. There are also some potential applications that must exist in order to support the scenario, but are not part of TEGRID itself. For example, there is a requirement that new incidents (e.g., traffic accidents) would be reported to the local sheriff's stations before they are able to propagate through the system. Such a reporting application is assumed to exist, and has been simulated in order to produce the dynamic behavior called for in the demonstration scenario.

TEGRID features several kinds of web service providers. Each of these implements a set of operations that allows exchange of the information that makes the functioning of the system possible. These operations such as subscription, information transfer, warning and alert generation, discovery, and assignment, are the minimum necessary to provide the functionality described in the demonstration. More operations can be easily added as TEGRID's capabilities increase in the future.

In addition, TEGRID includes software agents with automatic reasoning capabilities. Some of these agents could conceptually be seen as services. For instance, the Station Monitor Agent is able to publish alerts that the local stations can subscribe to, and at the same time the Station Monitor Agent is able to subscribe to notifications of planned power outages. The relationship between agents and services is perhaps a fertile field for further investigation: When is it more useful to implement functionality as an agent, and when as a service? Are the two orthogonal? Is it reasonable to think that the same set of functions might be an agent from one point of view, but a service from another? Does an agent consume services, provide services, or both? Since it seems likely that the answers to these questions depend on the nature of the individual agent, the definition of a conceptual framework for making such determinations might be a productive future goal.

The Fundamental Web Service Elements

Within the Internet context of web services, TEGRID builds on a number of standard protocols and elements. These elements are combined into an executing software entity, capable of seeking and discovering existing web services, extending its own information model through the information model of any discovered web service, and automatically reasoning about the state of its internal information model. As shown in Fig.1, this entity or Cyber-Spider consists of three principal components: a web server; a semantic web service; and, an information-centric application.

The web server, utilizing standard Hypertext Transfer Protocol (HTTP), serves as the gateway through which the Cyber-Spider gains access to other existing web services. Web servers primarily provide access to Hypertext Markup Language (HTML) data sources and perform only simple operations that enable access to externally programmed functionality. However, these simple operations currently form the building blocks of the World Wide Web.

The second component of a Cyber-Spider is a semantic web service (i.e., a web service with an internal information model). A web service is accessed through a web server utilizing standard protocols (e.g., UDDI, SOAP, WSDL, SML) and is capable of providing programmed functionality. However, clients to a standard web service are usually restricted to those services that implement specific predefined interfaces. The implementation of web services in the Internet environment allows organizations to provide access to applications that accept and return complex objects. Web service standards also include a limited form of registration and

discovery, which provide the ability to 'advertise' a set of services in such a way that prospective client programs can find services that meet their needs. The addition of an internal information model in a semantic web service allows the storage of semantic level descriptions (i.e., information) and the performance of limited operations on these semantic descriptions. In other words, the semantic web server component of a Cyber-Spider is capable of reasoning.

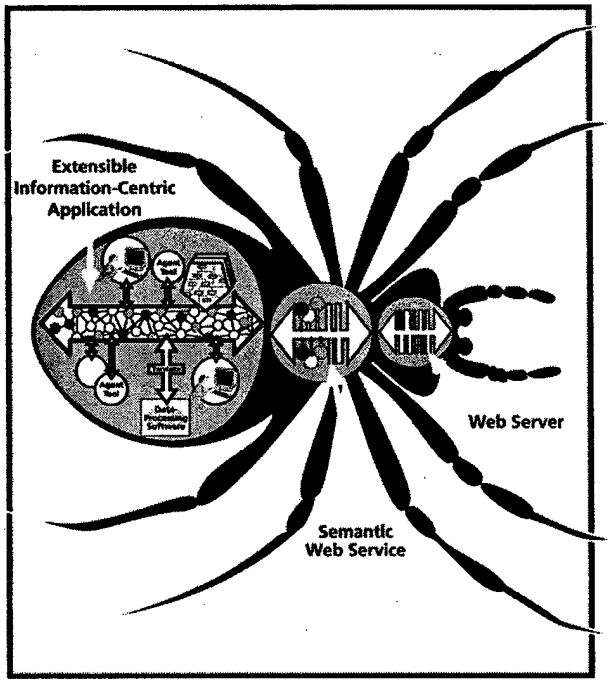


Fig.1: Anatomy of a Cyber-Spider

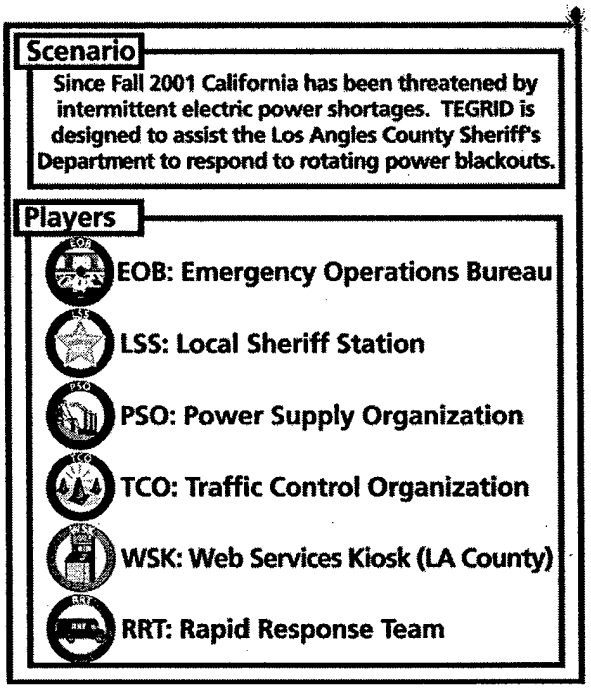


Fig.2: Cast of TEGRID players

The third component of a Cyber-Spider is one or more information-centric applications. These applications are designed to take advantage of the resources provided by a number of semantic web services, enabling them to reason about the usefulness of each service and support more sophisticated discovery strategies. Moreover, the application component is able to construct relationships among the information models of different services, with the ability to integrate services without requiring agreement on a common information model.

With these three components Cyber-Spiders are at least minimally equipped to operate in an Internet environment as autonomous software entities, capable of: discovering needed services; accepting services from external offerers; providing services to external requesters; gaining context through an internal information model; automatically reasoning about available information; extending their information model during execution; extending their service capabilities during execution; and, learning from their collaborations.

The TEGRID Players

The cast of players in the current TEGRID proof-of-concept demonstration includes six players or existing web services (Fig.2): the Emergency Operations Bureau (EOB) of the Los Angeles Sheriff's Department; several Local Sheriff Stations (LSS); a Power Supply Organization (PSO); a Traffic Control Organization (TCO); several Rapid Response Teams (RRT); and, a Los Angeles County Web Services Kiosk (WSK).

Fundamental to each player are three notions. First, each player operates as an *autonomous* entity within an environment of other players. Most, but not all of the other players are also autonomous. This requires the autonomous players to be able to discover the capabilities of other players. Second, each autonomous player has a sense of *intent* to accomplish one or more objectives. Such objectives may range from the desire to achieve a goal (e.g., maintain situation awareness, coordinate the response to a time critical situation, or undertake a predetermined course of action following the occurrence of a particular event) to the willingness to provide one or more services to other players. Third, each player (whether autonomous or not) is willing to at least *cooperate* with the other players. In some cases the level of cooperation will extend to a collaborative partnership in which the partnering players contribute to the accomplishment of a common objective. In other cases the cooperation may be limited to one player providing a service to another player, without any understanding or interest in the reason for the service request.

To operate successfully in such an autonomous Internet-based environment a Cyber-Spider player should be endowed with the following capabilities:

1. Subscribe to information from external sources (e.g., alerts, ontology extensions).
2. Accept subscriptions from external clients.
3. Dynamically change its subscription profile.
4. Extend its internal information representation.
5. Extend its own service capabilities.
6. Generate new agents for its own use.
7. Describe its own service capabilities to external clients.
8. Seek, evaluate and utilize services offered by external clients.
9. Provide services to external clients.
10. Describe its own (intent) nature to external clients.

The Cyber-Spiders in TEGRID are currently capable of demonstrating eight of these ten desirable capabilities. The ability of a Cyber-Spider to dynamically change its subscription profile, while technically a fairly simple matter, has not been implemented because it is not used in the demonstration scenario. The ability of a Cyber-Spider to describe its own nature to external clients, on the other hand, is technically a much more difficult proposition. It will require a Cyber-Spider to have an understanding of its personality as a collective product of its internal information model and the relationship of that model with the external world. At best this must be considered a challenging research area that is beyond the current capabilities of information-centric software systems.

The TEGRID Agents

Most of the reasoning capabilities available in TEGRID are performed by software agents that are components of the players (e.g., Cyber-Spiders). In other words, agents are predefined clients within player systems (i.e., information-centric applications) and perform internal functions that are necessary for the particular player to deliver its services and/or accomplish its intent. The following agents (i.e., collaborative tools) are available in the current TEGRID implementation:

Name of Agent	Owner	Description of Agent Capabilities
Risk Agent	EOB	Identifies high risk entities in the jurisdictional region of an activated LSS.
Deployment Agent	EOB	Determines whether RRT support is required for a particular activated LSS.
Power Level Agent	PSO	Determines if electric power demand has exceeded supply.
Situation Agent	EOB	Prepares and updates the 'EOB Situation Status Report'.
Station Monitor Agent	EOB	Identifies all LSSs that will experience power blackouts during the current and next blackout cycle.
Status Agent	LSS	Prepares and updates the 'LSS Situation Status Report'.
Local Station Agent	LSS	Determines whether sufficient local resources are available to deal with current conditions.
Scheduling Agent	EOB	Assigns RRTs and equipment to situations requiring RRT involvement.
Incident Agent	EOB	Monitors the response to a particular situation supported by one or more RRTs.
Routing Agent	TCO	Determines alternative routes to a particular situation location.

Demonstration Objectives

Stated succinctly, the objective of the TEGRID scenario is to demonstrate the discovery, extensibility, collaboration, automatic reasoning, and tool creation capabilities of a distributed, just-in-time, self-configuring, collaborative multi-agent system in which a number of loosely coupled Web Services associate opportunistically and cooperatively to collectively provide decision assistance in a crisis management situation. Specifically, these capabilities are defined as follows:

Discovery: Ability of an executing software entity to orient itself in a virtual cyberspace environment and discover other software services.

Extensibility: Ability of an executing software entity to extend its information model by gaining access to portions of the information model of another executing software entity.

Collaboration: Ability of several Web Services to collaboratively assist each other and human users during time critical decision making processes.

Reasoning: Ability of a software agent to automatically reason about events in near real time under time critical conditions.

Tool Creation: Ability of a Web Service to create an agent to perform specific situation monitoring and reporting functions.

Players' Intent

The TEGRID players or Cyber-Spiders are initialized with intent or willingness to cooperate based on their role and operational responsibilities, as follows:

EOB (Emergency Operations Bureau): To be immediately informed of imminent power blackout conditions, to coordinate all assistance to LSSs, to maintain situation awareness, and to take over local command responsibilities when conditions require actions that cross the jurisdictional boundaries of two or more LSSs.

LSS (Local Sheriff Station): To activate a predefined response plan as soon as it receives notification (from the EOB) that a power blackout condition is imminent within its jurisdiction, to respond to new emergency missions in its jurisdictional area, to provide RRTs to the EOB, and to request assistance from the EOB.

PSO (Power Supply Organization): To share information relating to the current status of power demand and availability with subscribers, to provide subscribers with information relating to a predefined rolling power blackout schedule on request, and to alert subscribers whenever the schedule is intended to be implemented.

TCO (Traffic Control Organization): To share information relating to historical traffic flows under typical conditions with subscribers, to provide subscribers with information relating to traffic control capabilities (e.g., types and location of traffic signals, sensors, and web-cameras), and to provide subscribers with alternate traffic routes on request.

RRT (Rapid Response Team): To share information relating to its current mission and location with subscribers, to execute missions requested by the EOB, and to provide assistance to any assigned LSS, and to request assistance from the EOB.

The TEGRID Demonstration Scenario

Armed with their individual intent and intrinsic Cyber-Spider capabilities (i.e., ability to: discover useful web services; subscribe to information and accept subscriptions from external clients; extend their internal information models; describe and provide services to external

clients; seek, evaluate and utilize services offered by external clients; and, extend their own service capabilities by generating new agents) the players commence their partly intentional and mostly opportunistic interactions.

Orientation

The players orient themselves in the virtual cyberspace environment by accessing one or more directories of available services and registering an information subscription profile with those services that they believe to be related to their intent (Fig.3).

EOB (Emergency Operations Bureau): Accesses the WSK (Los Angeles County Web Services Kiosk) based on its predefined authorization level, and:

Subscribes to any service changes in the WSK.
Finds the PSO address which it was seeking.
Discovers the TCO.
Discovers all of the LSSs.

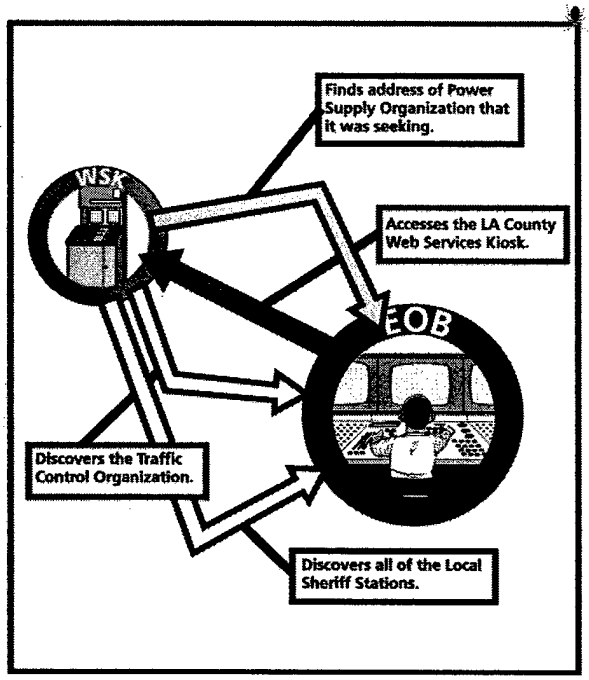


Fig.3: Orientation and discovery

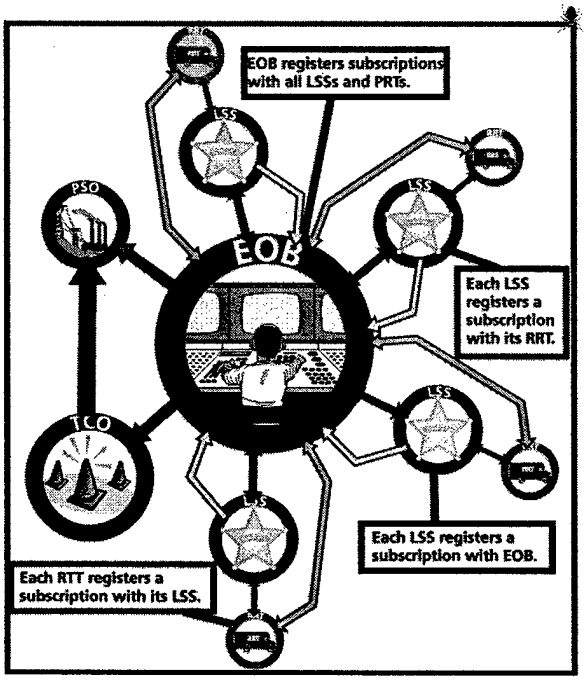


Fig.4: Information subscription

Subscription

The players access the services that they require to achieve their intent, register appropriate subscription profiles, and query for information that they believe to have a need for (Fig.4).

EOB (Emergency Operations Bureau): Registers a subscription profile with each LSS (Local Sheriff Station) that includes all current police unit locations, mission completion events, new mission events, and any information changes relating to the availability of its RRTs (Rapid Response Teams).

Queries each LSS (Local Sheriff Station) for all information relating to its RRTs (Rapid Response Teams) and extends its information model.

Registers a subscription profile with each RRT (Rapid Response Team) that includes its current location and mission.

Registers a subscription profile with the PSO (Power Supply Organization) that includes the current status of electric power demand and availability, and any change in its intention to implement the predefined rolling power blackout schedule.

Registers a subscription profile with the TCO (Traffic Control Organization) that includes any change in the status of traffic signals, sensors, and web-cameras.

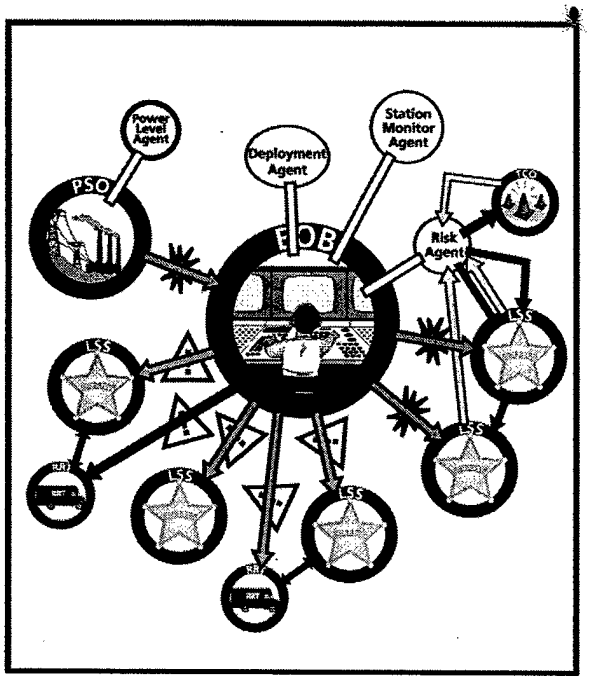


Fig.5: Power supply 'Warning'

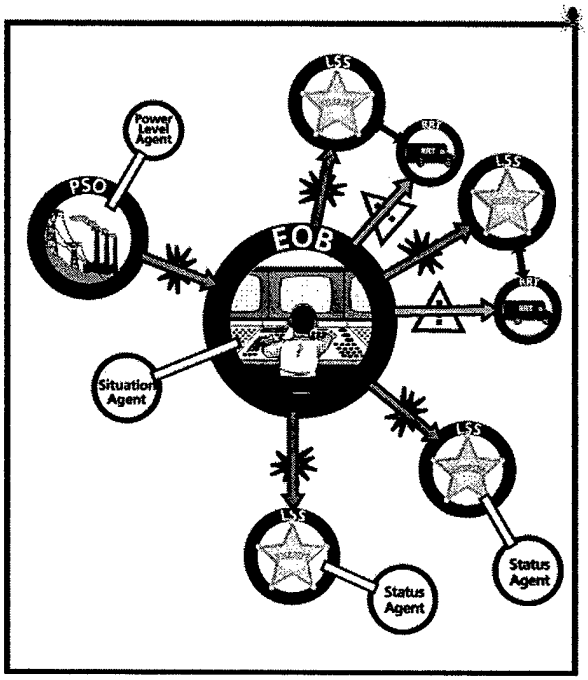


Fig.6: Power outage 'Alert'

LSS (Local Sheriff Station): Each LSS responds to the EOB (Emergency Operations Bureau) registration by registering a corresponding subscription profile with the EOB that includes the current mission and location of its RRTs (Rapid Response Teams), any EOB requests and orders to this LSS, and changes in the current 'situation status report' maintained by the EOB.

Each LSS (Local Sheriff Station) registers a subscription profile with its RRTs (Rapid Response Teams) that includes the current mission and location of the RRT, mission completion events, and new mission events (this duplication of its EOB (Emergency Operations Bureau) subscription profile allows the LSS to verify the accuracy of this portion of the 'situation status report' maintained by the EOB).

TCO (Traffic Control Organization): Registers a subscription profile with the PSO (Power Supply Organization) to include the location of all current power blackout areas.

RRT (Rapid Response Team): Registers a subscription profiles with the EOB (Emergency Operations Bureau) that includes any requests or orders to this particular RRT (Rapid Response Team), and any changes in conditions that impact the current mission and location of this RRT.

Registers a subscription profile with its home base LSS (Local Sheriff Station) that includes any request for information, and any 'situation status report' maintained by this LSS.

Power Outage Notification

The PSO (Power Supply Organization) alerts its subscribers that a rolling power blackout condition is imminent (i.e., will commence per predefined schedule within 15 minutes) (Fig.5).

PSO (Power Supply Organization): Utilizes its Power Level Agent to continuously monitor the relationship between power demand and supply. The PSO determines that demand is close to exceeding supply and sends an Alert to all appropriate subscribers.

EOB (Emergency Operations Bureau): Receives an Alert from the PSO (Power Supply Organization) that the predefined rolling power blackout schedule will be implemented within 15 minutes.

Utilizes its Station Monitor Agent to identify all LSSs (Local Sheriff Stations) that will experience power blackouts in their jurisdiction.

Warns all LSSs (Local Sheriff Stations) of imminent power blackout condition.

Alerts all LSSs (Local Sheriff Stations) in whose jurisdictions blackouts will occur and requests them to commence immediate implementation of their respective 'blackout response plans'.

Warns the RRTs (Rapid Response Teams) assigned to assist the LSSs (Local Sheriff Stations) in whose jurisdictions the first set of blackouts are scheduled to occur, to prepare for potential deployment.

Utilizes its Risk Agent to identify all high risk entities in the jurisdictions of the activated LSSs (Local Sheriff Stations). Utilizes its Deployment Agent to determine whether RRT (Rapid Response Team) involvement is anticipated under normal conditions.

LSS (Local Sheriff Station): Each LSS assumes 'alert' status. The LSSs in whose jurisdictions the first set of blackouts is scheduled to occur, prepare for deployment.

RRT (Rapid Response Team): The RRTs notified by the EOB (Emergency Operations Bureau) assume 'alert' status in preparation for potential deployment.

Power Outage Implementation

The PSO (Power Supply Organization) alerts its subscribers that the predefined rolling power blackout schedule has been implemented (Fig.6).

PSO (Power Supply Organization): Utilizes its Power Level Agent to determine that demand has exceeded the availability of electric power.

EOB (Emergency Operations Bureau): Receives an Alert from the PSO (Power Supply Organization) indicating that the predefined rolling power blackout schedule has been implemented.

Utilizes its Situation Agent to prepare the first version of the 'EOB Situation Status Report'.

Alerts all LSSs (Local Sheriff Stations) in whose jurisdictions the next scheduled set of blackouts will occur, to prepare for potential deployment.

Warns the RRTs (Rapid Response Teams) assigned to assist the LSSs (Local Sheriff Stations) in whose jurisdictions the next set of blackouts are scheduled to occur, to prepare for potential deployment.

LSS (Local Sheriff Station): All activated LSSs utilize their Status Agent to prepare the first version of their 'LSS Situation Status Report'.

The LSSs (Local Sheriff Stations) in whose jurisdictions the next set of blackouts is scheduled to occur, prepare for deployment.

Traffic Accident in Power Outage Area

A multi-car traffic accident occurs in a blackout area located within the jurisdiction of a particular LSS (Local Sheriff Station) (Fig.7).

EOB (Emergency Operations Bureau): Receives an Alert from a LSS (Local Sheriff Station) that a multi-car traffic accident has occurred on State Highway 5 south of Harbor Freeway.

LSS (Local Sheriff Station): Utilizes its Local Station Agent to determine that it has insufficient resources to deal with the multi-car traffic accident.

EOB (Emergency Operations Bureau): Receives a request for assistance from the LSS (Local Sheriff Station) to deal with the multi-car traffic accident.

Utilizes its Scheduling Agent to assign a RRT (Rapid Response Team) and equipment to the multi-car traffic accident.

Creates an Incident Agent to monitor the response to the multi-car traffic accident.

The new Incident Agent subscribes to the LSS (Local Sheriff Station) in whose jurisdiction the multi-car traffic accident has occurred (to obtain all information about this accident from now on).

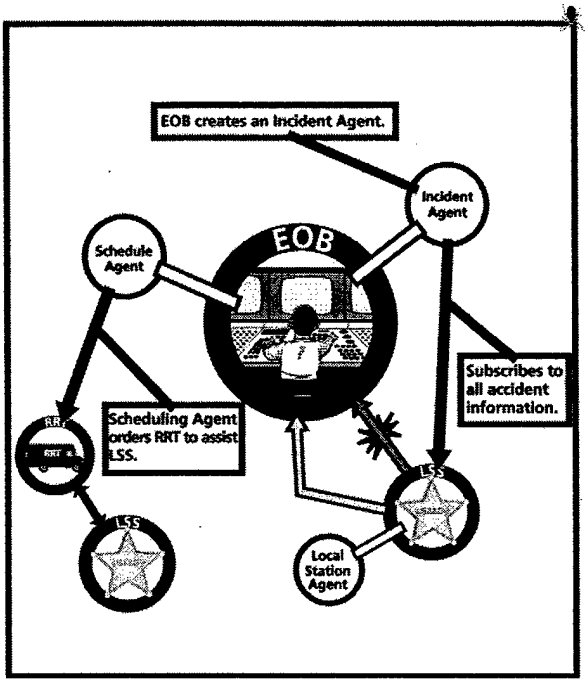


Fig.7: Traffic accident 'Alert'

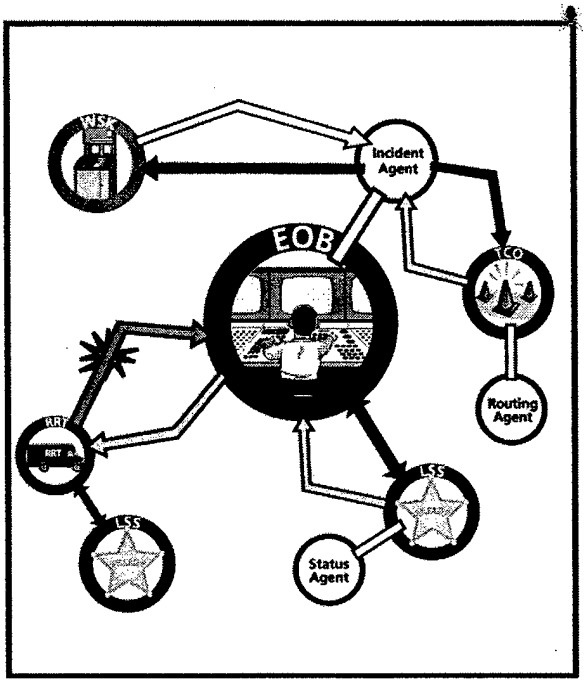


Fig.8: Routing assistance request

Routing Assistance Required

The dispatched RRT (Rapid Response Team) cannot reach the multi-car traffic accident due to traffic congestion and requests assistance in determining an alternative route (Fig.8) to the accident.

RRT (Rapid Response Team): Sends alert to the EOB (Emergency Operations Bureau) and requests assistance in determining an alternative route to the traffic accident.

EOB (Emergency Operations Bureau): Utilizes its Incident Agent to determine an alternative route. The Incident Agent accesses the WSK (Los Angeles County Web Services Kiosk) and discovers the TCO (Traffic Control Organization). It then registers a subscription profile with the TCO that includes routing information, and requests assistance in determining an alternative route to the traffic accident.

TCO (Traffic Control Organization): Receives the request for assistance from the EOB's (Emergency Operations Bureau) Incident Agent and utilizes its Routing Agent to determine an alternative route to the traffic accident.

Sends the alternate route to the EOB's Incident Agent..

EOB (Emergency Operations Bureau): Responds to the RRT (Rapid Response Team) by sending it the alternate route to the traffic accident.

Significance of the TEGRID Demonstration

The TEGRID proof-of-concept project was undertaken by the Collaborative Agent Design Research Center (CADRC) at Cal Poly (San Luis Obispo) as a small internally funded research endeavor with three objectives. The first objective was to explore the main capabilities that would be required of web service type entities (i.e., Cyber-Spiders) serving as largely autonomous decision-support components in a self-configuring, just-in-time, intelligent decision-assistance toolkit of collaborating software agents. Second, to determine if the currently available information-centric software technology could support at least basic (i.e., meaningful and useful) implementations of these required capabilities. And, third, to build a working experimental system that could serve as a test bed for longer term research studies focused on the behavioral characteristics of self-configuring intelligent systems in general, and the ability of such systems to deal with specific kinds of dynamic and complex problem situations.

The principal capabilities that are required by a Cyber-Spider to support the desired self-configuring, just-in-time, intelligent decision-support behavior have been identified and demonstrated in the TEGRID test bed environment, at least at a base level of functionality. These capabilities include the ability to: discover desired existing external services; accept and utilize services from external offerers; provide services to external requesters; gain understanding through the context provided by an internal information model; automatically reason about available information within the context of the internal information model; extend the internal information model during execution; spontaneously generate new agents during execution as the need for new capabilities arises; and, learn from the collaborations that occur within the cyberspace environment.

Viable Bandwidth Compression for Remote Sensing Applications

Project Investigator:

John A. Saghri, Ph.D.
Associate Professor
Electrical Engineering Department

Final Report

ONR-C3RP Funded Project

Viable Bandwidth Compression for Remote Sensing Applications

(Phase I)

John Saghri
Electrical Engineering Dept., Cal Poly, San Luis Obispo, CA 93407
jsaghri@calpoly.edu

December 2002

Summary

The major tasks of the project "viable bandwidth compression for remotely-sensed multispectral imagery data" were completed successfully. The results of this project are described in the attached two professional papers. The first paper "Analysis of JPEG-2000 versus JPEG for KLT-based Compression of Remotely-Sensed Imagery Data" (presented at the SPIE International Symposium, Seattle, July 2002) describes the implementation of our proposed JPEG-2000 based multispectral bandwidth compression system. It is shown that the proposed system generally outperforms the previous baseline JPEG-based system. The second paper "Class Prioritized Compression of Multispectral Imagery Data" (published in the April 2002 issue of Journal of Electronic Imaging) describes the added feature prioritization capability in our developed compression system.

Our previously developed multispectral compression system was based on a Karhunen-Loeve Transformation (KLT) for removing the spectral redundancy in the data, followed by the standard JPEG compression for removing the remaining spatial redundancy in the data. We successfully replaced the baseline JPEG module with the new standard JPEG-2000. As reported in the paper, the new JPEG-2000/KLT compression system offers superior performance compared to our former JPEG/KLT system. The JPEG 2000 compression utilizes a discrete Wavelet Transformation (DWT) as opposed to Discrete Cosine Transformation (DCT) which is used by the baseline JPEG. The new JPEG 2000 is relatively more sophisticated than the baseline JPEG. It is highly parameterized to fit the characteristics of various types of data. To incorporate a feature prioritization scheme in our design, we investigated a dual compression approach (see attached JEI paper for details). The dual system includes; (1) a primary unit (the baseline KLT/JPEG-2000 system described above) for conventional coding of multispectral image, and (2) an auxiliary unit to code the resulting compression-induced error incurring on features, or regions, of interest. The regions of interest are determined via a spectral unmixing procedure. The spectral unmixing transforms the original set of multispectral images into another set of images in which each images shows the concentration of each feature within the geographical area. We have demonstrated, via simulation, that this scheme does indeed provide a measurable feature discrimination capability in our compression system. That is, features of interest in the scene do maintain a relatively higher fidelity (precision level) in the compression process than the nonessential features.

Analysis of JPEG versus JPEG-2000 for KLT-based Compression of Multispectral Imagery Data[1]

John A. Saghri*[a], Andrew G. Tescher*[b], Frank E. Kozak*[c]
[a]Electrical Engineering Dept., Cal Poly, San Luis Obispo, CA. 93407, jsaghri@calpoly.edu
[b]Compression Science Inc., 901 Campisi Way, Ste 245, Campbell CA 95008, andytescher@attbi.com
[c]Lockheed Martin Corp., Santa Maria, CA 93455, frank.e.kozak@lmco.com

ABSTRACT

The performances of discrete-cosine-transform (DCT) JPEG and wavelet-transform (WT) JPEG-2000 for the Karhunen-Loeve-Transform (KLT) based lossy compression of multispectral imagery data are evaluated and compared. The evaluation is based on the measured amount of compression-induced root mean square error in the reconstructed imagery and, more importantly, the impact of compression on the classification of imagery data. We have opted to use classification to assess the impact on compression since it is one of the most widely used forms of machine exploitation procedures. An unsupervised classification via a thematic map is implemented. It is assumed that results for a supervised classification would be similar. The impact of compression is examined at various compression ratios for data obtained from two sensor platforms, LANDSAT TM satellite test imagery with a 30m footprint, and ERIM M7 Sensor aerial test imagery with a 4-6m footprint. Preliminary results, based on the selected test imagery and the selected multispectral bandwidth compression scheme, indicate that the JPEG 2000 generally outperforms the baseline JPEG by a small margin. The results are based on the root-mean-square (RMS) error and the classification accuracy and pertain to imagery with less than 50m footprints. For the 4-6m-footprint ERIM aerial test imagery, JPEG 2000 produces up to four percent higher classification accuracy while incurring up to twelve percent smaller RMS error. However, for the 30m-footprint LANDSAT test imagery, the performance of JPEG and JPEG 2000 are nearly the same. This study does not include imagery with greater than 50m footprint, e.g., NOAA's AVHRR with 1.1 km footprint. For this type of imagery, classification should be performed via a spectral unmixing procedure, instead of a thematic map, since the pixels do not represent pure species.

Keywords: Bandwidth compression, JPEG, JPEG-2000, Multispectral image compression, Spectral decorrelation, Classification, Confusion matrix

1. INTRODUCTION

Images acquired by sensors on board aircrafts or earth observation satellites represent large volumes of data. This data need to be stored on board and/or transmitted to ground stations for processing. The sheer volume of data creates challenging problems for both onboard storage (due to the stringent limitations on power, weight, and size) and transmission to ground stations (due to bandwidth limitation of the downlink channels). It is therefore imperative to reduce the volume of data to a minimum via bandwidth compression techniques. Lossless compression of data is ideal but the amount of compression achievable is inadequate. For this reason there have been numerous efforts by researchers to develop lossy or near lossless compression algorithms[1]. In this paper we study the impact of lossy compression on the machine-based exploitation of data. Specifically, we compare the performances of discrete-cosine-transform (DCT) JPEG versus wavelet-transform (WT) JPEG-2000 for the Karhunen-Loeve-Transform (KLT) based lossy compression of

[1]*Presented at SPIE's Applications of Digital Image Processing Conference, (Proc. Vol. 4790, No. 33), Seattle, July 2002*

multispectral imagery data. A statistical measure and a machine-based measure are used for performance evaluation. The impact of DCT-based JPEG/KLT compression on the classification was reported earlier[2]. For a statistical measure, we use the square root of the sum of the squares (RMS) of the difference between the original and compressed/reconstructed image sets. For a machine-based measure, we have opted to use classification since it is one of the most widely used forms of machine exploitation procedures. An unsupervised classification via a thematic map is implemented. It is assumed that results for a supervised classification would be similar.

For small-footprint (<50m) imagery, a conventional classification technique assigns a single label to each pixel. The label can be any one of the known categories such as water, forest, soil, and rock. The resulting thematic map can become a very useful aid for land cover interpretation provided that the imagery data is composed of pure pixels, meaning that each pixel represents the spectral signature of only one species. However, for large-footprint imagery such as 1.1km-footprint imagery from NOAA's AVHRR, classification via a thematic map is inappropriate since the pixels do not represent pure species. For this type of imagery, a spectral unmixing procedure should be used instead for classification[3,4].

In this paper, the impact of compression is examined at various compression ratios for small-footprint imagery data obtained from two sensor platforms; LANDSAT TM imagery with a 30m footprint (Montana images), and ERIM Airborne M7 Sensor imagery with a 4-6m footprint (Airfield images).

2. Performance Evaluation Methodology

Figure 1 depicts the methodology to evaluate the impact of compression on classification. The selected compression technique is the adaptive KLT/JPEG method previously reported[5]. Images are first spectrally decorrelated via KLT. The resulting spectrally-decorrelated images, i.e., eigen images, are then individually coded via either a standard DCT-based JPEG or a WT-based JPEG-2000. For either scheme, the coding bit rate for each eigen image is adjusted to achieve a uniform compression-induced error across all eigen images. For the baseline JPEG, this is accomplished via using the same quality factor Q for all eigen images. To change the overall compression ratio, a different Q factor within the range of 0-99 is selected. For the JPEG-2000, the effective coding bit rates obtained from the baseline JPEG are used to code the corresponding eigen images. An unsupervised classification via application of a modified ISODATA clustering procedure[6] is performed to obtain thematic maps corresponding to the original, and the compressed/reconstructed imagery. To reduce the dimensionality of the data, the ISODATA is applied to the first four principle-component images obtained from the two original test sets. An identical set of ISODATA parameters, e.g., initials centroids, number of classes, convergence threshold, was used for the original and compressed/reconstructed image sets. The resulting thematic maps are then compared and analyzed via a confusion matrix at various overall compression ratios.

3. Unsupervised Classification

The problem of allocating individual pixels to their most likely class (i.e. labeling the pixels) can be approached in one of two ways; supervised and unsupervised classifications. If we know the number of separable patterns that exist in the geographical area covered by the image, and if we can estimate the statistical characteristics of the values taken on by the features describing each of these

patterns, then a series of templates can be built up. This is referred to as supervised classification. Each template represents an ideal pattern. The individual pixels can be compared with each template in turn and the closest match found. Each pixel is therefore labeled as belonging to the class represented by the most similar template. In the alternative unsupervised classification, no knowledge of the number or character of the patterns present in the image is assumed initially. Instead, a method of allocating and reallocating the individual pixels to one of an initial set of arbitrarily chosen patterns is used. These will be called "basic patterns". At each stage, each pixel in turn is given the label of one of these basic patterns using some decision rule or classifier. At the end of the first round, when all pixels have been allocated, the basic patterns can be altered in character according to the nature of the pixels that have been associated with them. If necessary, some basic patterns can be dropped from the analysis if only a small number of pixels are allocated to them, or pairs or similar basic patterns can be combined by averaging. Also individual basic patterns can be split into two if they are thought to be too heterogeneous. The process of pixel labeling by association with one of the basic patterns is repeated using the updated basic patterns until the procedure converges, when the user will attempt to relate the basic patterns in the final cycle to some Earth-surface cover class. It is important to realize that these two methods differ in that the supervised methods attempt to relate pixel groups with actual Earth-surface cover types; the pixel groups are thus termed information classes. The unsupervised methods simply determine the characteristics of non-overlapping groups of pixels in terms of their spectral band values. These groups are therefore known as spectral classes, and their relationship with information classes must be worked out through fieldwork and/or map and air photograph interpretation. It is reasonable to assume that the impact of lossy compression will be nearly the same for the supervised and unsupervised classification of small-footprint imagery. For this reason, the supervised classification of small-footprint imagery is not considered in this paper. Unsupervised classification is carried out via the general ISOADATA clustering procedure outlined in the table 1.

4. The Confusion Matrix

A confusion matrix is an efficient and convenient method to assess the impact of compression-induced error on the classification. To evaluate the classification accuracy, the analyst selects a sample of pixels and then visits the sites (or vice-versa), then builds a confusion matrix to determine the nature and frequency of errors. The confusion matrix compares the relation between known reference data (ground truth) and the result of the automatic classification, so it tells how well the training samples of each class have been classified. The confusion matrix columns represent ground data (assumed to be correct), and the rows represent map data (classified by the automatic procedure). The diagonal elements represent agreement between ground and map, so ideally the matrix should have all zero off-diagonal elements.

The *errors of omission* (map producer's accuracy) are the incorrect in column divided by total in column, hence measures how well the map maker was able to represent the ground features, i.e. it indicates how well training set pixels of the given cover type are classified.

$$\text{Producer's Accuracy} = \frac{\text{Number of correctly classified pixels per category}}{\text{Number of reference pixels used for that category}}$$

$$= \frac{\text{Diagonal element}}{\text{corresponding column total}}$$

The *errors of commission* (map user's accuracy) are incorrect in row divided by total in row, hence measures how likely the map user is to encounter correct information while using the map, i.e. indicates the probability that a pixel classified into a category actually represents that category.

$$\text{User's Accuracy} = \frac{\text{Number of correctly classified pixels per category}}{\text{Number of pixels classified in that category}}$$

$$= \frac{\text{Diagonal element}}{\text{corresponding row total}}$$

The map *overall accuracy* is the total on diagonal divided by the grand total.

$$\text{Overall Accuracy} = \frac{\text{Total number of correctly classified pixels}}{\text{Total number of reference pixels}}$$

$$= \frac{\text{Sum of diagonal elements}}{\text{Sum of columns (or rows)}}$$

For our experiments, the pixels in the original thematic map are treated as being correctly classified, i.e., the ground truth. Thus, any difference between two thematic maps, i.e., one obtained from the original and the other from the compressed/reconstructed image sets, represents misclassified pixels, or classification error. The confusion matrix can efficiently capture the classification errors.

5. Experimental Results

Simulation experiments were carried out using test imagery from two sensor platforms. The first test imagery was from ERIM M7 airborne sensor and covered an aerial scene of an Airfield in sixteen unequal bands in the spectral range of 0.36-12.11 micrometers. Each spectral band was composed of 512x512 pixels with each pixel having a footprint of 4-6 meters. The second test imagery was from LANDSAT 5 TM space-borne sensor, and covered a rural/urban scene in Montana in seven unequal bands in the spectral range of 0.45-2.35 micrometers. Each spectral band was composed of 512x512 pixels with each pixel having a footprint of 30 meters (with the exception of band 6 with a 120m footprint). These sets of test imagery were selected because they contain a diverse range of natural and urban terrain and, as such, very challenging for classification.

The 14-class thematic maps corresponding to the original Airfield and Montana test imagery are shown in Figures 2 and 3, respectively. These thematic maps were obtained via applying a modified ISODATA clustering procedure[4] to the first four spectral bands of the original and the compressed/reconstructed sets. For the LANDSAT test set, only the first four of the seven bands were used in our experiment. Figure 4 shows a sample thematic map (Airfield) and the corresponding confusion matrix obtained from reconstructed imagery at a compassion ratio of 9-to-1, using JPEG-2000. Compared to the original thematic map of figure 2, the overall classification accuracy is measured to be 88 percent.

Figures 5 and 6 show the classification accuracy and the root-mean-square error versus the compression ratio for JPEG and JPEG-2000. The charts depicted in these figures indicate that the JPEG 2000 generally outperforms the baseline JPEG by a small margin. For the 4-6m-footprint ERIM aerial test imagery, JPEG 2000 produces up to four percent higher classification accuracy while incurring up to twelve percent smaller RMS error. However, for the 30m-footprint LANDSAT test imagery, the performance of JPEG and JPEG 2000 are nearly the same. These results pertain to imagery with less than 50m footprints only. For large-footprint imagery, e.g., NOAA's AVHRR with 1.1 km footprint, classification via a thematic map is inappropriate since the pixels do not represent pure species. For this type of imagery, a spectral unmixing procedure, instead of a thematic map, should be used for classification[3,4].

6. Conclusion

The performances of the baseline JPEG and JPEG-2000 for the KLT- based lossy compression of multispectral imagery data were evaluated and compared. The evaluation was based on the measured amount of compression-induced root mean square error in the reconstructed imagery and, more importantly, the impact of compression on the classification of imagery data. The impact of compression was examined at various compression ratios for data obtained from two sensor platforms; LANDSAT TM satellite test imagery with a 30m footprint, and ERIM M7 Sensor aerial test imagery with a 4-6m footprint. Preliminary results, based on the selected test imagery and the selected multispectral bandwidth compression scheme, indicated that the JPEG 2000 generally outperforms the baseline JPEG by a small margin. The results were based on the root-mean-square error and the classification accuracy and pertain to imagery with less than 50m footprints. For the 4-6m-footprint ERIM aerial test imagery, JPEG 2000 produced up to four percent higher classification accuracy awhile incurring up to twelve percent smaller root-mean-square error. However, for the 30m-footprint LANDSAT test im, the performance of JPEG and JPEG 2000 were nearly the same. This study did not include imagery with greater than 50m footprint, e.g., NOAA's AVHRR with 1.1 km footprint. For this type of imagery, classification should be performed via a spectral unmixing procedure, instead of a thematic map, since the pixels do not represent pure species.

Further experiments with different sensor imagery, including large-footprint imagery that requires spectral unmixing for classification, should be carried out to fully substantiate the preliminary results reported in this paper.

ACKNOWLEDGEMENT

The work reported in this paper was supported by a Lockheed Martin California State University partnership award (Eamon Barrett, technical representative) and, in part, by a grant from US Department of Navy, Office of Naval Research, award number N00014-01-1-1049 (George Solhan, technical representative)

REFERENCES

1. J. A. Saghri, A. G. Tescher, Class-prioritized compression of Multispectral Imagery Data," Journal of Electronic Imaging, 11 (2),246-256 , April 2002

2. J. A. Saghri, A. G. Tescher, and M. Omran, Impact of Lossy Compression on the Classification of Remotely-Sensed Imagery Data ", Proc. SPIE , 4115, July 2000

3. F. Maselli, Multiclass Spectral Decomposition of Remotely Sensed Scenes by Selective Pixel Unmixing, IEEE Trans. On Geoscience & Remote Sensing 36, (5), 1809-1819 (1998)

4. J. A. Saghri, A. G. Tescher, F. Jaradi, and M. G. H. Omran, "A Viable End-Member Selection Scheme for Spectral Unmixing of Multispectral Satellite Imagery Data," Journal of Imaging Science and Technology, Vol. 44, No. 3, 196-203, May/June 2000

5. J. A. Saghri, A. G. Tescher, and J. T. Reagan, Practical Transform Coding of Multispectral Imagery, IEEE Signal Processing Magazine 12, (1), 32-43 (1995).

6. M. K. Dhodi, J. A. Saghri, I. Ahmad, and R. Ul-Mustafa, "D-ISODAT: A Distributed Algorithm for Unsupervised Classification of Remotely-Sensed Data on Network of Workstations," Journal of Parallel and Distributed Computing, vol. 59, October 1999

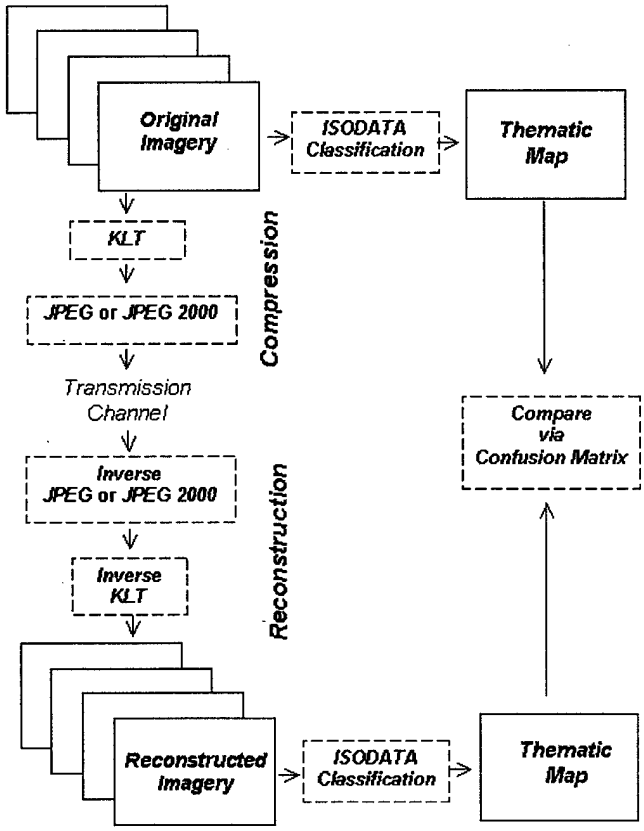


Figure 1. Methodology to evaluate the impact of compression on classification

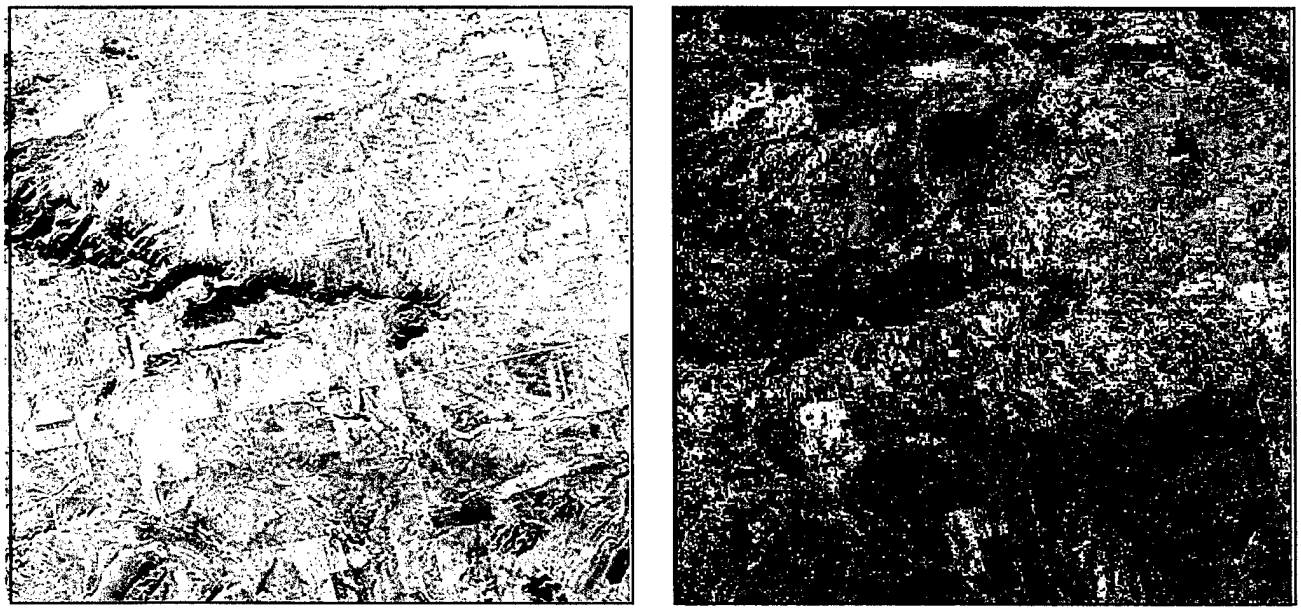
1. Pick k_0 arbitrary centroids *(mean vectors)* corresponding to k_0 classes
2. Classify the pixel vectors by assigning them to the class of the closest mean
3. Standard deviation for each feature axis is computed for each of the k_0 clusters • If a standard deviation > a prespecified threshold, cluster is split along that axis
4. The distance between cluster centers is found • If the distance is < a prespecified threshold, the two clusters are merged into one
5. The process is repeated with the new k_1 number of clusters until no clusters are split or merged

Table 1. ISODATA Clustering procedure



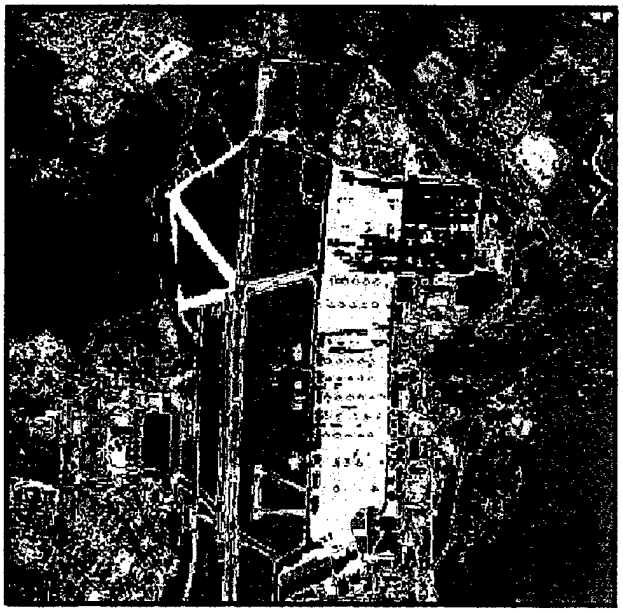
(a) *(b)*

Figure 2. ERIM M7 Airfield test imagery (16 bands of 0.36-12.11 micrometers, 512x512 pixels with 4-6 m footprint). ***(a)*** Band 1, ***(b)*** Thematic map with 14 classes



(a) *(b)*

Figure 3. LANDSAT 5 TM Montana test imagery (7 bands, 0.45-2.35 micrometers, 512x512 pixels with 30m footprint). ***(a)*** Band 1, ***(b)*** Thematic map with 14 classes



(a)

Class No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Row Sum	Producer Accuracy
1	**8808**	0	0	3	0	0	0	539	1	0	226	0	376	510	10463	**84%**
2	0	**10665**	0	1041	54	1172	0	329	0	0	0	1433	0	0	14694	**73%**
3	0	0	**14304**	0	0	0	0	0	0	272	0	0	0	0	14576	**98%**
4	4	1029	0	**13078**	0	1874	0	1123	0	0	0	5	0	0	17113	**76%**
5	0	45	0	0	**8973**	0	0	0	0	0	0	1204	0	0	10222	**88%**
6	0	965	0	1453	0	**15603**	0	0	0	0	0	1571	0	0	19592	**80%**
7	0	0	0	0	0	0	**25491**	0	736	373	0	0	0	0	26600	**96%**
8	629	335	0	1372	0	0	0	**11044**	0	0	331	0	0	0	13711	**81%**
9	1	0	0	0	0	0	645	0	**20463**	0	0	0	0	806	21915	**93%**
10	0	0	263	0	0	0	400	0	0	**29248**	0	0	0	0	29911	**98%**
11	182	4	0	0	0	0	0	383	0	0	**18452**	0	1446	0	20467	**90%**
12	0	1336	0	1	967	1918	0	0	0	0	0	**16137**	0	0	20359	**79%**
13	409	0	0	0	0	0	0	0	0	0	1454	0	**17437**	926	20226	**86%**
14	473	0	0	0	0	0	0	0	898	0	0	0	927	**19997**	22295	**90%**
Col. Sum	10506	14379	14567	16948	9994	20567	26536	13418	22098	29893	20463	20350	20186	22239	*262144*	***87%***
User Accuracy	**84%**	**74%**	**98%**	**77%**	**90%**	**76%**	**96%**	**82%**	**93%**	**98%**	**90%**	**79%**	**86%**	**90%**	***87%***	

Overall Accuracy	88%

(b)

Figure 4. Sample thematic map ***(a)*** and corresponding confusion matrix ***(b)*** obtained from the reconstructed imagery at 9-to-1 compression ratio (using JPEG-2000)

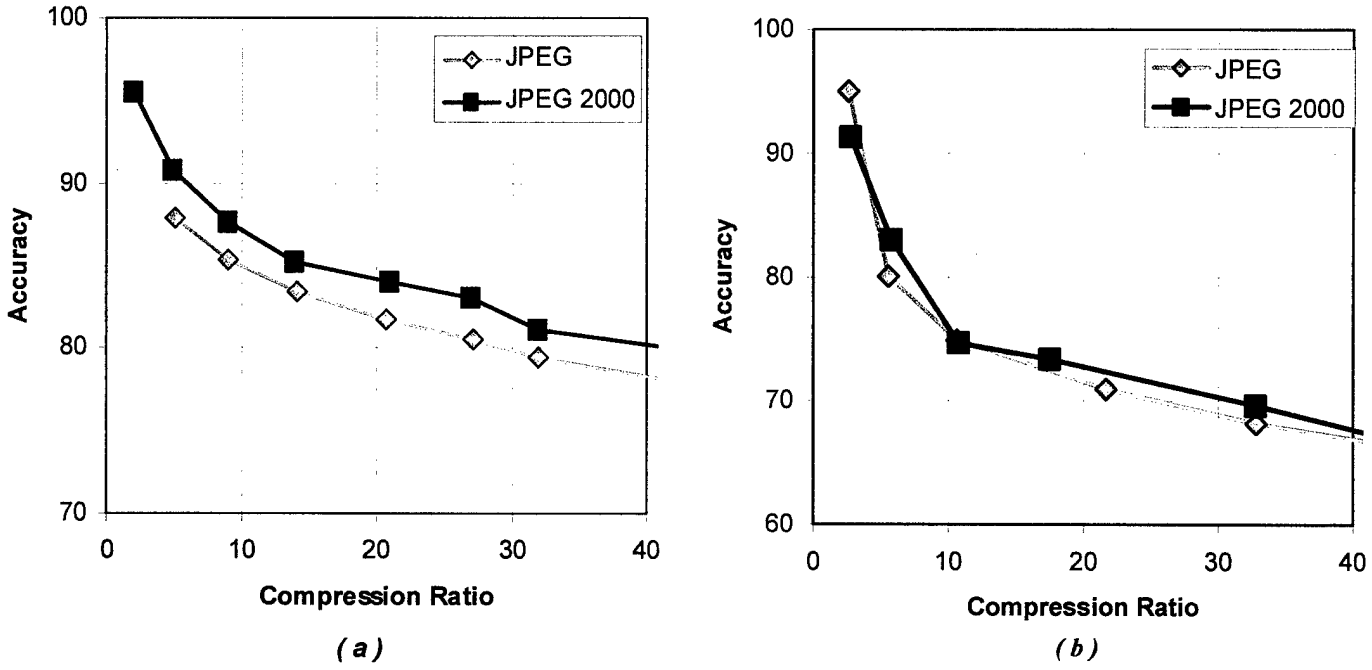


Figure 5. Classification accuracy versus the overall compression ratio (using base JPEG and JPEG-2000) for ***(a)*** Airfield test imagery, and ***(b)*** Montana test imagery

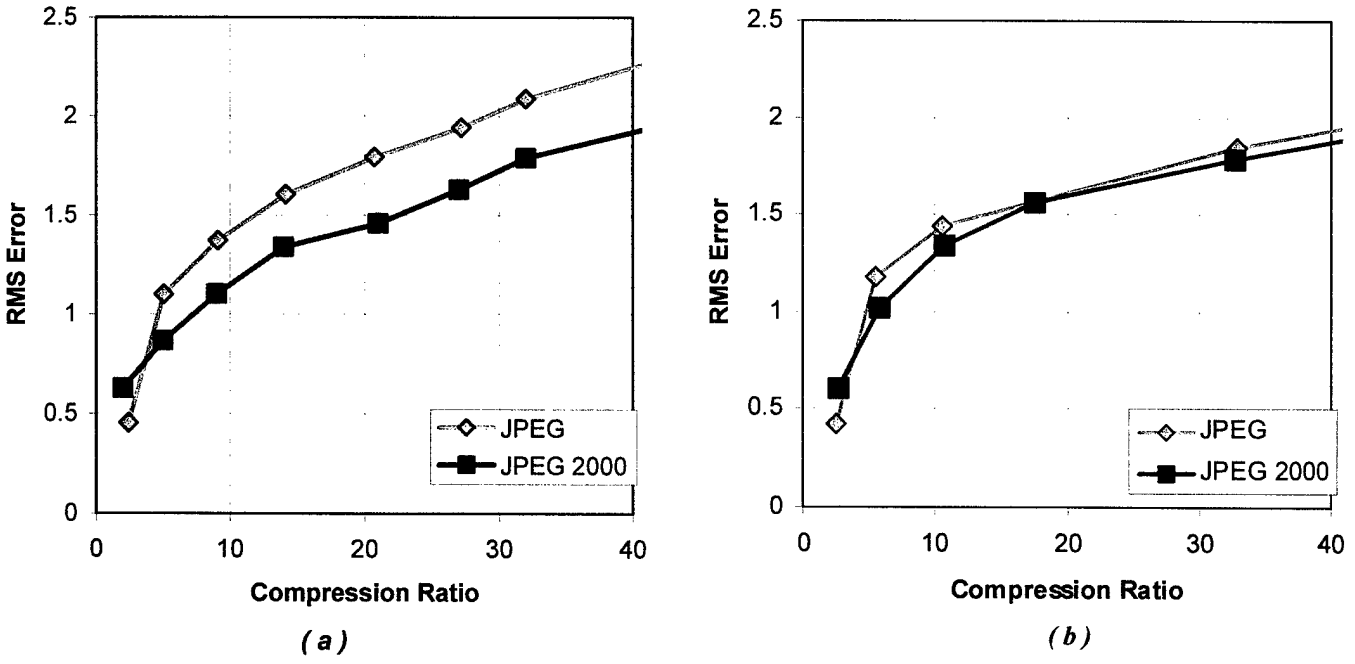


Figure 6. RMS error versus the overall compression ratio (using base JPEG and JPEG-2000) for ***(a)*** Airfield test imagery, and ***(b)*** Montana test imagery

Class-Prioritized Compression of Multispectral Imagery Data[1]

John A. Saghri
Electrical Engineering dept., California Polytechnic State university, San Luis Obispo, CA 93407
jsaghri@calpoly.edu

Andrew G. Tescher
Compression Science Inc., 901 Campisi Way, Ste 245, Campbell CA 95008
andytescher@home.com

ABSTRACT

A joint classification-compression scheme that provides the user with added capability to prioritize classes of interest in the compression process is proposed. The dual compression system includes a primary unit for conventional coding of multispectral image set followed by an auxiliary unit to code the resulting error induced on pixel vectors that represent classes of interest. This technique effectively allows classes of interest in the scene to be coded at a relatively higher precision level than the nonessential classes. Prioritized classes are selected from a thematic map or directly specified by their unique spectral signatures. Using the specified spectral signatures of the prioritized classes as endmembers, a modified linear spectral unmixing procedure is applied to the original data as well as the decoded data. The resulting two sets of concentration maps, which represent prioritized classes before and after compression, are compared and the differences between them are coded via an auxiliary compression unit and transmitted to the receiver along with conventionally coded image set. At the receiver, the recovered differences are blended back into the decoded data for an enhanced restoration of the prioritized classes. The utility of this approach is that it works with any multispectral compression scheme. This method has been applied to test imagery from various platforms including NOAA's AVHRR (1.1 km GSD), and LANDSAT 5 TM (30 m GSD), LANDSAT 5 MSS (79 m GSD).

Keywords: Multispectral data compression, image compression, bandwidth compression, Spectral unmixing, classification, spectral signature, spectral angle, end-member selection.

[1] ***Journal of Electronic Imaging / April 2002 / Vol. 11(2)***

1. Introduction

2.

Significant amount of environmental data, such as meteorological information is derived from satellite and airborne collection platforms. The utility of the information derived from this data is extensive and is applied in numerous essential applications including weather and climate analysis, sea surface temperature determination, land usage analysis, as well as agriculture and forestry applications. As sensor technology evolves into the twenty first century, the volume of data will increase rapidly due to utilization of sensors that, in addition to high spatial and dynamic resolutions, will have high spectral resolution covering up to several hundreds bands. Due to the downlink transmission constraint, the massive volume of radiometric data must be compressed via lossy techniques that yield high compression ratios.

Compression is achieved via exploitation of the inherent spatial and spectral correlation in the data. Lossless compression of the data [1-5] is ideal, but the entropy of the source bounds the amount of achievable compression. This entropy bound limits the obtainable amount of compression to the range of two to three, i.e., 3-5 bits per pixel. This level of compression ratio does not alleviate the downlink transmission constraint for many collection platforms. For this reason there have been a considerable efforts by the researchers to develop lossy or near-lossless compression algorithms. List of recent works in the area of multispectral and hyperspectral bandwidth compression includes transform-based techniques [6-12], predictive techniques [13-17], vector quantization-based schemes [18-23], and hybrid techniques [24-27]. In general all these new techniques exploit the inherent spatial and spectral correlation in the data to achieve compression.

For low to moderate compression ratios, the compression-induced distortion is quantifiable and is shown to be within the range of other distortions inherent in the data and prediction models [20, 29, 30, 31]. However the dependence on the processed data by a diverse "user community" has reasonably created a strong influence to maintain the satellite collected information at the highest possible accuracy. Thus, the lossy algorithms have not yet received overall acceptance by the end users of the final product, partly because the demonstration of the impact of lossy compression techniques in the "final product domain" has not been broadly available. In general, to minimize the overall compression-induced error, the compression technique must fully exploit the spectral and spatial correlation in the data. However, to minimize the impact of compression on the final product, the distribution of the inherent compression-induced error becomes a major factor as well. A possible solution is to apply prioritization in the compression process where features of interest in the scene are coded at a higher precision than the nonessential components. A scheme that may be used to achieve prioritization is the modified Karhunen-Loeve / Discrete Cosine Transform (KLT/DCT) algorithm reported in reference [9]. This spectral screening based algorithm is intended to preserve the spatial and spectral fidelity of small, statistically insignificant, objects in the compression process. It artificially increases the statistical significance of small objects, via spectral screening, to allow them to appear in the first few principal component images that are coded at a relatively higher bit rate. Via simple modifications, this method can achieve similar prioritization for other classes regardless of size.

In general to achieve prioritization the compression should be carried out jointly with classification. One approach is to mathematically transform the image set into another set in which classes of interest appear individually as concentration maps or images. Provided that the transformation is reversible with negligible residual error, the compression can be performed in the transform domain instead, where prioritization on a class-by-class basis is feasible.

Initially, the joint classification and compression approach was implemented via a modified spectral unmixing procedure [28]. Figure 1 shows the block diagram of this intuitive and simple class-prioritized compression system. The scheme involved, 1) applying spectral unmixing to the original image set, 2) coding the resulting prioritized concentration maps at a relatively high bit rate, 3) coding the resulting non-prioritized concentration maps at a relatively low bit rate, and 4) remixing the decoded class concentration maps at the decoder to form the reconstructed class-prioritized multispectral image set. However, experimental result revealed that although this technique allows class prioritization, it does not achieve an impressive overall compression ratio due to the fact that there is a considerably larger number of class concentration images, compared to the original set of images, that need to be individually coded.

In this paper we propose a modified version of initial intuitive design that alleviates the shortcoming of the initial design. The two-stage system includes a primary compression unit for conventional coding of the set of multispectral images followed by an auxiliary compression unit for coding the compression-induced error on the pixel vectors representing the prioritized classes. The added utility gives the user control over the distribution of the inherent compression-induced error over various components of the scene. This technique effectively allows classes of interest in the image to be coded at a relatively higher precision level (bit rate) than the nonessential components. This approach minimizes the impact of lossy compression on the "final product" domain.

The proposed class prioritization capability can adapt to any multispectral compression scheme. The novelty of this technique is that it provides restoration of the prioritized classes in two complementary forms; 1) it produces restored prioritized class concentration maps and, 2) it enhances the reconstructed quality of the pixel vectors that represent the prioritized classes. Since the number of end-members selected is independent of the number of channels [32], the algorithm works well with any platform regardless of the number of spectral channels. Also, although spectral unmixing is generally intended for large footprint imagery, the size GSD does not impact the efficiency of the proposed algorithm. A smaller GSD, i.e, a higher spatial resolution such as SPOT, merely results in generation of concentration maps with saturated (pure) pixels resembling a thematic map. This technique has been successfully applied to test imagery from various platforms including NOAA's AVHRR (1.1 km GSD), and LANDSAT 5 TM (30 m GSD), LANDSAT 5 MSS (79 m GSD). This technique, in its present form, is not suitable for SAR imagery due to the presence of inherent speckle noise.

2. Compression and Decompression Block Diagrams

2.1 Compression Block Diagram

Figure 2 shows the block diagram of the proposed compression system. The two-stage compression system is based on an initial conventional coding of the multispectral image set followed by the standard JPEG coding of the resulting compression-induced error incurred on the classes of the interest. Selection of JPEG for the secondary compression unit was to reduce the complexity of the overall algorithm (by leveraging on the standard technology). The adopted algorithm for the primary compression unit is the previously developed three-dimensional Karhunen-Loeve discrete cosine transform technique [6]. Classes of interest are determined via a modified linear pixel unmixing procedure that entails a very small residual error, i.e., less than 0.5 percent [34]. The end-members for spectral unmixing can be obtained manually or automatically from the scene and/or the user can specify them externally [33-39]. We employed an end-member selection procedure

based on a modified version of the ISODATA unsupervised classification procedure [34]. A method of successive projections proposed by Maselli [32] was used to determine the optimum subset, and order, of end-members to unmix each pixel vector.

In our experiments, the pool of end-members was obtained from the original image set. However, the unique subset of end-members, and their respective order, for each pixel vector were obtained from the reconstructed image set. The reason is that unlike the original set of end-members, information regarding the unique subset of end-members selected for each pixel vector is not transmitted to the decoder since that would increase the overhead bits substantially. This information must be regenerated again from the image set at the decoder. Since only the reconstructed image set is available at the decoder, we must therefore use the same image set at the coder level as well to obtain the end-members for each pixel vector. Hence, using an identical subset and order of end-members, spectral unmixing is applied to each original and reconstructed pixel vector at the coder. The compression-induced error incurring on the classes of interest, i.e., prioritized classes, is obtained via obtaining the differences between the resulting original (before compression) and reconstructed (after compression) prioritized class concentration maps. Note that since each concentration map is produced as an array of floating point numbers, the difference map between each pair of them is also an array of floating point. numbers. The floating-point difference maps are then nonlinearly mapped to an eight bits per pixel image format and coded individually via the auxiliary compression unit (standard JPEG) and transmitted to the decoder.

2.2 Decompression Block Diagram

Decompression block diagram is shown in Figure 3. The compressed multispectral image set and the difference images are first decoded. The difference images are then re-mapped to their original floating-point scale via the recovered scaling parameters. Prioritized class concentration maps are extracted from the reconstructed multispectral image set via an identical linear pixel unmixing procedure used at the coder level. Restoration of the extracted concentration maps is achieved by adding each recovered floating-point difference map to the corresponding class concentration map. In the next module, partial restoration of the prioritized pixel vectors is performed. A prioritized pixel vector is defined as a pixel vector that, after unmixing, will have a non-zero fractional component corresponding to a prioritized class concentration map. The restored class concentration maps can thus be used to identify the prioritized pixel vectors. The identified prioritized pixel vectors may also have non-zero fractional components for non-prioritized classes as well. However, only the fractional components associated with the prioritized classes are restored. As such, the restoration of the prioritized pixel vectors is labeled as partial. Compensation of each prioritized pixel vector is achieved by adding to it an error compensation vector obtained from the difference maps and end-members in accordance with the procedure discussed in section 5. The prioritized pixel vectors in the reconstructed image set are then replaced with their restored version.

3. Conventional Multispectral Compression Module

The primary compression unit can be any conventional lossy multispectral compression scheme [6-27]. Figure 4 depicts the block diagram of the adopted primary compression scheme [6]. It consists of four modules; 1) Data partitioning, 2) Karhunen-Loeve Transformation, 3) Mapping eigen planes to 8-bit eigen images, and 4) JPEG compression. In the data partitioning module the set of multispectral images are partitioned into sets of non-overlapping images, i.e., sub-block sets, which are sequentially fed to the KL transformation module for spectral decorrelation. In the KL transformation module the multispectral sub-block set is spectrally decorrelated to produce a set of eigen planes. The basis functions for the KL transformation are the eigenvectors of the cross-

covariance matrix associated with the multispectral sub-block set. The covariance matrix is estimated first and then quantized to only two bytes to reduce the required overhead information transmitted to the decoder. The eigen planes are formed by matrix multiplication of the sub-block set and the basis functions [6]. The eigen planes are in floating point format and assume both positive and negative values. In the next module the eigen plane set is converted into the 8-bit eigen images set via linear/nonlinear mapping of each plane into the 0-255 range. For high dynamic range imagery consisting of 10-12 bits, the eigen plane set may be mapped to 0-1023 range instead. The spectrally decorrelated eigen images are then compressed in the next module using the JPEG algorithm. The quantized covariance matrix and mapping information are transmitted along with the compressed bit stream as overhead information. This three-dimensional transform-based compression algorithm efficiently exploits the spectral and spatial correlation in the data. A significant practical advantage of this algorithm is that it is leveraged on the highly developed JPEG compression technology. Because of the significant compaction, i.e., spectral decorrelation, of the data resulting from the initial KLT process, an 8-bit JPEG can be used for coding the eigen images associated with 8, 10, or 12 bits multispectral data. However, for a dynamic range of 12 bits or higher, a 12-bit JPEG and a non-linear mapping function is recommended. The algorithm adopts well to the local terrain variation since the covariance matrix, from which the transformation basis function are derived, is updated very frequently over each small sub-block set of multispectral data. The bit requirement for the sub-block covariance matrix and mapping information is negligible and, as such, there is no need to resort to a stored covariance look-up-table to minimize the overhead bit information.

4. Spectral Unmixing Module

In this module the set of multispectral images is transformed into a set of compositional maps known as abundance images or concentration maps. Each map shows the fractional concentration of one class throughout the scene. In linear spectral unmixing, the spectral signature of each original pixel in the scene is represented as a linear combination of a limited set of fundamental spectral components referred to as end-members. However, since the number of end-members obtained in the first module is likely more than the intrinsic true spectral dimension of the data N_c, the linear unmixing process faces the so-called "condition of identifiability". To alleviate this constraint we employ the recent method of successive projections proposed by F. Maselli [32].

4.1 Linear Pixel Unmixing

The conventional spectral unmixing is modeled as,

$$\mathbf{x} = \mathbf{M}\,\mathbf{f} + \mathbf{e} \tag{1}$$

where:

- $\mathbf{x}$ a pixel vector of N_b components
- $\mathbf{M}$ an N_b x N_c matrix formed by end-members $\mathbf{m}_{1,\ldots N_c}$
- $\mathbf{f}$ vector of N_c fractional components
- $\mathbf{m}_i$ i_{th} end-member of N_b components
- $\mathbf{e}$ residual error vector of N_b components
- N_b number of bands
- N_c true spectral dimension of the set of N_b multispectral images

The solution is,

$$\mathbf{f} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{x} \tag{2}$$

4.2 Selection of the optimum subset of end-members

The pool of all end-members for spectral unmixing can be obtained manually or automatically from the scene and/or the user can specify them externally [33-39]. Each end-member results in the generation of a corresponding class concentration map. Thus to prioritize a class, the user must ensure that spectral signature of that class is included in the set of available end-members. We employed an end-member selection procedure based on a modified version of the ISODATA unsupervised classification procedure [34]. For each pixel vector, a unique subset of end-members is selected from the available pool of end-members [32]. The unique subset is selected so as to minimize the residual error after decomposition of the corresponding pixel vector.

To determine the N_c optimum end-members for pixel vector $\mathbf{x}$, the pixel vector is projected onto all available normalized end-members. The most efficient projection, which corresponds to the highest value c_{max}, indicates the first selected end-member $\mathbf{m}_{max}$. It can be shown that this procedure is equivalent to finding the end-member with the smallest spectral angle with respect to $\mathbf{x}$ [34]. The residual pixel signature, $\mathbf{r}_x = \mathbf{x} - c_{max} . \mathbf{m}_{max}$ is then used to identify the second end-member by repeating the projection onto all remaining end-members. The process continues up to the identification of a prefixed maximum number N_c of end-members from the total of N_e available end-members.

5. Restoration of Prioritized Pixel Vectors

A prioritized pixel vector is defined as a pixel vector that, after unmixing, will have a non-zero fractional component corresponding to a prioritized class concentration map. The objective of the class-prioritized compression is thus an attempt to restore the integrity of prioritized pixels following compression. The restoration of a prioritized pixel vector can, and ideally should, be partial since it is only required to restore the fractional component that is used to form the prioritized class concentration maps. This objective is achieved via two complementary processes. In the first process, the auxiliary compression unit attempts to fully restore the prioritized class concentration maps via compensating them for the errors induced by the primary compression unit. In the second process, the compensations made to the prioritized class concentration maps are used again to partially restore the pixel vectors representing the prioritized classes. This is accomplished by remixing the restored fractional components to form partially restored prioritized pixels. Note that the restored prioritized class concentration maps can not be retrieved from the partially restored pixel vectors via a subsequent unmixing procedure. This is because the correction received by a fractional component distributes to all fractional components upon a subsequent unmixing. It is thus necessary to retain the restored class concentration maps separately. It is important however to note that only prioritized pixel vectors, i.e., those that have non-zero fractional components corresponding to the prioritized classes, are selected to receive compensations from the auxiliary unit. Hence prioritization is achieved in the second stage as well. It can be shown that, on the average, the fraction of pixel vectors that are selected to receive compensation is equal to N_c/N_e.

Let $\mathbf{x}$ and $\mathbf{x}_r$ represent the original and the reconstructed pixel vectors, respectively. Using an identical subset of N_c end-members, spectral unmixing is applied to both $\mathbf{x}$ and $\mathbf{x}_r$. Each pixel vector can thus be represented as a linear combination of N_c end-members $\mathbf{m}_1\ \mathbf{m}_2\ \mathbf{m}_3\ \ldots\ \mathbf{m}_{N_c}$, with respective fractional components $(f_1, f_2, f_3, \ldots f_{N_c})$ and $(f_1', f_2', f_3', \ldots f_{N_c}')$. We may then write,

$$\mathbf{x} = \mathbf{M}\,\mathbf{f} + \mathbf{e}$$
$$\mathbf{x}_r = \mathbf{M}\,\mathbf{f}' + \mathbf{e}' \qquad (3)$$

where: $\mathbf{M} = [\mathbf{m}_1\ \mathbf{m}_2\ \mathbf{m}_3 \ldots \mathbf{m}_{N_c}]$

$\mathbf{f} = [f_1, f_2, f_3, \ldots . f_{N_c}]^T$

$\mathbf{f}' = [f_1', f_2', f_3', \ldots . f_{N_c}']^T$

$\mathbf{e}$ and $\mathbf{e}'$ are the residual error vectors for $\mathbf{x}$ and $\mathbf{x}_r$, respectively

From equation 3 we obtain,

$$\mathbf{x} = \mathbf{M}\,[\mathbf{f}' + \mathbf{d}] + \mathbf{e} \qquad (4)$$

where: $\mathbf{d} = [\mathbf{f} - \mathbf{f}']$

Thus, pixel vector $\mathbf{x}$ can be represented in terms of $\mathbf{x}_r$ as,

$$\mathbf{x} = \mathbf{x}_r + \sum_{i=1}^{N_c} d_i\,\mathbf{m}_i + (\mathbf{e} - \mathbf{e}') \qquad (5)$$

where: $d_i = (f_i - f_i'),\ \text{for } i = 1, \ldots, N_c$

Hence, assuming the term $(\mathbf{e} - \mathbf{e}')$ is negligible, the term $\sum d_i\,\mathbf{m}_i$ must then be added to $\mathbf{x}_r$ to obtain $\mathbf{x}$. The term $\sum d_i\,\mathbf{m}_i$ represents the error induced on $\mathbf{x}$ by the primary compression unit. The d_i represents the difference between fractional components of $\mathbf{x}$ and $\mathbf{x}_r$ for class *i*. If *i* is one of the prioritized classes, then d_i can be retrieved from the reconstructed difference map associated with class *i*. Since the number of such retrievable d_i's is less than N_c, only a partial restoration of $\mathbf{x}_r$ is feasible. Denoting $\bar{\mathbf{x}}_r$ as an approximation of $\mathbf{x}_r$, we have,

$$\bar{\mathbf{x}}_r = \mathbf{x}_r + \sum_{i=1}^{N_p'} d_i\,\mathbf{m}_i \qquad (6)$$

where:

$\bar{N}_p'$ a subset of the total number of prioritized classes that receive nonzero fractional components from $\mathbf{x}$.

In our adopted unmixing implementation [32], each pixel is decomposed into its own unique subset of $N_c < N_e$ end-members. Thus the probability of a given prioritized class concentration map receiving a nonzero fractional component from a pixel vector $\mathbf{x}$ is N_c / N_e. The average value N_p' for all pixel vectors is,

$$\tilde{N}'_p \cong N_p (N_c / N_e) \tag{7}$$

where:

$\tilde{N}'_p$ average number of prioritized classes that receive nonzero fractional components from a pixel vector

N_p total number of prioritized classes

N_e total number of end-members

Assuming lossless coding of d_i in the auxiliary compression unit, the percent reduction of the compression-induced error for a pixel vector **x** is,

$$R_e = 100 \frac{\left| \sum_{k=1}^{N'_p} d_k \mathbf{m}_k \right|}{\left| \sum_{i=1}^{N_e} d_i \mathbf{m}_i + (\mathbf{e} - \mathbf{e}') \right|} \tag{8}$$

where:

R_e percent reduction of the error on a pixel vector by primary compression

$|.|$ norm of the vector

Assuming $(\mathbf{e} - \mathbf{e}')$ is negligible, the average percent error reduction $\check{R}_e$ can be obtained by further assuming that $|d_k \mathbf{m}_k| = |d_i \mathbf{m}_i|$ for all combinations of k and i. In that case equation 8 reduces to:

$$\check{R}_e \cong 100 (N'_p / N_e) \cong 100 (N_p N_c / N_e^2) \tag{9}$$

where:

$\check{R}_e$ percent average error reduction on a pixel vector **x**

Thus $\check{R}_e$ is directly proportional to N_p. However, depending on the selected coding bit rate, lossy compression of the prioritized difference images will reduce $\check{R}_e$.

6. Auxiliary Compression Unit

The floating-point differences between each pair of pre- and post-compression prioritized class concentration maps are mapped into eight bits per pixel (bpp) image format. For high dynamic range imagery consisting of 10-12 bpp, a 10 bpp image format may be used for the difference image. These difference images are then coded individually via the auxiliary JPEG standard compression unit and transmitted along with the conventionally coded image set. Selection of JPEG for the auxiliary compression unit is to reduce the complexity of the overall algorithm (by

leveraging on the standard technology). Note that, because of the high frequency content of the difference images, the compression-induced error will be relatively higher.

To reduce the quantization error in the formation of the difference images, the following nonlinear mapping was adopted.

$$M = \frac{255\ d^{exp}}{d_{max}\ |d_{max}|^{(exp-1)} - d_{min}\ |d_{min}|^{(exp-1)}} + 0.5 \qquad (10)$$

where:

- M mapped integer difference in the range of $0 \le M \le 255$
- d floating-point difference
- d_{min} minimum floating-point difference
- d_{max} maximum floating-point difference
- *exp* exponent parameter in the range of $0.0 \le exp \le 1.0$

Note that for *exp* = 1 the equation 10 reduces to a simple linear mapping. For our test image set, the optimum value of *exp* was empirically determined to be 0.6. The RMS quantization error for the optimum nonlinear mapping was about 30 percent lower than that for the linear mapping. The resulting eight bpp images may be coded individually using any available scheme. We used the standard JPEG coding algorithm for this purpose. To save on the coding bandwidth, a priori known null pixels may be removed from each difference image to form a smaller-size image that require fewer bits to code. Recall that every pixel vector is unmixed into a unique subset of N_c end-members where $N_c < N_e$. Therefore on the average a fraction $(N_e - N_c) / N_e$ of the total number of points in each difference image will be null, i.e., zero. The exact locations of the null points in a difference image are known for both the coder and the decoder.

Considering the reduced size of the packed difference images, the overall bit rate for the two-stage compression system can be shown to equal,

$$R_t = R_p + \frac{(N_p\ N_c)}{(N_b\ N_e)}\ R_a \qquad (11)$$

where:

- R_t overall bit rate for the compression system
- R_p bit rate for the primary compression unit
- R_a bit rate for the auxiliary compression unit

Note that the bit rate for the auxiliary unit is multiplied by the scale factor $(N_p\ N_c)\ /(N_b\ N_e)$. This implies that the additional bit rate requirement for prioritization is inversely proportional to both, the number of bands N_b and the number of available end-members N_e. For example, the additional

bit rate requirement to prioritize a given class using Landsat TM imagery with $N_b = 7$ will be 3/7 of the corresponding rate required by Landsat MSS with $N_b = 4$.

7. Simulation Results

The proposed compression algorithm was implemented in C language and successfully applied to test imagery from various platforms including NOAA's AVHRR (1.1 km GSD), and LANDSAT 5 TM (30 m GSD), LANDSAT 5 MSS (79 m GSD). In this paper however, only the results for the latter is presented. Since the number of end-members selected is independent of the number of channels [32], the algorithm works well with any platform regardless of the number of spectral channels. Also, although spectral unmixing is generally intended for large footprint imagery, the size GSD does not impact the efficiency of the proposed algorithm. A smaller GSD, i.e., a higher spatial resolution such as SPOT, merely results in generation of concentration maps with saturated (pure) pixels resembling a thematic map. However, this technique, in its present form, is not suitable for SAR imagery due to the presence of inherent speckle noise.

Figure 5 shows band 1 of the four-channel multispectral test image set corresponding to Lake Tahoe region in the United States. Each channel comprises of a 1300 by 1300, 8 bits per pixel (remapped from original 6 bits), image. The test data is one of the North American Landscape Characterization (NALC) Landsat Multispectral scanner data sets obtained from U.S. Geological Survey (USGS). The result of a preliminary principal component study on this data set indicates that its intrinsic true spectral dimension N_c is three. A total of six end-members were obtained from the data set [34]. Figure 6 shows the six concentration maps, i.e., abundance images, resulting from the application of the discussed spectral unmixing procedure. Water concentration in the scene appears as the first abundance image in this figure.
To assess the impact of the total number of end-members on the resulting unmixing residual error, N_e was increased by one from an initial value of four. As expected, the residual error decreased with increasing N_e , although it levels off at $N_e = 6$. The average absolute residual error was 0.65, 0.43, and 0.41 for four, five, and six end-members, respectively.

To test the effectiveness of the prioritization, *water* was selected as the prioritized class. The operating bit rate for the primary compression unit R_p was adjusted to 1 and 2 bpp for two separate case studies. For each case, the resulting floating-point difference map for water was packed to approximately N_c/N_e =3/6 of its original size via removing the null points embedded in them by the unmixing process. The packed difference maps were then mapped to a 0-255 range via the non-linear function given in equation 10. Figure 7 shows the mapped difference image for water for R_p=1. Ideally, the difference image has a homogeneous structure. The sharp peak in the center of the depicted histogram represents the null points in the difference map. For each case study, the packed difference image was coded via standard JPEG algorithm at various bit rates. Figure 8 shows the compression-induced error on the prioritized "water" concentration map for the two case studies. The error shown is due to both the primary and auxiliary compression units. Note that the depicted error at R_a= 0, results from disengaging the auxiliary compression unit, i.e., no prioritization. Once the prioritization mechanism is engaged the error decreases exponentially with R_a . The chart indicates that, for the purpose of lowering the compression-induced error on the prioritized classes, it is not necessary to operate the primary unit at a relatively high rate. Instead, it is only necessary to operate the auxiliary unit at a sufficiently high rate. This observation can also

be made from the redrawn rate-distortion charts shown in figure 9. The value of the total bit rate R_t is obtained from equation 11. It can be seen that operating the primary unit at R_p=1 bpp instead of R_p=2 bpp will reduce the required total bit rate by almost one half. Thus, with respect to lowering the compression-induced error on the prioritized classes, operating the primary unit at a lower rate is more advantageous.

As discussed before, restoration of the prioritized class concentration maps constitutes only the first part of the prioritization process. The second part involves reusing the same compensations used for concentration maps to partially restore the pixel vectors that represent the prioritized classes, i.e., the prioritized pixel vectors. Figure 10 shows the resulting percent reduction in compression-induced error on pixel vectors that represent the water. According to the estimate given in equation 9, this reduction in error would have doubled had an additional class been prioritized. Note that the error compensation received by a pixel vector distributes to all the classes it represents, not just the prioritized classes. However, only the prioritized pixel vectors, receive this partial error compensation. In that sense prioritization is being achieved in this stage as well.

The optimum combination of the coding bit rates for the primary and auxiliary compression units were obtained empirically. First, the bit rate for the primary unit was adjusted so as to arrive at the minimum acceptable quality level for the non-prioritized. The auxiliary compression unit was then engaged and its operating rate was progressively increased to enhance the quality of the prioritized classes, both at the concentration map and pixel vector levels, to the desired level.

8. Conclusion

A two-stage compression system that provides the user with added capability to prioritize classes of interest in the compression process was proposed. The added utility gives the user control over the distribution of the inherent compression-induced error over various components of the scene. This technique effectively allows classes of interest in the image to be coded at a relatively higher precision level than the nonessential classes. The novelty of this technique is that it provides restoration of the prioritized classes in two complementary forms; 1) it produces restored prioritized class concentration maps and, 2) it enhances the reconstructed quality of the pixel vectors that represent the prioritized classes. The utility of this approach is that it can adapt to any multispectral compression scheme. Although spectral unmixing is suitable for imagery with a relatively large ground sampling distance, i.e., having non-pure pixels, this approach worked well with imagery from all sensor platforms regardless the pixel footprint

The system included a primary unit for conventional coding of multispectral image set followed by an auxiliary unit to code the resulting error induced on pixel vectors that represent classes of interest. There is a tradeoff between the primary and auxiliary units coding bit rates in terms of the relative reconstructed quality of the prioritized and non-prioritized classes. In practice, the optimum combination of the two rates may be obtained via first adjusting the primary coding bit rate to arrive at the minimum acceptable quality for the non-prioritized classes, and then setting the auxiliary coding rate to boost the quality of the prioritized classes to the desired level.

Further study is needed to reduce the computational complexity of the proposed system through incorporation of shortcuts, alternatives, and compromises.

References

[1] M. J. Ryan and J. F. Arnold, "The Lossless Compression of AVIRIS Images by Vector Quantization," IEEE Trans. Geoscience and Remote Sensing, vol. 35, May 1997.

[2] R. Roger and M. Cavenor, " Lossless Compression of AVIRIS images,"IEEE Trans. Image Processing, vol. 5, May 1996.

[3] P. G. Howard and J.S. Vitter, "New Methods for Lossless Image Compression using Arithmetic Coding," Proc. Data Compression Conference April 1991.

[4] S. E. Qian, A. B. Hollinger, and Y Haniaux, "Study of Real-Time Lossless Data Compression for Hyperspectral Imagery," Proc. Of IEEE IGARSS, Hamburg, Germany, July 1999.

[5] R. Logeswaran and C Eswaran, "Neural Network Based Lossless Coding Schemes for Telemetry Data," Proc. Of IEEE IGARSS, Hamburg, Germany, July 1999.

[6] J.A. Saghri, A.G. Tescher, and J.T.Reagan, "Practical Transform Coding of Multispectral Imagery," IEEE Signal Processing Magazine, January 1995.

[7] J.A. Saghri and A.G. Tescher, "Near-Lossless Bandwidth Compression for Radiometric Data," Optical Engineering, vol. 30, no.7, Jul. 1991.

[8] J.A. Saghri, A.G. Tescher, and J.T.Reagan, "Terrain-Adaptive Multispectral Bandwidth Compression," Proc. of the Industry Workshop of Data Compression Conference, Snowbird Utah, April 1994.

[9] J. A. Saghri, A. G. Tescher, and A. Boujarwah, "Spectral-signature preserving Compression of Multispectral Data", Optical Engineering, Vol. 38, No. 12, pages 2081-2088, December 1999

[10] B. V. Bernard, D. H. Hadcock, and J. P. Reitz, " Spectrally and Spatially Adaptive Hyperspectral Data Compression," Proc. Of SPIE, vol. 2821, pp 55-63, June 1996.

[11] B.R. Epstein, R. Hingorani, J.M. Shapiro, and M. Czigler, "Multispectral KLT-Wavelet Data Compression for Landsat Thematic Mapper Images," Proc. of Data Compression Conf., Mar. 1992.

[12] G. P. Abousleman, " Adaptive Wavelet Coding of Hyperspectral Imagery, " Proc. Of SPIE, vol 2762, pp 545-556, 1996.

[13] A. Rao, S. Bhargava,"Multispectral Data Compression Using Interband Prediction," Proc. of Image Compression Applications and Inovations Workshop, March 1994.

[14] G. P. Abousleman, M.W. Marcellin, and B. R. Hunt ," Hyperspectral Image Compression Using Entropy Constrained Predictive Trellis Coded Quantization," IEEE Trans. Image Processing, vol 6. April 1997.

[15] B. Brower, B. Gandhi, D. Couwenhouven and C. Smith, "ADPCM for Advanced LANDSAT Downlink Applications, "Proc. of Twenty Seventh Asilomar Conf. on Signals, Systems & Computer, Nov. 1993.

[16] B. Aiazzi, L.Alparonie, and S. Baronti, "Advantages of Bidirectional Spectral Prediction for Reversible Compression of Multispectral Data," Proc. Of IEEE IGARSS, Hamburg, Germany, July 1999.

[17] J. Hu, Y. Wang and P. Cahill," Segmentation based Adaptive DPCM for lossy Compression of Multispectral MR Images," J. Vis. Comm. & Image Rep., vol. 8, pp. 69-82, March 1997.

[18] S. Gupta and A. Gersho, "Feature predictive Vector Quantization of Multispectral Images," IEEE Trans. Geoscience and Remote Sensing, vol.30, pp. 491-501, 1992.

[19] S. Gupta and A. Gersho, "Variable Rate Multistage Vector Quantization of Multispectral Imagery with Greedy Bit Allocation," Proc. Of SPIE, vol. 2094, pp. 890-901, 1993.

[20] B. Hu, S. E. Qian, and A. B. Hollinger, "Impact of Vector Quantization Compression on the Surface Reflectance Retrieval," Proc. Of IEEE IGARSS, Hamburg, Germany, July 1999.

[21] A. T. S. Ho, T. Yu, S. C. Tan, and L. T. Yap, "Improving Vector Quantization of Satellite Images through the Application of Bi-orthogonal Wavelets," Proc. Of IEEE IGARSS, Hamburg, Germany, July 1999.

[22] S. Jaggi," An Investigative Study of Multispectral Lossy Data Compression using Vector Quantization," Proc. Of SPIE, vol. 1702, pp. 239-249, 1992.

[23] R.J. Ryan and J.F. Arnold," Lossy Compression of Hyperspectral Data Using Vector Quantization," Remote Sens. Environ. Vol. 61, pp. 419-436, Elsevier Science Inc. 1997.

[24] G. P. Abousleman, M. W. Marcellin, and B. R. Hunt, 'Compression of Hyperspectral imagery using the 3-D DCT and Hybrid DPCM/DCT," IEEE Trans. Geoscience and Remote Sensing, vol. 33, pp. 26-34, 1995.

[25] S. Briles, "A real-time, Onboard Hyperspectral-image Compression System for a parallel Push-broom Sensor," Proc. Of SPIE, vol. 3071, pp. 182-190, 1997.

[26] G. P. Abousalem," Coding of Hyperspectral Imagery Using Adaptive Classification and Trellis-Coded Quantization," Proc. SPIE, vol. 3071, 1997.

[27] T. Yu, A. T. S. Ho, S. C. Tam, and L. T. Yap, "A Novel Hybrid Bi-Orthogonal Wavelet/ADPPCM Algorithm for Very Low Bit Rate Satellite Image Compressionl," Proc. Of IEEE IGARSS, Hamburg, Germany, July 1999.

[28] J.A. Saghri and A.G. Tescher, "Feature-Based Multispectral Bandwidth Compression,", Proceeding of IEEE International Geoscience and Remote Sensing Symposium, Hamburg, Germany, volume6/99, 28 June- 2 July 1999

[29] A.G. Tescher, J.T. Reagan, J. A. Saghri, K.D. Hutchinson, D. Paul, and P. C. Topping ",Enviromental Data Compression Issues", Proc. SPIE, vol. 3164, July 1997.

[30] J.A. Saghri, A.G. Tescher, and J.T. Reagan "Space-Based Data Compression Issues," Journal of Electronic Imaging, vol.8(3), July 1999.

[31] S. Shen, J. Lindgren, and P. Payton, "Effects of Multispectral Image Compression on Machine Exploitation," in Proceedings of the Twenty Seventh Asilomar Conference on Signals, Systems, and Computers, pp 1352-1356, 1993

[32] F. Maselli, "Multiclass Spectral Decomposition of Remotely Sensed Scenes by Selective Pixel Unmixing," IEEE Trans. On Geoscience & Remote Sensing, Vol. 36, NO. 5, pages 1809-1819, September 1998

[33] M.M. Baumback, J. A. Antoniades, J. H. Bowles, P. J. Palmadesso, L. J. Rickard, "Use of Filter Vectors in Hyperspectral Data Analysis," Infrared Spaceborne Remote Sensing III Conference, Proceedings of SPIE, volume 2553 , No. 148, pages 8194-1912, September 1995

[34] Saghri, J.A., Tescher, A.G., Jaradi, F. and Omran, M., "A Viable End-Member Selection Scheme for Spectral Unmixing of Multispectral Satellite Imagery Data", The Journal of Imaging Science & Technology, Vol. 44, No. 3, May/June 2000

[35] M. O. Smith, S. L. Ustin, J. B. Adams, and A. R. Gillespie, "Vegetation in Deserts: A Regional measure of Abundance from Multispectral Images,"Remote Sensing of Environment, vol. 31, pp 1-26, 1990

[36] N. A. Quarmby, R.G. Townshend, J. J. Settle, M. Milnes, T. L. Hindle, and N. Silleos,"Linear Mixture Modeling appled to AVHRR data for Crop Area Classification", Int. J. Remote Sensing, vol. 13, No. 3, pp 415-424, 1992

[37] J. J. Settle and N. A. Drake,"Linear Mixing and Estimation of Ground Cover Proportions", Int. J. Remote Sensing, vol. 14, No. 6, pp 1159-1177, 1993

[38] Y. E. Shimabukuru and J. A. Smith,"The Least Square Mixing Models to Generate Fraction Images Derived from Remote Sensing Multispectral Data," IEEE Trans. Geosci. Remote Sensing, vol. 29, pp16-20 Jan. 1991

[39] A. Bateson and B. Curtiss,"A Method for Manual Endmember Selection and Spectral Unmixing", Remote Sensing of Enviornment, vol. 55, pp 229-243

John Saghri received his BS degree in electronic engineering from California Polytechnic University, San Luis Obispo, in 1973, his MS degree in electrical and computer engineering from Oregon State University in 1975, and his Ph.D. in electrical and systems engineering from Rensselaer Polytechnic Institute, Troy, New York, in 1979. While at Rensselaer he was on the faculty of the electrical engineering department as an instructor ('78-'79). From 1979 to 1988 he was with the Aerospace Corporation, El Segundo, California, where he was an engineering specialist in the Signal Processing Department. He was a recipient of the Outstanding Accomplishment Award from the Aerospace Corporation. In 1988 he joined Lockheed Palo Alto Research Laboratories, now called Lockheed Martin Advanced Technology Center, in California as a senior staff scientist in the Electronic Sciences Laboratory. He was the lead investigator for a multispectral bandwidth compression techniques development study for LANDSAT Program Office. From 1994-1999 he was with Kuwait University as an associate professor of electrical and computer engineering. Since September 2000, he has been with the electrical engineering department at California Polytechnic State University, San Luis Obispo. Dr. Saghri's technical interests are in signal and image processing, bandwidth compression and coding, computer vision, and remote sensing. His publication record includes over fifty papers and several textbook chapter contributions. He is engaged in joint research work with Lockheed Martin in the areas of environmental remote sensing technology. His professional community services include technical reviewing for various journals and conference organization activities. He also served as an associate editor for book reviews for the journal of Optical Engineering. Dr. saghri is a fellow of the International Society of Optical Engineering (SPIE) and a senior member of the IEEE.

Andrew G. Tescher is currently affiliated with Compression Science Corporation where he is involved with modern video compression technologies. On behalf of Microsoft Corporation, he is the International Representative of NCITS L3 and the Head of Delegation for the U.S. to SC 29. (NCITS is the Subgroup of the Information Technology Industry Council, ITI, which is the principal US standards organization for information technology. SC 29 is the international plenary organization for multimedia technologies, including MPEG and JPEG). He is also a "Guest Researcher," at the Information Access Division of The National Institute of Standards and Technology, (NIST). He has been the chair of the Scientific Advisory Board of the Integrated Media System Center (IMSC) of the University of Southern California (USC) and he is Editor-in-Chief of the USC IMSC Press. Previously at Lockheed Martin, he was a technology advisor on compression /multimedia applications. Dr. Tescher received numerous major professional awards including the Edward Rhein Prize from Germany's Edward Rhein Foundation, recipient of the Gold Medal of SPIE, all in recognition of Dr. Tescher's contributions to the image and video compression fields. He is Fellow and Life Member of SPIE, Fellow of the Optical Society of America, Past President of SPIE. He is co-inventor of several teleconferencing systems and co-author of related key patents. He has written extensively on compression technologies for commercial and space applications and other areas of signal processing. Dr. Tescher received his Ph.D. in Electrical Engineering from the University of Southern California.

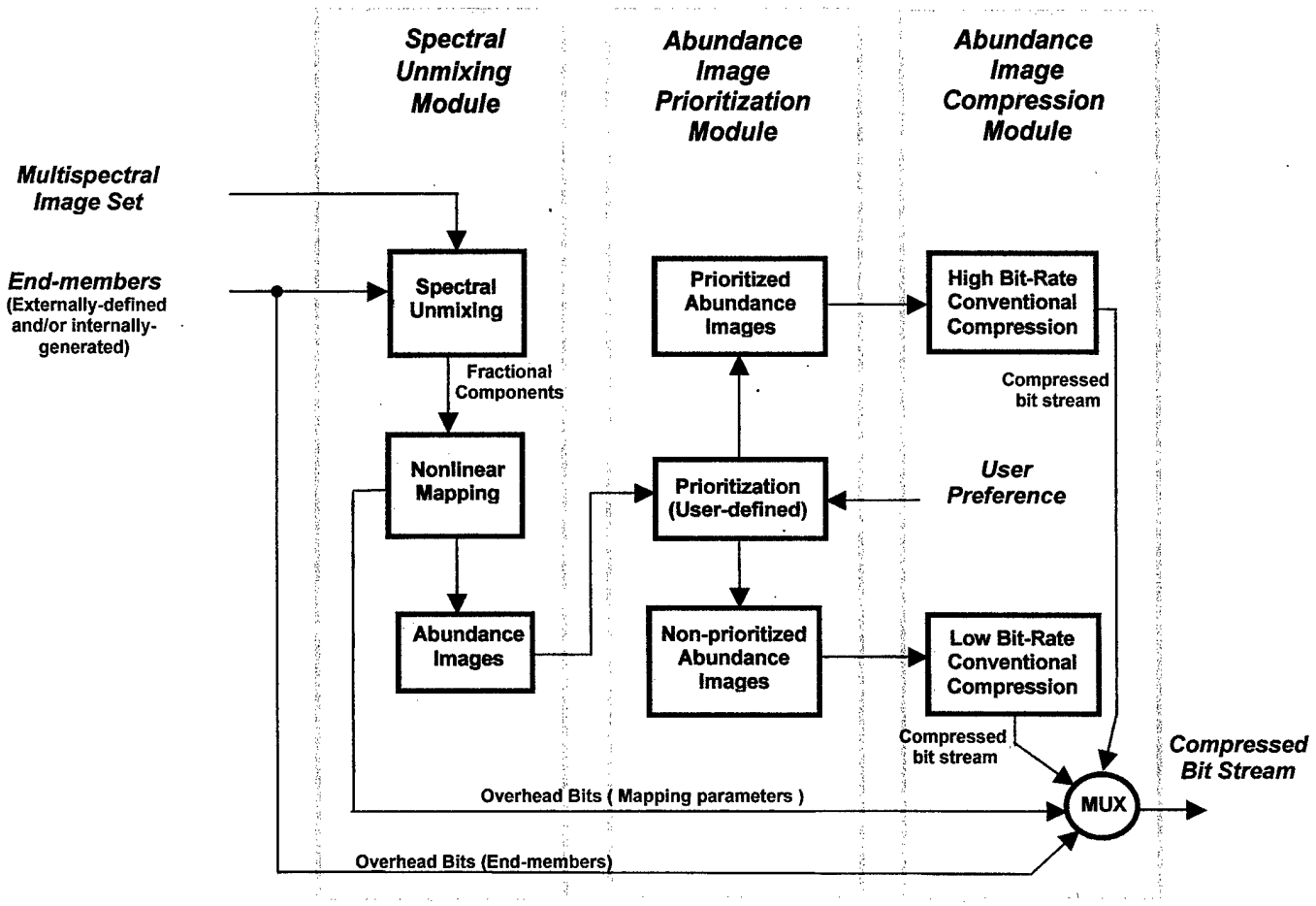


Figure 1. Initial conceptual design considered for class-prioritized compression

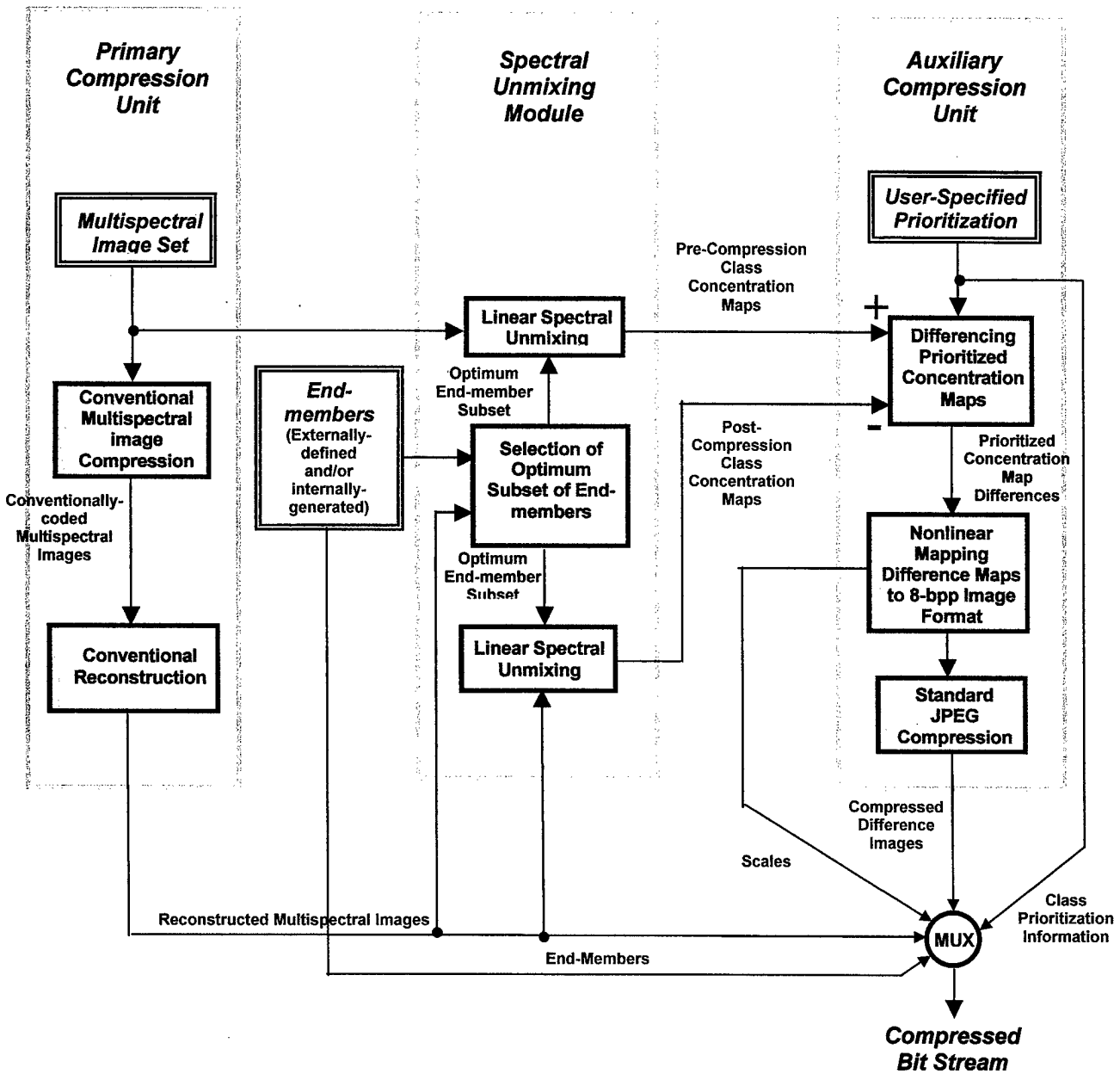


Figure 2. Block diagram of the proposed class-prioritized compression system

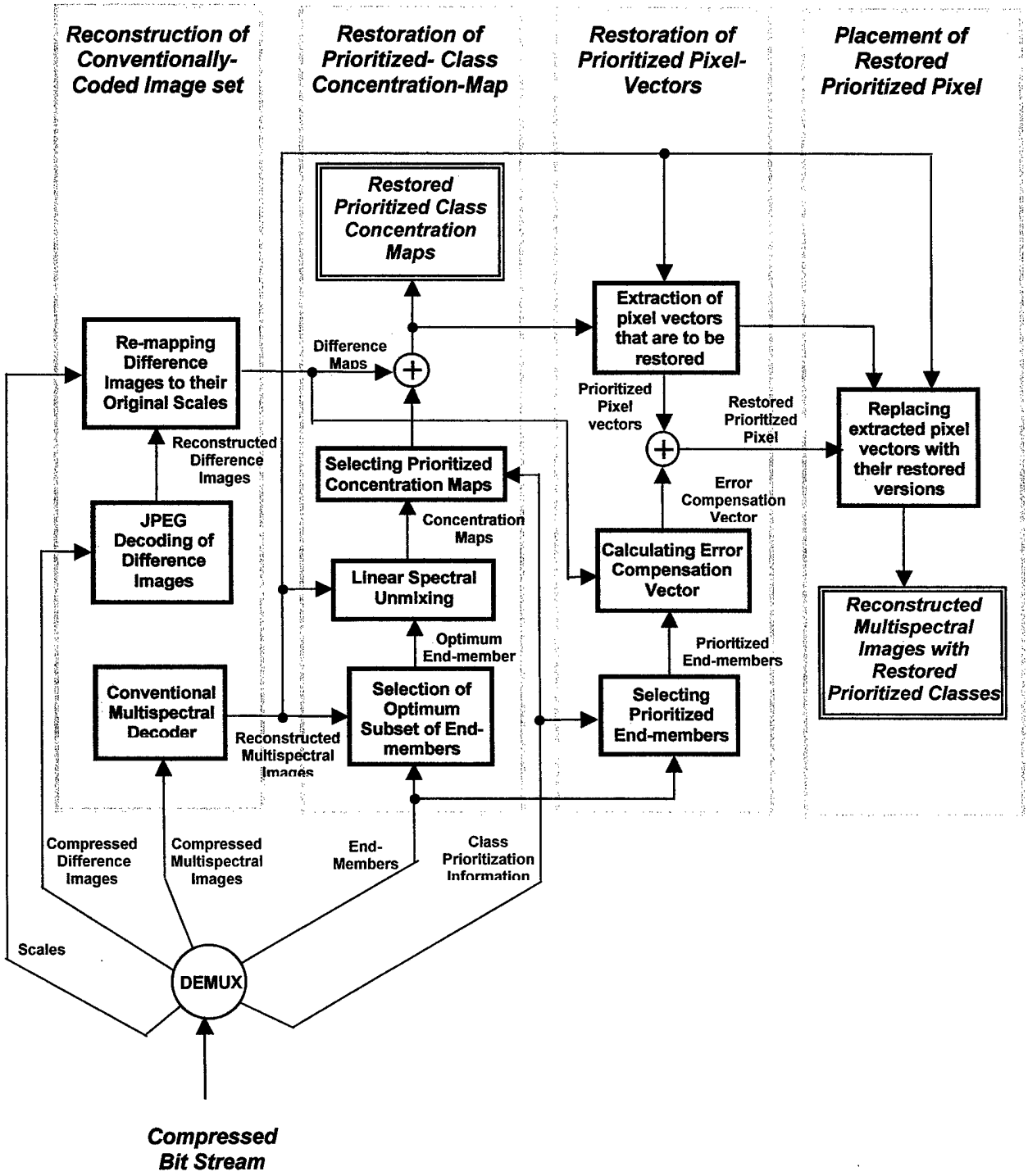


Figure 3. Decompression Block diagram

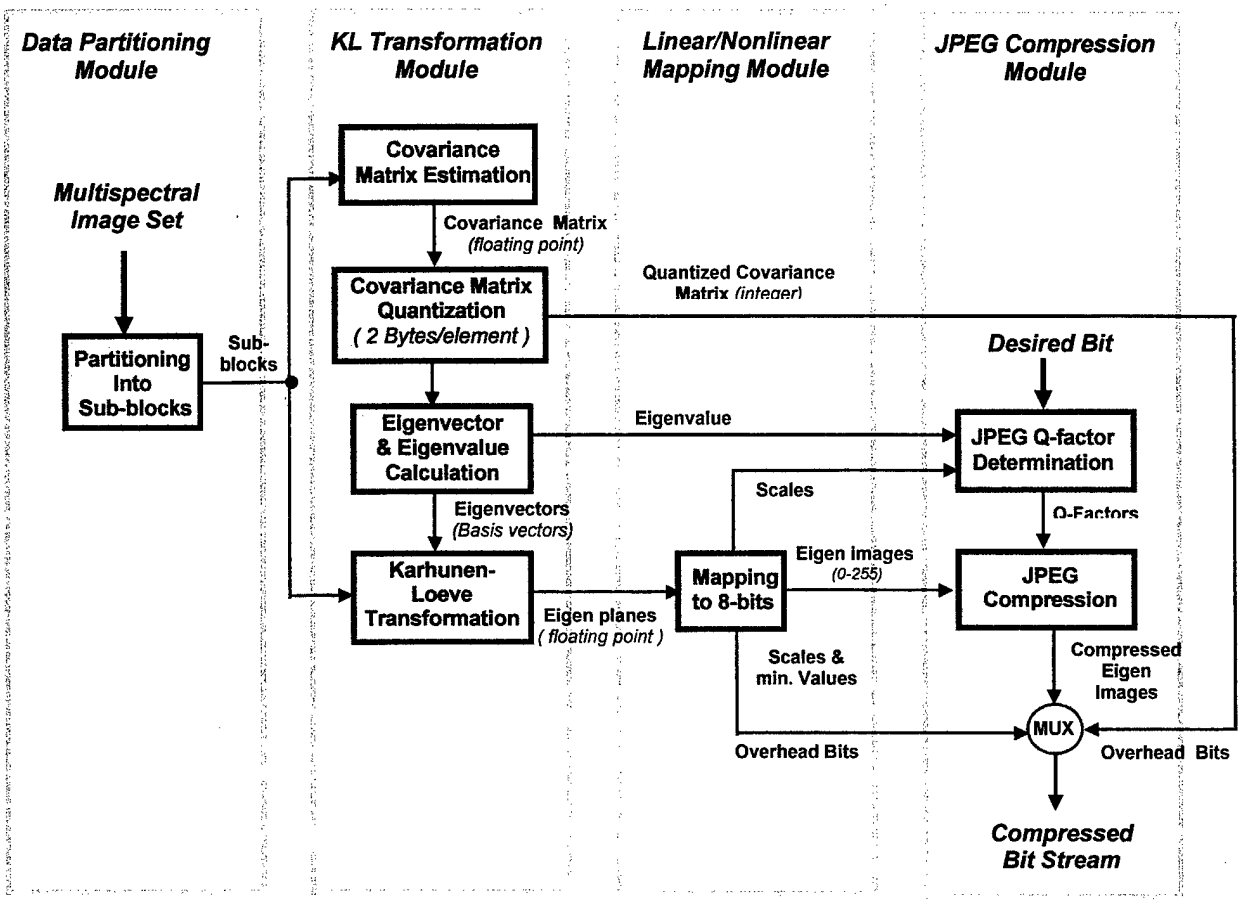


Figure 4. Terrain-Adaptive Multispectral Image Compression selected for the primary compression module

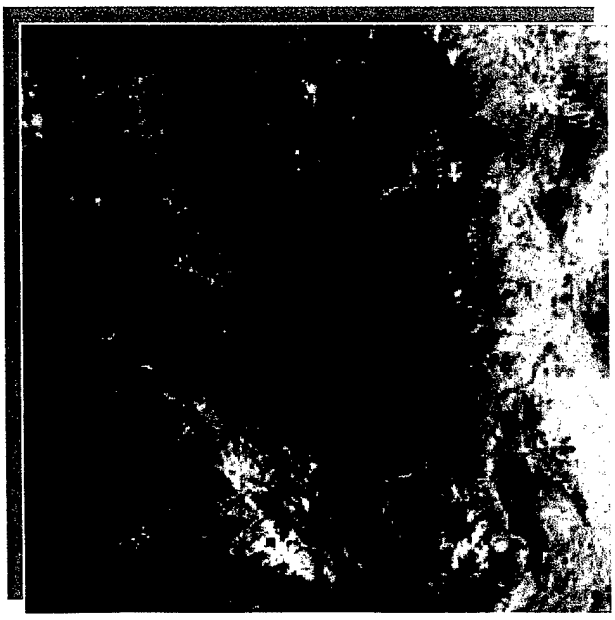


Figure 5. Band 1 of the Landsat MSS



Figure 6. The abundance images showing concentration of six different classes in the scene

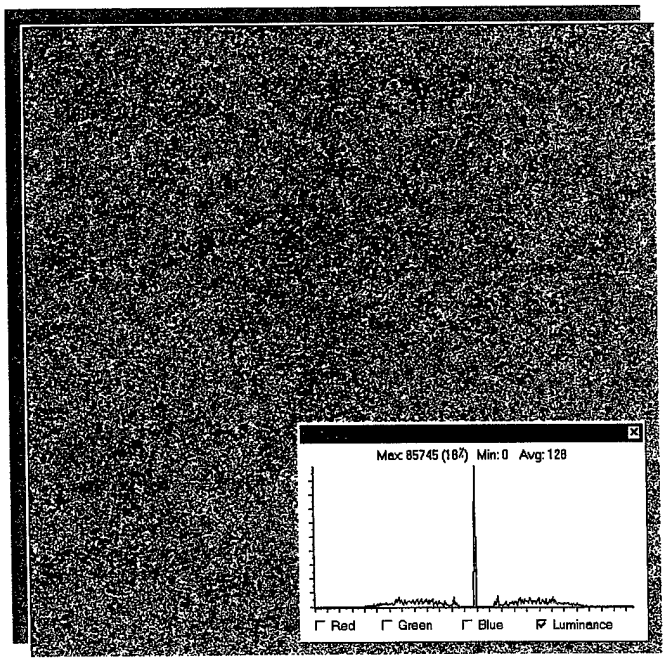


Figure 7. Resulting homogeneous difference image (representing compression-induced error on "water" concentration map) due to the primary compression unit at R_p =1 bpp

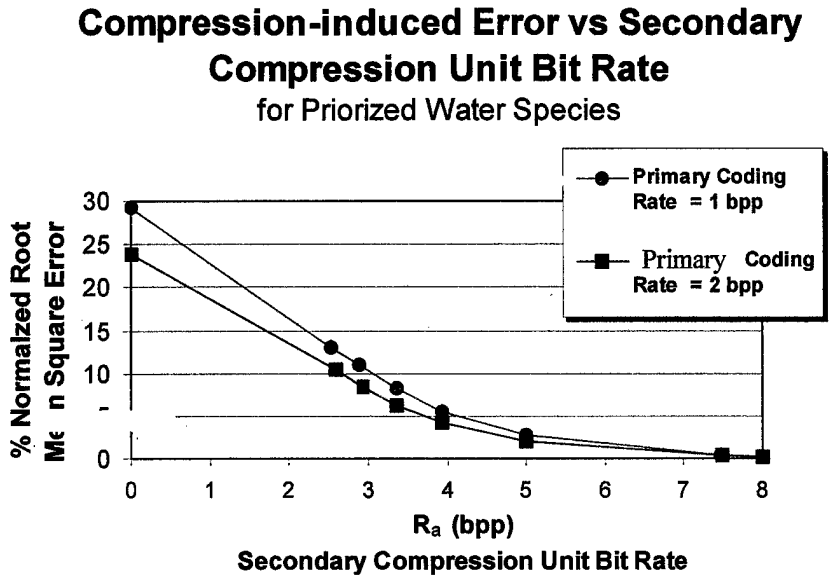


Figure 8. Compression-induced error versus secondary compression unit bit rate, for prioritized *Water* class, at primary compression unit bit rates of 1 and 2 bpp .

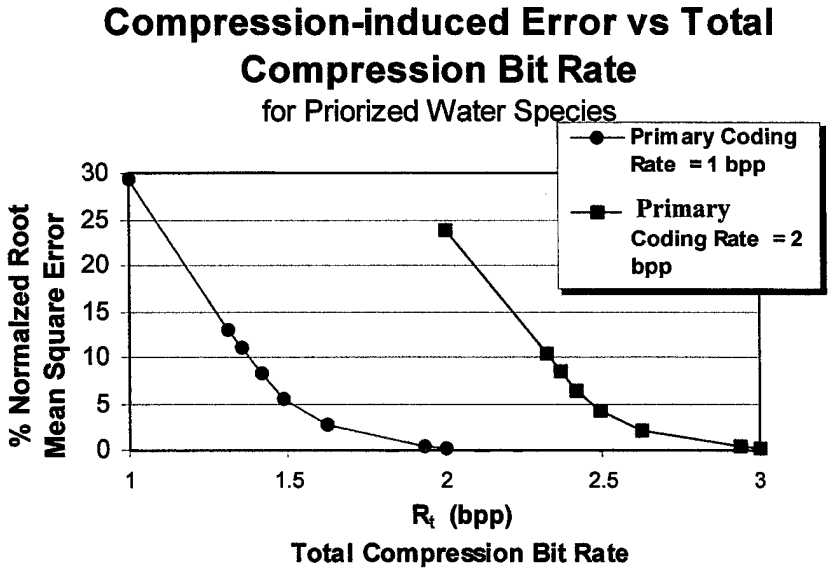


Figure 9. Compression-induced error versus total compression bit rate, for prioritized *Water* class at the primary compression unit bit rates of R_p =1 and 2 bpp

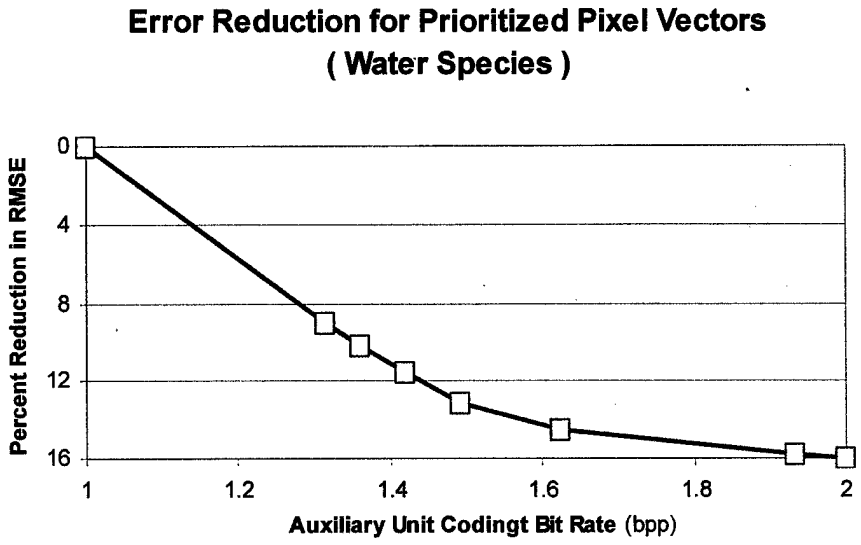


Figure 10. Percent error reduction at various auxiliary unit coding bit rates, for pixel vectors representing *Water*. Primary coding bit rate is 2 bpp