CESDIS

Center of Excellence in Space Data and Information Sciences

Annual Report

Year 9

July 1996 - June 1997

Dr. Yelena Yesha, Director

CESDIS
NASA Goddard Space Flight Center
Code 930.5
Greenbelt, MD 20771

phone: 301-286-4403
fax: 301-286-1777
e-mail: cas@cesdis1.gsfc.nasa.gov
http://cesdis.gsfc.nasa.gov

Operated by Universities Space Research Association in cooperation with the National Aeronautics and Space Administration

NAS5-32337

CESDIS

Center of Excellence in Space Data and Information Sciences

Annual Report

Year 9
July 1996 - June 1997
Dr. Yelena Yesha, Director

CESDIS
NASA Goddard Space Flight Center
Code 930.5
Greenbelt, MD 20771

phone: 301-286-4403
fax: 301-286-1777
e-mail: cas@cesdis1.gsfc.nasa.gov
http://cesdis.gsfc.nasa.gov

Operated by Universities Space Research Association in cooperation with the National Aeronautics and Space Administration

FOREWORD

This report summarizes the range of computer science-related activities undertaken by CESDIS for NASA in the twelve months from July 1, 1996 through June 30, 1997. These activities address issues related to accessing, processing, and analyzing data from space observing systems through collaborative efforts with university, industry, and NASA space and Earth scientists.

The sections of this report which follow, detail the activities undertaken by the members of each of the CESDIS branches. This includes contributions from university faculty members and graduate students as well as CESDIS employees. Phone numbers and e-mail addresses appear in Appendix D (CESDIS Personnel and Associates) to facilitate interactions and new collaborations.

TABLE OF CONTENTS

Applied Information Technology Branch

OVERVIEW

CESDIS, the Center of Excellence in Space Data and Information Sciences, was developed jointly by the National Aeronautics and Space Administration (NASA), Universities Space Research Association (USRA), and the University of Maryland in 1988. It is operated by USRA, under a contract with NASA. The program office and a small, core staff are located on-site at NASA's Goddard Space Flight Center in Greenbelt, Maryland.

USRA and the CESDIS Science Council

USRA is a nonprofit consortium of 80 colleges and universities, offering graduate programs in space sciences or related areas, which operates research centers and programs at several NASA centers. Most notable are the Lunar and Planetary Institute (LPI) at the Johnson Space Center in Houston, Texas, the Institute for Computer Applications in Science and Engineering (ICASE) at the Langley Research Center in Hampton, Virginia, and the Research Institute for Advanced Computer Science (RIACS) at the Ames Research Center at Moffett Field, California.

Oversight of each USRA institute or program is provided by a science council which serves as a scientific board of directors. Science council members are appointed by the USRA Board of Trustees for three-year terms. Members of the CESDIS Science Council during 1996-1997 were:

- Dr. Rama Chellappa
  University of Maryland College Park
- Dr. Burt Edelson
  George Washington University
- Dr. Richard Muntz
  University of California, Los Angeles
- Dr. David Nicol
  Dartmouth College
- Dr. Jacob Schwartz
  New York University
- Dr. Patricia Selinger
  IBM Almaden Research Center
- Dr. Harold Stone (Convener)
  NEC Research Institute
- Dr. Satish Tripathi
  University of Maryland College Park
- Dr. Mark Weiser
  Xerox PARC

The CESDIS Science Council meets annually at Goddard to review ongoing CESDIS research programs and new initiatives.

The CESDIS Mission

CESDIS was formed to focus on the design of advanced computing techniques and data systems to support NASA Earth and space science research programs. The primary CESDIS mission is to increase the connection between computer science and engineering research programs at colleges and universities and NASA groups working with computer applications in Earth and space science. Research areas of primary interest at CESDIS include:

- High performance computing, especially software design and performance evaluation for massively parallel machines,
- Parallel input/output and data storage systems for high performance parallel computers,
- Parallel hardware and software systems,
- Database and intelligent data management systems for parallel computers,
- Image processing,
- Digital libraries, and
- Data compression.

CESDIS funds multiyear projects at U.S. universities and colleges. Proposals are accepted in response to calls for proposals and are selected on the basis of peer reviews. Funds are provided to support faculty and graduate students working at their home institutions. Project personnel visit Goddard during academic recess periods to attend workshops, present seminars and collaborate with NASA scientists on research projects. Additionally, CESDIS takes on specific tasks for computer science research requested by NASA Goddard scientists.

A small, core staff is housed on-site at NASA Goddard. (A CESDIS organizational chart is included at the end of this introductory section.) This staff includes USRA employees and university research personnel attached to CESDIS via subcontracts who work in one of three branches: Computational Sciences, Applied Information Technology, or Administration. The bulk of this report describes the work of each branch in detail.

CESDIS World Wide Web Homepage

The CESDIS web site is fully indexed and can be located through:

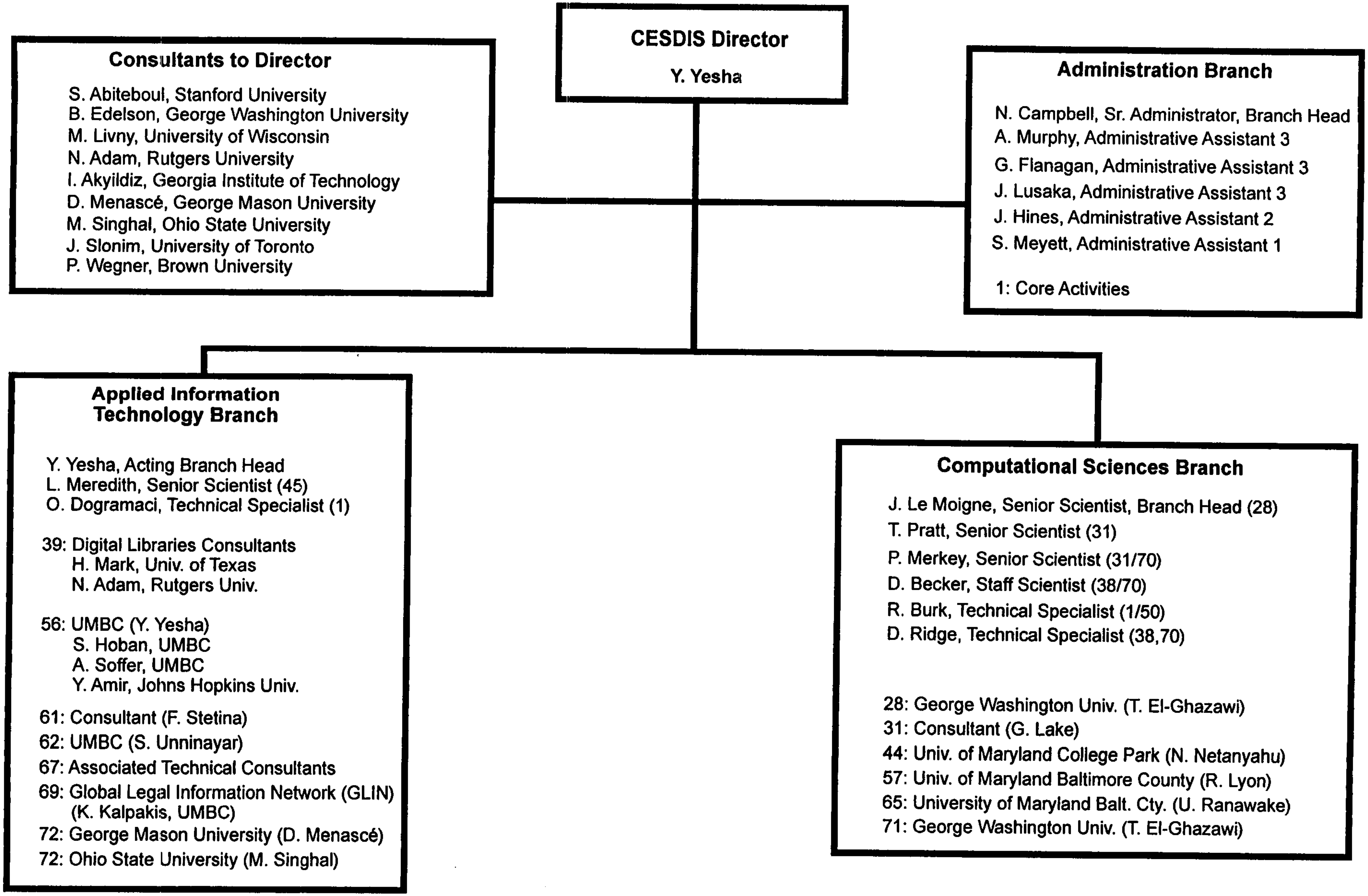
http://cesdis.gsfc.nasa.gov/

Contained in this home page are an overview of the CESDIS mission, special announcements, an explanation of the CESDIS organizational structure, and links to specific research projects and accomplishments.

The CESDIS home page is an active link to the heart of CESDIS activities. Feedback and comments are encouraged electronically to:

cas@cesdis.gsfc.nasa.gov

CESDIS ORGANIZATIONAL CHART



CESDIS Director

Y. Yesha

Consultants to Director

S. Abiteboul, Stanford University
B. Edelson, George Washington University
M. Livny, University of Wisconsin
N. Adam, Rutgers University
I. Akyildiz, Georgia Institute of Technology
D. Menascé, George Mason University
M. Singhal, Ohio State University
J. Slonim, University of Toronto
P. Wegner, Brown University

Administration Branch

N. Campbell, Sr. Administrator, Branch Head
A. Murphy, Administrative Assistant 3
G. Flanagan, Administrative Assistant 3
J. Lusaka, Administrative Assistant 3
J. Hines, Administrative Assistant 2
S. Meyett, Administrative Assistant 1

1: Core Activities

Applied Information Technology Branch

Y. Yesha, Acting Branch Head
L. Meredith, Senior Scientist (45)
O. Dogramaci, Technical Specialist (1)

39: Digital Libraries Consultants
H. Mark, Univ. of Texas
N. Adam, Rutgers Univ.

56: UMBC (Y. Yesha)
S. Hoban, UMBC
A. Soffer, UMBC
Y. Amir, Johns Hopkins Univ.

61: Consultant (F. Stetina)
62: UMBC (S. Unninayar)
67: Associated Technical Consultants
69: Global Legal Information Network (GLIN)
(K. Kalpakis, UMBC)
72: George Mason University (D. Menascé)
72: Ohio State University (M. Singhal)

Computational Sciences Branch

J. Le Moigne, Senior Scientist, Branch Head (28)
T. Pratt, Senior Scientist (31)
P. Merkey, Senior Scientist (31/70)
D. Becker, Staff Scientist (38/70)
R. Burk, Technical Specialist (1/50)
D. Ridge, Technical Specialist (38,70)

28: George Washington Univ. (T. El-Ghazawi)
31: Consultant (G. Lake)
44: Univ. of Maryland College Park (N. Netanyahu)
57: Univ. of Maryland Baltimore County (R. Lyon)
65: University of Maryland Balt. Cty. (U. Ranawake)
71: George Washington Univ. (T. El-Ghazawi)

DIRECTOR

DR. YELENA YESHA
(yesha@cesdis.edu)

Dr. Yelena Yesha is a tenured full professor in the Department of Computer Science and Electrical Engineering at the University of Maryland Baltimore County (UMBC), holds a joint appointment with the University of Maryland's Institute for Advanced Computer Studies (UMIACS) in College Park, and serves as the CESDIS Director through a memorandum of understanding between the University of Maryland and USRA.

Dr. Yesha received a Bachelor of Science degree in computer science from York University in Toronto, Canada in 1984, and a Master of Science and Ph.D. in computer and information science from Ohio State University in 1986 and 1989 respectively. She is a Senior Member of the IEEE Society, and a member of the ACM and New York Academy of Science. Her research interests include distributed databases, distributed systems, and performance modeling. She has authored more than 50 papers and edited six books in these areas.

Prior to joining CESDIS in December 1994, Dr. Yesha was on leave from the University to serve as the Director of the Center for Applied Information Technology at the National Institute of Standards and Technology. The Center's mission was to advance the goals of the National Information Infrastructure by identifying, developing, and demonstrating critical new technologies and their applications which could be successfully commercialized by U. S. industry.

ACTIVITIES

- Attended a meeting at the White House with Thomas Kalil (Executive Director to the National Economic Council), Burt Edelson and Neil Helm (George Washington University), and Jim Johnson (GSFC Code 833.1). The purpose of the meeting was to discuss CESDIS involvement in the G-7 Information Technology Program and its potential role in coordinating projects that are part of that program.

- Gave a presentation about CESDIS technical work to the participants in the Visiting Student Enrichment Program sponsored by the USRA,s Goddard Visiting Scientist Program.

- Attended several Maryland Technology Alliance meetings at UMBC.

- Served as a guest editor (with Nabil Adam, Rutgers University) of the *IEEE Transactions on Knowledge and Data Engineering* special issue on digital libraries which appeared in August 1996.

- Hosted the annual meeting of USRA's CESDIS Science Council at GSFC in September 1996.

- Phil Merkey, Don Becker and I met with Mr. Koob, a program manager from ARPA, to discuss the proposal that CESDIS submitted jointly with computer science professors from Johns Hopkins University in the area of metacomputing.

- Served as one of the hosts for the Global Legal Information Network (GLIN) workshop for project directors from 12 countries.

- Attended a meeting with Abe Har'Aven, the Director of the Israel Space Agency, and Mr. Joe Rothenberg, the Director of NASA Goddard. The topic of the discussion was the scientific collaboration between the two agencies.

- Attended a workshop at Rutgers University and gave a lecture on CESDIS research.

- Visited Columbia University and met with Professor Al Aho, the Chairman of the Computer Science Department.

- Hosted a two-day meeting of the U.S.-Israel Science and Technology Commission. We presented the GSFC proposal for five projects in the area of information technology that was submitted to the Commission for possible funding.

- Met with Kathy Nado, the special assistant for outreach to the Director of Goddard, and Mr. Joe Rothenberg, the Director of Goddard. The topic of the meeting was the new presidential initiative in the Internet area.

- Visited Rome, Italy, where I attended a G-7 meeting on "The Global Marketplace for Small and Medium Enterprises" as a member of the U. S. delegation. The first three days were dedicated to meetings and the rest of the time was spent visiting Italian companies that are specializing in the area of electronic commerce.

- Hosted the Workshop on Data Matching and Mapping organized by Stanley Zdonik of Brown University and held at GSFC November 7, 1996. Presentations were made by Milt Halem (GSFC Code 930), James French (University of Virginia), Marc Postman (Space Telescope Science Institute), Susan Davidson (University of Pennsylvania), Yannis Ioannidis (University of Wisconsin), H. V. Jagadish (AT&T Laboratories), Serge Abiteboul (Stanford University), Peter Buneman (University of Pennsylvania), David Maier (Oregon Graduate Institute), Stanley Zdonik, George Lake (University of

Washington), Raghu Ramakrishnan (University of Wisconsin), Peter Wegner (Brown University), Bob Grossman (University of Illinois), Mariano Consens (University of Waterloo, Ontario), Nabil Adam (Rutgers University), and Miron Livny (University of Wisconsin). The text of the workshop report is included in the Consultants to the Director section of this report.

- Attended the annual international CASCON conference in Toronto which is sponsored by IBM. The conference was attended by 1500 scientists and engineers. I gave a tutorial on Challenges in Global Electronic Commerce and also chaired a workshop on electronic commerce.

- Met with Usia Galil, CEO of Elron Industries, several times to discuss the U. S.-Israel Information Technology Program, the involvement of Elron companies in pilot projects, and the NASA/CESDIS proposal to start a program in information technology that has been submitted to the U. S.-Israel Commission.

- Gave an invited lecture at the New Jersey Institute of Technology and spent some time with computer science faculty members in an effort to develop collaboration between CESDIS and faculty at NJIT.

- Visited Shamim Naqvi, Bellcore Chief Scientist, to meet with Bellcore scientists who are working on developing technology for the Internet.

- Met with Avi Silberschatz when he visited GSFC to explore collaboration possibilities between his research group at Lucent Technologies at Bell Laboratories and CESDIS in the area of mass storage.

- Visited IBM Toronto Labs and met with scientists, developers, and Mr. Robert Leblanc, the Director of IBM Toronto Labs. The topic of the meetings was the joint research in the area of global electronic commerce.

- Traveled to France and England to establish collaborations with faculty at INSEAD in Paris, France, THESEUS in Sophia, Antibe, and the London Business School. Gave an invited presentation and chaired a panel at the conference on digital cash held at THESEUS.

- Visited the Department of Computer Science at Johns Hopkins University with Nabil Adam (Rutgers University) to meet with Yair Amir to review his research activities in the areas of networking and multimedia that are currently funded under the CESDIS contract.

- Traveled to Bonn, Germany to attend a G-7 conference on Global Marketplaces for Small and Medium Enterprises. Held numerous meetings with the representatives of business and government sectors from all over the world to discuss the role of GLIN (Global Legal Information Network) in the booming arena of global electronic commerce. Gave a presentation on the GLIN work conducted by CESDIS and the Library of Congress.

- Attended the Advances in Digital Libraries Conference held at the Library of Congress in May 1997.

- Met with Bill Howard (Director of USRA's Division of Astronomy and Space Physics) and several Goddard scientists to discuss the involvement of CESDIS in the SOFIA project.

NEW INITIATIVES

Intercomparison of Automated Registration Algorithms for Multiple Source Remote Sensing Data

Proposal submitted in response to NASA Research Announcement NRA-97-MTPE-03, Satellite Remote Sensing Measurement Accuracy, Variability, and Validation Studies.

PI: Jacqueline Le Moigne (CESDIS)

Co-Investigators: James Tilton (GSFC Code 935), Samir Chettri (Global Science Technology, Inc.), Tarek El-Ghazawi (George Washington University), Emre Kaymaz, Bao-Ting Lerner, and John Pierce (KT-Tech, Inc.), Manohar Mareboyana (Bowie State University), David Mount and Nathan Netanyahu (University of Maryland College Park), and Srinivasan Raghavan and Wei Xia (Hughes STX).

Abstract: Many of the analysis techniques which will be utilized by the Mission to Planet Earth program will necessitate multiple data integration, which will require accurate registration of these data. Currently, the most common approach to registration is based on the manual extraction of a few outstanding characteristics of the data, called ground control points, from which a geometric transformation is computed. But such a point selection represents a repetitive, labor- and time-intensive task which becomes prohibitive for very large amounts of data. Also, since this interactive choice of control points is sometimes difficult, too few points, inaccurate points, or ill-distributed points might be chosen, thus leading to registration errors. For these and other reasons, automatic registration is becoming an important data analysis and production requirement.

Given the diversity of the data sources, it is unlikely that a single registration technique will satisfy all different applications. Although automated registration has been developed for a few Earth science applications, there is no general scheme which would assist users in the selection of a registration tool. In this work we propose to 1) develop an operational toolbox which consists of some of the most utilitarian registration techniques, and 2) provide a quantitative intercomparison of the different methods, which will allow a user to select for his/her application the desired registration technique based on this evaluation and the visualization of the registration results. The intercomparison will be based on accuracy, applicability, level of automation, and computational requirements criteria. These criteria will be computed for each algorithm utilizing several datasets which are relevant to the MTPE program, such as NOAA/AVHRR, Landsat/TM and MSS, GOES, and MODIS Airborne Simulator data.

The Khoros environment has been chosen as the framework for the implementation of these techniques. Since Khoros is an open software system, it will enable us to widely distribute the toolbox along with the results of our evaluation and to get feedback from a variety of users.

Integrating Environmental and Legal Information Systems

Proposal submitted in response to NASA Cooperative Agreement Notice CAN-97-MTPE-02, Extending the Use and Applications of Mission to Planet Earth (MTPE) Data and Information to the Broader User Community.

PI: Konstantinos Kalpakis, Assistant Professor, UMBC; CESDIS subcontractor on GLIN project.

Project Members: Susan Hoban (UMBC/CESDIS), Durwood Zaelke, David Hunter, and Gary Cook (Center for International Environmental Law), Steven Jamar (Howard University), Rubens Medina and Nick Kozura (Library of Congress), William Campbell (GSFC Code 935), and Pat Gary (GSFC Code 930).

Abstract: This project is a cooperative effort among the Universities Space Research Association, the University of Maryland Baltimore County, the Center for International Environmental Law (CIEL), the U. S. Library of Congress, and NASA's Goddard Space Flight Center. It is proposed to expand the use of MTPE science products to the field of environmental law, in much the same way as forensic technologies are now being applied to criminal law. One facet of the proposed effort will integrate remotely sensed and in situ data with two existing legal systems: the Global Legal Information Network (GLIN), under development by the Library of Congress and NASA, and the Environmental Law Information Network Exchange (E-Line), constructed by CIEL. A second aspect of the proposal entails the development of a Model Environmental Legislation (MEL) system to teach law students how to craft environmental legislation which takes advantage of MTPE remotely sensed data. MEL will also be used as a teaching tool for comparative studies of existing environmental treaties and legislation.

RESEARCH

Tools for Analyzing the Performance of Hierarchical Mass Storage Systems (with Odysseas Pentakalos [NRC Research Fellow] and Daniel Menascé [George Mason University].)

Hierarchical mass storage systems are becoming more complex each day and there are many possible ways of configuring them. The options range from the type and number of devices to be used to their connectivity. Furthermore, the demands placed on the mass storage systems are continuously increasing in intensity. This forces system managers to constantly monitor the system, evaluate the demand placed on it, and tune it appropriately using either heuristics based on experience or analytic models. This procedure involves two steps. First, workload characterization must be used to understand and describe in a concise manner the load imposed on the system. Then, a model of the system must be constructed and solved to detect the bottlencks and evaluate various reconfigurations. Generating an accurate model of the workload through workload characterization is a laborious and time consuming process. Once the workload model has been obtained, constructing an accurate performance model of the system requires understanding the application, solution techniques, and limitations of analytic modeling. To automate both of these tasks we developed two tools: Pythia/WK, a tool for automated clustering-based workload characterization, and Pythia, an extensible object-oriented performance analyzer.

The main features of Pythia/WK are:

- Automatic support for peak-period determination: histograms of system activity are generated and presented to the user for peak-period determination.
- Automatic clustering analysis: the data collected from the mass storage system logs is clustered using clustering algorithms and tightness measures to limit the number of generated clusters.
- Reporting of varied file statistics: the tool computes several statistics on file sizes such as average, standard deviation, minimum, maximum, frequency, as well as average transfer time. These statistics are given on a per cluster basis.
- Portability: the tool can easily be used to characterize the workload in mass storage systems of different vendors. The user needs to specify through a simple log description language the way in which a specific log should be interpreted.

Pythia was designed and implemented to allow users to easily investigate the most cost-effective configurations for a given workload. One of the most important reasons to build such a tool is to provide a simple way through which queueing analytic models can be used for performance prediction and system sizing of mass storage systems. The tool incorporates a modeling wizard component that is capable of automatically building a queueing network model from a mass storage system representation defined through a graphic editor. Thus, the user of the tool does not need to know queueing network modeling techniques to use it.

The design of the analyzer was based on the following requirements:

- Architecture and system independence: the analyzer should not be tailored to any architecture even one as broad as the IEEE Mass Storage Systems Reference Model or be tailored to a specific system design such as the Unitree or IBM's DFSMShsm.

- Extensibility: the analyzer should be easily adaptable to new media technologies and devices.

- Graphical user interface: the analyzer should provide a graphical user interface for the specification of the particular mass storage system to be analyzed.

Figure 1 shows the main screen of the tool. Pythia provides a graphic editor for describing the architecture of the hierarchical mass storage system to be analyzed. The main screen consists of five components: the menu bar, the toolbar, the mode selection button, the canvas, and the status bar. The menu bar provides access to the major functions of the tools such as file loading and saving through the file menu, canvas editing functions through the edit menu, performance solutions through the tools menu, and performance experiment plotting through the view menu. Keeping with the extensibility design requirement, the toolbar is dynamically created based on the objects specified within the backend database. Adding support to the tool for a new type of storage media or interconnect can be done by simply adding a record to the backend database.

One of Pythia's main features is a modeling wizard that builds a queuing network from the graphical description of the system using a set of heuristics. In what follows, we give an example of the heuristic used to model a robotic tape library placed at the nearline level. First we introduce the notation used to describe the heuristics. The lower case letters *o*, *w*, and *q* denote elements of the sets of mass storage system objects, workloads, and queuing network objects. The expression *o* ==> *q* means that object *o* generates a single device *q* in the queuing network. The expression *a.b* evaluates to the value of attribute

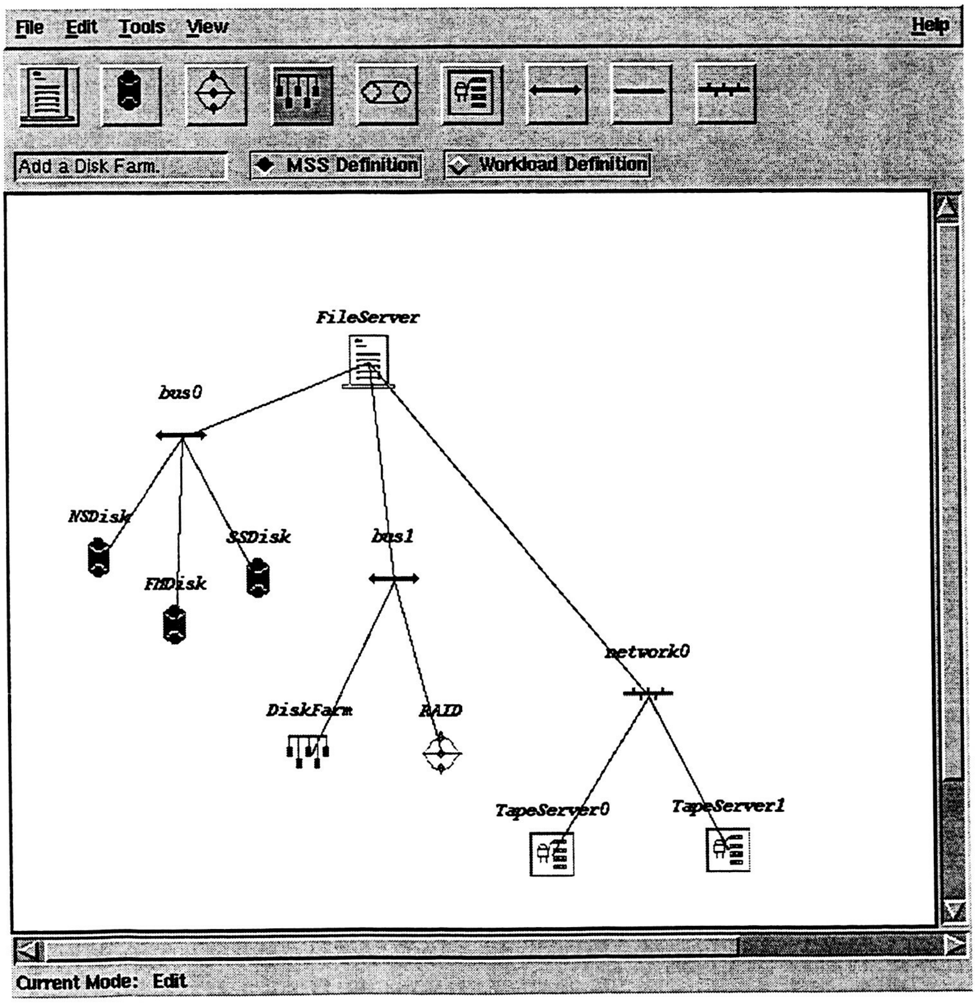


Figure 1: Main Screen of the Analyzer

b of element *a*. The function *HitRatio(o,w)* returns the effective hit ratio of workload *w* on object *o*. It is defined as

$$HitRatio(o,w) = \begin{cases} 1 & \text{if } o.numl = 1 \\ (1\text{-}w.hr_o) & \text{if } o.level = \text{n} \\ & \text{and } o.numl = 2 \\ (1\text{-}w.hr_o) \times w.hr_n & \text{if } o.level = \text{n} \\ & \text{and } o.numl = 3 \\ (1\text{-}w.hr_o) \times (1\text{-}w.hr_n) & \text{if } o.level = \text{f} \\ & \text{and } o.numl = 3 \end{cases}$$

The function evaluates to 1 if the number of levels in the hierarchy is 1, it evaluates to $(1\text{-}w.hr_o)$if the number of levels is 2, and the storage object is at the nearline level, it evaluates to $(1\text{-}w.hr_o) \times w.hr_n$ if the number of levels is 3, and the storage object is at the nearline level, and evaluates to $(1\text{-}w.hr_o) \times (1\text{-}w.hr_n)$ if the number of levels is 3 and the storage object is at the offline level. The following heuristic is used by the modeling wizard to convert a robotic tape library object *o* into components q_{robot} and q_{tape} of the QN.

if *o.level* = n and *o* ∈ w.dev then *o* ==> q_{robot} and *o* ==> q_{tape} where

$$q_{robot}.sd_w = w.vr_o \times HitRatio(o,w) \times o.mnt$$

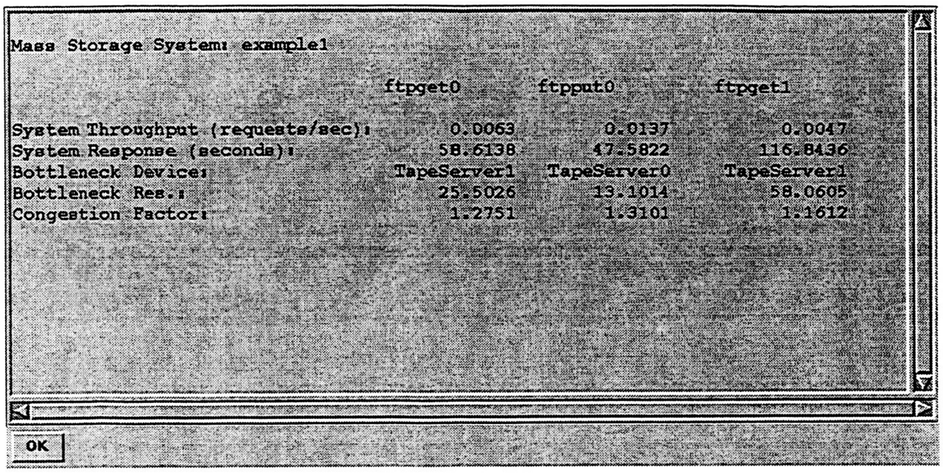
and

$$q_{tape}.sd_w = w.vr_o \times HitRatio(o,w) \times [o.seek + [w.fs/o.blks] \times (o.blks/o.trate)]$$

The above heuristic indicates that two devices, q_{robot} and q_{tape} should be generated. The equations for $q_{robot}.sd_w$ and $q_{tape}.sd_w$ indicate how the service demands should be computed for these two QN devices. The different workload and object parameters that appear in the equations are obtained from the user during the specification of a mass storage system.

The generated QN models are then solved using the approximate multi-class MVA algorithm along with a set of approximations developed by the authors specifically for the domain of hierarchical mass storage systems. This approximations include techniques for dealing with the fork and join synchronization found in RAID-5 devices, and for dealing with the simultaneous resource possession exhibited by requests to transfer files from a network-attached tape drive to a network-attached disk drive. These approximations were validated against discrete-event simulations and also actual measurements at the mass storage system at NASA's Center for Computational Sciences.

The tool solves the queueing network generated internally to provide the user with performance information such as the throughput per workload measured, the response time of the system, the name of the bottleneck device, the residence time at the bottleneck device per workload, and the congestion factor (defined as the ratio of residence time over service demand) per class at the bottleneck device. Figure 2 shows the screen the presents the results of solving the analytic model.



```
Mass Storage System: example1

                                    ftpget0       ftpput0       ftpget1

System Throughput (requests/sec):    0.0063        0.0137        0.0047
System Response (seconds):          58.6138       47.5822      116.8436
Bottleneck Device:              TapeServer1   TapeServer0   TapeServer1
Bottleneck Res.:                    25.5026       13.1014       58.0605
Congestion Factor:                   1.2751        1.3101        1.1612
```

Figure 2: Sample System Solution

In addition to solving the analytic model for the specified parameters, the tool also allows the user to perform a number of experiments. The experiments are throughput and response time versus workload intensity and throughput and response time versus hit ratio. The user may go back and modify the configuration of the mass storage system and see the effects of the change by running the experiment again. The automatic generation of the analytic model from the graphical description along with the efficiency of the approximation algorithms make Pythia a valuable tool for capacity planning of mass storage systems.

The following papers were published as a result of this research effort.

Odysseas I. Pentakalos, Daniel Menascé, and Yelena Yesha. Pythia and Pythia/WK: tools for the performance analysis of mass storage systems, to appear in *Software Practice and Experience.*

Odysseas I. Pentakalos, Daniel Menascé, and Yelena Yesha. Pythia: A Performance Analyzer of Hierarchical Mass Storage Systems, 9th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation, St. Malo, France, June 2-6, 1997.

Odysseas I. Pentakalos, Daniel Menascé, and Yelena Yesha, Automated Clustering Based Workload Characterization, *Proceedings of the 5th Goddard Conference on Mass Storage Systems and Technologies*, September 1996, College Park, Maryland.

Strategies for Maximizing Seller's Profit Under Unknown Buyer's Utility Values
(with Konstantinos Kalpakis and Bella Bellagradek, UMBC)

We studied a simple market model with very restricted information available to the participants. The main incentive to our work is the lack of reliable or up-to-date information in the existing markets, especially in the electronic markets where the situation tends to change rapidly.

Our model employs the following protocol of sales: at time moment (we assume that time is divided into some units, not necessarily uniform, numbered by non-negative integers). The seller posts the current price per unit of a product together with the number of units available for sale. At the same time the buyer reads the seller price and makes a decision to buy or not to buy, based on the comparison of the posted price and the buyer's utility value unknown to the seller initially. This process repeats a given number of times or until the product supply lasts or indefinitely.

The main question is "Is there an algorithm (pricing strategy) that allows the seller to maximize the collected profit, using mainly the history of his/her own sales?" Some additional questions:

1. How do we measure the optimality of the algorithm?
2. What additional assumptions are to be made to get meaningful results? and
3. What are the computational complexities of the proposed algorithms?

We obtained the following results in assumptions that the buyer's utility value is a constant in a given interval and only single unit of the product can be sold at any moment of time:

1. There are polynomial time/space dynamic programming algorithms that maximize the average case profit or the expected profit under a given distribution of buyer's utility values. This method works when both supply and time to sell are unbounded or one or both of them are finite.

2. We also studied the worst case when both supply and time to sell are unbounded. We got the lower and upper bounds for the loss of a pricing algorithm that differs by factor of order logN. The loss of an algorithm is a difference between the best achievable profit for a given input and the actual profit made. It is a convenient measure of "goodness" of an algorithm, often used in computational learning theory.

Future work lies in the direction when the buyer's utilities are varying in time rather than constant.

Publication:

Starategies for Maximizing Sellers' Profit Under Unknown Buyers' Utility Values (with Kostas Kalpakis and Bella Bellagradek), submitted to CASCON97.

Evolving Databases: An Application to Electronic Commerce (with Serge Abiteboul [Stanford University], Brad Fordham [NIST], and Konstantinos Kalpakis [UMBC].)

The evolving database represents a universe of discourse by capturing arbitrarily many semantic dimensions of its constituent entities. Here a "semantic dimension" is defined by the knowledge engineer by specifying two things: 1) the syntax(es) for values which populate that semantic dimension and 2) an evolving algebra specification of the "semantics" implied by each syntax.

This renders ALL semantics of an entity into a common form, the evolving algebra, which can be executed automatically -- allowing the user to experience the interactions of the various entity semantics. In effect, the evolving database has two powers. First, it captures static entity representations and relationships. Second, it maintains known semantics over time as stimuli external to the system change the evolving database's state.

The term "semantic dimension" in very broad. This is possible because the underlying evolving algebra formalism is a Turing-complete formal specification methodology. Some examples of semantic dimensions may be:

1. DATA defined with a syntax "?attribute=?value" and instantiated as color=green, weight=1.5.

2. CONSTRAINT defined with syntaxes "?attribute BETWEEN ?min AND ?max", "?attribute < ?value" and instantiated as weight BETWEEN 1 AND 5, height < 12.

3. I/O-BEHAVIOR defined with a syntax "?inputs ==> ?outputs" and instantiated as paper, toner, power ==> hardcopy

This broadness in the semantic dimension, the key to our data modeling and manipulation approach, permits a clean integration between many traditional semantic notions found in current DBMSs including: relationships like inheritance, constraints, behaviors of various flavors, inference, and so on. Thus, the major objective of the evolving database is to provide a DBMS in which very semantically complex and dynamic entities can be fruitfully modeled and managed.

Publications:

Evolving databases: an application to electronic commerce, (with Serge Abiteboul and Brad Fordham), to appear in the *Proceedings of the International Database Engineering and Applications Symposium*, Montreal, Canada, August 25-27, 1997.

Electronic commerce: current limitations and future visions (with Kostas Kalpakis and Brad Fordham), invited for IEEE *Transactions on Knowledge and Data Engineering.*

SELECTED PUBLICATIONS

Publications in refereed journals

Adam, N., Awerbuch, B., Slonim, J., Wegner, P., & Yesha, Y. (1997). Globalizing business, education, culture through the Internet. *Communications of the ACM*, 40(2), 115-121.

Adam, N., El-Ghazawi, T., Halem, M., Kalpakis, K., & Yesha, Y. (1996). The Global Legal Information Network. *The American University Law Review.* 46(2), 477-491.

Adam, N., Halem, M., Holowcazak, R., Lal, N., & Yesha, Y. (1996). Digital Libraries Task Force, IEEE *Computer.*

Adam, N. & Yesha, Y. (Editors). (1996). Special section on digital libraries, IEEE *Transactions on Knowledge and Data Engineering.*

Yesha, Y., & others. (1996). Strategic directions in electronic commerce and digital libraries. *ACM Computing Surveys.* 28(4), 818-835.

Articles/Papers Accepted

Kalpakis, K., Awerbuch, B., and Yesha, Y. Towards free information markets. *Mathematical Modeling and Scientific Computing.*

Kalpakis, K., Fordham, B., & Yesha, Y. Electronic commerce: Current limitations and future visions. IEEE *Transactions in Knowledge and Data Engineering.*

Pentakalos, O. I., Menascé, D., & Yesha, Y. Pythia: a performance analyzer of hierarchical mass storage systems. PNPM/Tools Conference, San Malo, France, June 3-6, 1997.

Pentakalos, O. I., Menascé, D., & Yesha, Y. Pythia and Pythia/WK: Tools for the performance analysis of mass storage systems. *Software Practice and Experience.*

Books

Adam, N., and Yesha, Y. (Eds.). (1996). *Electronic commerce: Current research issues and applications.* Lecture Notes in Computer Science. Berlin, Germany: Springer-Verlag.

Adam, N., Halem, M. & Yesha, Y. (Eds.). (1996). *Proceedings of the Third Forum on Research and Technology Advances in Digital Libraries (ADL '96).* Los Alamitos, CA: IEEE Computer Society Press.

Chapters in books

Adam, N., & Yesha, Y. (1996). Electronic commerce: An overview. In Adam, N., and Yesha, Y. (Eds.). *Electronic commerce: Current research issues and applications.* (pp. 5-12). Lecture Notes in Computer Science. Berlin, Germany: Springer-Verlag.

Publications in proceedings

Pentakalos, O. I., Menascé, D., & Yesha, Y. (1996). An analytic model of hierarchical mass storage systems with network-attached devices. In ACM SIGMETRICS '96, Philadelphia, PA.

Pentakalos, O. I., Menascé, D., & Yesha, Y. (1996). Automated clustering-based workload characterization. In *Proceedings of the Fifth NASA Goddard Conference on Mass Storage Systems and Technologies,* College Park, Maryland.

CONSULTANTS TO THE DIRECTOR

Task 1 on the CESDIS contract (the general administrative task) allows the Director to bring to CESDIS consultants who are not funded by specific task originators. CESDIS entered into agreements with the individuals reported upon in this section for the purpose of program development.

Serge Abiteboul
Stanford University, Department of Computer Science

Nabil Adam
Rutgers University, Center for Information Management, Integration, and Connectivity

Ian Akyildiz
Georgia Institute of Technology, Broadband and Wireless Networking Laboratory

Data Mapping and Matching Group
Serge Abiteboul (Stanford University), Peter Buneman (University of Pennsylvania),
David Maier (Oregon Graduate Institute), Stanley Zdonik (Brown University)

Burt Edelson
George Washington University, Department of Electrical Engineering and Computer Science

Miron Livny
University of Wisconsin, Department of Computer Science

Daniel Menascé
George Mason University, Department of Computer Science

Mukesh Singhal
Ohio State University, Department of Computer and Information Science

Jacob Slonim
University of Toronto (Ontario), Computer Science Department

Peter Wegner
Brown University, Department of Computer Science

Serge Abiteboul

Stanford University, Department of Computer Science
(abitebou@db.stanford.edu)

Statement of Work

Dr. Abiteboul has worked with the CESDIS Director to model the state of an electronic commerce transaction as an active database with an emphasis on communication with the external world. Rules and constraints are used to describe the agreed upon laws that govern the transaction and the protocols that describe how participants interact with the database. The goal of the effort is to determine which portion of database technology is well suited in this context, which portions have to be extended, and which features (such as process modeling tools) are missing. Providing tools with formal semantics to describe electronic commerce transactions has also been considered.

Results

Most of this work was performed while at Stanford. I also visited CESDIS for two weeks in July 1996 and for shorter visits. The work on electronic commerce (EC) was performed with Prof. Yelena Yesha (CESDIS) and Brad Fordham (NIST).

The goal was to investigate a formal approach to electronic commerce modeling based on Gurevich evolving algebras.

There is a real need for more formal foundations for definitions in that domain are often unclear and this generates confusion and a lot of redundancy in efforts. Evolving algebra is a formalism originally proposed to formally specify program semantics. Some work was needed to make it suited for describing EC applications. We developed the needed concepts. A system based on these ideas was implemented and the approach validated with some particular applications. Although more work is needed, we can already say that the work proved to be quite successful.

What has been achieved:

1. Evolving algebras are well-suited for capturing EC applications. However, they do not provide a user-friendly specification language. We developed "evolving databases", a general purpose model convenient for specifying such applications and in particular suited to describe their active component (via rules). The general approach is described in [1].

2. We developed a first prototype to validate the ideas. The developer and prime architect is Brad Fordham. Some applications were implemented.

3. We set up the basis for a formal study of customizable EC models based on simple rules (extending DATALOG) [2]. The main focus is on adding "active features", while keeping the simplicity of the language. This is in order to provide automatic help (e.g., what the user should do next) and customization checking. The work was initiated from discussion with Al Aho (Columbia University) and Alberto Mendelzohn (University of Toronto) in the summer.

The following future efforts are considered:

a. Develop a complex application that would in particular highlight the possibilities of the model with respect to customization, and on-line modification of the EC model.

b. Understand better how the model can integrate standards such as EDI or integrate tools such as available payment mechanisms.

c. The integration of Behavioral Datalog [3] in the prototype.

Efforts in related themes:

(i) semistructured data: how to query data that is very irregular [3,4,5,6,7]

(ii) distributed query optimization: how to optimize and restructure data [5,8,9]

(iii) theory of the Web: trying to capture the essential aspects of Web computation and complexity vs. more standard kinds of computations. [10,11]

References:

[1] Fordham, B., Abiteboul, S., & Yesha, Y. (1997), Evolving Databases: An Application to Electronic Commerce, International Database Engineering and Applications Symposium (IDEAS), Montreal.

[2] Abiteboul, S., Fordham, B., & Yesha, Y., Behavioral Datalog, in preparation.

[3] Abiteboul, S. (1997) Querying semistructured data, *Proceedings of the International Conference on Database Theory*, Delphi, Greece. (invited paper)

[4] Abiteboul, S., Quass, D., McHugh, J., Widom, J., & Wiener, J. (April 1997) The lorel query language for semistructured data, *International Journal on Digital Libraries*, 1(1):68-88.

[5] Abiteboul, S., Cluet, S., & Milo, T. (1997) Correspondence and translation for heterogeneous data, *Proceedings of the International Conference on Database Theory*, Delphi, Greece.

[6] Abiteboul, S., Goldman, R., McHugh, J., Vassalos, V., & Zhuge, Y. (1997) Views for semistructured data, International Workshop on Management of Semistructured Data, Tucson.

[7] Nestorov, S., Abiteboul, S., & Motwani, R. (1997) Inferring Structure in Semistructured Data, International Workshop on Management of Semistructured Data, Tucson.

[8] Abiteboul, S., Cluet, S., Christophides, V., Milo, T., Moerkotte, G., & Simeon, J. (April 1997) Querying Documents in Object Databases, *International Journal on Digital Libraries*, 1(1):5-19.

[9] Abiteboul, S. & Vianu, V. (1997) Regular Path Queries with Constraints, *Proceedings of ACM Principle of Database Systems*, Tucson.

[10] Abiteboul, S. & Vianu, V. (1997) Queries and Computation on the Web, *Proceedings of the International Conference on Database Theory*, Delphi, Greece

[11] Papakonstantinou, Y., Abiteboul, S., Garcia-Molina, H. (1996) Object Fusion in Mediator Systems, *Proceedings of the International Conference on Very Large Data Bases*, Bombay.

Nabil Adam

Rutgers University,
Center for Information Management, Integration, and Connectivity
(adam@adam.rutgers.edu)

Statement of Work

Dr. Adam worked as a CESDIS subcontractor to do the following:

- Provide technical management of CESDIS research projects and help nurture an environment of interactive supervision of the CESDIS branch heads.

- Assist the CESDIS Director and technical staff members in developing a proposal in response to the next CESDIS contract procurement notice.

- Assist the CESDIS Director and technical staff in developing new initiatives.

- Help increase the visibility of the CESDIS research staff by developing stronger ties to and improving communication with the NASA scientific community at Headquarters, GSFC, and the other NASA centers.

- Help increase the visibility of CESDIS within the scientific community by encouraging and facilitating the acceptance of articles describing CESDIS-supported research by journals and papers by conference committees.

Results

- I held meetings with and provided technical oversight and guidance to Drs. Yair Amir (Johns Hopkins University) and Aya Soffer (CESDIS/UMBC) regarding the digital libraries project. I accompanied Dr. Yesha on a visit to the facilities of the Johns Hopkins Computer Science Department. In addition, I explored with Dr. Soffer a possible collaboration between CESDIS and Rutgers University on the Regional Validation Center project. I also met with Dr. Susan Hoban (CESDIS/UMBC) to discuss increasing the visibility of the digital libraries program through the Advances in Digital Libraries '97 conference.

- I met with Jim Fischer (Code 930) to discuss CESDIS's role in the HPCC program. I also met with Karen Moe (Code 522) to discuss CESDIS's role in the Mission to Planet Earth (MTPE) Program. I attended Goddard reorganization meetings and studied the new organization at GSFC as well as the strategic direction of NASA. I identified CESDIS in-house expertise as well as outside expertise that complement that of CESDIS which will enable us to build a world class team that is needed to meet GSFC's new environment and NASA's long term strategic direction. At this point it seems that MTPE and Regional Validation Centers have strong potentials and CESDIS should increase its future i nvolvement in these areas. By their very nature, problems related to these areas are complex and require a team that is multidisciplinary in nature. For CESDIS to be able to effectively compete for the new contract, it is critical to build on its current strength. This includes the work by Yelena Yesha, Jacqueline Le Moigne, Phil Merkey, Don Becker, and Rick Lyon.

- In an attempt to explore opportunities within GSFC, I held meetings with the SEWP team to identify possible CESDIS involvement including the development of innovative algorithms and software. Furthermore, Dr. Yesha and I met with Douglas Norton of NASA HQ (Director, Program Integration

Division) to explore creating a pilot project for on-line NASA awards. I also began and continue to work on the following initiatives.

1. Maximum Entropy and Maximum Likelihood Restoration of Atmospherically Degraded Imagery - Rick Lyon (CESDIS/UMBC) and N. S. Kopeika (Ben-Gurion University, Israel)

   Both Mr. Lyon and Dr. Kopeika have much to offer on this rich and exciting problem. A collaborative effort between them would be fruitful for both and potentially beneficial to NASA and the general scientific community for optimal information extraction from both satellite-based Earth sensing systems and ground-based remote sensing systems. Dr. Kopeika brings to the effort an understanding of atmospheric contributions and a model of the atmospheric blurring process; Mr. Lyon brings the contributions due to the telescope, detector, and noise statistics as well as a suite of high fidelity maximum entropy and maximum likelihood image restoration algorithms developed, coded, and implemented on a MasPar MP2. Mr. Lyon has a number of publications in both sensor modeling and algorithm development.

2. Possible CESDIS involvement in the NASA/USRA SOFIA Project - Joe Bredekamp, NASA HQ Office of Space Science, Research Program Management Division

   The Stratospheric Observatory for Infrared Astronomy (SOFIA) will be a 2.5 meter, optical/infrared/ submillimeter telescope mounted in a Boeing 747, to be used for advanced astronomical observations performed at stratospheric altitudes. More than 160 science flights per year are planned with data collected 60 times faster than by the preceding flying observatory, the Kuiper Airborne Observatory. I believe that CESDIS involvement in managing the data from the SOFIA project is a good idea. SOFIA is expected to generate massive amounts of data. CESDIS's ability to handle massive amounts of data and its position at Goddard make it a strong candidate.

3. Held several meetings and discussions with Karen Moe (Goddard Mission Operations and Data Systems Directorate, Data Systems Technology Division, Software and Automation Systems Branch) about possible collaboration with CESDIS in FY98. Potential projects include:

   a. Video conferencing: Dr. Yair Amir's work is a natural fit.

   b. Risk management: Dr. Al Aho (Columbia University) was a CESDIS visitor in the summer of 1996. His research interests and practical experience in software engineering for large systems would be valuable for such a project.

   c. CORBA (Common Object Request Broker Architecture): CORBA is not only capable of handling platform heterogeneity of underlying information systems, but, more importantly, the heterogeneity of applications used at different information systems. With CORBA-standard products, an application can communicate with another totally different application through a common interface such as the Interface Definition Language (IDL). Each of the application programs has to maintain an IDL interface (for client and server). A change in the server application program requires all the IDL interfaces (of each client as well as the server) to be updated (recompiled) accordingly. The work at Rutgers related to integrating heterogeneous and autonomous information sources would be of relevance here.

- I held regularly scheduled (every 2-3 weeks) technical staff meetings and individual meetings with the technical staff. The following is a summary of some of these meetings.

1. Dan Ridge made a brief presentation on Linux which generated a very lively discussion with a number of interesting questions. One of the questions raised was whether we could explore ways for combining the current CESDIS efforts in the ACTS, Linux, and Beowulf projects. Dan, Tarek El-Ghazawi, and Burt Edelson will explore further.

2. Phil Merkey made a presentation on Beowulf.

3. Yair Amir presented his work on videoconferencing and multicasting.

4. Kostas Kalpakis (CESDIS/UMBC) discussed some of the technical challenges facing the GLIN team. Commonalities between some of the GLIN-related search concepts and some of Nathan Netanyahu's (CESDIS/University of Maryland College Park) work in the context of image registration was discussed.

- I served as the General Chair of the Forum on Research and Technology Advances in Digital Libraries (ADL '97) sponsored by the IEEE Computer Society that was held May 7-9, 1997 at the Library of Congress in Washington, D. C.

1. Ihelped secure the sponsorship of the IEEE Computer Society, the National Library of Medicine, and the Library of Congress in addition to CESDIS and NASA Goddard Space Flight Center.

2. I helped secure the participation of Robert Price (Director of the Mission to Planet Earth Program Office at Goddard), Bruce Lehman (Assistant Secretary of Commerce and Commissioner of Patents and Trademarks), as well as Larry Irving (Assistant Secretary of Commerce for Communications and Information).

3. Karen Moe chaired a panel on "Very Large Digital Libraries" and John Dalton (Deputy Associate Director, Goddard Mission Operations and Data Systems Directorate, Earth Science Data and Information System Project) was one of the participants of the panel.

- Publication and Presentations

1. Published/submitted the following papers which include acknowledgment of CESDIS support:

   Adam, N., Awerbuch, B., Slonim, J., Wegner, P., & Yesha, Y. (1977). Globalization and the future of the Internet. *Communications of the ACM*. February.

   Adam, N., Atluri, V., & McKeown, K. (1997). Clinical information systems: making use of research in digital libraries. NSF Workshop on Research and Development Opportunities in Federal Information Services. May.

   Adam, N., Atluri, V., & Huang, W. Modeling and analysis of workflows using Petri Nets. Submitted April 1997.

   Adam, N., Adiwijaya, I., Atluri, V., & Yesha, Y. EDI Through A Distributed Information Systems Approach.

2. Published the following book/proceedings which include acknowledgment of CESDIS support.

   Adam, N., & Gangopadhyay, A. (1997). *Database Issues in Geographic Information Systems.* Kluwer Academic Publishers. June.

   Adam, N., & Aho, A. (Eds). (1997). *Proceedings of the IEEE International Forum on Research and Technology Advances in Digital Libraries (ADL'97).* Los Alamitos, CA: IEEE Computer Science Press. May.

3. Made several presentations on CESDIS-supported work including the following:

   IEEE Computer Society Board meeting, May 1997.
   NSF Workshop on Research and Development Opportunities in Federal Information Services, May 1997.
   Department of Computer Science, SUNY at Buffalo, April 1997.
   GCDIS, March 1997.

- Other

1. Next Generation Information Technologies and Systems '97 conference: Co-organized and co-chaired a panel on digital libraries with Dr. Yesha as part of the NGITS'97 conference that was held June 30-July 3, 1997 in Neve Ilan, Israel. Panel participants included researchers from Bell Laboratories, Carnegie Mellon University, University of Maryland College Park, and George Mason University.

2. Chairing the IEEE Computer Society Task Force on Digital Libraries which is a collaboration among academia, industry, and NASA. Members of the task force include Dr. Milton Halem (Code 930), Dr. Harold Stone (NEC Research Institute), and Dr. Yesha.

3. Lead a CESDIS-supported research project that resulted in a Ph.D. dissertation in the area of digital libraries, May 1997.

4. Served as Editor-in-Chief with Dr. Yesha for the *International Journal of Digital Libraries.* The first issue was published in May 1997.

Ian F. Akyildiz

Georgia Institute of Technology,
Broadband and Wireless Networking Laboratory
(ian@ee.gatech.edu)
http://www.ee.gatech.edu/users/ian

Statement of Work

Dr. Akyildiz was tasked with advising the CESDIS Director on any or all of three proposed areas: 1) an efficient traffic/congestion control mechanism in ATM over satellite enviroment, 2) A LANs/MANs interconnection architecture using ATM over satellite, or 3) mobility management for multi-tier personal communication systems.

Results

1. Introduction

During my very productive stay at CESDIS during the summer of 1996, I prepared survey reports, wrote proposals, participated in several meetings, seminars, and group research activities. The detailed activities are summarized below:

2. High Performance Networking

The purpose of this report [1] was to point out the state-of-the-art problems and short-term and long-term research needs. The areas covered are ATM networks, satellite ATM networks, wireless networks, mobile computing. This report should serve NASA CESDIS a basic framework to start a new research direction.

3. Wireless Networks

I also led the effort to develop a research proposal [2]. The proposal had the objective to research highly mobile wireless multimedia architectures for the needs of the digital battlefield with capability for intelligent survivability and adaptive connectivity in a hostile environment. The project will contribute to the foundation of intelligent highly mobile wireless multimedia network design, in particular, network architecture, intelligent distributed database management, routing, multicasting, location management, channel allocation, information security, digital information storage, and power control.

4. Satellite Networks

I investigated satellite networks in two directions: satellite ATM networks and mobility management.

4.1. Satellite ATM Networks: A Survey

The survey in [3] points out the key issues for interconnecting satellite and ATM networks. ATM technology is useful in multiplexing various traffic types such as data, voice, video and still images. There are several open issues which need to be overcome in order to achieve a seamless integration of ATM and satellite networks. First the requirements for the interconnection are described in [3], followed by a discussion of recent research issues, challenges, and possible solutions. Finally, an overview of current projects about ATM over satellite is given and future directions are discussed.

4.2. Handover Management over LEO Satellite Networks

Low Earth Orbit (LEO) satellite systems have been proposed in recent years to provide global coverage to a more diverse user population. In contrast to geostationary (GEO) satellites, LEO satellites circulate the Earth at a constant speed. Because of this non-stationary characteristic, the coverage area of a LEO satellite changes continuously. The serving satellite for a particular connection may change over time resulting in a handover. Thus, LEO satellite networks require a reliable handover protocol that is critical for connections with multihop intersatellite links (ISLs). I introduced the Footprint Handover Re-route Protocol (FHRP) in [4] that maintains the optimality of the initial connection route without performing the routing algorithm after satellite handovers. Furthermore, the FHRP handles the inter-orbit handover problem which has been neglected in the existing literature.

References:

[1] Akyildiz, I. F. (September 1997), A Handover Management Protocol for LEO Satellite Networks, Technical Report, August 1996. Also to appear in ACM/IEEE Mobicom'97 Conference.

[2] Akyildiz, I. F., & Jeong, S. H. (July 1997), Satellite ATM Networks: A Survey, IEEEE *Communications.*

[3] Akyildiz, I. F., Gelenbe, E., Singhal, M., Wiederhold, G., Wolfson, O., & Yesha, Y. (September 1996). Intelligent Agents for Adaptable Wireless Networks. (Technical Report)

[4] Akyildiz, I. A., El-Ghazawi, T., & Yesha, Y. (August 1996). Research Challenges in High Performance Networking Area. (Technical Report)

A CESDIS – University Collaboration on Data Mapping and Matching: Languages for Scientific Datasets

Serge Abiteboul, Stanford University, Department of Computer Science (abitebou@db.stanford.edu)

Peter Buneman, University of Pennsylvania, Department of Computer and Information Science (peter@cis.upenn.edu)

David Maier, Oregon Graduate Institute, Department of Computer Science and Engineering (maier@cse.ogi.edu)

Stan Zdonik, Brown University, Department of Computer Science (sbz@cs.brown.edu)

Statement of Work

These four professors proposed a one-year collaboration between CESDIS and a group of computer scientists specializing in the area of database languages and systems. Goals included:

- New interactions between computer scientists, as developers of a next generation of database technology, and NASA scientists, as primary users whose applications should both specify the state-of-the-art and test its limits.
- Development of a study that would do the following:
  1. Identify the technical challenges of "mapping and matching scientific datasets",
  2. Evaluate the extent to which these problems can be addressed by current database technology and what new technological capabilities are required, and
  3. Identify promising approaches to address the new challenges.

Results

1. Introduction

This is a report of a working group[1], initiated by Paris Kannelakis, on the topic of data mapping and

[1] The work of this group was supported by NASA CESDIS.

matching. The starting point for the group's discussions was the observation that most of the world's data is not in databases, but exists in the form of structured text and hyper-text (e.g., the World Wide Web) and in a wide variety of formats designed for the interchange and archiving of data (e.g., scientific data). If databases are to continue to have the success that they have enjoyed over the past twenty-five years and provide services such as secondary storage management and concurrency control to a larger and larger community of users, they must accommodate these "non-database" sources of data. The group believes there are two major challenges here.

- Finding the right data model and languages for exchanging and restructuring these new forms of data. This is the "mapping" part of the title.

- Combining the well-developed querying and optimization techniques for databases with the equally well-developed pattern matching techniques for these new forms of data. This is the "matching" part of the title.

In this report we will elaborate on these two questions and provide a series of challenges and research problems that need to be solved in order to extend traditional database technology to the needs of the wider community of technical users, especially scientific users. The overall purpose of this report is to raise issues that could form the basis for a research program in data mapping and matching for scientific data. The authors believe that this is a very rich and promising field and that such a research effort would go a long way toward making scientific data more useful and accessible.

2. Data Models

2.1 Background

The topic of data models is almost as old as that of databases, and the number of proposed models is very large. A data model is typically captured in two distinct linguistic parts – the data definition language (DDL) and the data manipulation language (DML). The DDL allows us to express typing information as well as a limited set of constraints, and the DML allows users and programs to interact and manipulate data that has been described in terms of the DDL.

Query languages (DML) for the relational data model, and in particular SQL, addressed the problem of data matching for tabular data. These query languages were also the basis for data mapping within the limited world of relations. With the query language, it is possible to define declarative *views* that map a set of relations to a different set of relations. Thus, it is possible to restructure data as long as the desired mapping can be expressed with the primitives of the query language.

In some sense, the mapping and matching of data provided by relational systems can be thought of as a model of what we would like to achieve within the context of more complex data types. The extent to which this is realizable is an open research problem and forms the basis for the research program advocated in this report.

Often, advanced data models (e.g., ER, SDM) were used as "conceptual models" primarily targeted at the database design phase. For this purpose the variety and the lack of precision of these models was not really a problem because they were only used as conceptual aids in the design of, say, a relational database. Once the relational schema had been formulated, the model could be discarded. There was relatively little need to formalize – or compute with – these conceptual models.

The type systems of object-oriented databases have brought us closer to an agreed upon, formal data model for complex data, but the need to cope with new data sources places new demands on the model. While the data abstraction capability of object models is quite flexible and allows us to construct arbitrary behaviors for complex new types, it does not fully address the problems of constructing an efficient query facility for those types. Such a query facility requires that we have a formal model over which a query processor (optimizer) can reason.

In what follows, we give a more detailed account of the need for a formal model for complex data? Solutions to these problems have immediate practical payoff to the area of scientific data management.

- The major problem in data mapping is transforming data from one format to another. For *N* data formats it is unrealistic to build and maintain N^2 translators, however translating into, and out of, a common format makes this task feasible. Moreover, many sources have an incomplete specification, and the data model can work as a type system for such sources – an essential part of the software needed to query them.

- In many scientific domains, the task of understanding another scientist's data is a major undertaking. It is unrealistic to expect computer scientists to learn enough about the scientific domain to be of much help in this process. A simple and precise model is needed as a medium of communication in which ideas about data can be exchanged and through which queries can be formulated.

2.2 Mediating Data Models

A common data model can, then, be viewed as the *lingua franca* that binds together data from multiple sources. We will refer to such a model as a *mediating data model.* It guides both the design and the implementation of data exchange mechanisms, as well as the integration of data from different sources. At present the relational data model works well for mediating data that can be easily viewed relationally. For example, the EDA/SQL product provides an environment where the mediating data model is the relational model. It currently supports mapping among dozens of databases, file formats, and desktop tool formats. It can make use of existing data manipulation facilities when they exist, but has its own internal capabilities as well. However, the tabular nature of its mediating model is not general enough for most scientific data, which exhibits features such as nested values and multidimensional arrays.

Object-oriented data models are starting to emerge as a new standard. The ODMG standard provides a starting point, but there are other closely related models that have been used for various aspects of data mapping, PDES/EXPRESS is a standard for engineering design data exchange, OPM for representing biological experiments, and there are many more proposed exchange formats and a number of object-oriented design tools that are used for database design.

Here is a list of questions that must be addressed in the development of a mediating data model that would serve the role of the relational model for complex data:

1. **Types and constraints** What types should be in the model? Recent work has indicated that models, like the ODMG, based on the object paradigm with a free combination of tuples, and collection types (i.e., lists, arrays bags and sets) can serve as an interface to a variety of structured data sources. Should there be direct support for other types such as variants?

   Also, one may argue that data models are not simply type systems. For example, they might also include constraints on data. Constraints can become arbitrarily complex, but simple constraints such as inclusions among extents and existence of inverse relationships are ubiquitous. Also new types may sometimes be expressed through existing types and constraints. For example, variant types can, unsatisfactorily, be expressed via inheritance and constraints on extents. We must answer the question of exactly what constraint types are useful for scientific data and which of these can be practically handled by a data mapping and matching engine.

2. **Optimization** It makes sense to use the high level concept of a mediating data model only if appropriate optimization of data access and translation is supported. A question is what optimizations should be supported by the software for this model. For example:

    - What optimization techniques hold for the language primitives of the mediating model?
    - In translating between formats *A* and *B*, having specified both *A* and *B* in the data model, do we literally need to convert into, and out of, the common data model or can we "optimize out" some or all of this translation?
    - How do we exploit constraints in optimizing queries or data translation?
    - How do we "offload" optimizations in the mediating software to optimizations that can be performed by the data sources (e.g., when they are SQL servers)?

3. **Change control and transactions** By the very nature of these applications, we have to deal with multiple representations of the same data in different formats. Standard techniques for transaction management and concurrency control have to be revisited since the applications will be running on autonomous systems and, furthermore, since the atomic pieces of information may be different in the various representations. The issues here are:

    (a) Should the data exchange protocol handle notions such as transaction, locking?
    (b) Should it know about versions, incremental updates? What else for controlling change?
    (c) Many data sources are *archival*. Once placed in the database, data items are never forgotten. How does this impact on the data model and on version control?

2.3 Semi-structured and Self-describing Data

In the past two years a variety of proposals have emerged for databases built on dynamic type systems. In such a system, each value carries a description of its structure. There is no schema or static type system that is required for the compilation of query languages. This allows, among other things, for data to be highly irregular. Sets, for example, do not have to be built out of similarly typed elements. The study of languages and other tools for such data is in its infancy, but a number of important questions are beginning to emerge.

1. In converting from structured to semi-structured data, there are usually a number of equally satisfactory representations. How is one chosen? Also, it seems useful to try to carry some of the intended semantics of the original structure, but there are, as yet, no well-developed techniques for doing this.

2. Converting in the opposite direction (type discovery) is difficult and because of the data irregularities may be ambiguous or require searching for approximating types.

3. The previous two questions are important because they ask whether semi-structured data is a good idea for data mapping and integration. In some sense, it is trivial to integrate heterogeneous source after some translation to a mediating semi-structured data model. However, a lot of the semantics have been lost in the translation phase which may lead to an integration of poor quality. This seems to suggest investigating hybrid approaches and mediating models where structured and semi-structured data can coexist.

4. Because the distinction between schema and data is lost, semi-structured data should be a good model for browsing. But how does one build good browsing tools for data when the sheer size of the data will interfere with the discovery of the structure? Can ad-hoc (non browsing) queries be specified and evaluated efficiently on semi-structured data?

5. Should we convert the data physically to semi-structured format (i.e., materialize the result of the transformation) in a data warehousing style? Or is it possible to consider the translation purely virtual and use rewriting techniques as done for mediating relational databases?

3. Data Matching

The basic matching primitives found in database query languages are based primarily on the equality on atomic types and on matching *relational patterns*, i.e., sets of tuples possibly containing variables. This has to be extended to provide a richer set of patterns for more complex (graph) data and include matching techniques to be found in the underlying domains. For example, we might want to augment our query language with primitives for classifying time series, image matching, or some simple natural language recognition. In a trivial sense, such extensions are already allowed by object-oriented databases and by object-relational systems which allow arbitrary abstract data types (ADTs) to be added. However the ADT approach has serious shortcomings:

- Some ADTs may contain structures that are similar, say, to relations; and one would like to adapt the query language primitives to such ADTs. Exploiting the stored structures directly in the query language can often lead to improved performance or ease of expression.

- Database query languages have a very rich equational theory that provides essential optimization techniques. ADTs often have equally rich equational theories (array algebras, for example). The conventional ADT approach provides no support for including these equational theories into a query optimizer or for combining multiple theories in a simple way.

- In many cases, a boolean approach (i.e., truth values) to filtering data is not appropriate for the type of data that is encountered in scientific applications. In these cases, approximate matching is required; e.g., a picture may be more or less recognized as that of a plane.

3.1 Combining databases with other domains

Scientific databases (as in many other disciplines) require that we be able to efficiently manage and search collections of complex and widely-varying data. Currently, a separate technology exists for each category of data. The information retrieval (IR) community has techniques for doing keyword in context searching in large textual databases. The pattern matching community supports retrieval of images. Others have investigated retrieval of data like time-series or audio.

Unfortunately, each of these technologies is currently quite separate. There is no easy way to form a query that "crosses" data domains. Moreover, even if the query could be posed, there is little or no support for query optimization techniques that span domains. In large part, we are looking for a more general theory of data retrieval – an umbrella under which any type of information request can be posed. Moreover, we would like such a theory to allow us to compose any of these techniques within a single paradigm.

The general problem of combining new domains with database organization and languages leads to a large number of practical problems. Below, we briefly outline a representative sample of such problems.

- *Query Languages*

  - Does a clean combination of database query languages and information retrieval languages exist that can form a suitable basis for expressing search patterns for scientific data? On the one hand, database query languages are quite precise in what they specify. On the other hand, information retrieval languages are an imprecise characterization of what the user is looking for.

This leads to notions of *precision* and *recall.* What new features are needed in a formal query language to deal with scientific data (e.g., to deal with approximation or imprecision in retrieval patterns)?

- How should structure be extracted from complex data items (e.g., maps, time series) to make pattern matching tractable? As an example, we might use line segments to approximate the rising and falling of a waveform (i.e., time series). In this example, the point that we choose to start the line fitting operation in both the data and the pattern specification (i.e., query) can make a big difference in the slope of the line and, therefore, in the accuracy of the match. Having a robust way to find these pattern components is a very important issue.

- Can querying of scientific and multimedia data types be based on more general bulk types? For example, can a time-series be modeled as a list; can an image be modeled as a 2-dimensional array; can a document be modeled as a tree? If so, the equational theories for these more general types can be exploited.

- Data is retrieved on the basis of patterns. In the relational setting a pattern is a simple boolean predicate. Can we develop a set of *requirements* for patterns in other domains? For example, it seems reasonable to assume that patterns could be composed (and, therefore, be decomposable). Current query optimizers rely heavily on this characteristic of simple patterns (i.e., predicates). By analogy, it would be desirable to explore how patterns can compose in order to better understand the optimization process.

- *Query processing and optimization*

  - To what extent can standard database and IR access methods (spatial indexes, key word indexes) be adapted to more general pattern matching searches?

  - Can standard database query facilities be used to pre-filter data before more computationally intensive pattern matching takes place? In other words, can an optimizer be "smart" about how it chooses to order the evaluation of the parts of a pattern?

  - How would the architecture of a query optimizer be affected by the introduction of more complex and less structured data? For example, would our standard notions of logical and physical level optimization hold up?

  - What cost models are appropriate for deciding on alternative access strategies? Scientific data poses new challenges in this context. For example, if the query requires a large matrix operation, performance will heavily depend on how rows/columns of the matrix are accessed on disk.
  - In the scientific environment, the data is often captured by a remote sensing device. This could be an experimental probe in the laboratory, a patient monitoring device in a hospital, or a data gathering device in the space shuttle. Like a more traditional database, these devices are capable of reporting various kinds of data, some of which is relevant to a query that is being processed. Imagine a query that wanted to know how many patients currently show signs of cardiac arhythmia. The query processor would have to poll various devices as a part of the query evaluation process. We would like to make this as efficient and seamless as possible. In other words, we would like to treat these instruments as part of the database. Research is needed to understand how to do this.

3.2 Data mining

Closely related to this discussion of data matching is the issue of *data mining.* The most advertised aspect of data mining is the search for unanticipated patterns in existing data. Most attempts to do this have been directed at finding simple correlations or implications in tabular data. Mining data with more complex structure has been little attempted, nor have many attempts been made to combine tabular data mining with other pattern recognition techniques.

First, it should be recognized that in real applications, the main effort to mine data often goes into massaging (transforming and cleaning) data into a form that can be mined, which is a major concern in our framework of data matching and mapping.

Also, the issue of understanding the structure of data sources is essential in data mining. Often, data found in scientific applications is little documented and its structure is unknown. This is, for instance, the case if the data is the output of programs that perform complex operations of a large set of raw data (e.g., satellite images), if it is in a format that is not known by the system, or if it consists of text with weak formatting information (e.g., html pages). A data mining task has to be performed to understand the implicit structure of the data, which relates to the issue of type discovery that was previously mentioned.

For these reasons, much of the work we have advocated under both data mapping and matching is an essential prerequisite for any successful data mining project. However, the data mining theme adds the challenge of figuring out what to look for.

Burt Edelson

George Washington University,
Department of Electrical Engineering and Computer Science
(edelson@seas.gwu.edu)

Statement of Work

Dr. Edelson worked with the CESDIS Director on digital libraries issues and on developing new initiatives using high data rate communication satellites.

Results

My work over the past year involved the following activities:

- Developed plan to get GSFC and CESDIS involved in several experiments in the G-7 Global Interoper ability for Broadband Networks (GIBN) project. Worked with Pat Gary (930) to develop plans for experiment to connect the U. S. Library of Congress in Washington, DC with the Japanese National Library in Tokyo. Will try to include the Smithsonian and the National Library of Medicine in the project.
- Worked on the Global Legal Information Network (GLIN) project with the Law Library of Congress. Met frequently with Dr. Rubens Medina, Law Librarian, and other Library of Congress representatives. Participated in GLIN Project Directors Meeting. Outlined and supervised preparation of GLIN

Development Plan to be coordinated by Susan Hoban (UMBC/CESDIS). Arranged for GSFC to develop a "mirror site" for the GLIN central server to be developed by Kostas Kalpakis (UMBC/CESDIS). Arranged for meetings with U. S. Department of State to set up international GLIN organization.

- Worked on plans for the U.S.-Israel Information Technology Development Program. Organized a steering committee meeting on October 10 with Lee Bailey of the U. S. Department of Commerce, Prof. Danny Dolev of Hebrew University, Dr. Milt Halem of Code 930, Yelena Yesha, and others. As a result, the proposed Implementation Plan was rewritten for submission to the U. S.-Israel Commission. Ultimately it was not approved for funding. Contacted IT Task Force members and Mary Good, Under Secretary of Commerce, in an attempt to keep some parts of the program alive.
- Worked with Al Aho (Columbia University), Nabil Adam (Rutgers University), and Susan Hoban (UMBC/CESDIS) on plans for the Advances in Digital Libraries 97 conference to be held at the Library of Congress in May 1997. Served as session chair. Gave introduction for keynote speaker Peter House of the Smithsonian.

Miron Livny

University of Wisconsin, Department of Computer Science (miron@cs.wisc.edu)

Statement of Work

Dr. Livny was tasked with working with the CESDIS Director and the HPCC ESS Project Scientist (Dr. George Lake, University of Washington) to identify ways in which to transfer high throughput computing technology developed by Dr. Livny's group at the University of Wisconsin to NASA and develop approaches for demonstrating its potential using existing processing resources.

Results

For many experimental scientists, scientific progress and quality of research are strongly linked to computing throughput. In other words, most scientists are concerned with how many floating point operations per week or per month they can extract from their computing environment rather than the number of such operations the environment can provide them per second or per minute. Floating point operations per second (FLOPS) has been the yardstick used by most high performance computing (HPC) efforts to evaluate their systems. Little attention has been devoted by the computing community to environments that can deliver large amounts of processing capacity over long periods of time. We refer to such environments as high throughput computing environments.

The key to HTC is effective management and exploitation of all available computing resources. Since the computing needs of most scientists can be satisfied these days by commodity CPUs and memory, high efficiency is not playing a major role in an HTC environment. The main challenge a typical HTC environment faces is how to maximize the amount of resources accessible to its customers. Distributed ownership of computing resources is the major obstacle such an environment has to overcome in order to expand the pool of resources it can draw from. Recent trends in the cost/performance ratio of computer hardware have placed the control (ownership) over powerful computing resources in the hands of individuals and small groups. These distributed owners will be willing to include their resources in an HTC environment only after they are convinced that their needs will be addressed and their rights protected.

We believe that HTC can improve the productivity of a wide spectrum of NASA scientists and would like to demonstrate it. The object of the proposed effort is to identify ways to transfer HTC technology developed by our group to NASA and to develop approaches to demonstrate its potential using existing processing resources.

Daniel Menascé

George Mason University, Department of Computer Science
(menasce@cs.gmu.edu)

Statement of Work

Dr. Menascé was tasked with developing analytic models for hierarchical mass storage systems and designing tools that implement these models.

Results

During the period July 1, 1996 to June 30, 1997 my activities focused on the general topic of performance models of mass storage systems. The following papers resulted from this activity [copies available through CESDIS administrative office.]:

Analytical performance modeling of hierarchical mass storage systems, O. I. Pentakalos, D. A. Menascé, M. Halem, and Y. Yesha, to appear in the IEEE *Transactions on Computers.*

Pythia and Pythia/WK: tools for the performance analysis of mass storage systems, O. I. Pentakalos, D. A. Menascé, and Y. Yesha, to appear in *Software Practice and Experience.*

Pythia: A Performance Analyzer of Hierarchical Mass Storage Systems, O. I. Pentakalos, D. A. Menascé, and Y. Yesha, Performance Tools'97, 9th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation, St. Malo, France, June 2-6, 1997.

Automated Clustering Based Workload Characterization for Mass Storage Systems, O. I. Pentakalos and D. A. Menascé, Fifth NASA Goddard Space Flight Center Conference on Mass Storage Systems and Technologies, College Park, MD, September 17-19, 1996.

Another paper available through CESDIS was submitted for the *IEEE Transactions on Parallel and Distributed Systems*:

Analytic modeling of distributed hierarchical mass storage systems with network-attached storage devices, O. Pentakalos, D. A. Menascé, and Y. Yesha, submitted to the IEEE *Transactions on Parallel and Distributed Systems.*

The paper went through the first round of reviews and we were asked to do some modifications before the paper is accepted for publication. We are currently working on these modifications.

Besides the activities reported in the papers mentioned above, I have been working on the general issue of how to deal with convergence problems in Queuing Networks (QNs) with load dependent devices. Load dependent devices are very useful to model multiple servers, such as multiple robots in an automated tape library. Using load dependent devices works well for QN models with one class of customers only since the algorithm is non-iterative and does not lead to convergence problems. The multiple class case has to be solved through approximate iterative algorithms that fail to converge sometimes and exhibit an oscillatory behavior.

To overcome this problem, I have been investigating the accuracy of an approximation proposed by Seidmann, Schweitzer, and Shalev-Oren (in *Large Scale Systems Journal*, North Holland, Vol 12, 1987, pp. 91-107) in which a device with *J* identical servers with service demand equal to *D* is replaced by a sequence of two devices: a load independent (LI) queuing server and a delay server (see figure 1). The service demand for the LI device is D/J and the service demand for the delay device is (*J*-1) x *D / J*. The approximations give very accurate results at light and heavy loads. At intermediate loads, the maximum relative error observed in the response time 11% for the stress case when the multiple server device is a bottleneck with respect to others by a factor of 10.

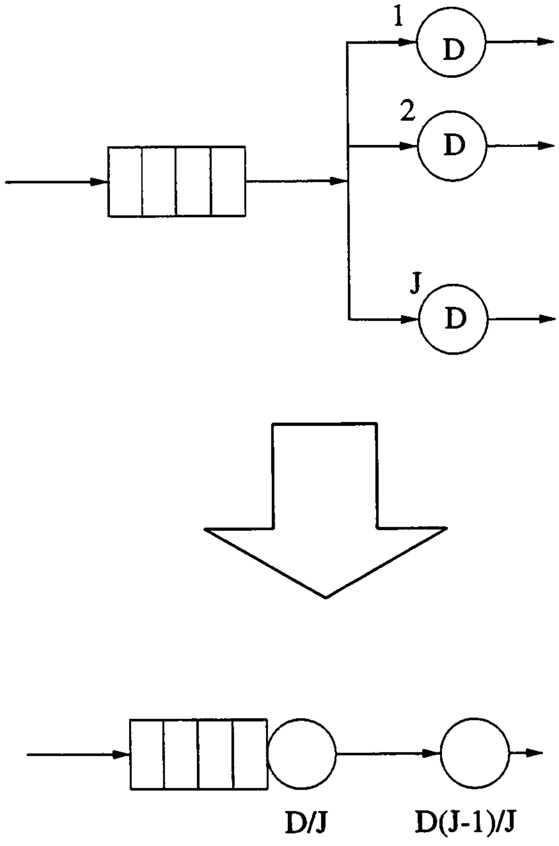


Figure 1: Approximation for Multiple Servers in QNs.

We will be investigating how this approximation applies to the models of mass storage systems that we have been developing.

Mukesh Singhal

Ohio State University, Department of Computer and Information Science
(singhal@cis.ohio-state.edu)

Statement of Work

Dr. Singhal was tasked with advising the CESDIS Director on new initiatives in analyzing and evaluating massive video data and images.

Results

First, major activity of the consultant focused on a study of the architecture of NASA mass storage server so that it can be mathematically modeled and its scalability properties can be studied. EOSDIS at NASA GSFC has been developing a massive data storage and would like to determine its scalability. The consultant made several trips to CESDIS and participated in discussions with the researchers at CESDIS, GSFC, and George Mason University to understand the operations and architecture of the mass data archive. The consultant studied important operations in the mass storage server at EOSDIS at NASA so that it can be mathematically modeled and its scalability properties can be studied. The consultant studied several documents, reports, and published articles that describe the operations and the architecture of the mass data archive. The consultant developed an understanding of the "Ingest Data" and the "Retrieval and Processing" operations for a performance characterization and analytic model of the mass storage. Special emphasis was placed on the completeness and correctness of the understanding of the operations, hardware mapping, and assumptions that will be made during the modeling of the system. In a meeting with NASA personnel on March 24, the operation was fine tuned for a realistic performance characterization and analytic modeling of the mass storage server.

Second, with regard to the broader goals on research on databases as it applies to digital libraries, the consultant investigated fast parallel/distributed techniques to retrieve, analyze, store, and display data. The consultant searched existing literature for organizing data for efficient retrieval for fast query processing. The consultant investigated digital signature based techniques for developing fast parallel/distributed techniques to retrieve and display of data. He investigated an approach wherein a meta database layer is used to locate the data of interest. This helps considerably narrow down the data search in the main database, expediting the data retrieval substantially.

Jacob Slonim

University of Toronto (Ontario), Computer Science Department
(jslonim@cs.toronto.edu)

Statement of Work

Dr. Slonim was tasked with advising the CESDIS Director on the state of the industry in information technology, specifically network-based applications such as digital libraries, electronic commerce, and tele-education.

Results

The following material is taken from a document prepared for CESDIS by Dr. Slonim entitled *Project Managment--Industrial Perspective: Focusing on Industrial Software Engineering; Best Practice for Developing High Quality Software Products.* The entire document is available as technical report TR-97-188.

Abstract

The emphasis of this report is to explore ways in which research organizations can help to improve the process of managing developing software, whether for their own use or for the use of industrial colleagues. The concerns that most research organizations are raising involve ways in which industrial partners can take advantage of the innovative ideas that the research community is developing and convert them to economic benefits. It seems that there is a great complement between the research organization and industrial practitioners, where the researchers are extremely powerful in creating new ideas, the practitioners in the industrial setting are strong in the execution.

This report concentrates on trying to bridge the communication gap by highlighting areas the industrial practitioners are concerned with. The report puts emphasis on industry practices and the importance of software engineering, and stresses the importance of the disciplinary part of the software management process. The report covers topics such as project management, business plans, software risks, resource estimations, configuration management tools, process life cycle models, reuse and object-oriented technology, software requirement/specification engineering, software testing, documentation, software measurements and metrics, and software engineering ethics.

In this report, we will not concentrate on either the design or the programming because we feel that these areas are well understood by the research community. The report goes into great detail on each of the areas and the issues that are raised. We highlight the issues, and where we can identify directions which we recognize as workable solutions, we discuss them. Otherwise it is left for the researcher to acquire more information from his partners in industry. There are no bulletproof solutions and there is no magic in developing software. This report attempts to share some of the author's experience in the hope that it will help researchers to collaborate with industry colleagues on understanding the pressure cooker development environment, or as many of my colleagues in industry like to call it, real world issues and problems.

Executive Summary

Software engineering is becoming a major challenge to industry practitioners and academic researchers. The software industry is seeing a paradigm shift from being a host computing environment to what is now called network centric. The major change in the next five years in the software industry will be the development of systems and applications that are transparent to the end user which will be based on a combination of software that is reused from many different sources. The software will need to be scaled up to serve hundreds of millions of end users, exhibit high performance and quality, and be fully integrated.

In this report we examine areas in product development that are considered to be weak in research organizations.

In the past few years, the software industry has found itself fighting to introduce new products in a relatively short timeframe, i.e., nine to twelve months from the time the idea was created to the time that it is introduced to the market, while at the same time increasing the quality of the products that are introduced. Being second or third to get to the market, in most cases, will lead to loss of market opportunity. It seems that there is a great complement between research organizations and industrial practitioners. The strength of the researcher is in creating new concepts while the strength of the practitioner is in the execution. The lack of communication between researchers and practitioners results in the loss of great ideas and, more importantly, some of the solutions that we are finding in products lack the creativity that we see in research prototypes. This report concentrates on trying to bridge this communication gap by introducing to researchers the areas that development organizations are concerned with, in the hope that if they are better understood by researchers, their contributions can be adapted in industrial settings. As a result, the collaborative work between the two communities will bring benefit to both groups.

The concern that industry has today with the down-sizing in R & D, especially in basic research efforts, is that it will put more pressure on researchers in universities to become the vehicle for generating new ideas. At the same time, universities themselves are suffering from reductions in support from government agencies and other funding organizations which could result in the opportunity for industry to get involved with researchers on a business need rather than for the sake of being good citizens.

This report concentrates on software engineering practices in industry and puts great emphasis on the disciplinary part of the software development process. In the report we will not cover either design or programming in the development life cycle. They are a very important part of the development, but because we feel that these areas are well understood by the research community, we will focus on areas that we feel are neglected by researchers, such as:

a) Project Management: The emphasis in this area is on the mandatory steps to take in order to be able to manage and control the outcome of the development of a product. In order for researchers to incorporate their prototypes into a product, they will need to understand the business plan of an organization. This plan includes product definition, market segment for which the product is targeted, and identification of the inhibitors that are foreseen by the development team.

b) Software Risks: Software risk metrics provide an indication of software risk viewed from several sources of information such as organization, estimation, monitoring, development methodology, tools, culture, usability, correctness, reliability, and personal preference. Unfortunately one of the major risks for which there is no clear solution has to do with software cost, time, and resources estimations. There is no clear answer to this problem, especially in an environment that requires constant change to the plan but is strongly motivated by the need for speed in reaching the market at the same time.

c) Software Configuration Management: This tool is the central focal point for the development of the product. In making sure that all different phases of the development are tightly controlled, consistencies and milestones are delivered on time. It would be wise for researchers to use configuration management to develop their prototypes since it will help them to understand how the development is managed and controlled.

d) Life Cycle Models: The experience that we have gained in the past 20 years has shown us that the waterfall model, which was adopted from hardware manufacturing, is too restricted. The newest process models that were introduced in the early 90's are more flexible in their approach to developing software. They are based on iterative processes such as the spiral, interactive enhancement, prototype, and water fountain, all of which exhibit more the dynamic of assessing risk during the life cycle process and being able to react to changes in requirements or other risks to the software development. The new life cycle models are more often taking into account the idea that software development processes are different from those used in hardware manufacturing processes.

e) Reuse and Object-oriented Technology: This area is gaining importance in developing software in the network centric environment. Contrary to the beliefs of many managers, the change to reuse and OO is one that requires the developers not only to understand the language in which they are writing (such as JAVA and C++), but the need to focus on requirements, specification, and design. In order to reuse a software component, it needs to be designed for reusability. Until now the promise that reuse and OO will increase people's productivity has not been utilized, in the author's opinion. The author believes that in the long run, reusability (especially with JAVA) will increase development productivity and at the same time increase software quality and the reliability of the product.

f) Software Requirement/Specification Engineering: If we do it correctly, this is the area that will remove many of the uncertainties and the risks that organizations are taking in developing new software. It is the author's opinion that if the development teams have a strong understanding of the requirement and a close relationship with their customers, it will lead to a clear, concise specification and low level design. The implementation using OO techniques will be straightforward.

g) Testing: Both industry and research are having great difficulties with testing. Most development organizations are spending great amounts of resources on software quality and reliability and have very little to show for it. In the author's opinion, the reason for the high cost is that most software is not designed for testability; testing is an afterthought process. Testing will become a greater and more urgent issue in the future because of the increase in complexity in building and testing network centric applications. We have little understanding of how to test the new applications, such as electronic commerce, which are developed by one vendor using many components from other vendors. There are no tools today in the market that will be able to debug that type of program. The challenge to the research community and software vendors is to create R & D activities to build new testing tools that will be able to execute the type of application that we will see in the future.

h) Software Measurements and Metrics: This area is made up of people measuring different software development objects ranging from products to processes. For example, products include source code components, software requirement/specification, software design, and software reliability. Examples of processes include the architectural design process, coding, unit testing, and the system test process. The Capability Maturity Model (CMM) describes the principles and practices underlying software process maturity and is intended to help organizations improve the maturity of their software processes through an evolution from ad hoc, chaotic to mature, and disciplined. Each key process is comprised of a set of key practices that indicate if the implementation and institutionalization of that area is effective, repeatable, and lasting. ISO 9001 is the standard most pertinent to software development and maintenance. Organizations use it when they must ensure that the supplier conforms to specific requirements during several stages of development including design, implementation, production, installation, and servicing.

i) Technical Communication (Documentation): We are graduating students who have brilliant minds but who have very few skills in communicating their ideas to customers or colleagues. In order for a development organization to take advantage of the creative work that researchers produce, they will insist that the appropriate technical documentation accommodate them. Industry appears to be hiring technical personnel not primarily for their technical expertise, but for their communication skills. To ensure that end user information is appropriate for its intended audience and that documentation is usable, we will need to improve education in technical communication.

j) Professional Ethics: Computer software developers are not software engineers, but the trend in the software industry is to move toward certifying their developers to be software engineers. Many software industry executives are complaining of the lack of what they consider ethical behavior by the software development organizations. We must be able to emphasize this area much more in courses taught at universities just as it is in other professions, such as engineering and medicine.

In this report, we will go into detail on each of these areas and try to highlight issues and identify any directions which the author recognizes as the best practice today. There is no bulletproof solution to successfully managing a project and there is no magic. Researchers should constantly pay attention to their industrial partners and try to understand the ramifications of their own collaborative work with them.

Peter Wegner

Brown University, Department of Computer Science
(pw@cs.brown.edu)

Statement of Work

Dr. Wegner was tasked with advising the CESDIS Director on interactive models of computing.

Results

Research has included work with Nabil Adam and Yelena Yesha on globalization with special emphasis on education and electronic commerce, resulting in [1]. This work also included contributions to the working group report on Strategic Directions in Electronic Commerce and Digital Libraries for the Computing Surveys 50th anniversary symposium [2].

During this year considerable progress was made on research on interactive models as a foundation for software technology, agent-oriented systems, and data mining. Publications in this area include a paper on interactive software engineering with a case study on the Earth observation system [3], as well as a paper on why interaction is more powerful than algorithms [4], and a paper on the interactive foundations of computing [5]. This research makes the ambitious claim that Turing machines are not the most powerful model of computing, and applies the new interactive modeling paradigm to problems of interoperability, coordination, frameworks, patterns, and data mining.

[1] Globalizing business, education, culture through the internet, *CACM*, February 1997.

[2] Special Issue on strategic directions in computing research, *Computing Surveys*, December 1996.

[3] Interactive software engineering, *Handbook of Computer Science and Engineering*, CRC Press, 1997.

[4] Why interaction is more powerful than algorithms, *CACM*, May 1997.

[5] Foundations of interactive computing, will appear in *Theoretical Computer Science*.

COMPUTATIONAL SCIENCES BRANCH

Remote Sensing Group

Jacqueline Le Moigne, Senior Scientist – Branch Head
Richard Lyon, University of Maryland Baltimore County
Nathan Netanyahu, University of Maryland College Park

Scalable Systems Technology Group

Phillip Merkey, Senior Scientist
Terrence Pratt, Senior Scientist
Donald Becker, Staff Scientist
Daniel Ridge, Technical Specialist

Udaya Ranawake, University of Maryland Baltimore County
Tarek El-Ghazawi, George Washington University
Joel Saltz, University of Maryland College Park

Principal Investigators Funded Under University Research Program in Parallel Computing

HPCC Earth and Space Science Project Scientist: **George Lake** (University of Washington)
Adam Frank, University of Rochester
Derek Richardson, University of Washington

Thomas Sterling left CESDIS in June 1996 to join the Center for Advanced Computing Research at the California Institute of Technology and the High Performance Computing Systems and Applications Group at the Jet Propulsion Laboratory. His vacant position as head of the Scalable Systems Technology Branch was not filled, so the remaining members of this branch merged with the Computational Sciences Branch.

Jacqueline Le Moigne, Senior Scientist – Branch Head (lemoigne@nibbles.gsfc.nasa.gov)

Profile

Dr. Le Moigne holds three degrees from the University of Paris VI, Paris, France: a Bachelor of Science in theoretical mathematics, a Master of Science in pattern recognition, and a Ph.D. in computer vision (1983). From 1983-1987 she served as a research scientist in the University of Maryland College Par, Center for Automation Research, Computer Vision Laboratory. She directed new software development for the Autonomous Land Vehicle project and studied a range sensor utilizing the principle of structured light by projection of grids.

From 1988-1990, Dr. Le Moigne worked as a scientist with Martin Marietta Laboratories. In this capacity she conducted research on the fusion of regions and edges by relaxation methods and studied texture analysis methods for safe Mars landings.

After two years as a National Academy of Sciences-National Research Council Senior Resident Research Associate with the Goddard Space Data and Computing Division (Code 930), Dr. Le Moigne joined CESDIS in October 1992 as a staff scientist. She was appointed to the position of Computational Sciences Branch Head in January 1995 and was promoted to senior scientist in June 1995. Professional memberships include the IEEE Geoscience and Remote Sensing Society for which she has been Chairman and Vice Chairman of the Washington/Northern Virginia Chapter and to which she was elected as senior member in 1996.

Dr. Le Moigne's current work involves the multi-sensor registration, fusion, and analysis of remotely sensed data. This research is of interest in many Earth science applications, such as GOES data landmark registration, the assessment of forested areas utilizing AVHRR and Landsat-TM data, as well as the validation and calibration of new sensor data (such as Modis) with already known data such as Landsat-TM data. This research is also very important for automatic multi-sensor integration when data is gathered at far-remote sites, such as for Mars exploration. All of the techniques involved in this research, especially wavelet-based image registration, have been developed as parallel algorithms on the MasPar MP-2.

Report

In studying how our global environment is changing, research in the Mission to Planet Earth (MTPE) program involves the comparison, fusion, and integration of multiple types of remotely sensed data at various temporal, radiometric and spatial resolutions. Results of this integration will be utilized for global change analysis, as well as for the validation of new instruments or of new data analysis. For example, the analysis of global coverage by low-resolution data can be validated by using local very high resolution data.

The first step in the integration of multiple data is registration, either relative image-to-image registration or absolute geo-registration, to a map or a fixed coordinate system. As this need for automating registration techniques is recognized, each new program involved in the development of a new instrument is independently developing another registration method. Very often, these methods are developed based on something quite similar existing for another sensor, without surveying all the possibilities. Therefore, we feel that there is a need to survey all the registration methods which may be applicable to MTPE problems and to evaluate their performances on a large variety of existing remote sensing data as well as on simulated data of soon-to-be-flown instruments.

Although automatic image registration has been extensively studied in other areas of image processing, it is still a complex problem in the framework of remote sensing. Given the diversity of the data sources, it is unlikely that a single registration technique will satisfy all different applications. We propose to: 1) develop an operational toolbox which consists of some of the most important registration techniques, and 2) provide a quantitative intercomparison of the different methods, which will allow a user to select the desired registration technique based on this evaluation and the visualization of the registration results.

1. General Approach - A Registration Toolbox

As was described in the 1996 CESDIS annual report (pps. 119-124) and in [Lem96], a large variety of techniques can be utilized for registration of remote sensing images, which leads us to the development of a toolbox of registration techniques, where each technique will be associated with a set of parameters and working conditions. Depending on the type of sensors, the desired registration accuracy, the computer availability, and the speed requirement, one or another of the techniques will be chosen. Multiple resolutions of the sensor data will be dealt with by the different approaches included in the toolbox. A multi-resolution approach such as the one taken by a wavelet-based technique deals with the multiple spatial resolution data. Other content-based approaches such as an object-oriented registration might be more appropriate for dealing with multi-temporal or multi-radiometric approaches. This toolbox will be implemented on several architectures, new parallel architectures as well as the DEC-Alpha workstation, under a widely-used software package such as the Khoros image processing package.

We have chosen the Khoros environment as the framework for the implementation of these techniques. Khoros is an object-based data analysis, data visualization, and application development environment. In Khoros, a "toolbox" is a collection of programs and libraries that is handled as a single object. In that sense, our registration toolkit is also composed of the various registration routines, each of which can be handled as an object. Khoros is also an open software system and will enable us to widely distribute the toolbox and to get feedback from a variety of users. Having the toolbox developed in Khoros also makes it compatible with the software developed by the Applied Information Sciences Branch at NASA/Goddard Space Flight Center for the Regional Validation Centers (RVC's) program (Branch Chief: W.J. Campbell, NASA/GSFC, Code 935, [Cam94]). Several members of our team are already involved in this software development. The RVC's will receive remote sensing data by direct readout from various satellites and the users will utilize this software to process in real-time data needed for their regional applications (e.g., monitoring regional change, storm prediction, etc.). Successive versions of our toolbox will be integrated with this software and will also provide us with feedback from the remote sensing community. This feedback will enable us to develop, or help the users develop, new tools which would be more adapted to their particular applications.

Under support of NASA/GSFC Code 935 and in collaboration with J.C. Tilton (NASA/GSFC, Code 935), B.T. Lerner, E. Kaymaz and J. Pierce (KTT Corp.), W. Xia and S. Raghavan (HSTX), T. El-Ghazawi (George Washington University), M. Mareboyana (Bowie State University), and S. Chettri (GST Inc.), a first version of the toolbox has been developed and includes the following techniques:

- Semi-manual registration where pairs of corresponding control points are manually selected, followed by the transformation computation (with choice of polynomial, rotation, translation, rigid or affine transformations),
- Correlation-Based Methods including phase correlation [Kug79] and spatial correlation [Gon87],
- Feature-Based Methods with edge-, corner-, region- and wavelet-based methods [Lem94,Lem95,Kay96,Til96,Rag96],
- Post-Processing Tools, such as interpolation of the matching functions, matching based on moment invariants, as well as robust matching of points given, for example, by a region-based method.

Figure 1 shows the top level of the Khoros graphic user interface of the current registration toolbox. The next versions of the toolbox will include a larger choice of techniques from the ones described in the 1996 CESDIS Annual report (Tables 1 and 2: pp. 122-123), as well as newly developed algorithms. Future methods will also consider cloud masking. Masking of the clouds is a very important issue, and the team is already collaborating with H. Stone (NEC Research Institute) to conduct a study on this issue.

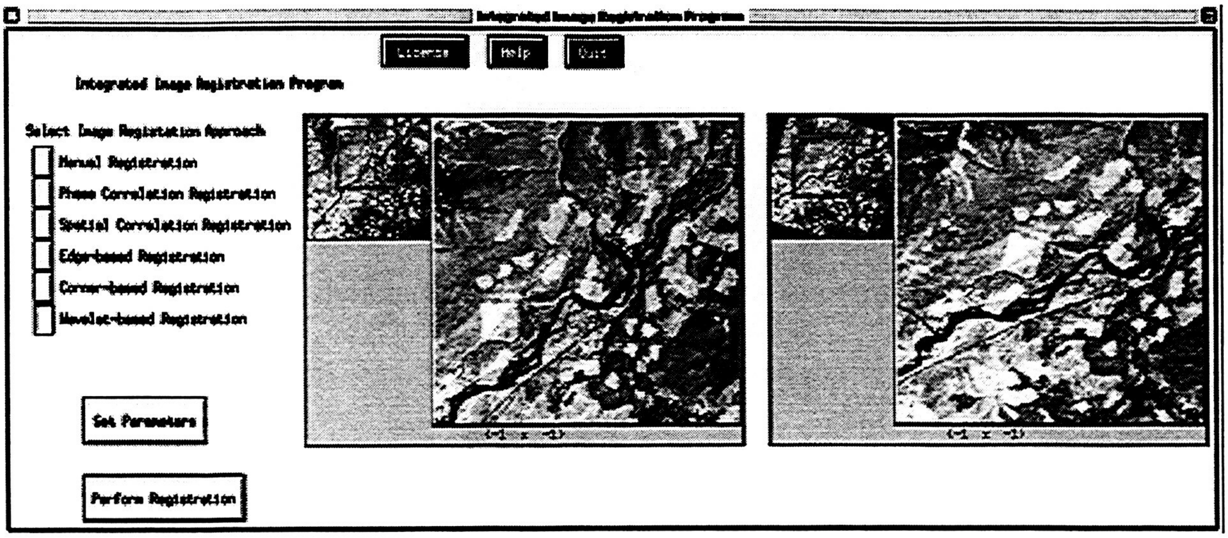

Integrated Image Registration Program
License
Help
Quit
Integrated Image Registration Program
Select Image Registration Approach
□ Manual Registration
□ Phase Correlation Registration
□ Spatial Correlation Registration
□ Edge-based Registration
□ Corner-based Registration
□ Wavelet-based Registration
Set Parameters
Perform Registration
(-1 x -1)
(-1 x -1)

Figure 1: Khoros Graphic User Interface of the Current Registration Toolbox

The semi-manual tool is similar to the method most commonly utilized for registration. A human operator "manually" selects Ground Control Points (GCP's) in two images, and these points become the input to compute the deformation model between the two datasets, often chosen as a polynomial transformation. In our implementation, users select GCP's from the displayed reference image and the image to be registered respectively. Zoom capabilities are available to help users choose the GCP's more accurately. Then a choice of transformations is provided to the user: rotation, translation (e.g., shift), rigid, affine, and polynomial transformations. The GCP's are then used to calculate the parameters of the chosen transformation: either the rotation angle, or the translation shifts, or the transformation coefficients for rigid, affine, or polynomial transformations. This method will be providing "ground truth" registrations against which automatic methods will be validated.

Most of the new technologies developed for the first version of the toolbox are feature-based. Since features are more reliable than intensity or radiometric values, feature-based methods are usually more accurate. In the current feature-based methods included in the toolbox, the features correspond either to edges, corners, regions, or wavelet characteristics. All these methods will be described in the first Image Registration Workshop being organized by our team for November 20-21, 1997 at NASA/Goddard (supported by CESDIS, NASA/GSFC Code 935 and the Washington/North Virginia chapter of the IEEE Geoscience and Remote Sensing Society). Two particular methods have been applied to geostationary image data and are described in section 3 below. Other tools included in the toolbox are utilities for determining subpixel accuracy, based on interpolation of the correlation function, as well as statistically robust point matching (see N. Netanyahu's report in this section).

Implementation issues will also be considered and will focus on the computational aspects and speed of processing of the proposed techniques, through algorithm enhancement, performance evaluations, and parallel implementations of the proposed methods (see T. El-Ghazawi's report).

2. Image Registration Algorithms Quantitative Evaluation

As stated earlier, it is unlikely that a single registration technique will satisfy all different applications. Although automated registration has been developed for a few Earth science applications, there is no general scheme which would assist users in the selection of a registration tool. By providing the results of an intercomparison of multiple registration techniques, we will enable the users to choose the method which is the most appropriate for their particular application. Having all the algorithms implemented in a single toolbox will reinforce their ease of use and will provide the visualization capabilities that will facilitate this choice.

Defining intercomparison criteria is a relatively difficult task, since each application might have different requirements and the importance of the criteria might vary from one application to the next. After a first evaluation, the toolbox and the results of the evaluation will be made available to the scientific community. From comments and feedback, new evaluation criteria as well as new registration techniques might be defined and will give rise to a new version of the toolbox.

The first criteria which will be implemented are the following:

a. *Accuracy*
Several methods can be thought of to quantify the accuracy of a given registration method.
- A first method consists of registering the same set of data manually and automatically. Then, considering the manual registration as our "ground truth", the error between manual and automatic registration characterizes the accuracy of the automatic registration.
- Another method requires a processing which corrects for the illumination variations from two scenes. This correction would be applied after transforming back the sensed image by the computed deformation model. Then, a Mean Square Error (MSE) would be computed between transformed sensed image and reference image.
- If the registration is performed between a remotely sensed image and a map, the MSE can be computed on selected ground features such as coastlines.
- Another way to quantify the accuracy of an automatic method is described in [Cra89]; it utilizes high-resolution data such as Landsat-TM or SPOT ("Satellite Pour l'Observation de la Terre") data which are degraded to lower spatial resolution. Then the lower resolution data are registered and accuracy can be measured at a subpixel level using the full high-resolution data.

b. *Computational Requirements*
The computational requirements of each method will be computed from two means:
1. the computational complexity of each algorithm will be evaluated,
2. each method will be implemented and timed on various architectures.

c. *Level of Automatization*
As was described in the executive summary, given the large amounts of data to process, the automatic techniques should be as free as possible of parameters to tune. Whenever possible, thresholds or other such parameters will be computed adaptively from within the programs. If necessary, training on large numbers of data will be performed and parameters will be chosen from this training.

d. *Applicability*
This last criterion intuitively corresponds to qualitative judgments, such as "if the scene includes a city grid, a corner-based method will work faster than a region-based method." A quantitative way to evaluate the "Applicability" criterion might be statistical; a large amount of sensor data over a large variety of scenes will be gathered and the results of the three previous criteria will be combined to compute a probability of the applicability of an automatic registration technique given a particular dataset and particular scene contents.

In future work, we could also consider (as was proposed by Rignot [Rig91] and Manjunath [Man96]) to couple the registration toolbox with a planning system which would use the above criteria to decide which algorithm to use depending on the application, the type of data, the requested accuracy, and the time and computational constraints. Such a planning/scheduling system has been developed at NASA/Goddard for the RVC's and is based on the work of N.M. Short and A. Lansky [Bod94,Lan95].

At the same time that we are developing the toolbox algorithms, we are collecting a large variety of NASA datasets, on which our toolbox will be evaluated. They will represent at least three main types of applications:

- Multi-temporal studies with multi-temporal datasets of one sensor over the same area collected at different times (various times of the day, various seasons, multiple years, etc.), such as AVHRR/LAC and GAC, Landsat/TM and MSS, GOES,
- Multi-instrument data fusion with multi-sensor datasets representing multiple spatial, temporal, and radiometric resolutions, such as AVHRR/LAC versus Landsat/TM or MSS, GOES versus Landsat/TM or AVHRR/LAC, Landsat/TM versus SPOT, and MAS versus Landsat/TM,
- Channel-to-channel co-registration with multiple radiometric and spatial resolutions of the different channels of one given sensor, such as GOES, MAS, and a hyperspectral instrument such as ASAS.

The next section presents some preliminary registration results using geostationary image data from the Geostationary Operational Environmental Satellite (GOES).

3. A Registration Example: Phase A Feasibility Study for the Image Registration of Future Advanced Geostationary Studies (AGS) Imager Data

This section describes a feasibility study which has been performed in the context of a Phase A investigation for the Geosynchronous Advanced Technology Environmental System (GATES) Imager Mission, now called Advanced Geostationary Studies (AGS). For this study, we assume that the registration system is part of the Ground Processing system. After each swath has been scanned, calibrated, and time corrected using the gyro and the star tracker information, the registration and resampling subsystem includes four main modules which are illustrated in Figure 2:

1. Correlation of Successive Swaths: This module considers two successive arrays of scene data, and correlates them using windows of interest located around each of the 175 gyro report locations. Depending on the accuracy and the precision of the Kalman filter model based on the star tracker and the gyro reports, this module might not be necessary.

2. Landmark Registration: Landmark registration is performed within each swath utilizing a database of predetermined landmarks such as coastlines, lake outlines, rivers, or even city grids; in general any well-structured ground feature. To avoid several resamplings which would be computationally expensive and time-consuming, we chose to distort the map database instead of resampling the image to a fixed grid. Which means that given the attitude, orbit, and ancillary data, a perspective map of the visible landmarks within the current swath will be computed in order to be matched to the image. A minimum of two landmarks up to 10 landmarks will be registered per swath. For each landmark, a window of size maximum 256x256 pixels will be considered and an absolute transformation (rotation, shift, scaling) will be computed for each landmark. Later on, landmark registration could also be used to refine the orbit computation.

3. Resampling: From the gyro and star tracker information, each pixel in each swath is "tagged" with the information relative to the attitude model. From the landmark registration, each pixel is also associated with an absolute location on the ground. The image resampling module uses these two pieces of information to compute the final resampled swath which will have been projected onto a coordinate system chosen by the user.

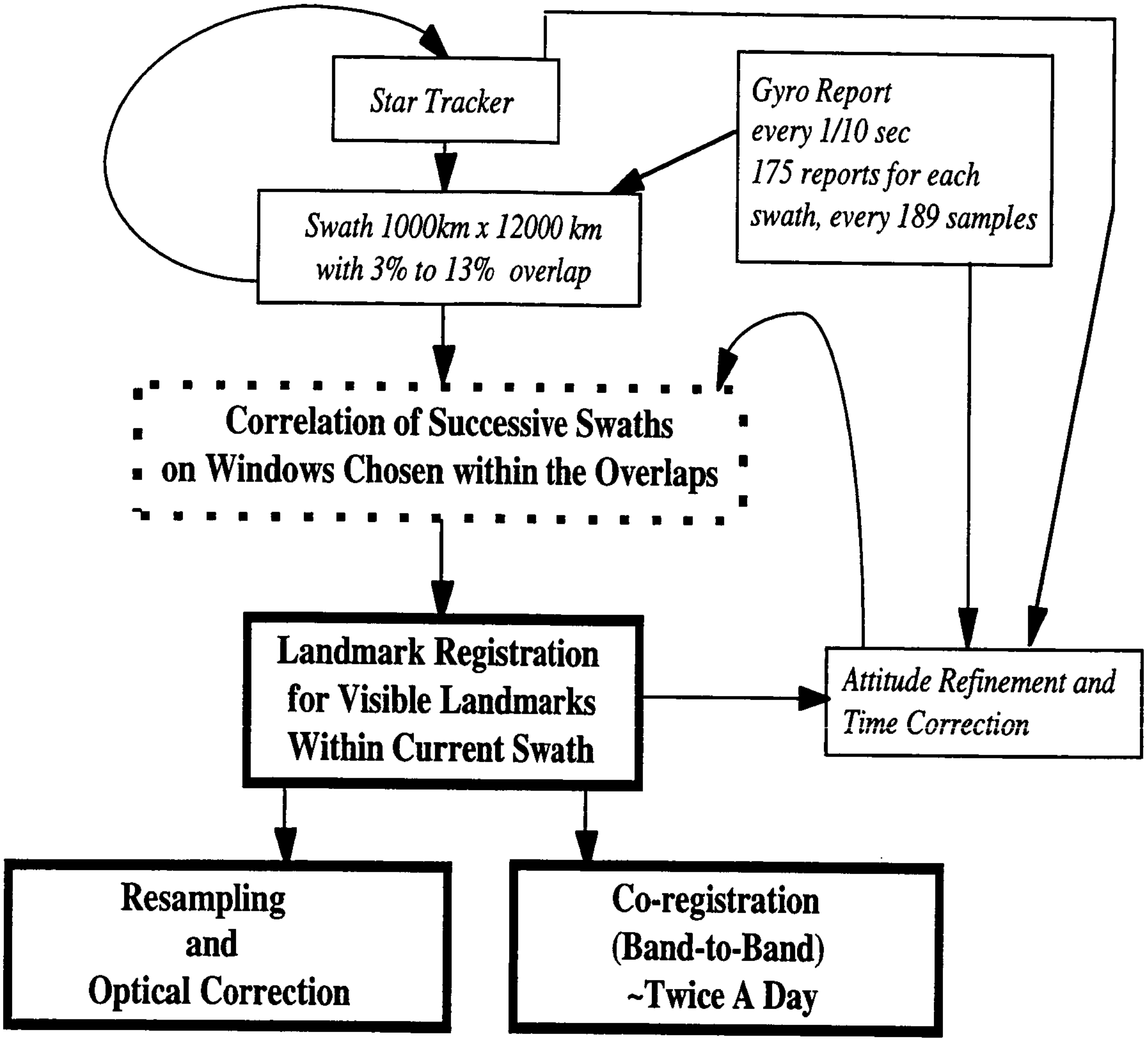


Figure 2: Image Registration and Resampling Subsystem

4. Channels Co-registration: Channel-to-channel co-registration is a calibration-type operation which will not be necessary to perform for every swath or even every Earth image. We assume that co-registration will be computed about twice a day and at each of these computations, a look-up-table will be updated with the five co-registration parameters (rotation, shift, and scaling). Instead of computing these parameters on every possible pair of channels, a reference channel from each focal plane will be selected and co-registration will be computed in two steps:

   S1: reference channels are co-registered to the highest-resolution channel visible at that time of the day,
   S2: every channel of a given focal plane is co-registered to the reference channel of this focal plane.

   For example, if Channels 1, 4, 7, and 15 are respectively the reference channels associated with the four focal planes, the co-registration is performed on the following pairs of channels:

   S1: (1,4), (4,7), (4,15)
   S2: (1,1a), (1,2), (4,3), (4,5), (7,6), (7,8), (7,9), (7,10), (7,11), (7,12), (15,13), (15,14), (15,16), (15, 17).

> Co-registration can be performed either for each swath or for every full Earth image. For each case, co-registration can be performed either on every pixel or utilizing special areas of interest (not necessarily landmarks which might not be seen in every channel). After launch, co-registration parameters could be carefully initialized utilizing a moon image, and after this first initialization, the system would assume that all channels of the same focal plane are somewhat stable relative to each other, thus reducing the search space of the transformations parameters.

Two recent studies dealing with the registration of satellite meteorological data show significant contributions in this domain. The first study [Bla95] deals with Meteosat data, while the second one [Epp96] concentrates on GOES data. Both studies consider a shift-only transformation, and obtain sub-pixel accuracy by up-sampling the data. The Meteosat study uses Normalized Cross-Correlation (NCC) on edges of landmarks such as coastlines. The GOES study uses only lakes and islands for landmarks, and evaluates six different matching methods. Among these methods, Cross-Correlation (CC) and NCC of enhanced gray levels as well as Edge Matching are evaluated as performing the best; Edge Matching is the least sensitive to cloud cover, while NCC provides a slightly more accurate position estimation. Both studies also utilize a masking of the clouds in order to increase the reliability of the registration. The GOES study also provides a good procedure description and well-defined requirements for the choice of the landmarks.

The previous studies seem to be very applicable to future GATES data. They will have to be adapted to generalize the search from a shift-only transformation to a more general transformation, such as a rigid transformation (composed of a rotation, translation-shift, and scaling). The impact of up-sampling the data to achieve sub-pixel accuracy will also have to be quantitatively studied with test data involving texture and shape resampling. Other landmarks such as rivers, roads, and city grids should also be investigated. Masking of the clouds is a very important issue and is being studied by H. Stone (NEC Research Institute).

On the problem of co-registration of remote sensing data, to our knowledge no systematic study has been conducted, in particular for meteorological satellite data. In this case, the issues are somewhat different from the general image registration problem:

1. The transformations to be considered are usually smaller. For example, we can assume that:
   - rotations will vary in the interval [-1degree,+1degree],
   - translation shifts in [-2pixels,+2pixels],
   - and scaling factors in [-0.9pixels,+0.9pixels].
2. Although the observed areas are about the same in each channel, the visible features can be quite different, which leads to two main differences with the general image registration problem:
   - coastline registration is not always applicable,
   - clouds should be used and not eliminated from the registration process.
3. The highest-resolution channel which is utilized as a reference will be different for different times of the day and its spatial resolution will be lower at night.

The initial proposed approaches for the registration of GATES data are based on some of our preliminary results [Lem94,Lem95] as well as on the conclusions of the two previous studies [Bla95,Epp96] and of an internal study [Tilton96] which compares manual registration, edge matching, and phase correlation. All these converge to conclude that edge or edge-like features are very appropriate to highlight regions of interest such as coastlines, and that Normalized Cross-Correlation seems to be the best matching measure. Therefore, we are proposing two methods based on the Normalized Cross-Correlation of edge and wavelet features.

An edge detection computes the gradient of the original gray levels and highlights the pixels of the image with higher contrast. Since edge values are less affected by local variations in the intensity or time-of-the day condition than original gray level values, edge features are more reliable in any registration scheme.

The features provided through a wavelet decomposition are of two different types: the low-pass features which provide a compressed version of the original data and some texture information, and the high-pass features which provide detailed information very similar to edge features. The advantages of using a wavelet decomposition are twofold; (a) by considering the low-pass information, one can bring various spatial resolution data to a common spatial resolution without losing any significant features, which is very useful for channel-to-channel co-registration, (b) by utilizing high-pass information, one can retrieve significant features which are correlated in the registration process, similarly to edge features.

Another aspect of our algorithm is to perform the registration in an iterative manner, first estimating the five parameters of the transformation and then iteratively refining these parameters. This approach presents a lower computational complexity as well as accuracy advantages.

The following is a description of the general registration algorithm which we propose to utilize in the three previous modules (1), (2) and (4), namely swath correlation, landmark registration, and co-registration. This algorithm includes three main steps:

R1. *Preprocessing Step:* This step mainly enhances the contrast of the features which are utilized to perform the registration. For some channels, gray levels will also have to be inverted to consider homogeneous computations.

R2. *Wavelet Decomposition:* Since the images to be registered might have three different spatial resolutions, the wavelet decomposition step brings these images to a common spatial resolution without degrading the image quality. Wavelet decomposition is pursued further down if wavelet coefficients are used in the registration step.

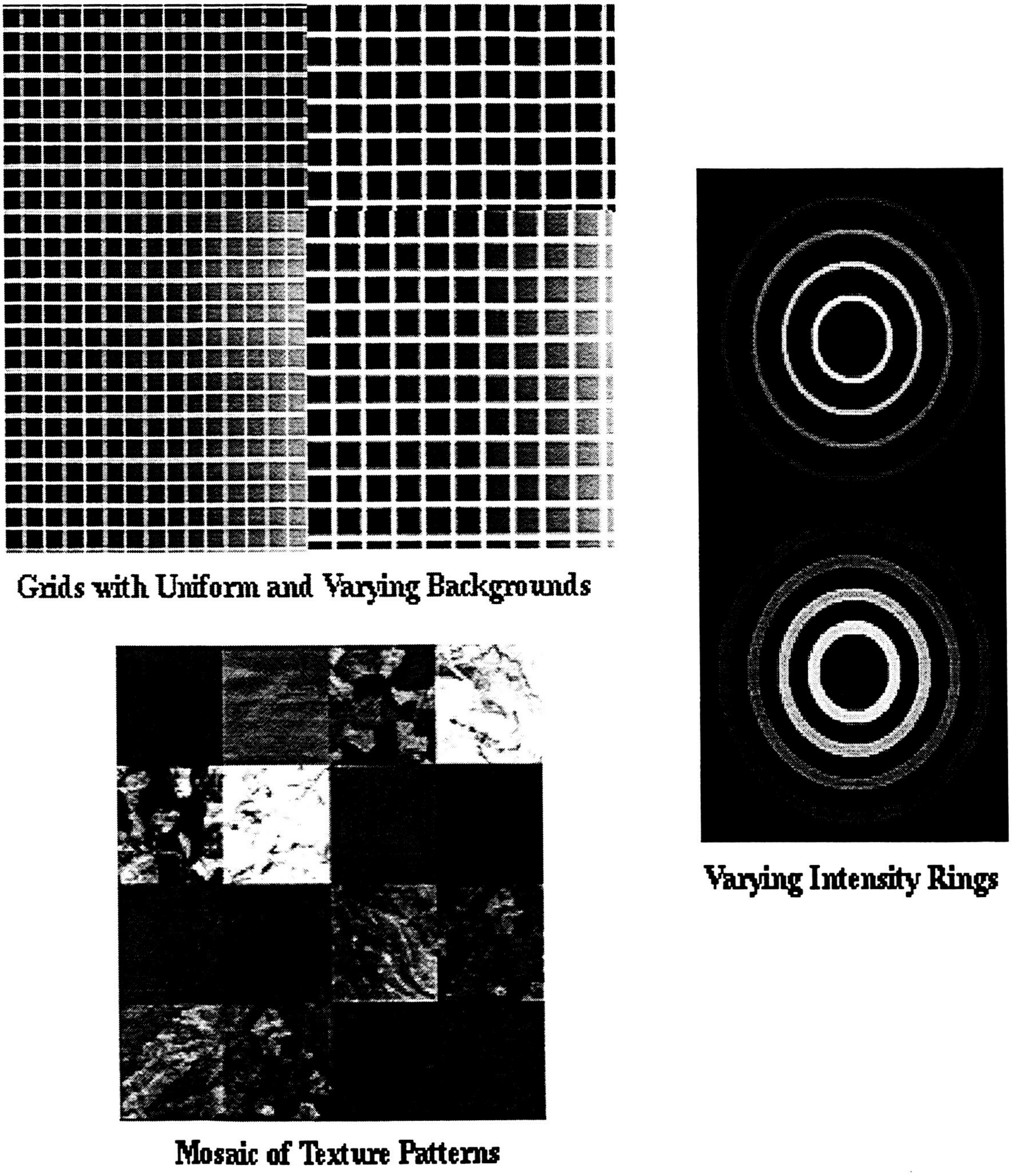
R3. *Registration:* This step can be performed by cross-correlating either edge features or wavelet features. If using edge features, an edge detector such as a Sobel edge detector is applied to each image to register. Then with either type of feature, the registration is performed by:

R3.1. Estimating independently the five parameters: rotation, shift in the x-direction, shift in the y-direction, scaling in x, and scaling in y. First the rotation and the scalings are assumed to be negligible and the shift in x and y are estimated. Then, taking into account the estimated shift, the two scaling parameters are neglected and the rotation angle is estimated. Then, after applying the previous shift and rotation parameters, the two scaling parameters are computed. These scaling parameters are kept constant for the rest of the search.

R3.2. The three previous rotation and shift parameters are iteratively refined, at better and better accuracies. For example, if the first step looks at an accuracy of 2 degrees rotation, and 2 pixels shift, four successive iterations will look at the respective accuracies of 1 degree/1 pixel, 0.5 degree/0.5 pixel, 0.25 degree/0.25 pixel.

The two differences between the edge-based registration and the wavelet-based registration reside in the type of features that are considered to perform the registration, edges versus wavelet coefficients, and in the size of the images on which the computations are carried out: for the edge-based registration, the full size images are utilized for every step. For the wavelet-based registration, the initial search is carried out on the lowest level of wavelet decomposition, i.e., the smallest size images, then each refinement is computed on the next size-up, with the final refinement being computed on the full size image. This last variation explains the difference in the number of operations needed for each algorithm, about 360 floating point operations per pixel for the edge-based registration versus about 80 floating point operations per pixel for the wavelet-based registration. One issue of the wavelet-based registration which will have to be studied is the accuracy of the initial estimate when computed on lower-resolution images.

The algorithms described in the previous sections have been tested on test data as well as GOES imagery. Since preliminary results of the wavelet-based registration were already reported in the 1996 CESDIS Annual Report, we will only present below the results related to the edge-based registration.


Grids with Uniform and Varying Backgrounds

Varying Intensity Rings

Mosaic of Texture Patterns

Figure 3: Test Patterns Including Grids, Rings, and a Texture Mosaic

Test Patterns

First, a series of test patterns was created to test the response of our algorithms to edge directions, local intensity variations, as well as texture variations. Figure 3 shows these seven different images. A few preliminary experiments were conducted by applying two known transformations to the original image data (in this case, a composition of a rotation and a translation), and then registering the transformed images to the original image using the edge-based automatic registration.

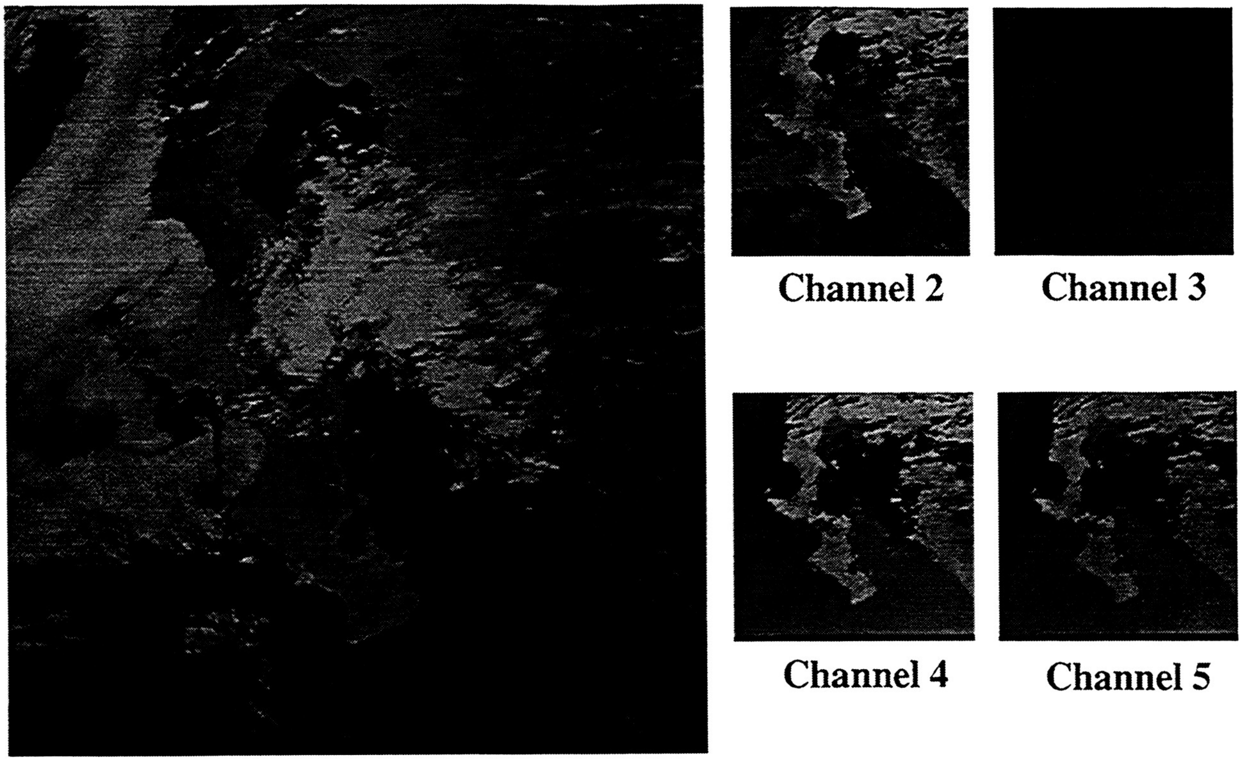
Results are shown in Table 1 and indicate average absolute errors of 0.12 degree in rotation and 0.28 pixel in translation. Although these first results are very encouraging, the present size of the dataset is not sufficient to give any conclusions about the accuracy or the precision of the system, but it will be extensively tested in Phase B experiments.

TEST PATTERN	"TRUE" TRANSFORM			COMPUTED TRANSFORM			ERROR=|true-computed|		
		(Rot: degrees			TranslX: pixels			TranslY: pixels)	
	(Rot	TX	TY)	(Rot	TX	TY)	(Rot	TX	TY)
GRID2.2.15.15	(0.5	1.8	0.5)	(0.5	2.2	0.85)	(0	0.4	0.35)
	(1	0.5	1.5)	(1	0.95	1.9)	(0	0.45	0.4)
GRIDG2.2.15.15	(0.5	1.8	0.5)	(0.5	1.85	0.5)	(0	0.05	0)
	(1	0.5	1.5)	(0.7	0.6	0.75)	(0.3	0.1	0.75)
GRID5.5.20.20	(0.5	1.8	0.5)	(0.5	0.75	0.55)	(0	1.05	0.05)
	(1	0.5	1.5)	(1	0.6	1.55)	(0	0.1	0.05)
GRIDG5.5.20.20	(0.5	1.8	0.5)	(0.5	2.2	0.45)	(0	0.4	0.05)
	(1	0.5	1.5)	(0.75	0.55	0.75)	(0.25	0.05	0.75)
RING10.15	(0	0.8	0.2)	(0.15	0.65	0.25)	(0.15	0.15	0.05)
	(0	2	1)	(-0.25	1.75	1.5)	(0.25	0.25	0.5)
RING2.20	(0	0.8	0.2)	(0.3	0.6	0.2)	(0.3	0.2	0)
	(0	2	1)	(-0.45	0.75	1.35)	(0.45	1.25	0.35)
MOSAIC	(0.6	1.5	0.8)	(0.6	1.55	0.8)	(0	0.05	0)
	(1	0.6	1.5)	(1	0.65	1.5)	(0	0.05	0)
	Average Error Rotation:				0.12 Degrees				
	Average Error Translation:				0.28 Pixels				

Table 1: Results of the Automatic Edge-based Registration on the Test Patterns of Figure 3

Co-registration of GOES Channels

Another set of experiments was performed utilizing multiple channels of a sequence of GOES images taken during 24 hours in two different sectors, "Baja" and "Florida". There are five GOES channels with the respective spatial resolutions of 1 km and 4 km. Figure 4a shows the five channels of a GOES scene of the "Baja" sector. After basic preprocessing of the data and in order to deal with similar spatial resolution data, a wavelet decomposition of Channel 1 is performed (see Figure 4b). After 2 decomposition levels, the spatial resolution of the decomposed Channel 1 is identical to the resolution of Channels 2 to 5. Then a Sobel edge detection is computed on the compressed Channel 1 and on Channels 2 to 5 (see Figure 4c), and the edge-based registration described above is applied to these edge features. We can notice that Channel 3 (Water Vapor Channel) does not present the same original or edge features as the other channels and this remark explains the following results relative to the registration of Channel 3.



Channel 2 Channel 3

Channel 4 Channel 5

Channel 1

Figure 4a: GOES Scene; "Baja" Sector; Five Channels (Images reduced for display purposes)

We tested 33 five-channel scenes corresponding to the "Baja" sector. According to the two steps described in section 1, we chose Channel 1 as the reference channel for the visible wavelength and Channel 4 as the reference channel for the infra-red wavelength; then the cascade of channels to register is: (1,4), (4,2), (4,3), (4,5). Three registrations were performed for each of the 33 scenes:

E1. All original channels are registered
E2. Channel 1 is translated by the vector (1,2) pixels, and Channels 2 to 5 are unchanged,
E3. Channel 1 is rotated by the angle 0.5 degrees and translated by the vector (0.5,1.2) pixels, and Channels 2 to 5 are unchanged.

From these experiments, we expect the automatic algorithm to compute the following transformations:

1. Channels (1,4):
   - Registration (E1): Rotation = 0 degrees Translation = (0,0) pixels
   - Registration (E2): Rotation = 0 degrees Translation = (1,2) pixels
   - Registration (E3); Rotation = 0.5 degrees Translation = (0.5,1.2) pixels

2. Channels (4,2), (4,3), (4,5):
   - Registrations (E1) to (E3): Rotation = 0 degrees Translation = (0,0) pixels

Table 2 shows some results of these experiments, detailed for the first three scenes, and then the accuracy and the standard deviation (or precision) of the rotation and translation parameters is computed over the 33 scenes for the four pairs of channels. The accuracy and the standard deviation of a parameter "a" are computed according to the following formulas:

$$\text{Accuracy (a)} = \frac{1}{33*3}\sum_{i=1}^{i=33*3}(a(i) - true\_a)$$

and

$$\text{StdDev (a)} = \frac{1}{33*3}\sum_{i=1}^{i=33*3}((a(i) - true\_a)^2) - (Accuracy(a)^2)$$

Correlation coefficients, which vary between 0 and 1, are also indicated for each registration. The results show a perfect registration of Channels 4 and 5 with a correlation coefficient close to 1, and a relatively good registration of Channels 4 and 2. As we noticed previously, the features extracted from Channel 3 do not permit a good registration, and it is illustrated by very low correlation coefficients (around or below 0.1). In this example, the registration of Channels 1 and 4 seems to present a bias of (0.95,0) in shift that is consistent among the 33 scenes of this sequence. In the absence of ground truth data and with correlation coefficients of average value, no conclusion could be drawn from this result. If additional information was available, such registration could indicate a shift in the geometric calibration of Channel 1 relative to the other channels.

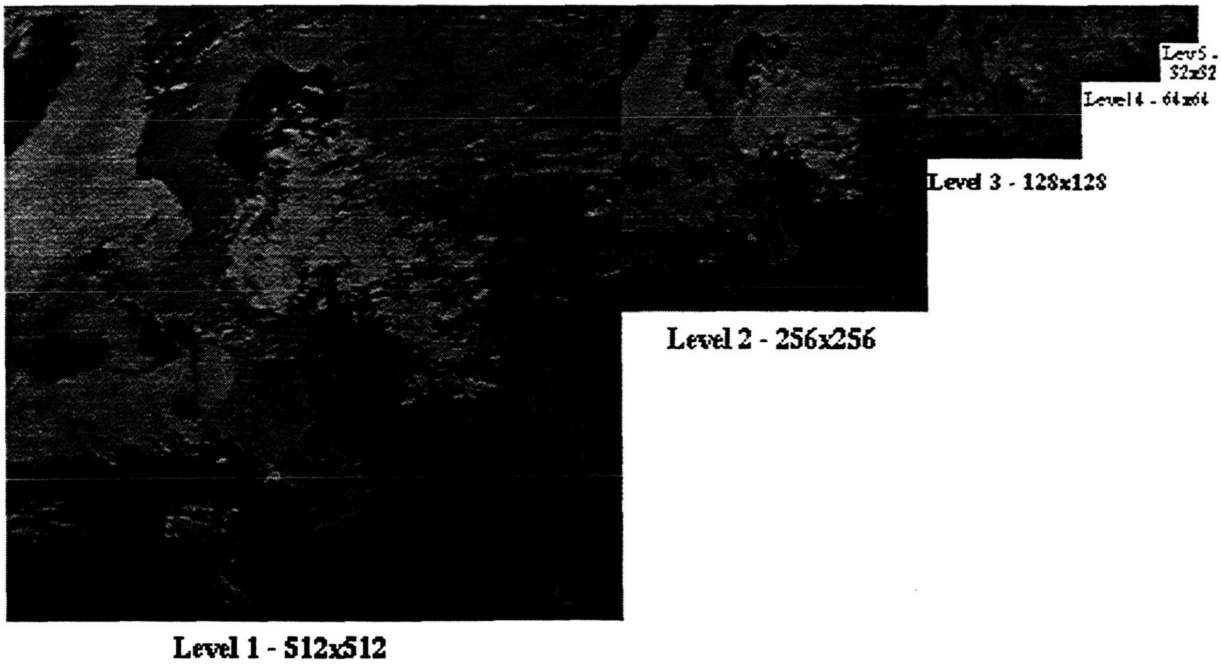
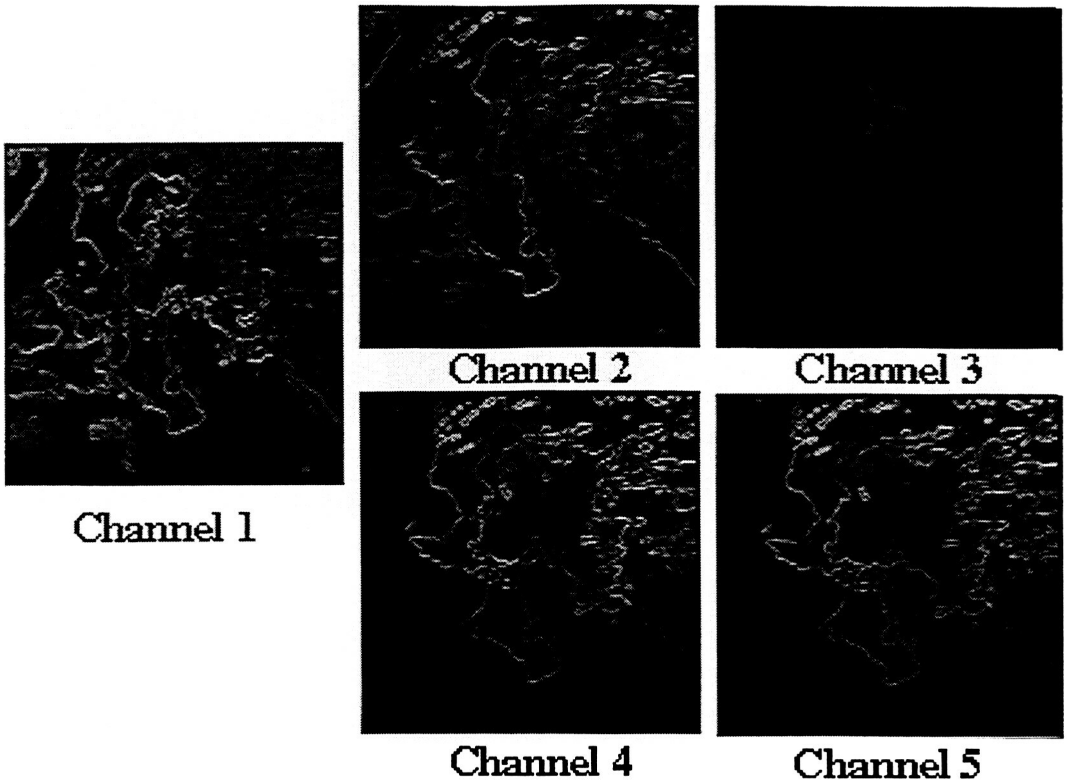


Figure 4b: Wavelet Decomposition of Channel 1 shown in Figure 4a



Channel 1

Channel 2

Channel 3

Channel 4

Channel 5

Figure 4c: Edge Detection on Level2-Wavelet of Channel 1 and on Channels 2 to 5 of Figure 4a

Image	Channels 1/4			Channels 4/2			Channels 4/3			Channels 4/5		
	Rot. (Deg.)	Transl. (Pix.)	Correl	Rot. (Deg.)	Transl. (Pix.)	Correl	Rot. (Deg.)	Transl. (Pix.)	Correl	Rot. (Deg.)	Transl. (Pix.)	Corr.
...101615...												
Original	0.15	0.95,0.35	0.44	0	0.2,0.2	0.65	0.15	-0.85,-0.1	0.12	0	0,0	0.94
R=0 T=(1,2)	0.1	1.9,2.3	0.44	-	-	-	-	-	-	-	-	-
R=.5 T=(.5,1.2)	0.5	1.55,1.25	0.44	-	-	-	-	-	-	-	-	-
...101632...												
Original	0.15	0.85,0.1	0.44	0	0,0	0.64	0.35	-1.6,0.65	0.05	0	0,0	0.94
R=0 T=(1,2)	0.05	1.95,2.05	0.44	-	-	-	-	-	-	-	-	-
R=.5 T=(.5,1.2)	0.5	1.55,1.25	0.44	-	-	-	-	-	-	-	-	-
...101645...												
Original	0	1,0	0.44	0	0,0	0.65	0.15	-1.1,0.15	0.11	0	0,0	0.94
R=0 T=(1,2)	0	2,2	0.44	-	-	-	-	-	-	-	-	-
R=.5 T=(.5,1.2)	0.5	1.55,1.25	0.44	-	-	-	-	-	-	-	-	-
After 33 Experiments:												
Accuracy	**-0.06**	**(-0.72,-0.18**		**0**	**(0.01,0.01)**		**-0.217**	**(0.62,-0.19)**		**0**	**(0,0)**	
Stand. Dev.	**0.07**	**(0.29,0.13)**		**0**	**(0.04,0.04)**		**0.115**	**(0.85,0.32)**		**0**	**(0,0)**	

Table 2: Results of Registration Experiments for the “Baja” Sector

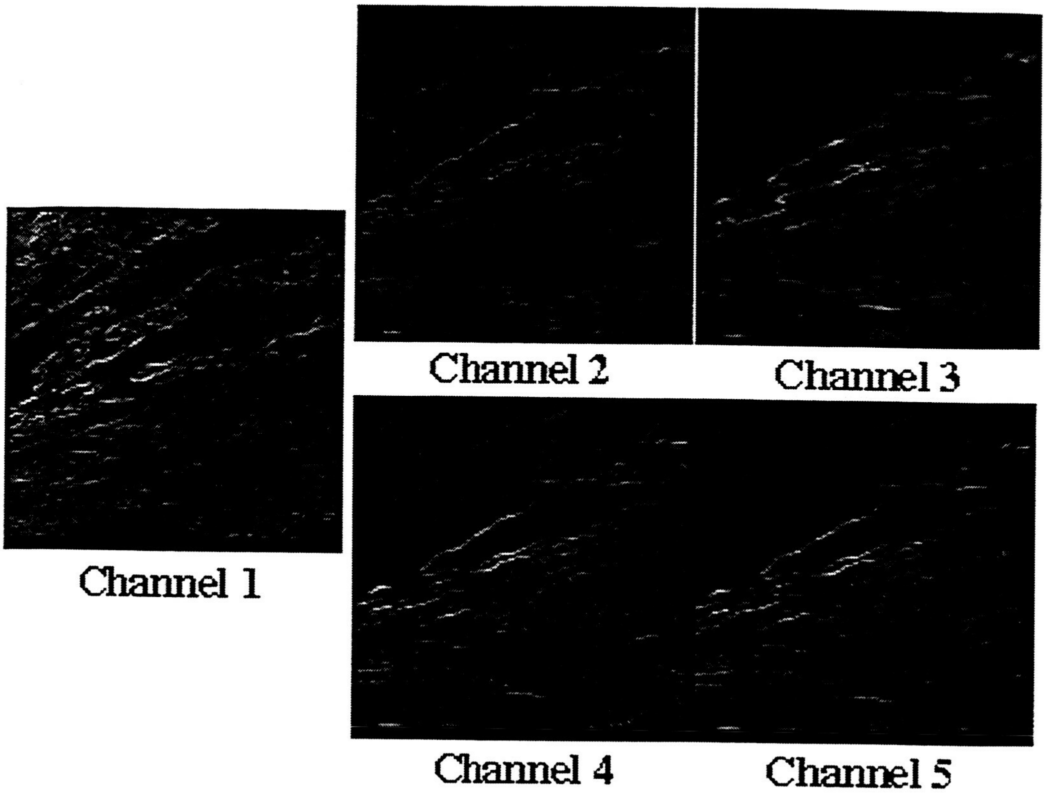


Figure 5a: GOES Scene; "Florida" Sector; Five Channels (Images reduced for display purposes)

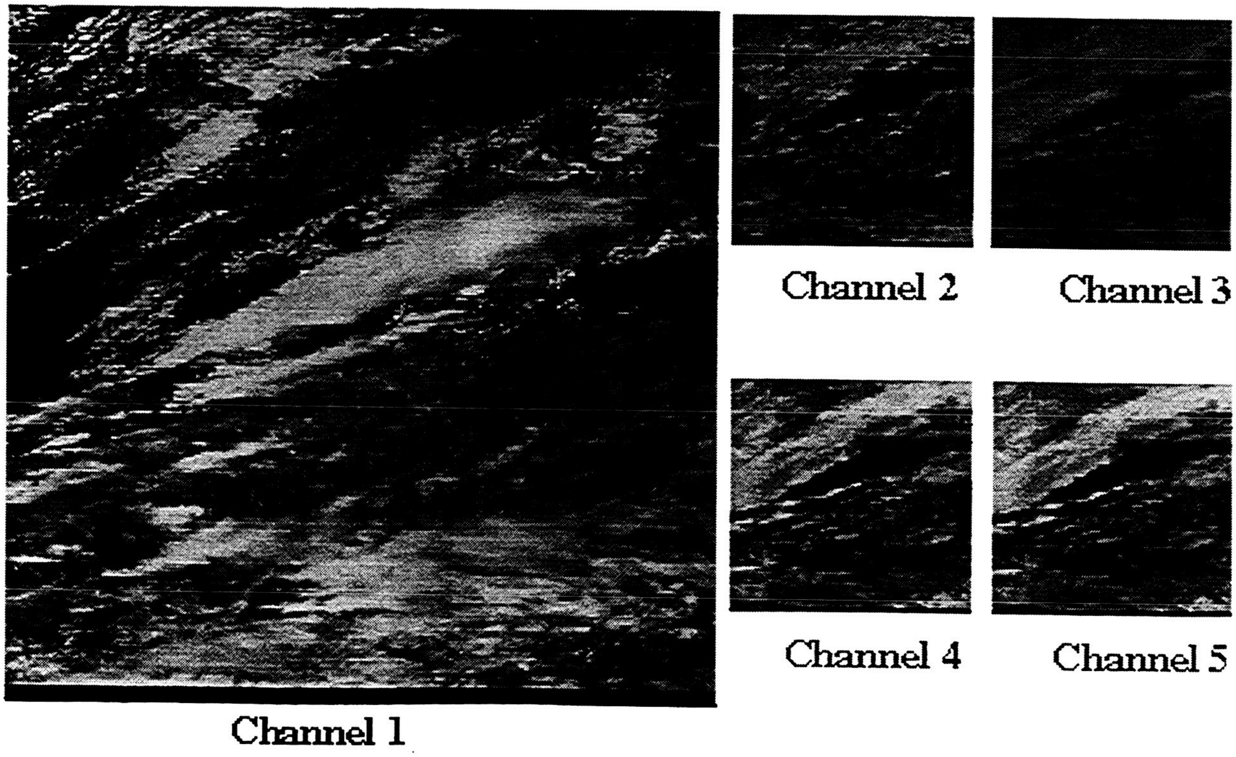


Figure 5b: Edge Detection on Level2-Wavelet of Channel 1 and on Channels 2 to 5 of Figure 5a

Similar experiments were conducted with the successive scenes of the "Florida" sector shown in Figure 5a. In this case, no land features are visible in any of the five channels and this example shows how channel-to-channel co-registration can be performed with only cloud features. Figure 5b shows the edges extracted from the five channels and Table 3 details the results of the registrations performed for the above experiments (E1), (E2) and (E3). These few results show good registrations of the five channels

Image	Channels 1/4			Channels 4/2			Channels 4/3			Channels 4/5		
	Rot.	Transl.	Corr.	Rot.	Transl.	Corr.	Rot.	Transl.	Corr.	Rot.	Transl.	Corr.
	(Deg.)	(Pix.)		(Deg.)	(Pix.)		(Deg.)	(Pix.)		(Deg.)	(Pix.)	
Original	0	(0,0)	0.47	0	(-.4,-.4)	0.68	0	(0,0)	0.72	0	(0,0)	0.98
R=0 T=(1,2)	0	(1,2)	0.47	-	-	-	-	-	-	-	-	-
R=.5 T=(.5,1.2)	0.5	(.5,1.25)	0.47	-	-	-	-	-	-	-	-	-

Table 3: Results of Co-registration Experiments for the Above Scene of the "Florida" Sector

References

[Bla95] Blancke, B., Carr, J. L., Lairy, E., Pomport, F. and Pourcelot, B. (1995, April) The Aerospatiale Meteosat Image Processing System (AMIPS), *1-st International Symposium on Scientific Imagery and Image Processing,* Cannes, France.

[Bod94] Boddy, M., White, J., Goldman, R. and Short, N. M. (1994, May 10-12) Planning for image processing, in *Proceedings of the Goddard Conference on Space Applications of Artificial Intelligence*, NASA Goddard, Greenbelt.

[Cam94] Campbell, W. J., Short, Jr., N. M., Coronado, P. and Cromp, R. F. Distributed Earth science validation centers for Mission to Planet Earth, *Methodologies for Intelligent Systems*, 8-th International Symposium, ISMIS'94, Charlotte, NC.

[Cra89] Cracknell, A. P. and Paithoonwattanakij, K. (1989) Pixel and sub-pixel accuracy in geometrical correction of AVHRR imagery, *International Journal of Remote Sensing, vol. 10,* nos. 4,5, pp. 661-667.

[Epp96] Eppler, W., Paglieroni, D., Louie, M. and Hanson, J. (1996, August 4-9) GOES Landmark Positioning System. SPIE Proceedings Vol. 2812, International Symposium on Optical Science, Engineering, and Instrumentation'96, *GOES-8 and Beyond* (2812-72), Denver, CO.

[Gon87] Gonzalez, R. and Wintz, P. (1987) *Digital Image Processing,* Addison-Wesley Co.

[Kug79] Kuglin, C. D., Blumenthal, A. F. and Pearson, J. J. (1979, May 22-24) Map-matching techniques for terminal guidance using Fourier phase information. *Proceedings of the SPIE Conference on Digital Processing of Aerial Images: Vol 186* (pp. 21-29) Huntsville, Alabama.

[Kay96] Kaymaz, E., Lerner, B-T., Pierce, J. F., Campbell, W. J. and Le Moigne, J. (1996, October 16-18), Multi-resolution geo-registration of remote sensing imagery employing the wavelet transform, *SPIE 25th AIPR Workshop*, Washington, DC.

[Lan95] Lansky, A. L., Getoor, L., Friedman, M., Schmidler, S. and Short, Jr., N. M. (1995, March) The COLLAGE/KHOROS link: Planning for image processing tasks. *AAAI Spring Symposium on Integrated Planning Applications,* Stanford, California: Stanford University.

[Lem94] Le Moigne, J. (1994, April 5-8) Parallel registration of multisensor remotely sensed images using wavelet coefficients. *Proceedings of SPIE Wavelets'94*, Orlando.

[Lem95] Le Moigne, J. (1995, July 10-14) Towards a parallel registration of multiple resolution remote sensing data. *IGARSS'95, International Geoscience and Remote Sensing Symposium* (pp. 1011-1013), Firenze, Italy.

[Lem96] Le Moigne, J., Campbell, W. J. and Cromp, R. F. (June 1996) An automated parallel image registration technique of multiple source remote sensing data, submitted to the *IEEE Transactions on Geoscience and Remote Sensing.*

[Man96] Manjunath, B. S. (1996) Registration techniques for multisensor sensed imagery. *Photogrammetric Engineering and Remote Sensing Journal.*

[Rag96] Raghavan et al, S. (1996) Real-Time Extraction of Meta-Data from Multi-Source Imagery. *Phase II SBIR Report.*

[Rig91] Rignot, E. J. M., Howk, R., Curlander, J. C. and Pang, S. S. (1991) Automated multisensor registration: Requirements and techniques. *Photogrammetric Engineering & Remote Sensing, vol. 57, no. 8, pp. 1029-1038.*

[Til96] Tilton, J. (1996, August 28) *Comparison of Registration Techniques for GOES-8 Visible Imagery Data. Notes Tea and Poster Seminar Sessions, NASA/GSFC, Atrium Building.*

Publications

Le Moigne, J. (January 6, 1997) Image Registration Subsystem. Report - GATES Phase A.

Gualtieri, A. J., Le Moigne, J., and Packer, C. V., 1997. (April 1997) Computing distances between images with a parallel Hausdorff distance metric. *International Journal of Computers and their Applications.*

El-Ghazawi, T. and Le Moigne, J. (1996, August) Wavelet decomposition on high-performance computing systems, *Proceedings of the 25-th International Conference on Parallel Processing (ICPP'96).* Bloomingdale, IL.

Kaymaz, E., Lerner, B-T., Pierce, J. F., Campbell, W. J. and Le Moigne, J. (1996, October 16-18) Multiresolution geo-registration of remote sensing imagery employing the wavelet transform. *SPIE 25th AIPR Workshop*, Washington, DC.

Szu, H., Le Moigne, J., Netanyahu, N., Francis, M. and Hsu, C. (1996, October 16-18) Wavelet index of texture for artificial neural network classification of Landsat images. *SPIE 25th AIPR Workshop*, Washington, DC.

Szu, H., Le Moigne, J., Netanyahu, N. and Hsu, C. (1997, April 21-25) Integration of local texture information in the automatic classification of Landsat images. *1997 SPIE's OE/Aerospace Sensing, Wavelet Applications Conference*, Orlando.

Xia, W., Le Moigne, J., Tilton, J.C., Lerner, B-T., Kaymaz, E., Pierce, J., Raghavan, S., Chettri, S., El-Ghazawi, T., Manohar, M., Netanyahu, N., Campbell, W. and Cromp, R. (1997, October) A registration toolbox for multi-source remote sensing applications. *1997 International Conference on Earth Observation and Environmental Information (EOEI'97)*, Egypt.

A Combined Phase Retrieval and Image Deconvolution Approach

Richard G. Lyon
University of Maryland Baltimore County
Department of Computer Science and Electrical Engineering
(lyon@jansky.gsfc.nasa.gov)

Profile

Mr. Lyon holds a Bachelor of Science in physics from the University of Massachusetts and a Master of Science in optics from the University of Rochester (New York) with work toward a Ph.D. in optics, also at the University of Rochester. He is a member of the Optical Society of America (OSA) and the Society for Photo-Instrumentation Engineers (SPIE).

From 1987 to 1992 Mr. Lyon was employed by Hughes Danbury Optical Systems as an optical systems engineer in the Space Sciences directorate. In that capacity he served as principal investigator of Hubble Space Telescope phase retrieval efforts to determine the on-orbit telescope error. During this period, he received a NASA Goddard Certificate of Recognition for Contributions to the Hubble Space Telescope Program, a NASA Goddard Group Achievement Award for the Hubble Space Telescope Mission Operations Team, and a NASA Award of a Flag flown on STS-31 for contributions to the Hubble Space Telescope Program.

From January 1993 to June 1994 Mr. Lyon worked as a research analyst for Radex Incorporated where he designed, developed, and implemented automated celestial image processing algorithms for the Mid-Course Space Experiment (MSX), a U.S. Air Force experimental radiometric satellite. In June 1994 he became a principal engineer with Hughes STX where he conducted research into the design and development of optical and image processing algorithms to operate in a massively parallel computational environment, including image restoration and image deconvolution algorithms to deblur Hubble Faint Object Camera images.

Mr. Lyon joined CESDIS through a subcontract with the University of Maryland Baltimore County in June 1995.

Report

Future space flight optical imaging systems will continue to increasingly rely on a combined optical and computational approach, with an increasing share of the image quality requirements being relegated to the computational domain. Thus, with this in mind, my previous research efforts have been twofold. The first concentrated on the development of model based image processing methods which deconvolve, or reconstruct imagery, and the second, on the use of phase retrieval methods to determine the wavefront error in an optical system. At first glance these might seem mutually exclusive, however, they are tightly coupled. Image deconvolution requires the systems optical response, i.e., the point spread function (PSF), and generally this is incompletely known. Thus one desires to determine the object signature without completely knowing the system that created it where the "system" includes an atmosphere, not necessarily the Earth's, the optics, and the focal plane detectors. Both the phase retrieval and deconvolution problem can be cast into a symbiotic hardware/software relation requiring the use of only one methodology, known as Phase Diversity. Thus one can recast the imaging problem as a tightly coupled phase retrieval and image deconvolution problem inherently coupled to the optical system either via phase retrieval or the use of active optics. It this coupling, and an overview of the methods, which will be briefly outlined here with current results and a research plan for exploring this rich research area in both the computational and laboratory environment.

1. Phase Diversity

Figure 1 shows a conceptual optical schematic of a phase diversity system. Both the telescope and the intervening medium contribute to the phase aberrations of the entire system. The telescope by deterministic design residuals, and the unknown, but fixed, misalignments, fabrication errors, surface scatter, and thermal/mechanical drift; the atmosphere by thermally induced, multi-layer, density changes contributing to stochastic index of refraction fluctuations. In Figure 1 the phase diversity is introduced via a beamsplitter to split the beam into two (or more) separate channels, each of which sees a deliberate, known, phase aberration such as focus. The telescope and atmospheric aberrations occur in both channels (common mode) while the diversity is dissimilar in each channel. Note that although focus is used in this simplified schematic, it is generally considered sub-optimal. The phase diversity method has not been shown to be either unique, convergent nor the results accurate, but, yet, it does appear to qualitatively work in practice. Why, how well, and what is the optimal approach? The optimal phase aberration is currently unknown, as is, which coded aperture technique. Thus the phase diversity problem is a rich research problem in terms of applied mathematics, physical optics, algorithms, computer science, and noise/estimation theory. Thus this problem has a rich future and will require many researchers, a multitude of technologies, and slightly beyond the state of the art in computer technology to keep up with stochastic atmospheric fluctuations. Towards this end, we have begun researching and enhancing different methods and applying our experience to simulate, optimize, and develop an experimental prototype benchtop system to research the problem in a controlled fashion. The goal being to model and simulate a multitude of real world effects, at first, in an open loop fashion and to ultimately develop a closed loop optical control system.

We have literature researched all the current methods and have made significant progress towards understanding this problem, and, have begun the process of optimization required to find the optimal diversity function. We have also begun developing a benchtop system to experimentally determine the real world accuracy and precision of each of the simulations and to try out each of the methods, first in an open loop,

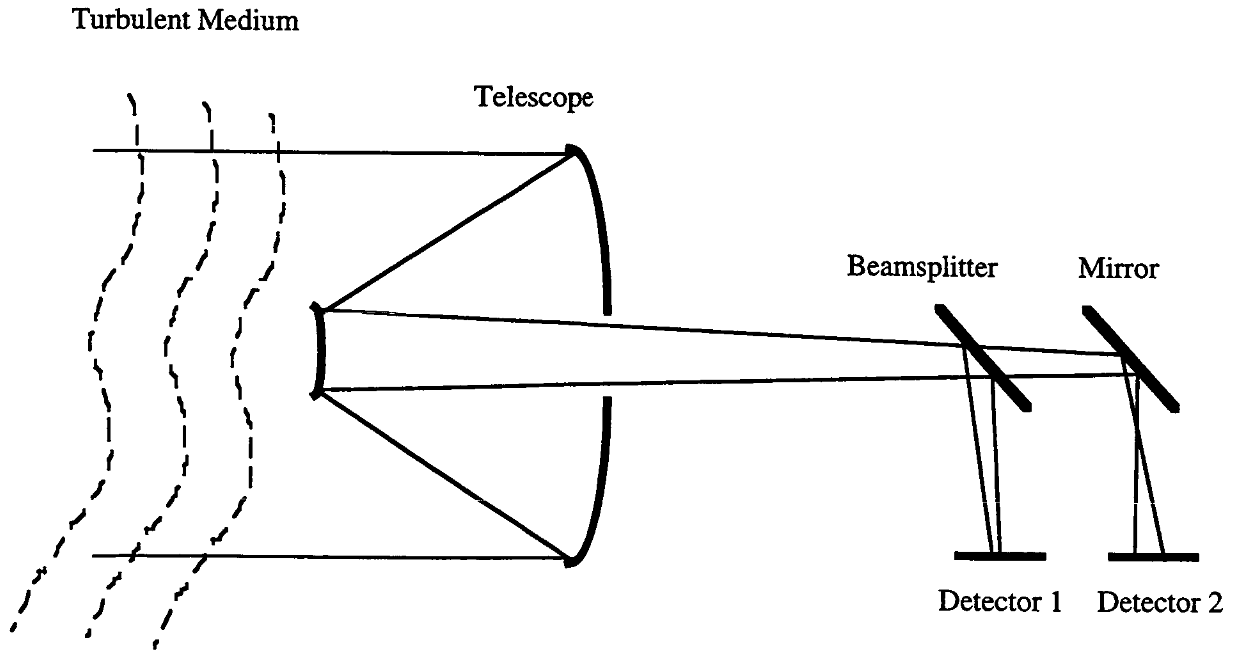


Figure 1: Conceptual Phase Diversity System

and subsequently in a closed loop environment. The primary limitations for closed loop control are a trade off between accuracy/precision and computational horsepower as are all methods of this nature, but with the advantage that these methods lend themselves well to massively parallel and subsequently high throughput techniques.

2. Phase Diversity Algorithm

For the two-channel phase diversity system shown in Figure 1, the image in channel 1 is given by:

$$d_1(x,y) = P_1(x,y) ** O(x,y) + \eta_1(x,y) \tag{1}$$

and for channel 2 by:

$$d_2(x,y) = P_2(x,y) ** O(x,y) + \eta_2(x,y) \tag{2}$$

where the object **O** is the same for both channels and the point spread functions, $\mathbf{P}_1$ and $\mathbf{P}_2$, differ for each of the bands. Note that ** denotes convolution. An additive noise model is assumed in this case where N_1 and N_2 are the channel 1 and 2 noise vectors respectively. Both channels "see" the same wavefront due to the atmosphere and/or telescope, however a wavefront diversity, *SW,* is introduced into the system via deliberately looking at the images at different foci. Note that focus error is only one form of phase diversity. The point spread functions (PSF) for a single plane diffraction model are given by:

$$P_1(x,y;W) = \frac{1}{\lambda^2 F^2}\left| \iint A(u,v) e^{ikW(u,v)} e^{-i\frac{2\pi}{\lambda F}(xu+yv)} \right|^2 \tag{3}$$

$$P_2(x,y;W+\delta W) = \frac{1}{\lambda^2 F^2}\left| \iint A(u,v) e^{ik(W(u,v)+\delta W(u,v))} e^{-i\frac{2\pi}{\lambda F}(xu+yv)} \right|^2 \tag{4}$$

Where λ is the wavelength and F the system focal length, A(u,v) is the aperture function, W(u,v) the combined atmosphere and telescope wavefront, δW(u,v) is the diversity wavefront, u,v are the exit pupil coordinates and x,y are the respective focal plane coordinates. The goal in phase diversity is to simultaneously estimate the wavefront W(u,v) and the object O(x,y) from a single dataset containing multiple images, each with a different diversity. Figure 2 shows the phase diversity method used here in a graphical flowchart form which will be discussed in more detail here.

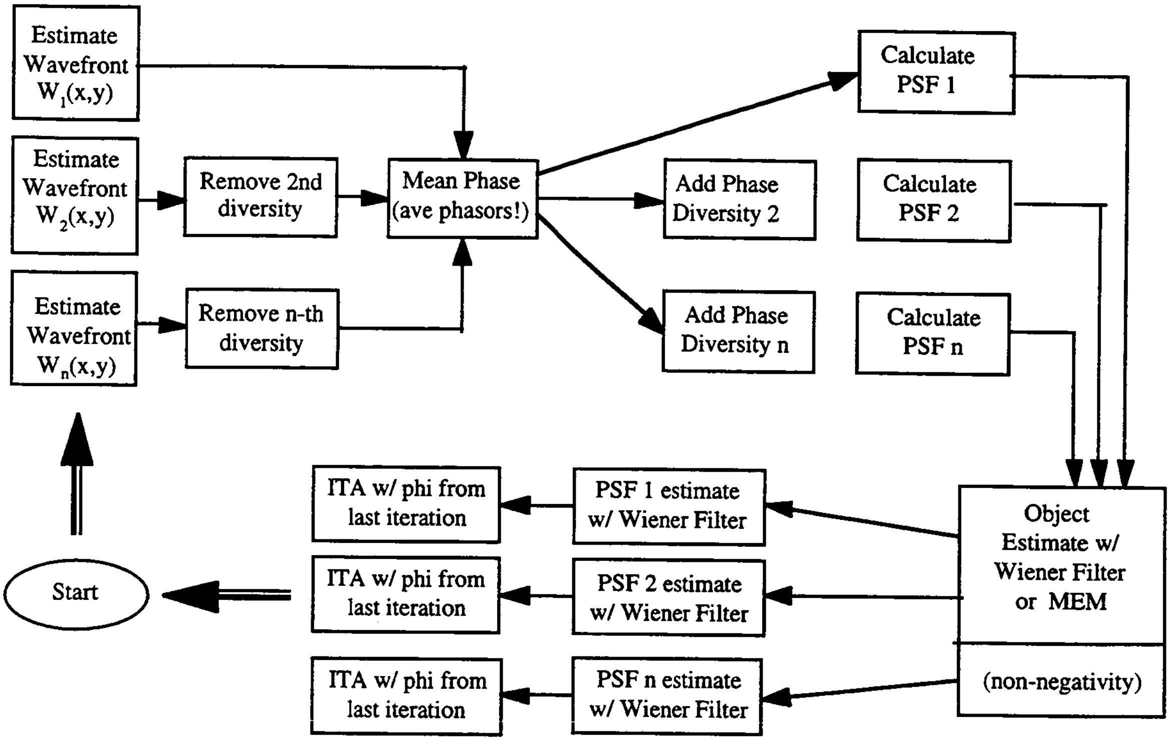

Estimate Wavefront $W_1(x,y)$
Estimate Wavefront $W_2(x,y)$
Estimate Wavefront $W_n(x,y)$
Remove 2nd diversity
Remove n-th diversity
Mean Phase (ave phasors!)
Calculate PSF 1
Add Phase Diversity 2
Add Phase Diversity n
Calculate PSF 2
Calculate PSF n
Start
ITA w/ phi from last iteration
ITA w/ phi from last iteration
ITA w/ phi from last iteration
PSF 1 estimate w/ Wiener Filter
PSF 2 estimate w/ Wiener Filter
PSF n estimate w/ Wiener Filter
Object Estimate w/ Wiener Filter or MEM
(non-negativity)

Figure 2: Multi-Channel Phase Diversity

One starts by assuming an initial guess for the wavefront W(u,v) and evaluating equations (3) and (4) via fast Fourier transform (FFT) techniques. An initial object estimate is then made, in the Fourier domain, via a Wiener filter of the form:

$$\tilde{O}(f_x,f_y)=\frac{\tilde{P}_1^*(f_x,f_y)d_1(f_x,f_y)+\tilde{P}_2^*(f_x,f_y)d_2(f_x,f_y)}{\left|\tilde{P}_1^*(f_x,f_y)\right|^2+\left|\tilde{P}_2^*(f_x,f_y)\right|^2+\beta} \tag{5}$$

The object spectrum is inverse FFTd and the positivity constraint is enforced, i.e., any points which are less than zero in the object are set to zero. The PSFs are then re-estimated from the positively constrained object via Wiener filtering:

$$\tilde{P}_1(f_x,f_y)=\frac{\tilde{O}^*(f_x,f_y)d_1(f_x,f_y)}{\left|\tilde{O}^*(f_x,f_y)\right|^2+\beta} \text{ and } \tilde{P}_2(f_x,f_y)=\frac{\tilde{O}^*(f_x,f_y)d_2(f_x,f_y)}{\left|\tilde{O}^*(f_x,f_y)\right|^2+\beta} \tag{6}$$

The PSFs are inverse FFTd and again the positivity constraint applied. An iterative transform algorithm (ITA) is then used to back propagate from the respective focal planes to the pupil plane; the focal plane phase used for the ITA is from the previous iteration. A value of zero is used as a starting point. In the pupil plane, the phase diversity is removed and the wavefronts averaged to yield an updated value for W(u,v) and then the entire process repeats. Typically it takes about 100 iterations before a solution is reached.

Figure 3 is an example of this phase diversity algorithm with synthetic data. The upper left of Figure 3 shows a synthetically grown fractal scene of land, water, and coastline as seen from low Earth orbit. This scene is then convolved with two blurring PSFs representing the atmospheric blurring and the telescope optics. These two images are akin to what an actual system would see looking through the atmosphere and are shown in the upper middle and upper right. The lower right shows the result after only 10 iterations of the algorithm and the lower middle shows the result after 100 iterations. The lower left shows the difference between the true and the 100 iteration solution. The difference image has been contrast stretched to visually enhance the detail. Note that most of the error occurs in the high spatial frequencies.

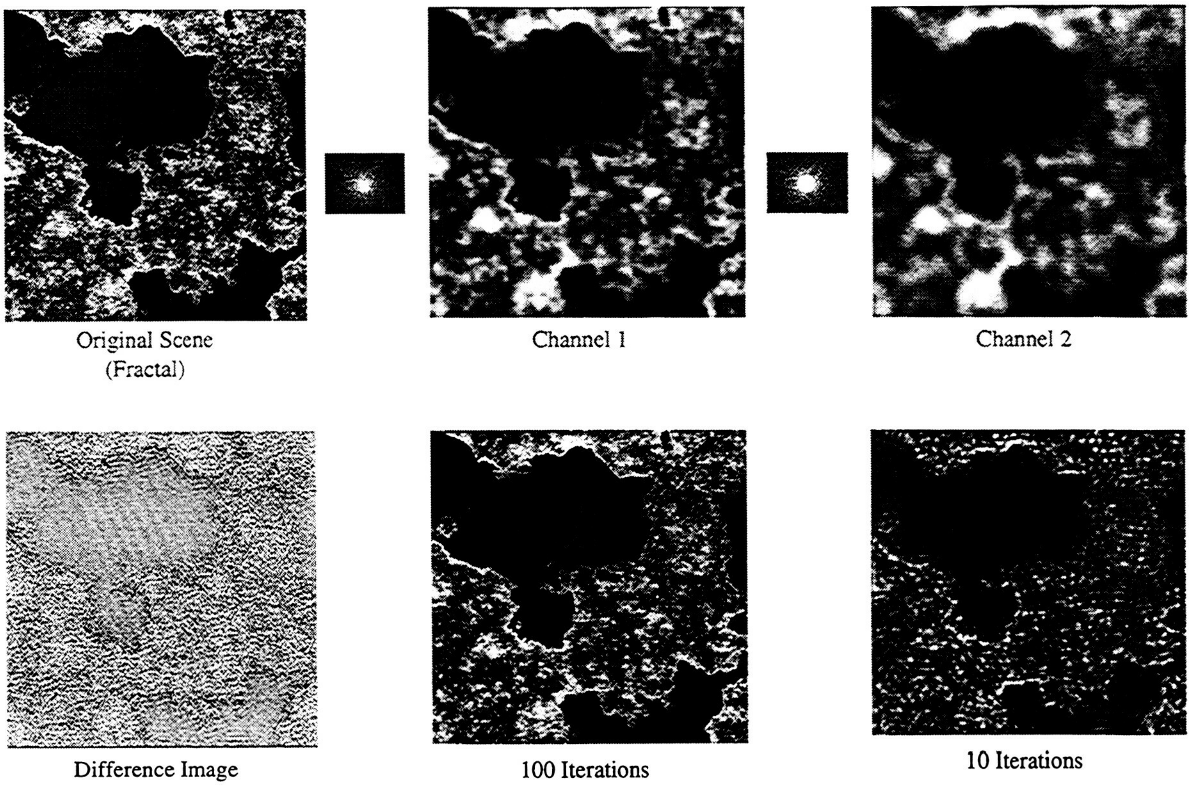


Figure 3: Phase Diversity Example

3. Summary and Future Plans

A phase diversity has been developed which simultaneously combines both image deconvolution and phase retrieval. The current algorithm uses a simple Wiener filter for the deconvolution and a modified iterative transform algorithm developed by this author. It is believed that a better deconvolution result would be obtained from the use of a maximum entropy with maximum likelihood constraint particularly for low signal to noise ratio data. This is currently in the process of being implemented. The phase diversity problem is a rich research problem with much potential use for NASA, both for ground-based telescopes, down-looking sensors for Earth remote sensing, and for remote sensing through other planetary atmospheres. Towards this end, an optics and imaging lab has been established at GSFC to build a research-oriented benchtop prototype phase diversity system.

Enhanced Metadata Extraction for the Regional Validation Center

Nathan S. Netanyahu
University of Maryland College Park
Center for Automation Research (CFAR)
(nathan@nibbles.gsfc.nasa.gov)

Profile

Dr. Netanyahu holds a Bachelor of Science and Master of Science in electrical engineering from the Technion, Israel Institute of Technology and a diploma in computer science from Tel-Aviv University. He also holds a Master of Science and a Ph.D. in computer science from the University of Maryland at College Park. He is a Member of the IEEE.

From 1973 to 1978, Dr. Netanyahu served as a technical officer and project engineer in the Intelligence Unit of the Israel Defense Forces where he designed and developed electronic communication subsystems. From 1978 to 1985 he served as a senior project engineer in the Electronic Research Department at the Israeli Ministry of Defense where he designed and developed electronic communication systems and computerized test modules for their automatic performance evaluation.

While working on his advanced degrees at the University of Maryland, Dr. Netanyahu was employed as a research assistant by the Center for Automation Research's Computer Vision Laboratory. From 1992 to May of 1994 as a National Research Council Associate attached to NASA Goddard's Space Data and Computing Division (Code 930), he worked on unsupervised methods for clustering air/spaceborne multispectral images, and derived computationally efficient algorithms for robust statistical estimation.

Dr. Netanyahu joined CESDIS through a subcontract with the University of Maryland at College Park in May 1994. He has been working on supervised classification of remotely sensed images, and has continued to pursue unsupervised (robust estimation-based) clustering of multispectral images and computationally efficient algorithms for robust estimation.

Current research interests include algorithm design and analysis, computational geometry, image processing, pattern recognition, remote sensing, and robust statistical estimation.

Report

1. Introduction

Metadata extraction is emerging as a powerful tool that will enable a large community of users (Earth scientists, for example) to perform efficient "data mining", i.e., search (by content) and analyze extensive data sets.

Indeed, to prepare for the challenge of handling the archiving and querying of tera-byte sized scientific spatial databases efficiently, the Applied Information Sciences Branch (AISB), Code 935, NASA Goddard, has pursued a number of methods for extracting image content (or "metadata") from remotely sensed images. The methods pursued rely mainly on supervised (image) classification techniques, and their derivation comprises an integral part of the development of the Intelligent Information Fusion System (IIFS). The IIFS serves as the main building block of the recently developed end-to-end Regional Validation Center (RVC) information system.

The task(s) reported below are aimed at enhancing some of these supervised techniques (e.g., neural network-based), plus providing additional modules (e.g., efficient feature matching for image registration) to further extend the functional capability of the RVC. (See AISB's internal report, "Prototype Regional Validation Center: Description Package, Revision 2, for a detailed description of the RVC.) The algorithms developed/incorporated under these tasks should be suited for use by the metadata extraction modules available at an RVC, and as such these algorithms must be robust (performance-wise) and computationally efficient.

2. Enhanced Metadata Extraction from Remotely Sensed Imagery

We have focused primarily on artificial neural networks to perform (supervised) image classification. Based on previous studies carried out at the ISTB, we have suggested that extracting metadata from remotely sensed images in a fast and (relatively) accurate manner can be adequately achieved through a combination of a probabilistic neural network (PNN) and a backpropagation-trained neural network. (See CESDIS annual report 1994-1995.) Given its fast training time, the PNN serves to establish initially an ("optimal") training set that is representative of all the classes in a given (set of) scene(s) (e.g., those compiled at a regional data center), after which a (trained) backpropagation network is invoked (in its feed forward mode) to classify these scenes.

The PNN's learning process is based on nonparametric density estimation techniques. Specifically, the network estimates the probability density function of a newly introduced test pattern by computing for each class (that is present in the training set), a sum of Gaussians centered at each individual training pattern that belongs to the class, and evaluated at the test pattern. The pixel (or test pattern) is assigned to that class for which the above computation is the highest.

Taking advantage of the fact that the PNN lends itself naturally to a single instruction multiple data (SIMD) parallelization, we have developed a parallel version of the PNN on the massively parallel machine, the 16K processing element MasPar, MP-2. This has reduced significantly the run-time of the PNN. Furthermore, to enhance the performance of the PNN, we have been exploring the following additional directions.

2.1 Integration with Wavelet Parameters

When using a straightforward classification scheme, most of the classification errors are expected to occur at the boundary between classes. These errors seem to emanate from the fact that pixel-based classification methods (e.g., neural network-based) do not incorporate local, spatial information in the classification process. Originating from this premise, we have conducted a preliminary study as to the impact that such information could have on the overall classification. Specifically, to improve the performance (i.e., accuracy) of the classification module(s) discussed previously – it is estimated that the PNN provides approximately 70-80% accuracy – we have attempted to exploit *texture* information (in addition to spectral intensities), in the overall framework of a neural network-based classification scheme. (See Szu et al. (1997), for a detailed discussion.)

In general, texture may be captured through statistical, spectral, structural, or model-based methods. In particular, statistical methods rely on the spatial distribution and the spatial dependence among local gray tones. Spectral methods, on the other hand, extract relevant texture information by computing the energy associated with different frequency bands (e.g., through the use of Fourier transform). Recently, there has been a growing interest in wavelet transforms for texture extraction. Wavelet-based texture analysis can be regarded as both a statistical and spectral technique since an isotropic wavelet exploits the localization of a wavelet transform in computing (local) energy properties, and since it also provides a spatial density function of the so-called co-occurrence texture. Indeed, various recent studies have demonstrated the usefulness of applying wavelets in texture analysis.

In Szu et al. (1997) we pursued the following combined approach. First the multiband image(s) were preprocessed, retaining the most significant bands obtained by a principal component analysis (PCA). (For example, Figures 1(a)-(b) depict the two most significant PCA bands obtained for a Landsat-TM image over Washington, DC. Then, an isotropic, composite wavelet filter (which preserves edge information while emphasizing texture components) was applied to these PCA bands. (The specific wavelet transform was realized as a combination of a Mexican Hat and a Morlet type wavelet. Also, a number of scale values, *a*, were experimented with for each PCA band. See Figures 1(d)-(f).) The augmented set of images, i.e., the PCA bands plus the additional images produced by the above wavelet transform, were then fed (as new input) to the PNN. (Training sets, too, were augmented accordingly.) Preliminary classification results (see, e.g., Figures 1(g)-(h)) suggest that additional wavelet information may enable to refine the classification in regions where texture is well structured, e.g., urban areas. On the other hand, this information might not be as useful in regions where spectral information is sufficient to determine the classification (e.g., water). At any rate, we propose to pursue a number of related issues, as part of future research.

2.2 Mixed Pixel Classification

Usually classification procedures assume a "winner take all" strategy with a single pixel being placed within one land use/land cover category. While this is a valid assumption over vast homogeneous regions, there are many cases where a pixel actually consists of several ground cover elements. For example, the National Oceanic and Atmospheric Administration (NOAA)/AVHRR platform series consists of sensors having a spatial resolution of 1.1[km/pixel]. Given this (low) resolution, it is quite likely that several surface cover classes will be found within each pixel in the image. (We refer to such a pixel as a "mixed pixel".) Thus, in classifying mixed pixels, an alternative approach would be to assume that a pixel is composed of all possible end-members (i.e., pure classes) of a given scheme, e.g., that its observed (spectral) reflectance comprises a linear combination of the (spectral) reflectances of the various end-members. Determining the abundance (i.e., the relative fraction) of each component is known as the (linear) "mixture modeling problem". Providing a solution for this problem is considered more appropriate, in many cases, for the extraction of metadata/content from remotely sensed images.

Indeed, there have been a number of approaches proposed in the (remote sensing) literature for the spectral unmixing problem. For example, Shimabukuro and Smith (1991) and Settle and Drake (1993) present solutions to the problem that are based on the conventional principle of (constrained) least squares. However, the solutions provided are not general (the number of end-members is assumed smaller (e.g., three) than the number of spectral bands) and they are ad hoc (in the way the desired coefficients are guaranteed to be non-negative). A more recent paper by Bosdogianni et al. (1997) presents a similar approach, but it, too, lacks a general solution to the problem. (Extending the technique beyond the 4-class case considered is not obvious, and maintaining non-negative coefficients in the general case could become highly inefficient.)

In Chettri and Netanyahu (1996), we have proposed an alternative approach which is based on the principle of "maximum entropy". We provide much justification, from an information-theoretic and a combinatorial perspective, as to the appropriateness of this approach in the specific context of mixed pixel classification. The methodology pursued yields a general solution to the (linear) mixture modeling problem by ways of optimizing the "entropy" associated with the abundance vector. Also, it guarantees (by definition) non-negative fractional distributions.

The paper first introduces the mathematical formalism of linear mixture modeling and the principle of maximum entropy. It then presents algorithms for the solution of the (linear) mixture model both in the conventional case and in the case where a more sophisticated technique, the penalty function method is employed to optimize the entropy function. (The issue of noise is dealt with in both cases.) In addition, the paper makes an interesting connection to Bayesian methods, and compares the maximum entropy-based approach with standard regularization methods. Finally, empirical results are presented for real and

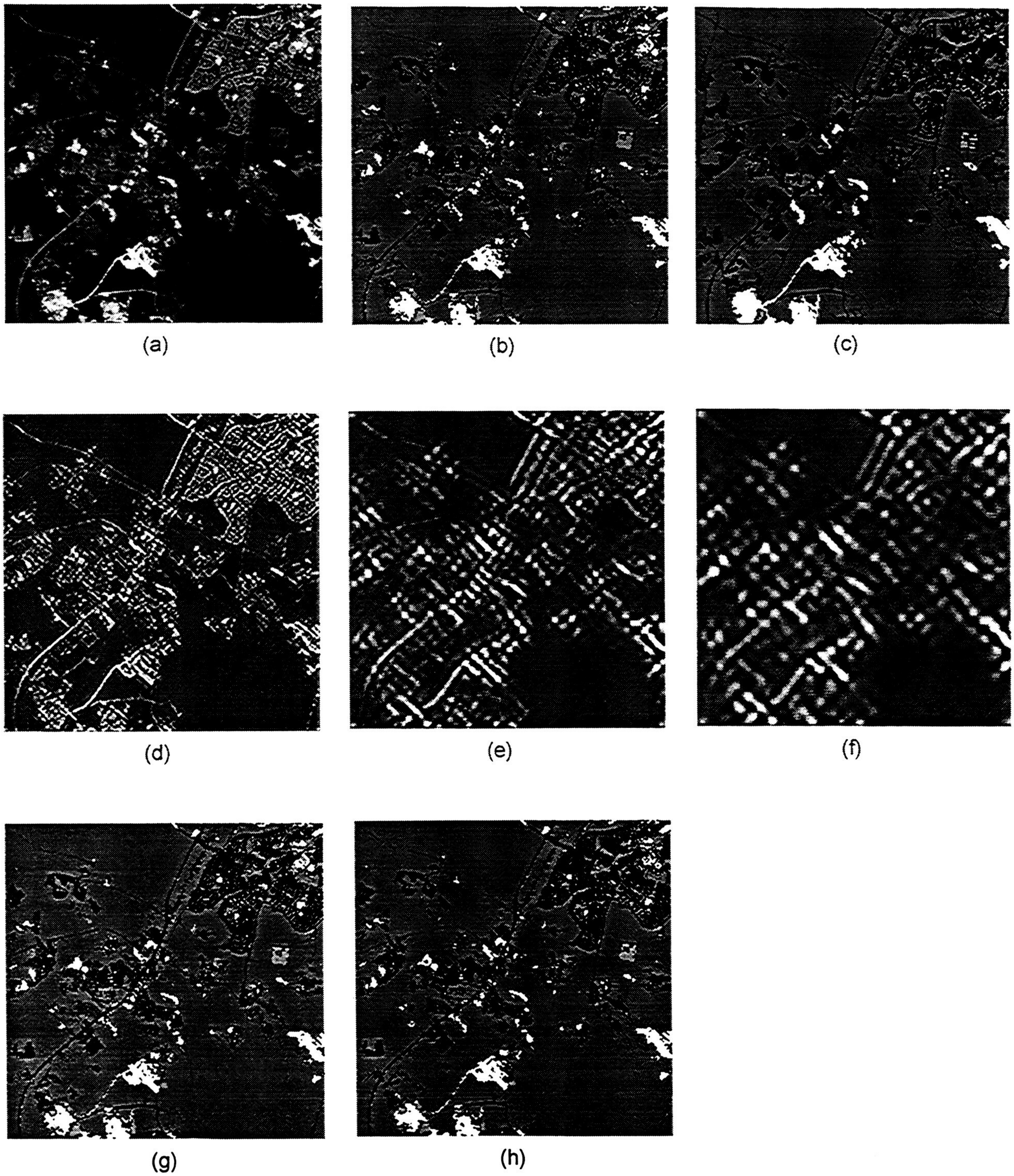


Figure 1: (a) Most significant band of a 256 X 256 Landsat-TM scene over Washington, DC; (b) second PCA band of the above scene; (c) ground truth image; (d) Wavelet transform (applied to most significant PCA band) for $a = 0.25$; (e) ditto, for $a = 0.5$; (f) ditto, for $a = 0.75$; (g) PNN-based classification; (h) ditto with spatial information (due to a wavelet transform).

simulated satellite data. (Real data were obtained from a large repository of AVHRR data and simulated data were obtained via realistic models of reflectance properties of different land surface types and the inclusion of atmospheric effects.) Our preliminary results suggest that the maximum entropy-based method yields higher accuracy, i.e., the method appears to be promising.

In an attempt to reduce the computational complexity associated with the maximum entropy method, we have proposed to extend the method in conjunction with the concept of multiresolution. The basic idea is to decompose (recursively) a given image to a set of lower resolution images (by applying, for example, well-known wavelet transforms, e.g., the Haar transform) and then carry out the maximum entropy unmixing technique with respect to one of these images (e.g., the so-called low/low (LL) image). Applying our technique to this smaller size image yields satisfactory fraction approximations (as compared to those obtained for the original image) while gaining a considerable speedup. (The resulting procedure is roughly 4 times faster.) See Chettri et al. (1997), for a detailed discussion.

3. Efficient Feature Matching

Many of the analysis techniques that an RVC is expected to utilize will involve the integration of multiple data sources. (For example, the analysis of global coverage by low-resolution data could be validated by using local, very high-resolution data.) Thus, to enable users to analyze large amounts of pertinent data sets more accurately and efficiently, a preliminary requirement is that the RVC contain a sound image registration scheme.

Various modules of such a scheme have been developed recently by a team of researchers at the AISB. The idea is to establish, essentially, a registration toolbox, i.e., a diverse set of tools for image registration which will be incorporated, eventually, into the RVC.

One of the fundamental building blocks in any (control point-based) registration scheme relies on matching features that are extracted from one image (the sensed image) to their counterparts in a second image (the reference image). (The extracted features could be points, edge segments, corners, etc.) Although feature-based methods tend to be relatively accurate (as features are more reliable than intensity or radiometric values), they could become computationally expensive.

Thus, to arrive at a sound registration that would be both accurate and fast, we have pursued an algorithmic methodology for a robust, feature-based matching module. The methodology is based largely on computational geometry techniques, e.g., geometric searching, hierarchical decomposition of spatial data structures, etc. An implementation of the methodology is currently underway, and is expected to yield an efficient, general purpose building block that could be incorporated into an end-to-end registration scheme, available under the RVC's registration toolbox. (See Mount, Netanyahu, and Le Moigne (1997), for details.)

4. Proposals/Grants

Co-investigator on NRA 97-MTPE-03, Proposal SRS/97-0082, "Intercomparison of Automated Registration Algorithms for Multiple Source Remote Sensing Data".

5. Awards

NASA/GSFC Group Achievement Award, in recognition of the development of a Regional Validation Center, May 1997.

6. Recent Publications

Robust Estimation of Parameters for Fitting Circular Arcs (with A.J. Stromberg, V. Philomin, and A. Rosenfeld), presented at a Symposium on Statistics and the Sciences, Halifax, Nova Scotia, Canada, August 12-16, 1996.

Robust Detection of Road Segments in Noisy Aerial Images (with V. Philomin, A. Rosenfeld, and A.J. Stromberg), to appear in *Pattern Recognition*; see also *Proceedings of the Thirteenth IAPR International Conference on Pattern Recognition*, Vienna, Austria, August 25-30, 1996, Vol. B, pp. 151-155.

Learning in Navigation: Goal Finding in Graphs (with P. Cucka and A. Rosenfeld), *International Journal of Pattern Recognition and Artificial Intelligence*, special issue, in memory of King-Sun Fu, Vol. 10, No. 5, pp. 429-446, August 1996.

Wavelet Index of Texture for Artificial Neural Network Classification of Landsat Images (with H.H. Szu, J. Le Moigne, and M. Francis), in *Proceedings of the Twenty Fifth Annual SPIE Applied Imagery Pattern Recognition Workshop*, Washington, DC, October 16-18, 1996, pp. 36-44.

Spectral Unmixing of Remotely Sensed Images using Maximum Entropy (with S. Chettri), in *Proceedings of the Twenty Fifth Annual SPIE AIPR Workshop on Emerging Applications of Computer Vision*, Washington, DC, October 16-18, 1996, pp. 55-62.

A Practical Approximation Algorithm for the LMS Line Estimator (with D.M. Mount, K. Romanik, R. Silverman, and A. Wu), in *Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, Louisiana, January 5-7, 1997, pp. 473-482.

Integration of Local Texture Information in the Automatic Classification of Landsat Images (with H.H. Szu, J. Le Moigne, and C. Hsu), in *Proceedings of the SPIE Conference on Wavelet Applications*, Orlando, Florida, April 22-24, 1997, pp. 116-127.

Multiresolution Maximum Entropy Spectral Unmixing (with S. Chettri, J. Garegnani, J. Robinson, P. Coronado, R.F. Cromp, and W.J. Campbell), to appear in *Proceedings of the International Symposium on Artificial Intelligence, Robotics, and Automation in Space*, Tokyo, Japan, July 14-16, 1997.

A Registration Toolbox for Multi-Source Remote Sensing Applications (with W. Xia, J. Le Moigne, J.C. Tilton, B-T. Lerner, E. Kaymaz, J. Pierce, S. Raghavan, S. Chettri, T. El-Ghazawi, M. Manohar, W.J. Campbell, and R.F. Cromp), submitted to the International Conference on Earth Observation and Environmental Information, Alexandria, Egypt, October 13-16, 1997.

An Efficient Algorithm for Robust Feature Matching (with D.M. Mount and J. Le Moigne), abstract submitted to the CESDIS Image Registration Workshop, NASA Goddard Space Flight Center, Greenbelt, MD, November 20-21, 1997.

Beowulf Parallel Workstation

Phillip Merkey, Senior Scientist (merk@cesdis.gsfc.nasa.gov)
Donald Becker, Staff Scientist (becker@cesdis.edu)
Daniel Ridge, Technical Specialist (newt@cesdis.edu)

Profiles

Phillip Merkey

Dr. Merkey holds a Bachelor of Science degree in mathematics from Michigan Technological University, and took a Ph.D. in mathematics in the area of algebraic coding theory from the University of Illinois (1986). He is a member of the AMS and SIAM.

Prior to joining CESDIS in 1994, Dr. Merkey was employed as a research staff member by the IDA Supercomputing Research Center in a classified working environment. His experience includes application of high performance computers to grand challenge problems, investigation of instruction level parallelism using the VLIW parallel computer, benchmarking experiments on the Multiflow Trace computer, algorithmic design for empirical solutions to problems in applied discrete mathematics, and innovative parallel implementations of advanced algorithms.

Dr. Merkey concluded the evaluation effort the Convex Exemplar 1000 by contributing to summary article in IEEE *Computer*. Since then he has assume the role of technical lead on the Beowulf Bulk Data Server project. He is responsible for the overall design and progress on the project. He also responsible for identifying and evaluating applications that will be suitable applications to demonstrate the machine capabilities and guide its development.

Dr. Merkey has also engaged in outside collaborations with the IDA Center for Computing Sciences and has served as an instructor at the University of Maryland Baltimore County.

Donald Becker

Mr. Becker holds a Bachelor of Science degree from the Massachusetts Institute of Technology in electrical engineering and has completed graduate computer science courses at the University of Maryland College Park. From 1987 to 1990 he was employed by Harris Corporation, Advanced Technology Department, Electronic Systems Sector as a senior engineer. He performed research and development work on the Concert multiprocessor, maintained and extended the Concert C compiler (based on PCC) and libraries, and wrote network software.

As a research staff member of the IDA Supercomputing Research Center from 1990 to 1994, Mr. Becker wrote a substantial proton of the low-level LINUX networking code, designed, implemented, and characterized an interfile optimization system for the GNU C compiler, implemented a peephole optimizer for a data-parallel compiler (DBC), and implemented several symbolic logic applications.

Since joining CESDIS in 1994, Mr. Becker has proven to be among the most important contributors to CESDIS. He has established a world class reputation in the operating system community with his contributions in networking software.

Mr. Becker is the principal investigator for system software on the Beowulf Parallel Workstation project, a program of research to study the potential for and the methods of harnessing commercial mass market grade computing for high end scientific workstation applications.

Daniel Ridge

Mr. Ridge is working on his undergraduate degrees in computer science and aerospace engineering at the University of Maryland College Park. He began working with Donald Becker at CESDIS in 1995 and in 1996 took a leave of absence from Maryland to work as a Technical Specialist on the Beowulf project in 1996.

Mr. Ridge has shown himself to be among the most important contributors to CESDIS and Beowulf project. In addition to developing system software for the Beowulf Workstation and the Beowulf Bulk Data Server, Mr. Ridge has been active in the worldwide Beowulf user community. He is also available to provide advise and tutorials for more than 10 Beowulf class systems being constructed worldwide. On a strictly experiment basis, Mr. Ridge has collaborated with UMCP to bring Beowulf/Linux up on their principal parallel platform (40 Digital Alpha CPUS).

Report

CESDIS continues to maintain a leadership role in the cluster computing community. The Beowulf project was presented (by Donald Becker) as a keynote session at IEEE Aerospace '97 and a CESDIS/JPL/ Caltech collaboration produced a tutorial for the Cluster Computing Conference in Atlanta. In addition to these presentations and published articles the Beowulf project has built and maintains a significant Web presents.

The Beowulf/LINUX WWW site hosted by CESDIS has registered in excess of 45,000 hits. CESDIS WWW server statistics are always available on-line at <http://cesdis.gsfc.nasa.gov/web-stats/overview.html> We started the Web based Beowulf University Consortium to make it easier to collaborate with others in the academic Beowulf community. We have established and moderate several majordomo mailing lists on Beowulf related topics.

Beowulf system software is increasingly integrated into the RedHat LINUX distribution. RedHat has an installed base (conservatively) of 2 million users worldwide. We have cultivated a close working relationship with the development team at RedHat to ensure that future Beowulf software technologies are deployed as rapidly as possible.

The Beowulf Bulk Data Server project is a three year effort to develop a built to scale prototype secondary storage system. The system will be a 64 processor cluster with an hierarchical interconnection network that will provide a aggregate volume of a terabyte of disk storage and will deliver bulk data packets at the rate of a gigabyte/sec.

The Beowulf Bulk Data Server project is a rare project in the sense that it is jointly funded by NASA and DARPA. Dr. Sterling, former Head of the Scalable Systems Technology Branch of CESDIS deserves credit for identifying and securing the opportunity to develop a project that directly addresses the concerns of both organizations. The project provides a trivially replicatable solution to the problem of bandwidth I/O required by many high performance applications. NASA applications that involve large data sets and large simulations, which in effect produce large data sets, will directly benefit from this work.

In spite of difficulties securing appropriate staff and unexplainable delays in procurement, the Beowulf Bulk Data Server project has made significant progress and currently enjoys strong momentum. The low-level software development, for example, global process control and prefetch for virtual memory, will play an important role in the Bulk Data Server as well as the Workstation. We have constructed the Phase one system; the figures show the cluster as it was being assembled. This skeletal system has 60 CPU with 250 GBytes of disk connected with Fast Ethernet. This will be fully populated with disk in phase two and with the high bandwidth network in phase three.

Publications

How to Build a Beowulf: A Tutorial, 3/9/97, Cluster Computing Conference (CCC'97), Emory University

Daniel Ridge, Donald Becker, Phillip Merkey, Thomas Sterling, Beowulf: Harnessing the power of parallelism in a pile-of-PCs. *Proceedings, IEEE Aerospace*, 1997

Sterling, T., Merkey, P., Savarese, D., Improving application performance on the HP/Convex Exemplar, IEEE *Computer*, Vol 29, No. 12, pp. 50-55, Dec 1996.

Bergman, K., Burdge, G., Carlson, D., Coletti, N., Jordan, H., Kannan, R., Lee, K., Merkey, P., Prucnal, P., Reed, C., Straub, D., Optical sorting, the fast fourier transform, and data packing, *Proceedings of MPPOI 96*.

HPCC/ESS Evaluation Project

Terrence Pratt, Senior Scientist
(pratt@cesdis.edu)

Profile

Dr. Pratt earned B.A., M.A., and Ph.D. degrees in mathematics and computer science at the University of Texas at Austin. He is a member of the ACM, the IEEE, and SIAM. In 1972-73 he served as an ACM National Lecturer, and in 1977-78 a SIAM Visiting Lecturer. His research interests include parallel computation, programming languages, and the theory of programming.

Prior to joining CESDIS, Dr. Pratt held teaching and research positions at Michigan State University in East Lansing, the University of Texas at Austin, and the University of Virginia. At the latter he was one of the founders of the Institute for Parallel Computation and served as its first director.

During the 1980s, Dr. Pratt worked with scientists at USRA's ICASE and NASA Langley on the development of languages and environments for parallel computers. He is the author of two books: Programming Languages: Design and Implementation (Prentice-Hall, second edition, 1984) and Pascal: A New Introduction to Computer Science (Prentice-Hall, 1990).

Dr. Pratt joined CESDIS as the Associate Director in October 1992 and was appointed Acting Director in October 1993 upon the retirement of Raymond Miller. He served in that capacity until November 1994 when he left CESDIS to pursue other interests, but maintained ties with CESDIS as a consultant on high performance Fortran. He rejoined CESDIS as a Senior Scientist early in 1996.

Report

This research project is part of the NASA HPCC Earth and space science (ESS) project centered at Goddard. The ESS project funds nine "grand challenge" science teams at various universities and federal research laboratories. In addition, through a cooperative agreement with SGI/Cray, a 512 processor SGI/Cray T3E parallel computing system has been placed at Goddard to serve as a testbed system in support of the science team projects. Each science team is responsible for developing large scale science simulation codes to run on the T3E and meet specified performance milestones (10 Gflop/sec in 1996, 50 in 1997, 100 in 1998). The codes are provided to an in-house science team at Goddard for performance verification, and ultimately the codes are submitted to the National HPCC Software Exchange for general distribution. For an overall view of the NASA HPCC/ESS project and its current status, visit the web page at http://sdcd.gsfc.nasa.gov/ESS/.

1. Research Goals

The CESDIS Evaluation Project is concerned with the large scale science simulation codes produced by the nine Grand Challenge science teams, their behavior on the massively parallel testbed computer system, and to a lesser extent their behavior on other parallel systems such as the CESDIS Beowulf systems. We expect to work with about 10-15 different science codes in total.

Our interest is in understanding how these large science codes stress the parallel system and how the parallel system responds to these stresses. In particular, we wish to find ways to:

- Quantify the stresses produced by the science codes on the testbed hardware and software.
- Quantify the performance responses produced by the system.
- Determine the causes of the observed responses in the codes and systems.
- Use the results to improve codes and systems.
- Develop new performance evaluation and prediction methods and tools as needed.

Ultimately the goal is publication of the results of this work in various journals and conference proceedings.

2. Approach

Our approach is to work directly with the science codes as they are submitted by the science teams to meet performance milestones. We use various measurement tools to understand the static structure of each code and its dynamic behavior when executed with a typical data set (also provided by the science team). Typically, a code is "instrumented" to collect the desired statistics and timings, and then run on the testbed system using various numbers of processing nodes. The results are analyzed, and if more data is required, the instrumentation is modified and the code rerun.

The insights gained from this research on a particular code often lead to understandings about how to improve the performance of the code. These insights are fed back to the science team to aid them in further development of the code. Results may also be useful to SGI/Cray in improving their hardware and software systems, so results are often forwarded to the in-house SGI/Cray team and the in-house science team.

3. Measurements of Interest

Part of the research effort is to determine what aspects of science code structure and behavior have the greatest effect on performance. To this end, we are measuring some of the following elements in each code:

- Flops counts and rates.
- Timings and execution counts of interesting code segments.
- Data flows between code segments.
- MPI/shmem/PVM message passing and synchronization profiles.
- I/O activity profiles.
- Cache use issues.
- Storage allocation sizes and use profiles.
- Scaling with problem size and number of processors.
- Load balance.

4. Tools Used

These studies use a variety of tools for instrumenting and measuring various characteristics of the science codes and their behavior. The primary tool to date has been a software system called Godiva (Goddard

Instrumentation Visualizer and Analyzer) developed by this project. We also use the SGI/Cray Apprentice software tool on the T3E and are investigating other tools from universities and national laboratories that might prove of use, such as Pablo from the University of Illinois and AIMS from the NASA Ames Research Center.

5. Current Status and Results

The 1996-97 year was primarily spent on development of the Godiva software system and on preparatory work to understand the research issues involved in this study. The Godiva software now runs on the SGI/ Cray T3E and T3D, the CESDIS Beowulf cluster machines, and Sun workstations.

The first science codes became available in the Fall of 1996, and a preliminary SGI/Cray T3D system was installed in October 1996. The T3E system arrived with 256 processors in March 1997; it was upgraded to 512 processors in June 1997.

Initially, we are making a pass through all of the science codes, as submitted to meet the 10 Gflop/sec milestone, in order to understand the various code designs and the instrumentation and measurement issues involved. Four codes have received serious study to date, with a more cursory look at two others. Issues studied have included code size (codes range up to 50,000 lines in size, making many forms of instrumentation difficult), language used (codes so far have included Fortran 77, C, and Fortran 90), cache use in key loops, parallel communication and synchronization (using MPI, PVM, and shmem libraries), and flops rates in selected code segments. In several cases, the Godiva software was extended to allow new forms of measurement and display.

Several studies of aspects of the NAS Benchmarks (Version 2 using MPI) have also been made in order to develop methodology and to understand research issues in these well-known benchmark codes.

No major studies have been completed, but these preliminary studies have resulted in some useful insights and results that have been fed back to the science teams and the SGI/Cray in-house team. Of particular note:

- As a result of study of the T3E cache behavior, and in collaboration with S. Swift of SGI/ Cray, we found how to speed up the key computation loop in the TERRA code of J. Baumgardner (LANL) from 70 megaflop/sec to 210 megaflop/sec, leading to an overall 20% decrease in the execution time of the code.

- As a result of study of the MPI profile, we suggested an improved parallel transpose algorithm to R. Dahlburg (NRL) for use in his CRUNCH-3D code.

- As a result of studying the timings of key segments in the treecode of K. Olson (Goddard), we suggested minor changes that resulted in an 82% speedup of the code.

We hope that a flow of these small scale results can be continuously fed to the science teams as the evaluation project proceeds, with larger, more general insights and understandings packaged as journal and conference papers.

6. Godiva Software Instrumentation Tool

The Godiva software system, developed as part of this project, has proven to be a useful new tool for the study of large science codes. Using Godiva, a wide variety of aspects of a code may be instrumented so that the dynamic behavior may be observed as the program executes. Of particular importance to date

have been the ability to study cache behavior on the T3E, computation (flop/sec) rates in selected code segments, parallel communication and synchronization profiles
using MPI, PVM, or shmem library calls, and load balance among processors.

Godiva has been developed as a personal research tool, not intended for general distribution, but it has been made available to T. Clune of the SGI/Cray in-house team, who has used it in studies of the loop behavior of the FCT code of R. DeVore (NRL). It will be made available to other researchers as appropriate. Because it is a personal research tool, it undergoes frequent change to meet the demands and new directions of the evaluation project.

The approach to code instrumentation used in Godiva is as follows. First, selected parts of the code are annotated to study whatever characteristics are of interest. These annotations use a syntax specified in the Godiva Users Manual. Annotations appear as comments to a Fortran or C compiler. The annotated code is fed through the Godiva preprocessor, which generates Fortran or C source code with calls to the Godiva run-time library inserted at appropriate points. The generated source program is then compiled and linked with the Godiva run-time library. Execution of the program generates a trace file on each processor. The trace file contains statistics collected on-the-fly during execution. Tracing overhead is generally quite low for typical statistics of interest (less than 5% additional execution time on most runs, but dependent on the user's choice of data to be gathered). After execution is complete, a Godiva postprocessor is used to generate tables, graphs, and histograms from the trace files produced by the processing nodes.
Currently Godiva supports about 30 different annotation types in the source program. These annotations may be used to generate about 20 different forms of output tables and graphs.

For more information about the CESDIS evaluation project and the Godiva software system, visit the web site at http://cesdis.gsfc.nasa.gov/people/pratt/atlanta.html. Included at this site are samples of Godiva annotations and output
tables and graphs.

7. Conclusion

The evaluation project is proceeding well. The Godiva software tool is proving useful, and good access to large scale science codes and to the T3E has been provided by the NASA HPCC/ESS Project. Collaborations with several members of the SGI/Cray in-house team, the Goddard ESS in-house team, and the members of the science teams have begun to develop. Useful small-scale results have been produced and disseminated. The outlines of more general insights and results are beginning to emerge.

Architecture Adaptive Computing Environment (aCe)

Udaya A. Ranawake
University of Maryland Baltimore County
Department of Computer Science and Electrical Engineering
(udaya@neumann.gsfc.nasa.gov)

Profile

Dr. Ranawake received a B.Sc. degree in electrical engineering from the University of Moratuwa in Sri Lanka in 1982, and an M.S. in electrical engineering and Ph.D. in computer engineering from Oregon State University in 1987 and 1992 respectively. Prior to joining CESDIS on a subcontract with the Department of Computer Science and Electrical Engineering at the University of Maryland Baltimore County, he was a Senior Member of the technical staff at Hughes STX Corporation where he was the leader of the massively parallel processing (MPP) research task at NASA GSFC.

While a graduate student at Oregon State, Dr. Ranawake worked as a teaching assistant in the Department of Electrical and Computer Engineering, served as the Micro Computer Lab manager, and was a research associate on a two-year software project for implementing efficient parallel algorithms for the Monte Carlo simulation of a semiconductor device.

Dr. Ranawake is a member of the IEEE. His research interests include parallel and distributed computing, algorithms for scientific computation, compiler technology, computer architecture, and computer networks.

Report

1. Introduction

Parallel computers are playing an important role in satisfying the growing computational needs of scientific and commercial applications. The research community has recognized the need for efficient and easy to use languages for programming such computer systems. Data parallel languages which express parallelism through the simultaneous application of a single operation to a data set have received considerable attention in recent years and several such languages are under development. This report describes the ongoing activities in the aCe compiler project at NASA GSFC.

2. The aCe Programming Language Overview

aCe is an extension of the C programming language designed to help users develop portable parallel programs for massively parallel computer systems. These extensions include:

- A method for defining a virtual architecture specific to an algorithm and for declaring parallel variables belonging to that virtual architecture.

- Overloading of standard C operators and introducing several new operators to specify operations on parallel data and communication between parallel variables.

- Methods for selecting the parallel variables, and a specific set of elements of a parallel variable, upon which aCe code is to act.

- Definitions for partitioning and aligning of parallel variables.

aCe translates a source program into an ANSI C-based high level language. For example, on MIMD computers, an aCe program is translated into an ANSI C program augmented with message passing constructs to a runtime library routine based on PVM 3.3 to handle inter-processor communications.

3. Accomplishments

A source level debugger was implemented to aid in the debugging of parallel programs written in aCe. It is implemented by modifying the back end of the aCe compiler to output a C program that executes each source statement conditionally in a true data parallel fashion. The conditional execution is used to check whether break points are set and to take appropriate actions based on user commands. The debugger currently supports a basic set of user commands such as stepping through a program, setting/resetting of break points, printing information about identifiers, displaying values of variables, and performing xy plots or pixel maps. As the debugger works by interpreting each source statement in a true data parallel fashion, the inner working of an aCe program compiled for debugging will be different from the same program compiled for normal execution. Nevertheless, it provides the user with a valuable software tool to easily find the bugs of an aCe program. A program that works correctly under the debugger should also work correctly when compiled for normal execution.

The design goals for the debugger were ease of use, portability, and scalability. To ensure ease of use, the debugger is equipped with a graphical user interface and data visualization/animation facilities using xy plots and pixel maps. In addition, the single-threaded program image and the global name space provided by the aCe language adds to the ease of use of the debugger. The debugger is scalable because the latency on user interactions (for displaying the value of a variable for example) scales with the machine size and the intrusiveness in the debugger instrumentation (for conditional execution of each source statement and entering variables into the symbol table) adds only a fixed overhead. It could be easily ported to different platforms by modifying the software module responsible for gathering data for visualization and animation.

Currently, the debugger is available for serial workstations. Also, the manual "dbaCe: A Source Level Debugger for the aCe Programming Language" was written and made available on the Web at http://newton.gsfc.nasa.gov/aCe. This manual provides a complete overview of the debugger.

The aCe compiler was ported to the HAT (heterogeneous architecture testbed) workstation cluster at NASA GSFC. HAT consists of a six node Intel Pentium processor connected by a 100Mbps Fast Ethernet internal network. Porting the compiler involved modifying the aCe run-time library implemented using PVM3.3 to work with the new version of the aCe compiler. The run-time library is comprised of routines for handling data partitioning, handling communication primitives such as scatter, gather, broadcast, and reduce operations, and performing input/output operations.

Wavelet-Based Image Registration on Coarse-Grain Parallel Computers*

Tarek A. El-Ghazawi
The George Washington University
Department of Electrical Engineering and Computer Science
(tarek@seas.gwu.edu)

Statement of Work

Dr. El-Ghazawi has been tasked with supporting the development of high performance implementations of wavelet-based processing for NASA Earth science imagery by:

- Investigating and implementing wavelet decomposition and wavelet-based registration on architectures such as the Cray T3D and T3E, the J90, the Convex SPP, and the SGI Power Challenge, and
- Beginning the evaluation of Field Programmable Gate Array (FPGA) reconfigurable architectures for the purpose of wavelet decomposition and registration.

Related work calls for the evaluation of the performance of the Beowulf technology and a demonstration of the scalability of implementing automatic image registration using wavelets on the Beowulf platform. Affordability and real-time processing are two major requirements.

Report

*(This work has been conducted in collaboration with J. Le Moigne and P. Chalermwat. Workload characterization has been conducted in collaboration with T. Sterling and A. Meajil. The T3D and T3E parallel implementation work was funded by CESDIS Task 71; performance evaluation work was supported by Task 31.)

Given the tremendous amount of data generated from the MTPE remote sensing satellite systems and the number of images to be registered, the time required to register input images to existing reference images is quite large. This prompted the need for parallel processing to conquer the lengthy computation time.

In this report, we present the computational savings in image registration resulting from using the Wavelet-based technique which exploits the multiresolution property of Wavelet. This iterative refinement Wavelet-based method avoids exhaustive searches for the relative orientation of images in terms of rotation and 2-D translation. We also present our coarse-grained parallel image registration, which has been developed and evaluated using data from the Landsat/Thematic Mapper TM on the Cray T3D, Cray T3E, and the Beowulf network of Pentium workstations. Our parallel mapping of the algorithms uses a hierarchical domain decomposition in order to allocate each image correlation to one or more processors. For fewer number of processors, each processor performs one or more image registrations, followed by reduction to find the best estimates of rotation and 2-D translation. For larger number of processors, each processor becomes concerned only with computations that belong to a row subimage, and a two step hierarchical reduction is used to integrate the correlation data over local groups of processors working over the same image, followed by a global reduction step to find the best registration parameters.

It will be analytically shown that the Wavelet-based iterative refinement method, provides a great deal of computational savings and, therefore, represents an efficient starting point for our parallel image registration, and that our hierarchical decomposition of the problem provides a scalable parallel implementation.

Furthermore, performance measurements from the Cray T3D, Cray T3E, and the Beowulf (cluster of 16 Pentium-based PC's connected by a dual 100 Mbps Ethernet) show scalability characteristics as well as the overhead. It will be shown that the implementation is indeed scalable and the measurements expose a number of interesting comparative architectural characteristics for the underlying hardware architectures.

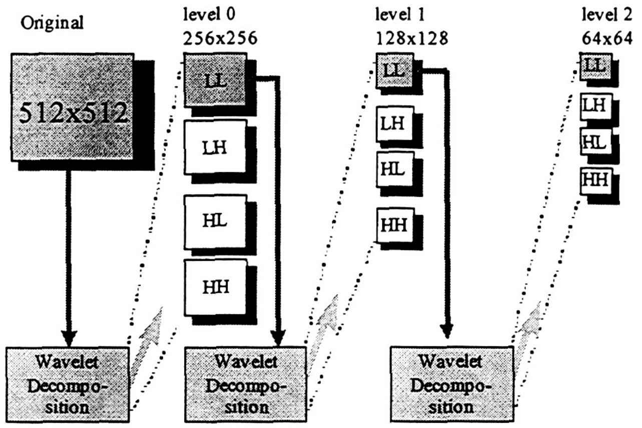
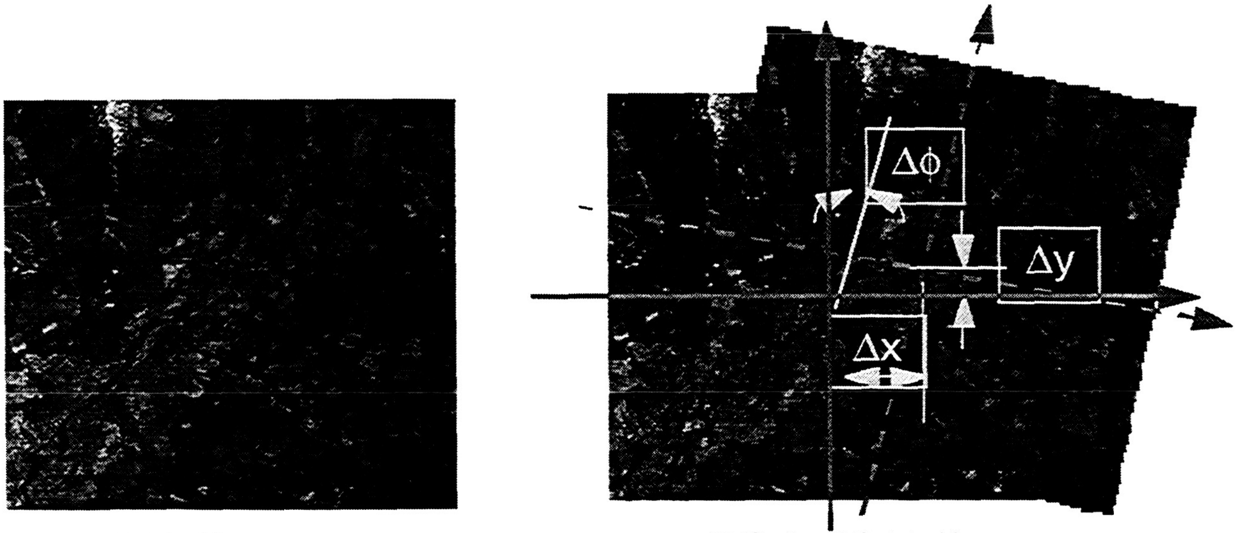


Figure 1: Wavelet Decomposition at Different Levels

Wavelet decomposition applies low and high filters along the rows and columns, followed by decimation operations to extract the low and high frequency contents of the image along the rows and columns with a lower resolution (figure 1). For example, LL is the filtered version of the original image where low pass filters are applied along the rows and the columns and each dimension is decimated by two, where the lower resolution allows seeing low frequency variations more easily. Thus LL has one fourth of the size of the original image. The process is repeated for several levels of such decomposition.



Original Image

Shifted and Rotated Image

Figure 2: Image registration

Image registration determines the best relative orientation of two images in terms of rotation and 2-D translation in the x and y directions (figure 2).

Instead of searching linearly for the best orientation by computing the image correlation for all possible rotation, x and y shifts, the iterative refinement algorithm searches for these parameters at the highest level of decomposition (the smaller image) with a coarse resolution. The result from the highest level of decomposition is used as an input to the previous level (higher resolution image) which is searched with a higher resolution to further refine the search result, and so on (see table 1).

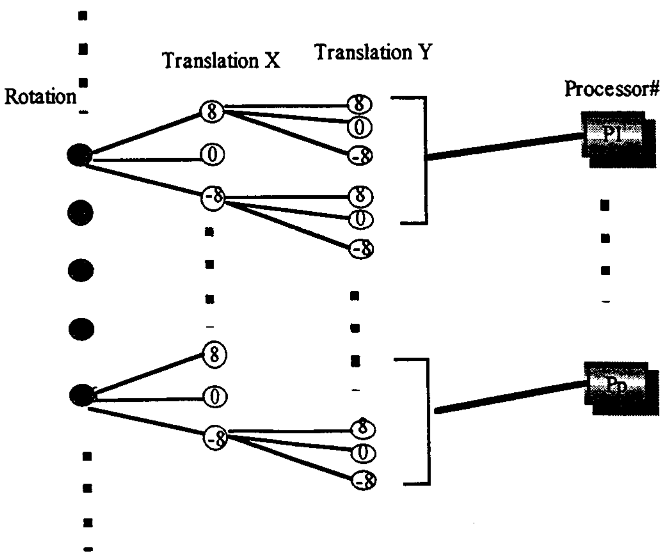
Level	Size	Rotation	Increment $\Delta\theta$	Tx, Ty	Increment Δx, Δy	Best θ, Tx, Ty
3	32x32	0±16	8	0,0	8,8	θ3, Tx3, Ty3
2	64x64	θ3±8	4	Tx3±8, Ty3±8	4,4	θ2, Tx2, Ty2
1	128x128	θ2±4	2	Tx2±4, Ty2±4	2,2	θ1, Tx1, Ty1
0	256x256	θ1±2	1	Tx1±2, Ty1±2	1,1	θ0, Tx0, Ty0

Table 1: Iterative Refinement of Image Registration

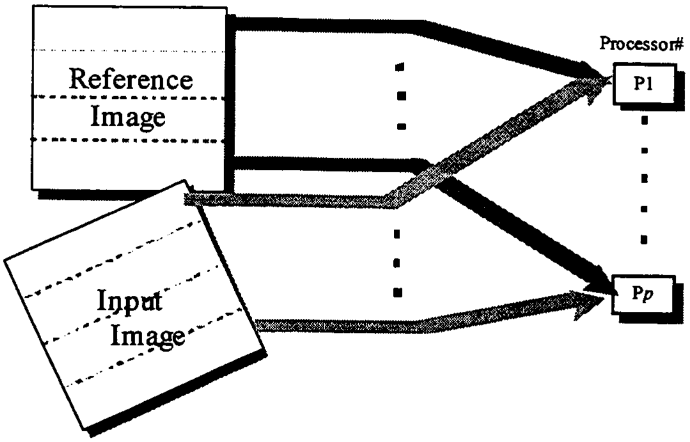
Consider the case of the iterative refinement algorithm as in the table with search starting at ± 16 points for theta, x, and y. Then, a linear search would yield a total of 35,937 correlations. The iterative refinement algorithm, however, uses four levels and in each level would compute only 125 image correlations, a speed of roughly 72 due to the selection of the sequential algorithm. In general, it can be shown that a speedup factor of

$$\text{Speedup} = (2R + 1)^a / (L\,5^a)$$

is achieved by using the Wavelet-based iterative refinement versus linear search, where the search region is ± R points for each of the a attributes (e.g., rotation degrees, x-translation pixels, and y-translation pixels), and where L levels of decomposition are used.



3a. Computation decomposition when number of processors is less than number of registrations.



3b. Data decomposition when number of processor is greater than number of registrations.

Figure 3: Mapping the Image Registration to Parallel Processors

For the parallel implementation, a hierarchical domain decomposition will be used to allocate images to a group of one or more processors. This part will be implemented in the next year. For fewer numbers of processors, each processor performs one or more image registrations, followed by reduction to find the best estimates of rotation and 2-D translation (figure 3a). For larger numbers of processors, each processor becomes only concerned with computations that belong to a row subimage, and a two step hierarchical reduction is used to integrate the correlation data over local groups of processors working over the same image, followed by a global reduction step to find the best registration parameters (figure 3b).

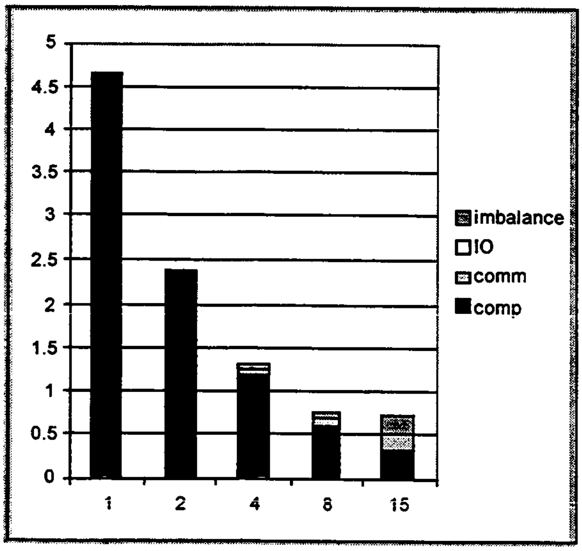


Figure 4: Execution Budget on the Beowulf Architecture

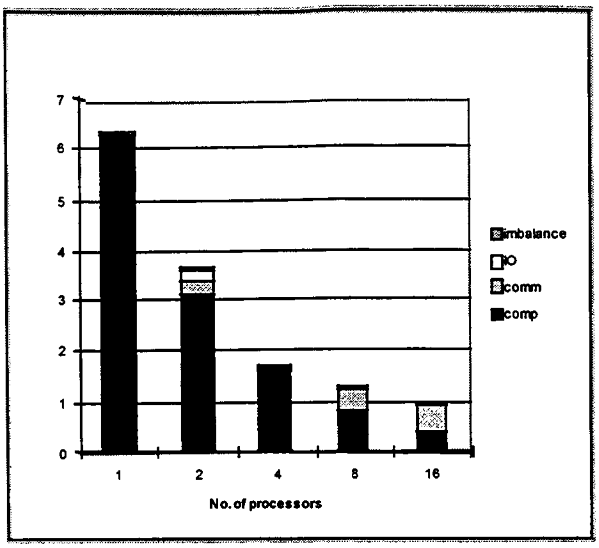


Figure 5: Execution on Cray T3D

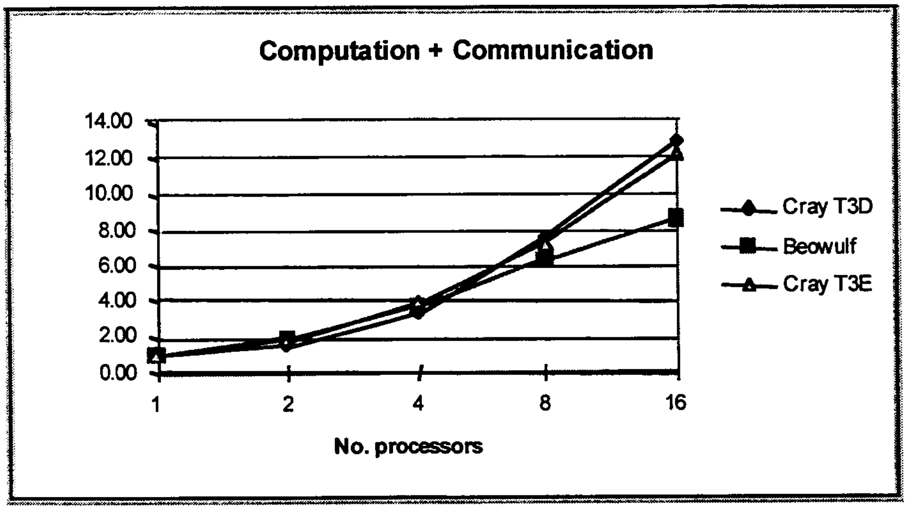


Figure 6: Comparison of Scalability on Beowulf, Cray T3D, and Cray T3E

The experimental measurements (figures 4-6), show that the selected parallel mapping scales well on these architectures. Only 16 nodes were used, to allow fair comparison with the Beowulf architecture. Beowulf shows worst scalability and efficiency is nearly 50% at the full configuration of 16 processors due to the limited communications bandwidth. Both Cray T3D and T3E show similar scalability characteristics, in spite of the fact that the T3E has a much faster processor. However, the scalability of the T3D is better than that of the T3E. This is because the network remained almost the same, while the processor of the T3E is much faster and, therefore, requires more communications. Due to the high integer processing requirements of the underlying application, Beowulf (100 MHz Pentium based) has been generally faster than the T3D (150 MHz DEC Alpha Processors), but was not faster than the T3E on absolute wall clock timing basis (see tables 2 and 3). Beowulf, however, has a clear performance cost advantage for such small configurations.

PEs	Cray T3E	Cray T3D	Beowulf
1	2.382	6.709	4.6635
2	1.303	4.284	2.39
4	0.604	2.022	1.279
8	0.33	0.896	0.739
16	0.196	0.52	0.545

Table 2: Wall Clock Timing in Seconds for the Test Problem

System	PEs	SPECint92	SPECfp92
Beowulf	Pentium 90 MHz	90	70
Cray T3D	64-bit DEC 21064 (Alpha EV4) 150 MHz	77	110
Cray T3E	64-bit DEC 21164 300 MHz & 450MHz	161.	194

Table 3: Speed Rates of the Underlying Processors

In addition, a workload characterization model for massively parallel computer architectures was developed and used for characterizing the NPB and predicting its performance on massive parallel systems. [6] - [11]

References

[1] El-Ghazawi, T., Charlemwat, P., and Le Moigne, J. (November 1997) Wavelet-Based Image Registration on Parallel Computers. *Supercomputing'97*, San Jose. (Accepted).

[2] Le Moigne, J. (1995, July 10-14). Towards a Parallel Registration of Multiple Resolution Remote Sensing Data, *Proceedings of IGARSS'95*, Firenze, Italy,

[3] El-Ghazawi, T. and Le Moigne, Jacqueline. (August 1994) Multiresolution Wavelet Decomposition on the MasPar Massively Parallel System. *Journal of Computers and Their Applications, Vol. 1*, NO. 1.

[4] Berry, Mike and El-Ghazawi, Tarek. (1996, April) Parallel Input/Output Characteristics of NASA Science Applications, *Proceedings of the International Parallel Processing Symposium (IPPS'96), IEEE Computer Society Press*, Honolulu.

[5] Le Moigne, J. and Tilton, J. C. (May 1995) Refining Image Segmentation by Integration of Edge and Region Data. *IEEE Transactions on Geoscience and Remote Sensing, Vol. 33*, No. 3, pp. 605-615.

[6] Meajil, A., El-Ghazawi, T. and Sterling, T. Characterizing and Representing Workloads for Parallel Computer Architectures. *Journal of Systems Architecture*. (Accepted)

[7] Meajil, A., El-Ghazawi, T. and Sterling, T. Performance Prediction Based on Workload Characterization. *The Supercomputer Journal*. (Accepted)

[8] Meajil, A., El-Ghazawi, T. and Sterling, T. (1997, March) An Architecture-Independent Workload Characterization Model for Parallel Computer Architectures, *Proceedings of the Aizu International Symposium on Parallel Algorithms and Architecture Synthesis (PAS-97), IEEE Computer Society Press*, Aizu, Japan.

[9] Meajil, Abdullah I. and El-Ghazawi, Tarek. (1997, March 14) A Framework for Performance Prediction of Parallel Systems Based on Workload Similarity, *The Eighth SIAM Conference on Parallel Processing for Scientific Computing, PP '97*, Minneapolis, Minnesota.

[10] Meajil, Abdullah, El-Ghazawi, Tarek, and Sterling, Thomas. (1996, August) A Quantitative Approach for Architecture-Invariant Workload Characterization, *Lecture Notes in Computer Science, Springer-Verlag, Berlin, November 1996. Presented at (PARA'96) Applied Parallel Computing*, Copenhagen.

[11] Meajil, Abdullah. (April 1997) Workload Characterization for Parallel Computer Architectures. *GWU Ph.D. Thesis (Directed by Tarek El-Ghazawi).*

Community Climate Model on Goddard Computing Facilities

Jules Kouatchou
The George Washington University
Department of Mathematics
(kouatcho@math.gwu.edu)

Statement of Work

Mr. Kouatchou was tasked with installing, testing, validating, and benchmarking parallel climate models on scalable computing systems such as the Cray T3D.

Report

To predict future climate, researchers rely on climate models that involve large systems of equations and require massively parallel computers to solve these systems. The National Center for Atmospheric Research (NCAR) has developed a climate model known as Community Climate Model (CCM3) that is a stable, efficient atmospheric general circulation model designed for climate research on high-speed computers.

The goal of this study is to port the CCM3 code to the Cray T3E and to provide an in-depth insight for researchers on the use and performance of the CCM3 Model at NASA's Center for Computational Sciences (NCCS) computing facilities at Goddard Space Flight Center.

During this study, the following accomplishments were made.

- CCM3 code was ported on the Cray T3E;
- Specific information was given to familiarize potential users with the overall execution procedures of CCM3, as well as how to access, compile, load, execute the code, and store the model output on a Cray J90 and Cray T3D and T3E systems;
- Some performance analysis of CCM3 was provided for a better understanding of the execution of the model under various configurations.

In addition, we have been providing technical assistance to users who employ CCM3 in its present form and to users who want to add new modules to the model.

Additional information is available in the following publication, available at the URL cited below:

Hudson, A. & Kouatchou, J., "*Implementing CCM3 with NASA Goddard Space Flight Center Parallel Computer Systems*". (Available at http://farside.gsfc.nasa.gov/RIB/repositories/PI_Kouatch)

ESDIS Project on High-Performance I/O Techniques

Joel Saltz, Anurag Acharya, Alan Sussman

University of Maryland College Park
Department of Computer Science
(saltz@cs.umd.edu, acha@cs.umd.edu, als@cs.umd.edu)

1. Overview

The goals of this project were two-fold: to understand the I/O requirements of algorithms for data product generation and to develop techniques that help meet these requirements on suitably configured machines. Towards the first goal, we have analyzed a variety of existing data product generation programs and have successfully parallelized two of them: Pathfinder and Climate which generate the Level 2 and Level 3 data products for the AVHRR Land Pathfinder effort. We have also developed a parallel I/O library suitable for parallel data generation programs. We will describe our experiences in Section 2. We will also present our suggestions regarding the structure of EOSDIS data product generation programs, the organization of the files used to store the data products and the runtime support needed for effective parallelization of data product generation applications.

Based on our understanding of the I/O and processing requirements of these applications, we have developed several techniques to help meet them on suitably configured machines. These techniques deal with (1) declustering multi-dimensional datasets on large disk farms, (2) partitioning satellite data products for efficient retrieval, (3) overlapping I/O, computation and communication to perform data retrieval and processing on the same processors, and (4) interprocedural analysis to automate the placement of asynchronous I/O operations. We describe these techniques in Sections 4 and 5. Based on these techniques, we have developed Titan, a high-performance database for remote sensing data. The computational platform for Titan is a 16-processor IBM SP-2 with four fast disks attached to each processor. Titan is currently operational and contains about 24~GB of AVHRR data on the NOAA-7 satellite. Titan supports interactive queries over its data and supports full-globe queries as well localized queries. Experimental results show that Titan provides good performance for global queries, and interactive response times for localized queries. We describe the design and evaluation of Titan in Section 3.

Based on our experience with Titan, we are currently in the process of developing an extensible framework for managing extremely large multi-dimensional datasets. We plan to implement this framework both as a stand-alone system for efficient storage, retrieval, and processing of large data repositories and as a database extender which allows multi-dimensional datasets to be integrated with commercial relational databases which store other forms of data, in particular metadata associated with the datasets.

2. Analysis and parallelization of data product generation

We collected data product generation programs from four groups: (1) Pathfinder and Climate from the GSFC DAAC (Code 902); (2) gaps and extract from the GIMMS group (Code 923); (3) the SeaWiFS processing chain from the SeaWiFS project; and (4) LAS and ADAPS from the EROS Data Center. From each source, we acquired multiple programs that formed a processing chain which took Level 1 data as input and generated Level 2 and Level 3 data products.

In spite of significant differences in the science algorithms and the organization of the code, all these applications have a common structure. Programs that generate Level 2 products process Level 1 files which contain information from a single satellite orbit and generate a single file that contains a composited multiband image for the area of interest. Before composition, individual values are corrected for various distortions and are navigated to the desired projection and resolution. The composition operation is a complex max-reduction operation - the specific predicate used to determine when one pixel is preferable to another depends on the program and the dataset. The reduction operation is performed by: (1) creating a temporary image; (2) processing all the inputs with a fixed chunk size;(3) processing and navigating the IFOVs in a chunk; (4) performing the max-reduction. Given the similarities, we selected Pathfinder and Climate as the prototypical programs for the generation of Level 2 and Level 3 data products.

We parallelized Pathfinder by partitioning the output image in equal-sized horizontal strips. Each processor is responsible for all processing needed to generate its partition of the output image. We chose to partition the output image (instead of the input data) as this allows all combination operations to be local to individual processors. No inter-processor communication is needed. We chose a horizontal partitioning scheme to take advantage of the row-major storage format used in all files (input, ancillary as well as output files). Horizontal striping allows I/O to be performed in large contiguous blocks.

Each processor computes the map from the input data set to the output image by subsampling (one scan line per chunk) all input files. It then reads the chunks that intersect with its partition of the output image. For each chunk, it maps each input pixel into the output image. Pixels that map into its partition are processed further, others are ignored. The individual partitions of the output image are also too large to be stored in main memory. Therefore, the composition operation is still out-of-core. Once all processing is completed, the final result is produced by concatenating the individual partitions.

In Climate, the mapping between the pixels of the input image and those of the output image is data-independent and can be computed a priori. The amount of computation to be done is proportional to the amount of input data. We parallelized Climate by horizontally partitioning the output image. Each processor reads the data that maps to its partition of the output image. Load balance is achieved by ensuring that all processors read approximately equal amounts of data.

The total I/O performed by Pathfinder is over 28GB and the total I/O performed by Climate is 75.5MB. The original version of Pathfinder ran for 18,800 seconds on a single processor of an IBM SP-2. Of this, about 13,600 seconds (76% of the time) was spent waiting for I/O. The final version took 1200 seconds using 12 processors. Of this, 10-15% time was spent waiting for I/O – Pathfinder is now computation-bound. The maximum aggregate application level I/O rate was 644 MB/s. For Climate, the execution time was reduced from 200 seconds to 32 seconds (on eight processors) of which 4-5% was spent waiting for I/O. The maximum aggregate application-level I/O bandwidth for Climate was 36 MB/s. These experiments were conducted on an IBM SP-2 which has been configured with 16 thin nodes, two Fast/Wide SCSI adaptors per node, and three IBM Starfire 7200 disks (7 MB/s application-level I/O bandwidth) per SCSI adaptor. More details about this tuning and evaluation effort can be found in our paper titled "Tuning the Performance of I/O-intensive Parallel Applications" which appeared in the *Proceedings of Fourth Annual Workshop on I/O in Parallel and Distributed Systems* (IOPADS'96).

3. Titan: a high-performance database for remote sensing data

We have designed, implemented, and evaluated Titan, a parallel shared-nothing database designed for handling remote sensing data. Titan consists of two parts: (1) a front-end that interacts with querying clients, performs initial query processing, and partitions data retrieval and computation; and (2) a back-end that retrieves the data and performs post-processing and composition operations. The front-end consists of a single host which can be located anywhere on the network. The back-end consists of a set of processing nodes on a dedicated network that stores the data and does the computation. The current implementation of Titan uses one node of the 16-processor IBM SP-2 as the front-end and the remaining 15 nodes as the back-end. No data is stored on the disks of the node used as the front-end.

Titan partitions its data set into coarse-grained data blocks and uses a simplified R-tree to index these chunks. This index is stored at the front-end which uses it to build a plan for the retrieval and processing of the required data blocks. The size of this index for 24 GB of AVHRR data is 11.6 MB, which is small enough to be held in primary memory.

Titan queries specify four constraints: (1) temporal bounds (a range in universal coordinated time), (2) spatial bounds (a quadrilateral on the surface of the globe), (3) sensor type and number, and (4) resolution of the output image. The result of a query is a multi-band image. Each pixel in the resulting image is generated by composition over all the sensor readings for the corresponding area on the Earth's surface.

When the front-end receives a query, it searches the index for all data blocks that intersect with the query. It uses the location information for each block (which is stored in the index) to determine the set of data blocks to be retrieved by each back-end node. In addition, the front-end partitions the output image among all the back-end nodes. Currently, the output image is evenly partitioned by blocks of rows and columns, assigning each back-end node approximately the same number of output pixels. Under this partitioning scheme, data blocks residing on the disks of a node may be processed by other nodes; each back-end node processes the data blocks corresponding to its partition of the output image. The front-end distributes the data block requests and output image partitions to all back-end nodes.

Each back-end node computes a schedule for retrieving the blocks from its disks. This schedule tries to balance the needs of all nodes that will process these data blocks. As soon as a data block arrives in primary memory, it is dispatched to all nodes that will process it. Once a data block is available for processing (either retrieved from local disk or forwarded by another node), a simple quadrature scheme is used to search for sensor readings that intersect with the local partition of the output image. After all data blocks have been processed, the output image can either be returned to the front-end for forwarding to the querying client, or it can be stored in a file for later retrieval.

Data layout decisions in Titan were motivated by the format of AVHRR data and the common query patterns identified by NASA researchers and our collaborators in the University of Maryland Geography Department. We distributed the AVHRR data on a large disk farm. We used the declustering algorithms described in Section 3 to compute the data distribution.

Titan is currently operational on a 16-processor IBM SP-2 with four IBM Starfire 7200 disks attached to each processor. It contains about 24 GB of AVHRR data from the NOAA-7 satellite.

We have run a sequence of experiments on Titan to evaluate our techniques for partitioning the images into chunks, declustering the chunks over a large disk farm, and placing the chunks assigned to individual disks. Experimental results show that Titan provides good performance for global queries and interactive response times for local queries. A global query for a 10-day composite of normalized vegetation index takes less than 100 seconds; similar queries for Australia and the United Kingdom take four seconds and 1.5 seconds respectively. Our data distribution techniques improved the disk parallelism, the number of disks active for individual queries, by 48 to 70%. The total estimated retrieval time was reduced by between 8 and 33%. We also evaluated schemes for placement of data blocks assigned to a single disk. We found that the average length of a read (without an intervening seek) can be improved by about a factor of two. Design, implementation, and evaluation of Titan has been described in our paper titled "Titan: A High-Performance Remote-sensing Database" which appeared in the *Proceedings of the International Conference on Data Engineering*, 1997.

Based on our experience with Titan, we are currently in the process of developing an extensible framework for managing extremely large multi-dimensional datasets. We plan to implement this framework both as a stand-alone system for efficient storage, retrieval, and processing of large data repositories and as a database extender which allows multi-dimensional datasets to be integrated with commercial relational databases which store other forms of data, in particular metadata associated with the datasets.

4. Declustering algorithms for multi-dimensional range queries

We investigated data declustering techniques for multidimensional datasets with the primary goal of minimizing response time and the secondary goal of maximizing disk space utilization. First, we extended the three best-known index-based schemes (Disk-Modulo, Fieldwise-XOR, and Hilbert-Curve) for declustering Cartesian product files to grid files which allow better utilization of disk space. Using simulation experiments, we showed that the scalability of Disk-Modulo and Fieldwise-XOR for multidimensional range queries is limited. That is, as the number of disks is increased beyond a threshold, the response time no longer decreases. This result is corroborated by an analytical study. The response time for Hibert-Curve scales better than Disk-Modulo or Fieldwise-XOR, but the difference between its performance and the best possible performance increases with the degree of skew in the data distribution. As an alternative to the index-based schemes, we developed a declustering algorithm based on a proximity measure. To evaluate this algorithm, we compared its performance with that of the three index-based schemes mentioned above as well as other proximity-based schemes in the literature. Our evaluation was based on two real datasets and three synthetic datasets. The real datasets were: (1) a sequence of snapshots from a Direct Simulation Monte Carlo code (from NASA Langley) and (2) stock price data for a basket of stocks over a period of time. This algorithm has also been used in the Titan database described in Section 2. Results from our simulation experiments indicate that the proposed algorithm achieves better declustering than the algorithms we compared it to, particularly for configurations with large number of disks. This research has been described in our paper titled "Study of Scalable Declustering Algorithms for Parallel Grid Files" which was presented at the Tenth International Parallel Processing Symposium (IPPS'96).

5. Interprocedural analysis for placement of I/O operations

We developed an Interprocedural Balanced Code Placement technique for compiler placement of I/O calls. The goal of this technique is to maximize the overlap between I/O and computation. Each synchronous I/O operation is replaced by a balanced pair of asynchronous operations. The asynchronous operations are placed so as to achieve overlap with the computation, while maintaining correctness and safety. To be able to overlap disk accesses with computation, it is important for the compiler to analyze code across procedure boundaries. We implemented a Fortran source-to-source transformation tool which performs the IBCP analysis. We used this tool to compile "satellite", a satellite-data processing template based on Pathfinder which repeatedly modifies an out-of-core image. Our results show that use of compiler-placed asynchronous write operations can reduce the I/O overhead for this template by 25%-40% and improve the overall performance by 13.3%-14.7%. Performing interprocedural analysis for placement was critical for getting better performance; almost no overlap would have been possible if the analysis was restricted intraprocedurally. These results have been included in our paper on "An Interprocedural Framework for Placement of Asynchronous I/O Operations" which was presented at the 1996 ACM International Conference on Supercomputing (ICS'96).

6. Lessons learned and suggestions

In this section, we briefly present the lessons learned from our experience. We believe they would be useful to science algorithm developers as well as to people working on configuring the hardware and software systems.

- Satellite data generation programs are relatively easy to parallelize.

  They are easy to parallelize given the common structure of different data product generation programs. We believe that the parallelization scheme described in this report should be suitable for most, if not all, data product generation programs. Since communication between peers is needed only for putting together the final output, this scheme should work as well on shared memory as on distributed memory machines.

- Proper code restructuring is important.

  As far as possible, I/O should be done in the outermost nest of a nested loop. Embedding I/O calls in inner nests of a nested loops usually results in a sequence of small requests interleaved with seeks. It is usually possible to restructure the loop nests so that the I/O is performed in the outermost nest and only computation done in the inner nests. This restructuring is illustrated by the following example which is based on the composition module of Pathfinder. For the applications we studied, this was not a difficult operation.

- Original code.

  ```
  for (i = 0; i < num_input_pixels; i++)
  for (j = 0; j < num_of_bands; j++)
  map input pixel to output image
  seek to output pixel in this band
  write pixel value for this band
  ```

- Restructured code.

  ```
  determine the bounding box output pixels involved

  for (j = 0; j < num_bands; j++)
  read in the bounding box for this band
  for (i = 0; i < num_input_pixels; i++)
  map input pixel to output image
  update output image pixel for this band
  write out the bounding box for this band
  ```

- Information about future requests is usually available.

  In the parallelization scheme described above, processors subsample the input files in the partitioning phase. At the end of this phase, every processor has complete information about its future requests for input reads. For the modified version of the out-of-core max-reduction (where modification consisted of a pair of simple loop-splitting and loop-reordering transformations), information about updates to all frequency bands of the output image is known before any updates are performed.

- It is possible to partition the intermediate data so that each processor reads and writes to its own local disk(s).

  Bandwidths for local disk access are substantially higher than the bandwidths for non-local accesses. In addition, local accesses are guaranteed not to interfere with I/O requests from other processors. This increases the utility of the file cache and makes the overall behavior of the application more predictable.

- stdio is usually not suitable for satellite data processing.

  Many programs use the fread/fwrite interface for I/O which introduces an extra level of buffering and requires one extra copy. Since individual requests are usually large enough, the buffering performed by fread/fwrite does not provide a performance advantage and the read/write interface is likely to provide additional benefit.

- Geo-location information should be placed at the top of file.

  In the absence of information that can be used to map the IFOVs contained in a file, our parallelization scheme is forced to have each processor subsample all the files. This is inefficient and limits scalability.

Providing geo-location information at the beginning of the file would allow each processor to read data proportional to the number of files.

- Diskful machines are important.

  Diskful machines (machines with local disks) allow problems to be partitioned such that most of the I/O requests are satisfied by local disks. As noted above, local disk accesses have a higher application-level bandwidth with the associated benefit of guaranteed lack of contention for the disk and the file cache. In combination with code restructuring to exploit locality, diskful machines can improve both the I/O performance and the overall execution time for out-of-core applications.

- Complex I/O interfaces are not required.

  After code restructuring, most requests in the studied applications were large. For large requests, the interface is usually less important. Small strided requests were a recurrent pattern in the original versions of Pathfinder and Climate. However, we found that these patterns were caused by the embedding of small I/O requests in the innermost loops. Relatively straightforward loop restructuring, including loop splitting, interchanging the order of nested loops, and fusing multiple requests were sufficient to coalesce these requests into large block I/O requests. None of the applications studied required collective I/O. This is not surprising given the size of the requests after code restructuring. All of the applications are parallelized in SPMD fashion. In our Earth science applications, all processes are independent (apart from initial and possibly final synchronization). Independent I/O requests were able to utilize the servers when they would have been idle in a collective-I/O model.

- Compiler-directed placement of I/O operations can be eliminate I/O waiting time.

  Our experiments showed that complete overlap of the write operations with computation can be achieved through flow-sensitive interprocedural analysis. Note that almost no overlap would have been possible if the analysis was restricted to within single procedures.

- The declustering algorithm mentioned above scales well.

  This is true as the number of disks is increased and consistently achieves a better response time compared to all the other algorithms (with a few exceptions when the number of disks is small). It also achieves perfect data balance and maximizes the disk space utilization. Furthermore, it rarely maps buckets that are close in the data space to the same disk indicating that the distributions it generates are probably quite close to the optimal distribution.

UNIVERSITY RESEARCH PROGRAM IN PARALLEL COMPUTING

In December 1992, CESDIS issued a call for proposals in parallel computing to be sponsored with funding ($50K per year for two to three years) from the NASA High Performance Computing and Communications program. The research program was intended to bring together under CESDIS sponsorship, university researchers and NASA applications scientists to connect on-going basic computer science research programs in selected areas of high performance computing with NASA applications in Earth and space sciences.

The goal was to build a national group of researchers interested in collaborating on key problem areas in parallel computing that affect NASA's efforts to collect, manage, store and process massive Earth and space science data sets. Participants have been provided access to a range of high performance parallel computer systems at Goddard including the Cray MPP, the TCM CM-5, and the Intel Paragon. Research topics of particular interest were the high performance I/O systems and storage systems attached to these testbed parallel machines.

Ten proposals in three major areas were selected for funding:

- Parallel I/O System Design

  Clemson University, Walter Ligon
  George Washington University, Tarek El-Ghazawi
  Syracuse University, Geoffrey Fox
  University of Illinois, Daniel Reed
  University of Minnesota, Matthew O'Keefe

- High Performance Scientific DBMS

  University of Florida, Theodore Johnson
  University of Virginia, James French
  University of Washington, Linda Shapiro
  University of Wisconsin, David DeWitt

- Intelligent Data Management

  University of Texas at Arlington, Diane Cook

Final reports from the 10 principal investigators follow.

CLEMSON UNIVERSITY

High Performance Input/Output for Parallel Computer Systems

Walter B. Ligon III
Department of Electrical and Computer Science
(walt@eng.clemson.edu)

1. Highlights

- Developed TPAW parallel system simulator for studying parallel architecture performance.
- Developed PVFS parallel file system including several distinct user interfaces useful for target applications.
- Parallelized AVHRR calibration and navigation code.
- Parallelized PNN image classification code.
- Developed out-of-core numerical routines.
- Studied algorithm performance on TPAW, Clemson DCPC, and CESDIS Beowulf.
- Explored using parallel I/O for distributed objected-oriented database.

2. Project Goals

The goal of our project is to study the I/O characteristics of parallel applications used in Earth science data processing systems such as Regional Data Centers (RDCs) or EOSDIS. Our approach is to study the runtime behavior of typical programs and the effect of key parameters of the I/O subsystem both under simulation and with direct experimentation on parallel systems. Our three year activity has focused on two items: developing a test bed that facilitates experimentation with parallel I/O, and studying representative programs from the Earth science data processing application domain. The Parallel Virtual File System (PVFS) has been developed for use on a number of platforms including the Tiger Parallel Architecture Workbench (TPAW) simulator, the Intel Paragon, a cluster of DEC Alpha workstations, and the Beowulf system (at CESDIS). PVFS provides considerable flexibility in configuring I/O in a Unix-like environment. Access to key performance parameters facilitates experimentation. We have studied several key applications from levels 1, 2, and 3 of the typical RDC processing scenario including instrument calibration and navigation, image classification, and numerical modeling codes. We have also considered large-scale scientific database codes used to organize image data.

The stated goals for this projects are as follows:

- Develop an understanding of the performance potential of I/O architectures with respect to the requirements of EOS applications,
- Provide a mechanism to integrate this understanding into the RDC or DAAC processing environments at NASA,
- Conduct an evaluation of parallel I/O architectures relative to the requirements of parallel applications, and
- Integrate the results of this study into the IIFS test bed being developed by NASA Code 930.1.

Our efforts have focused on three issues:

1) Developing an experimental environment for the study,

2) Studying the performance of several applications, and

3) Integration with NASA GSFC Code 935 IIFS software.

Our experimental environment consists of two major components. The first and primary one is a simulation environment for studying the effects of various architectural parameters on I/O performance. For this component we are building on the existing infrastructure in the Reconfigurable Architecture Workbench, which was developed under a previous research program. The second is a small-scale parallel computing system utilizing clusters of high-performance workstations built through an equipment grant from the NSF. This second facility is much less flexible in its capabilities, but does provide a means for validating some of the results obtained under simulation. This second approach also provides a view of the potential to be had in low-cost systems for Earth science computations.

Another critical component of our experimental environment is the means to configure and access a complex parallel I/O subsystem from a parallel application. To this end we have developed the Parallel Virtual File System (PVFS). PVFS controls access to files that are distributed across multiple I/O nodes in a parallel system by multiple compute nodes in the system. PVFS is not a true file system in that it uses the native file system of the host operating system for disk block allocation and management. PVFS manages only those aspects that pertain to parallel files and parallel access to a file. In addition to providing simple file access, we have explored a number of different interfaces to a parallel file system including a new type of interface called the scheduled I/O interface. Because PVFS sits on top of a standard Unix system interface, it is easily moved between a number of host platforms and parallel system simulators.

In our study we consider parallel systems as a collection of processors each with their own local memory and a network that provides basic message passing capability. This network can be a simple as an Ethernet, or as complex as cross bar, torus, or hypercube networks. I/O in these systems are provided by some number of I/O processors, where each I/O processor is a processor with local memory and directly attached storage devices such as disks and tape drives. I/O processors can be dedicated or used for computation. The number of I/O processors can be the same or different than the number of compute processors. I/O devices are attached to the I/O processors via a device bus (such as a SCSI bus) and a device controller. One or more devices can be connected to the device bus, and one or more device controllers can be attached to a given I/O processor.

Our performance models include device behavior (focusing on Winchester disk and digital audio tape (DAT) devices), I/O bus performance, and interconnection network performance. The load placed on these facilities is determined by the behavior of software both at the system and application levels. System software consists primarily of device driver codes, message passing libraries, and file system software. Of these, the parallel file system code is unique to this system model. In addition, key design choices in the message passing software may have an impact on I/O performance. Key issues include the amount of data copying performed and details of the networking protocol.

The main focus of our application study is on processing level zero telemetry data to levels 1 and 2, data product generation, and automatic metadata extraction. Level 1 and 2 processing algorithms include sensor calibration, georegistration, correction, and enhancement. Data product generation is highly application specific, but would include such things as vegetation index, snow and ice cover, sea surface temperature, atmospheric ozone content, etc. Metadata extraction is the process of preparing data sets for inclusion in an Earth science database by generating browse products and summary data. These activities would comprise a large share of the constant processing requirements for a typical RVC (Regional Validation Center) or DAAC (Distributed Active Archive Center) scenario in both preprocessing and reprocessing modes. Additional applications include out-of-core numerical methods used in processing high-level general circulation models (GCMs) and distributed object-oriented database codes used to organize and access Earth science data in an RVC or DAAC.

In order to explore integration of our results with ongoing efforts of NASA/GSFC Code 935, we have built an RVC consisting of a GOES satellite ground station, a remote sensing data product database, a mass storage system, and a Beowulf cluster of 18 Pentium PCs. We are expecting to add an HRPT ground station in the near future. Using these facilities, we are experimenting with storing data received at the ground station in a parallel file system on the Beowulf cluster, and issuing processing requests to the Beowulf cluster for parallel execution. The object database is maintained on the existing RVC platform, although we are considering a parallel implementation in the near future.

3. Experimental Environment

Development activities for our experimental environment included the development of the TPAW simulation environment and the PVFS parallel file system. In addition, these systems were adapted to the study of parallel architectures based on clusters of workstations (COWs) or piles of PCs (POPCs) such as the Clemson Dedicated Cluster Parallel Computer (DCPC) and the CESDIS Beowulf architecture. The following sections describe these development activities.

3.1 TPAW Simulation Environment

A key drawback to RAW as a tool for studying I/O systems is that RAW's processor simulator uses instruction-level simulation, which is to say target programs are compiled to an abstract assembly language and interpreted by the simulator. This was designed as such in order to facilitate the study of reconfigurable processing elements in a previous research program. In order to study the I/O system, it is important that considerably longer programs that are practical to study with such a system (due to the long simulation time) are used. Since processor instruction set is not a key issue in this study, the processor simulator has been replaced with a new module that executes the target application compiled to binary code suitable for the host on which the simulation is to be run. This results in a simulation that is several orders of magnitude faster than under the old system. Available systems that utilize such a technique are limited in that they only work for one host processor type and are limited to MIMD, and in some cases only shared memory architectures. In order to maintain RAW's flexibility in these areas, a new simulation tool was developed. This tool uses a source-code augmentation technique that maintains a high degree of portability.

All simulator code is written in C, and few vendor-dependent system calls are used. Where possible, development utilized POSIX compliant calls. This system simulates both SIMD and MIMD architectures and focuses on message passing systems (though shared memory is supported as well). Control and Network modules from the RAW simulator transfer readily to the new platform and the PVFS file system and I/O device models have been developed with the new simulator in mind. The details of TPAW have been documented in a Masters thesis by Mr. R. Agnew and a paper co-authored by Mr. Agnew and the Principle Investigator.

3.2 PVFS - Parallel Virtual File System

There are four purposes behind the design and implementation of PVFS. First, the system is designed to avoid common bottlenecks. TCP connections are used to pass data instead of a message passing library. Files used to store the parallel data are mmap()'d or accessed with low-level Unix read() and write() calls. Second, the software is designed to be portable. While the system will not run in a heterogeneous environment, the software was written using standard Unix system calls. Currently the software will run on both the DEC Alpha running OSF/1 and the Intel 80x86 processor running Linux. Third, the system is designed with a familiar user interface. The calls used to open, close, read, and write parallel files mimic their Unix counterparts, and modifications to a file's metadata are made with a call similar to an ioctl(). Because of this, many applications can be made to take advantage of PVFS simply by changing the file I/

O calls in the program to the PVFS versions. Finally, the system is designed to be flexible. The user has a great deal of control over the distribution of a file across the partitions in the file system and the view that an application has of this file. PVFS consists of four major components: the pvfsd, the pvfsmgr, the iod, and the library of calls used by applications. Each of these will be discussed in turn.

3.3 The PVFS Daemon

The purpose of the pvfsd is twofold; it serves as a link between applications on a single machine and the pvfs manager (pvfsmgr), and it provides information on mounted parallel file systems to these applications. The pvfsd runs on all machines where applications might be run. It stores data on mounted file systems and their associated managers. It also passes on requests to mount or unmount file systems, and open, close, or unlink files, to the manager. The pvfsd accepts connections from applications at a specified port. Applications use library calls to connect to this port and make requests using only the full path to a parallel file. No information about the location of the file on other machines is necessary; all mapping is handled by the system.

When a request is made to mount a parallel file system, the pvfsd will first determine the address of the appropriate manager. This is done using information passed on by the mount program if available. If not, a file similar to the fstab file used by mount is searched for an entry corresponding to the mount point. It then connects to the correct manager and requests that the file system be mounted. If successful, the address of the manager is cached along with the name of the file system.

When a file is opened, the pvfsd determines the file system being accessed by matching part of the path to the file with a pvfs mount point. Once a match is made, a connection is re-established with the manager, and the addresses of the I/O daemons for that file are passed back from the manager to the application. In addition, a file ID is also returned. The file ID is used by the application when referring to a parallel file instead of the filename. At this point, the application can directly access the I/O daemons via the addresses returned from the PVFS manager.

3.4 The PVFS Manager

The pvfs manager (pvfsmgr) is responsible for all functions related to iods. This includes starting and stopping the iods on remote systems when necessary, and servicing file access requests (except read and write) from applications by passing the necessary commands on to the appropriate iods. In order to keep track of the iods associated with a given file system, the iodtab file is used.

The iodtab file is similar to the fstab file used by mount. For each file system managed, a set of permissions, a path to the metadata file, and a list of iods is kept in this file. For each iod, the path to the parallel files themselves is also recorded. This path may be different for each iod listed for a file system, and an iod may be listed any number of times, both for a given file system and for other file systems. This allows multiple iods to run on the same machine, possibly serving parallel files off of different disks or serving separate parallel file systems altogether.

When a request to mount a file system is made the pvfsmgr will start an iod on each machine listed for that file system. Only if all iods are successfully started will the mount itself complete. Otherwise, an error will be reported. Requests to open a file result in messages being passed to all the iods from the pvfsmgr to open the parallel file. If all of the iods successfully open the file, a file ID is passed back to the application (through the pvfsd) along with the addresses of all the iods involved. The pvfsmgr will again be contacted when the file needs to be closed. It will then report this to the iods.

3.5 Metadata

It is the responsibility of the PVFS manager to distribute the metadata describing a file when it is opened. All metadata pertaining to files on a file system is kept in a single file. This metadata file describes the distribution of files across the partitions in a parallel file system. It contains the standard information on a file, the user id of the owner and permissions, as well as additional information specific to a parallel file system. This information includes:

- number of nodes (partitions) the file is spread across
- the base node
- the stripe size

These parameters are used by both the iods and the application to determine the location of data in a file. For each physical partition in a file system an iod is running whenever the file system is mounted.

3.6 The I/O Daemon

It is the responsibility of the I/O daemon (iod) to handle all file I/O for its partition. This includes reading, writing, creating, and removing parts of the parallel files contained on its partition. The iod is started by the pvfsmgr when a file system is mounted, who passes on information telling the iod about the file system for which it will provide service. Requests to read or write a file are sent directly to the iod from the application, and use a file ID, not the parallel file name. This file ID is obtained by the application from the pvfsmgr when the file is opened. The iod uses memory mapped I/O to speed file access when possible. This memory mapped region grows and shifts with file access, and if the file is opened for writing an area at the end of the file is preallocated in expectation of a growing file. This is especially useful when a file is initially created, as it eliminates extraneous mapping during file writing. The maximum size of this window can be set to prevent the iod from hogging too much memory. Using the metadata for a parallel file, the iod takes a request from an application to access data and maps that access to the data on its partition. In the case of a read request, this data would then be returned to the application through the TCP connection. In the case of a write, the data sent from the application would be written in to the correct locations in the parallel file based on this mapping.

The library routines used by the application are responsible for sending the correct data to each iod (in the case of a write), or reading the data from the iods back in the correct order (in the case of a read).

3.7 User Interface

PVFS gives the user a great deal of flexibility with regard to the distribution of a parallel file as well as allowing the user to set a logical "view" of the file. Each file in the parallel file system is striped across the partitions with a user-defined stripe size. This allows the user to tailor the stripe size to match characteristics of the data in the file. In addition, a separate "stripe offset" is allowed as well. This defines a point in the file where the striping should start. This feature allows the user to easily map files with header blocks, which would normally destroy the clean mapping of data into stripes. Users may also want to view only certain parts of a file. PVFS allows the user to specify a stripe size, stride, and offset that define a logical view of the file. Unlike most other parallel file systems, this logical view is completely independent of the physical partitioning of the data.

3.8 The PVFS system call library

A library of calls is available to application programmers in order to make pvfs easy to use. Basic file access routines (open, close, read, write, ioctl, lseek, unlink) are all available. This should make it possible to switch to pvfs by simply dropping in pvfs routines in the place of standard Unix calls.

The pvfs_ioctl call is used to get and set parameters relative to an open parallel file. This includes the current "view" of the file from the user's point of view, as well as the parameters that define how the file is distributed across partitions on the parallel file system.

At the same time, the Unix file system interface [RI78], which has been a standard for a number of years, is proving to be difficult to use and inefficient as a parallel application interface for a number of reasons. The difficulties in programming with this interface arise from a number of areas:

- Multiple accesses are needed to access multidimensional data,
- Explicit seeks are needed to access partitioned data,
- External synchronization is needed to manage access to shared data.

These difficulties in turn lead to two types of inefficiencies. First, additional overhead is caused by the number of system calls needed to perform the necessary seeks and accesses for partitioned or multidimensional data sets. Second, caching and prefetching by the file system is often impaired by the access patterns of these applications, which often do not match the patterns of sequential applications.

New interfaces and mechanisms for I/O accesses are being developed in order to address these problems, including methods to describe I/O access patterns, support for complex data types, and collective I/O. It has been shown that these mechanisms can be effective for parallel I/O in at least some environments, but it is not clear how these interfaces impact application development and performance in many environments – especially in a distributed network environment. There are examples of parallel I/O systems that have not performed well for some applications, even on the machine they were designed for.

We have developed several parallel I/O interfaces in response to perceived needs encountered while developing parallel out of core applications. In examining these interface options, we have also developed a new paradigm for applying collective I/O in SPMD programs which we call scheduled I/O. Scheduled I/O promises to provide the following features:

- Collective I/O without implicit synchronization,
- Managed prefetching and caching,
- Data-flow synchronization support for shared data, and
- Irregular and unbalanced I/O support.

By providing such a set of primitives and a simple method for describing accesses to the IODs that make up the parallel file system, we have made developing new interfaces for PVFS relatively simple. Thus far a Unix-style interface (described above), a multi-dimensional block interface, and two scheduled I/O interfaces have been developed. The Unix-style interface provides extensions to allow for strided accesses by partitioning a file. The last three interfaces are described below.

3.9 Multi-dimensional Block Interface

The PVFS block interface is designed to help in the development of out-of-core algorithms operating on multi-dimensional data sets. It allows the programmer to describe an open file as an N dimensional matrix of elements of a specified size and partition this matrix into a set of blocks. Blocks can then be read or written by specifying their indices in the correct number of dimensions. The dimensions and position of the block are used to create a single request to access the block stored on the file system. Accesses are made independently by all processes, so no constraints are placed on order of access and there is no implicit synchronization.

In the two dimension case, accesses are converted into a strided access. This allows the entire block to be read with a single request. When the data is defined to be of more dimensions, a batch request or nested strided request mechanism must be used. In all cases, the data in the file is currently stored in row major matrix order.

3.10 Scheduled I/O

Many types of applications have a regular I/O access pattern. For example, histogram equalization, calibration and navigation, and a number of other image processing algorithms process image data on a line-by-line basis. In a parallel implementation, scan lines are often doled out in a round-robin or block fashion. One approach to improving performance in this situation is to use collective I/O.

A collective-I/O interface is one in which all compute processes cooperate to present a single request to the file system, retaining the information that the individual requests make up a whole and allowing the file system to use this information to provide better performance. An obvious choice for a programming model for applications using collective I/O is the SPMD model. An inherent problem with many implementations of collective I/O is the necessity for an implicit synchronization at the point where I/O must take place or a message being passed from each compute process. However, if I/O for a parallel application is done in a regular pattern, then a single process should be able to provide the description for the entire group with a single request. This reduces the number of control messages and control connections while at the same time providing a logical grouping of the accesses. In addition, if this process is designed appropriately and not directly involved in all the data transfer or computation, it is free to schedule this I/O ahead of time. This allows request processing and data transfer to the compute processes to overlap computation.

Our scheduled I/O design uses this approach to provide the following advantages over typical collective I/O implementations:

- No strict boundary,
- All I/O specified with a single call, allowing IODs to optimize disk access, and
- I/O request and data transfer can be overlapped with I/O and computation.

Instead of synchronizing, all compute processes simply pass a message to the scheduler informing it that they have completed the access. It is unnecessary for them to wait for all processes to finish the I/O access in order to begin computation with current data.

3.11 Scheduled Block I/O

In applications with regular access patterns to a multi-dimensional data set, it makes sense to provide a means for describing these access patterns and simplifying the process of writing programs that access this data. The scheduled block I/O interface is an attempt to provide a mechanism for writing SPMD programs that operate on multidimensional data sets distributed to compute processes in a row, column, checkerboard, or broadcast manner. The matrix dimensions are specified using the same block I/O interface, and the distribution mechanism is specified for each open file descriptor. Blocks of the matrix can then simply be read or written, and a mapping function, transparent to the application programmer, moves the file pointer to the beginning of the next block after each access.

Instead of using a set of parameters that specify a partition of the matrix, a set of mapping functions are implemented for each distribution. These functions are used to reposition the file offset based on the number of compute processes accessing the file in tandem and the distribution mechanism being used for the file. New types of distributions can be implemented simply by adding the appropriate mapping functions to the interface.

As in the scheduled I/O interface, a separate process schedules the accesses to the file for the compute processes, except that a batch mode is used to specify to the IODs the set of accesses for each process. The accesses are then organized by the IODs to optimize disk access patterns before the data is transferred.

4. Applications

The key characteristic of this project has been studying the I/O behavior of real-life applications. The applications used in this study were taken from different levels of the typical RVC/DAAC processing cycle and include both image processing routines and database search routines. In each case the applications were taken and run in their entirety. In several cases the codes were partial rewritten to improve their I/O characteristics. This approach was taken because in many cases the original programmers had not considered I/O performance but were focusing on the program's functionality. Each application was studied using both parallel and sequential I/O systems. The following sections briefly describe each program studied.

4.1 AVHRR Calibration / Navigation

The processing algorithms we expect to comprise the bulk of a typical DAAC scenario would consist of those algorithms that process raw instrument telemetry data into specific data products needed by the scientific community. Some of these algorithms would be run as a standard processing suite during data ingest, others would be the result of a specific request for data. In each case, large amounts of data would need to be output to and input from archival storage, in addition to transfers between secondary staging devices and main memory. During the last year we have established contacts with NASA Code 930.2 (the International Data Systems Office) and have been working with them on codes for calibration and registration of radiospectroscope data, specifically AVHRR. We expect to continue with algorithms for standard data products such as vegetation index and sea surface temperature. This group is also looking towards the MODIS sensor as a natural follow-on to AVHRR. These algorithms are representative of those used on other similar radiospectroscopes, and could be adjusted to account for different spatial and spectral resolutions. We believe this is a good start at looking at critical and well-proven algorithms. In the future we hope to focus on more exotic and experimental algorithms.

4.2 PNN Satellite Image Classifier

This program uses a probabilistic neural network to perform classification on a multi-spectral satellite image. This routine uses established ground truth to create a set of training vectors that are then compared against image pixels in order to classify them for land cover/land use or to identify image content such as cloud cover. This routine can be used early in the processing cycle to extract image content vectors for inclusion in the metadata to enable content-based search. Alternatively, this routine can be used later in the processing cycle to identify any number of image content features. This routine can also be used as a precursor to more sophisticated classification methods by performing a fast classification used to define regions of interest for slower but more powerful classification routines. This particular classifier is relatively quick, and thus is more dependent on I/O performance.

4.3 Out-of-Core Numerical Codes

High level Earth science codes include numerical models of Earth systems including atmospheric models, ocean models, and Earth magnetic field models. At the heart of many of these applications is the solution of systems of linear equations. We have worked with a number of codes designed to model large electromagnetic fields using Method of Moments techniques that relay on large dense matrix equations. In order to process some of the larger codes out-of-core techniques are required on many of the more cost-effective parallel architectures due to memory constraints. We have developed a number of parallel out-of-core numerical codes based on matrix multiplication, matrix-vector multiplication, and matrix decomposition including Gaussian elimination, Jacobi and Gauss/Siedel iterative methods, domain decomposition, and relatively new reduced current fidelity techniques. These codes have been tested within a number of parallel numerical applications related to Earth science.

4.4 Object Databases

One of the least understood applications areas in RVC/DAAC processing is the area of intelligent data management. Of primary interest is the creation and access to an object database that records metadata needed to find specific earth science data in a terabyte sized archive. In addition, the object database must record knowledge on processing techniques in order to transform raw data into the desired end data products. Such a system would typically require significant amounts of storage in its own right, and must be able to support hundreds of simultaneous users. Considerable work in this area is being done by Code 935 at GSFC in the context of RVC system software. We are exploring ways to utilize parallel I/O to improve throughput in database operations.

THE GEORGE WASHINGTON UNIVERSITY

Understanding and Improving High-Performance I/O Subsystems

Tarek A. El-Ghazawi, Gideon Frieder, Mike R. Berry, and Sorin Nastea
School of Engineering and Applied Science
(tarek@seas.gwu.edu)

This research program has been conducted in the framework of the NASA Earth and Space Science (ESS) evaluations. The contributions of this program are drawn from three experimental studies conducted on different high-performance computing testbeds/platforms, and therefore presented in three different segments as follows.

1. Evaluating the parallel input/output subsystem of a NASA high-performance computing testbed, namely the MasPar MP-1 and MP-2;

2. Characterizing the physical input/output request patterns for NASA ESS applications which used the Beowulf platform; and

3. Dynamic scheduling techniques for hiding I/O latency in parallel applications such as sparse matrix computations. This study has also been conducted on the Intel Paragon and has also provided an experimental evaluation for the Parallel File System (PFS) and parallel input/output on the Paragon.

The summary of findings discusses the results of each of these three studies. The reader is encouraged to refer to the full final report (available through the CESDIS administrative office) and the list of publications for more details.

In addition to the research findings and publications produced under this effort, the program has helped orient the doctoral research program of two students towards parallel input/output in high-performance computing. Further, the experimental results in the case of the MasPar were very useful and helpful to MasPar and were shared with the technical management of MasPar.

1. Summary of Findings

1.1 MasPar Evaluations

This work has shown that programmers of I/O-intensive scientific applications can tune their programs to attain good I/O performance when using the MasPar. They should be at least aware of their I/O configuration, the specific I/O RAM size, and how it is locally partitioned in an attempt to partition data into files that can fit into the I/O RAM. The work further establishes that system managers are also encouraged to understand the I/O resource requirements of the applications running on their machines and tune the I/O RAM configuration for best performance. In specific, partitioning the I/O RAM among disk reads, disk writes, data processing unit (DPU) to front end communications, and interprocessor communications should be based on an understanding of the most common needs of the local application domain. Finally, the work has demonstrated that a full MasPar configuration with MPIOCTM and a full I/O RAM has potential for delivering scalable high I/O performance. However, for this to happen, the I/O RAM management should make good attempts at prefetching anticipated data. Further, the I/O RAM partitioning strategy should be more flexible by using cache blocks for different purposes as dynamically needed by the applications. At the least, files smaller than the I/O RAM size should be cacheable. Finally, the sustained performance of the disk arrays remains as the clear bottleneck, and is likely to limit the overall performance of parallel I/O systems for some time to come.

1.2 Physical I/O Requests of ESS Applications

This work has clearly shown that device driver instrumentation has the ability to distinguish among the different activities in the system, small explicit requests (less than page size), paging (4KB each in this case), and large objects (such as images). Further, it was shown that ESS codes have high spatial I/O access locality, 90% of accesses into 10% of space. On the other hand, temporal locality was measured as frequency of accesses and observed to be as high as six repeated accesses per second. In general, astrophysics simulation codes (PPM and Nbody) have similar I/O characteristics and have shown very little I/O requirements for the used problem sizes. Wavelet code, however, required a lot of paging due to the use of many different files for output and scratch pad manipulations, and could benefit from some tuning to improve data locality. It is therefore advised that a strategy for file usage and explicit I/O requests for this code be developed to do so. On the system side, Linux tends to allow larger physical requests when more processes are running, by allocating additional blocks for I/O. It is therefore recommended that Linux file caching should be further investigated and optimized to suit the big variability in the physical requests of the NASA ESS domain.

1.3 Dynamic I/O Scheduling and PFS Evaluations

Using the worker-manager paradigm, we have introduced a dynamic I/O scheduling algorithm which maximizes I/O latency hiding by overlapping with computations at run-time and is also capable of balancing the total load (I/O and processing).

Using sparse matrix applications as a test case, we have shown empirically that such scheduling can produce performance gains in excess of 10%. Much higher improvement rates are expected when the nonzero elements are distributed in a skewed manner in sparse matrix applications. The end-to-end (including I/O) scalability of such applications was studied and was shown to be very satisfactory under these scheduling schemes. In addition, the Paragon parallel file system (PFS) was evaluated and its various ways of performing collective input/output were studied. It was shown that the performance of the various calls depend heavily on the way of managing the file pointer(s). Calls that allow concurrent asynchronous access of processors to their respective blocks, but provide ordering at the user level performed better than the rest. This study was conducted in collaboration with Sorin Nastea (GMU) and Ophir Frieder (GMU).

2. Publications

2.1 Papers

Frieder, G., & El-Ghazawi, T., *Input/Output. The Encyclopedia of Computer Science*, 4th Edition. International Thomson Computer Press. (To appear in 1998, invited)

Nastea, S., El-Ghazawi, T., & Frieder, O. (June 1997), *Performance Optimization of Combined Variable-Cost Computations and I/O*, the 4th International Symposium on Solving Irregularly Structured Problems in Parallel (IRREGULAR-97), Paderborn, Germany. Lecture Notes in Computer Science, Springer-Verlag, Berlin

Nastea, S., Frieder, O., & El-Ghazawi, T., *Load-Balanced Sparse Matrix-Vector Multiplications on Highly Parallel Computers*, Journal of Parallel and Distributed Computing. (Accepted)

Berry, M., & El-Ghazawi, T. (April 1996), *Parallel Input/Output Characteristics of NASA Science Applications*, Proceedings of the International Parallel Processing Symposium (IPPS'96), IEEE Computer Society Press. Honolulu.

Nastea, S., Frieder, O., & El-Ghazawi, T. (April 1996), *Parallel Input/Output Impact on Sparse Matrix Compression.* Proceedings of the Data Compression Conference (DCC'96), IEEE Computer Society Press. Snowbird.

El-Ghazawi, T. (February 1995), *Characteristics of the MasPar Parallel I/O System*, Frontiers'95, IEEE Computer Society, McLean, VA.

2.2 Technical Reports

Berry, M., & El-Ghazawi, T. *MIPI: Multi-level Instrumentation of Parallel Input/Output.* (CESDIS TR-96-176)

Nastea, S., El-Ghazawi, T., & Frieder, O. *Parallel Input/Output Issues in Sparse Matrix Computations.* (CESDIS TR-96-170)

Berry, M., & El-Ghazawi, T. *An Experimental Study of the I/O Characteristics of NASA Earth and Space Science Applications.* (CESDIS TR-95-163)

El-Ghazawi, T. *Characteristics of the MasPar Parallel I/O System.* (CESDIS TR-94-129)

El-Ghazawi, T. *I/O Performance Characteristics of the MasPar MP-1 Testbed.* (CESDIS TR-94-111)

2.3 Submitted Manuscripts

Nastea, S., El-Ghazawi, T., & Frieder, O. Parallel Input/Output Issues in Sparse Matrix Computations. Submitted to Parallel Computing

Nastea, S., El-Ghazawi, T., & Frieder, O. Impact of Data Skewness on the Performance of Parallel Sparse Matrix Computations. Submitted to the IEEE Transactions on Parallel and Distributed Systems.

SYRACUSE UNIVERSITY

High Performance Input/Output System for High Performance Computing and Four-Dimensional Data Assimilation

Geoffrey C. Fox, Chao-Wei Ou
Northeast Parallel Architectures Center
(gcf@nova.npac.syr.edu)

1. Project Overview

The Northeast Parallel Architectures Center of Syracuse University is applying basic computer science research in high performance input/output systems for parallel computers to the NASA Grand Challenge applications of four-dimensional data assimilation. Our approach is to apply leading parallel computing research to a number of existing techniques for assimilation, and extract parameters indicating where and how input/output limits computational performance. Using detailed knowledge of the application problems, we are:

- developing a parallel input/output system specifically for this application;
- extracting the important input/output characteristics of data assimilation problems; and
- building these characteristics as parameters into our runtime library (Fortran D/High Performance Fortran) for parallel input/output support.

2. Research Activities

2.1 PASSION: Parallel And Scalable Software for Input-Output

I/O for parallel systems has drawn increasing attention in the last few years as it has become apparent that I/O performance rather than CPU or communication performance may be the limiting factor in future computing systems. Large scale scientific computations, in addition to requiring a great deal of computational power, also deal with large quantities of data. At present, a typical Grand Challenge application could require 1Gbyte to 4Tbytes of data per run. These figures are expected to increase by orders of magnitude as teraflop machines make their appearance. Although supercomputers have very large main memories, the memory is not large enough to hold this amount of data. Hence, data needs to be stored on disk and the performance of the program depends on how fast the processors can access data from disks. Unfortunately, the performance of the I/O subsystems of MPPs has not kept pace with their processing and communications capabilities. A poor I/O capability can severely degrade the performance of the entire program. The need for high performance I/O is so significant that almost all the present generation parallel computers provide some kind of hardware and software support for parallel I/O.

In order to develop a successful assimilation system for Earth science, there will be a continual need to process and reprocess data sets with an ever-improving and more complete assimilation system. There will also be a requirement to diagnose the quality of the data sets. Data assimilation provides the most compute-intensive as well as I/O-intensive undertaking in NASA Earth science research, and, therefore, high-performance I/O capability will be essential to new generation data assimilation systems. The objective of designing the PASSION software is to develop software support for parallel I/O that permits scalable I/O operations to match the growing computational power of the new parallel supercomputer.

At Syracuse University, we consider the I/O problem from a language, compiler, and runtime support point of view. We are developing a compiler and runtime support system called PASSION: Parallel And Scalable

Software for Input-Output. PASSION software support is targeted for I/O intensive out-of-core loosely synchronous problems. The PASSION Runtime Library provides routines to efficiently perform the I/O required in out-of-core programs. The goal of the PASSION compiler is to translate out-of-core programs written in a data-parallel language like High Performance Fortran (HPF) to node programs with calls to the PASSION Runtime Library for I/O. Other components of the PASSION project include a Portable Parallel File System (VIP-FS), integrating task and data parallelism using parallel I/O and file servers for multimedia applications.

2.2 PASSION Runtime Support for Parallel I/O

In out-of-core computations, data is stored in files on secondary storage such as disks. During program execution, data needs to be moved back and forth between disks and main memory. The PASSION Runtime Library provides routines to efficiently perform the I/O required in out-of-core programs. It provides support for loosely synchronous out-of-core computations which use a Single Program Multiple Data (SPMD) Model. PASSION uses a simple high-level interface, which is a level higher than any of the existing parallel file system interfaces. For example, the user needs only to specify what section of the array needs to be read in terms of its lower-bound, upper-bound and stride in each dimension, and the PASSION Runtime Library will fetch it in an efficient manner. PASSION thus provides a simple and portable level of abstraction above the native parallel file system provided on the machine. PASSION is designed to be either directly used by application programmers, or a compiler can translate out-of-core programs written in a high-level data parallel language like High Performance Fortran (HPF) to node programs with calls to the PASSION Runtime Library for I/O. A number of optimizations such as Two-Phase I/O, Data Sieving, Data Prefetching, and Data Reuse have been incorporated in the library for improved performance.

2.3 PASSION Compiler Support for Parallel I/O

The goal of the PASSION compiler is to compile out-of-core data parallel programs written in a language such as High Performance Fortran (HPF). The PASSION compiler has to perform the following two main tasks:

- Read and write distributed arrays. The compiler obtains distribution information from the HPF directives.
- Perform automatic program transformations to improve I/O performance.

The PASSION compiler takes an HPF program as an input and generates an node+MP+I/O program, with calls to the PASSION Runtime Library.

The PASSION compiler uses two distinct models for compiling an out-of-core program. The first model is called the Local Placement Model (LPM) and the second model is called the Global Placement Model (GPM).

3. TCE: Thread-based Communication Environment

TCE employs light-weight multi-threading. It assumes each I/O event is independent (a thread) and that scalable I/O can be accomplished by managing a parallel/distributed thread queue. The integration between PASSION and TCE can offer a more robust and unified view toward using meta-computing for scalable heterogeneous I/O for four-dimensional data assimilation.

TCE is designed:

- to provide an efficient, thread-based communication library capable of supporting distributed and parallel processing on variety on platforms;

- to ensure interoperability between different types of architectures with different CPUs and operating systems;
- to make the environment as simple as possible without compromising the performance or functionality; and
- to assist the programmer in choosing computational nodes and style of interactions between his processes.

By abstract communication objects called ports and channels it is possible to build the client-server connection as well as peer-to-peer parallel relations. By mapping application processes onto computational nodes, both data parallelism and functional parallelism can be exploited. These two paradigms can even be mixed in one application. The multithreaded support is based on a user-level thread package, thus TCE can be easily ported to different processors. Different architecture are supported:

- clusters of heterogeneous workstations – through ports and channels
- shared memory parallel machines – through multithreading
- distributed memory parallel machines – through ports and channels

The differences in data formats (byte ordering) is taken care of internally by the library. Machine-specific IPC operations are masked through higher level operations on channels and ports.

4. SPRINT: Scalable Partitioning, Refinement and INcremental partitioning Techniques

Load balancing of the distributed heterogeneous system is vital to scalable I/O. Our use of meta-computing for large-scale four-dimensional data assimilation problems makes this more critical. The load balancing problem can be viewed as a graph-partitioning problem. Graph-partition problems belong to the class of NP-complete problems; hence exact solutions are computationally intractable for large problems. However, good suboptimal solutions are sufficient for effective parallelization of most applications. SPRINT (Scalable Partitioning, Refinement and INcremental partitioning Techniques) collects three important partitioning methods based on physical information (e.g., recursive inertial bisection, recursive orthogonal bisection, and index-based partitioning). Index-based partitioning methods can not only be used to partitioning graphs, but also can be used to improve the disk allocation.

Efficient methods for graph partitioning and incremental graph partitioning are important for parallelization of a large number of unstructured and/or adaptive applications. The key problem in efficiently executing irregular and unstructured data parallel applications is partitioning the data to minimize communication while balancing the load. Partitioning such applications can be posed as a graph-partitioning problem based on the computational graph. We have developed a library of partitioners (especially based on physical optimization) which aim to find good suboptimal solutions in parallel. This initial target use of these partitioning methods are for runtime support of data parallel compilers (HPC, HPC++, HPF, etc.).

SPRINT software focuses on a subclass of applications in which the computational graph is such that the vertices correspond to two- or three-dimensional coordinates, and the interaction between computations is limited to vertices that are physically proximate. Examples of such applications include finite element calculations, molecular dynamics, particle dynamics, particle-in-a-cell, region growing, and statistical physics. For these applications, partitioning can be achieved by exploiting the above property. Essentially proximate points are clustered together and form a partition such that the number of points attached to each partition are approximately equal. Most of the interactions are local and the amount of interprocessor communication is low if proximate points are clustered together.

The SPRINT library provides software for parallel graph-partitioning using coordinate information such as index-based partitioning, recursive coordinate bisection, and recursive inertial bisection. SPRINT also provides incremental partitioning techniques based on the index-based method.

5. A Real-time Terrain Rendering Application on a PC Cluster

The goal of this terrain rendering project is to provide the user a real time interactive viewing environment for the available terrain data. We envision that the user starts out in the solar system, where Mars, Earth, and Venus are visible. Then the user chooses to visit one of the three planets.

This journey can be broken into four main viewing stages: stage one, from solar orbit to high planetary .orbit; stage two, from high planetary orbit to lower planetary orbit; stage three, from lower planetary orbit to high altitude flight path; and stage four, from high altitude flight path to low altitude fly-by. Each of these stages has distinct viewing characteristics that the terrain viewer must respect. As a result, multiple rendering techniques and data sets are needed to generate the images.

We focused on a distributed PC cluster pioneered by Beowulf system developed by NASA. We have experimented this application on a 4-node 486/100MHz PC Cluster with LINUX. This effort was limited because of short funding.

6. Achievements Summary

- PASSION can either be directly used by application programmers or a compiler.
  http://www.cat.syr.edu/passion.html

- TCE can offer a more robust and unified view toward using meta-computing for scalable heterogeneous I/O for four-dimensional data assimilation.
  http://www.npac.syr.edu/users/gcf/cps616threads/

- SPRINT can be used to partitioning graphs but also can be used to improved the disk locality.
  http://dante.npac.syr.edu:1996/SPRINT/index.html

- A Real-time Terrain Rendering Application has been conducted on a PC Cluster.
  http://www.npac.syr.edu/users/alvin/papers/terrain/terrain.html

7. References

Bordawekar, R., Choudhary, A., & Ramanujam, J (May 1996). Automatic Optimization of Communication in Out-of-core Stencil Codes. *In Proc. of 10th ACM Intl. Conference on Supercomputing*, May. (Scalable I/O Technical Report 114, November 1995) *http://www.cat.syr.edu/~rajesh/ics96.ps*

Bordawekar, R., Choudhary, A., & Ramanujam, J. (November 1995). *Compilation and Communication Strategies for Out-of-core programs on Distributed Memory Machines*. Syracuse University. (Scalable I/O Technical Report 113) *http://www.cat.syr.edu/~rajesh/jpdc.ps*

Bordawekar, R. & Choudhary, A. (July 1995). Communication Strategies for Out-of-core Programs on Distributed Memory Machines. *In Proc. of 9th ACM Intl. Conference on Supercomputing.* *http://www.npac.syr.edu/techreports/html/0650/abs-0667.html*

Thakur, R. (May 1995). Runtime Support for In-Core and Out-of-Core Data-Parallel Programs. Ph.D. Thesis, Dept. of Electrical and Computer Eng., Syracuse University.
ftp://ftp.npac.syr.edu/pub/users/thakur/papers/phd_thesis.ps.Z

Thakur, R. & Choudhary, A. (June 1995). *An Extended Two-Phase Method for Accessing Sections of Out-of-Core Arrays.* Center for Advanced Computing Research, Caltech. (Scalable I/O Initiative Technical Report CACR-103). *ftp://ftp.npac.syr.edu/pub/users/thakur/papers/ext2ph.ps.Z*

Bordawekar, R., Choudhary, A., Kennedy, K., Koelbel, C., & Paleczny, M. (July 1995). A Model and Compilation Strategy for Out-of-Core Data Parallel Programs. *In Proc. of the Fifth ACM SIGPLAN Symposium on Principles and Practices of Parallel Programming.*
http://www.npac.syr.edu/techreports/html/0650/abs-0696.html

Choudhary, A., Bordawekar, R., More, S., Sivaram, K., & Thakur, R. (June 1995). PASSION: Runtime Library for the Intel Paragon. *In Proc. of the Intel Supercomputer User's Group Conference.*
ftp://ftp.npac.syr.edu/pub/users/thakur/papers/isug95-passion.ps.Z

Bordawekar, R. & Choudhary, A. (March 1995). *A Framework for Representing Data Parallel Programs and its Application in Program Reordering.* (NPAC Technical Report SCCS-698)
http://www.npac.syr.edu/techreports/html/0650/abs-0698.html

Choudhary, A., Bordawekar, R., Harry, M., Krishnaiyer, R., Ponnusamy, R., Singh, T., & Thakur, R. (Sept. 1994). *PASSION: Parallel and Scalable Software for Input-Output.* (NPAC Technical Report SCCS-636)
ftp://ftp.npac.syr.edu/pub/docs/sccs/papers/ps/0600/sccs-0637.ps.Z

Thakur, R., Bordawekar, R., Choudhary, A., Ponnusamy, R., & Singh, T. (October 1994) PASSION Runtime Library for Parallel I/O. *In Proc. of the Scalable Parallel Libraries Conference.*
ftp://ftp.npac.syr.edu/pub/users/thakur/papers/splc94_passion_runtime.ps.Z

Thakur, R., Bordawekar, R., & Choudhary, A. (July 1994) Compiler and Runtime
Support for Out-of-core HPF Programs. *In Proc. of 8th ACM Int. Conf. on Supercomputing,*
pp. 382-391. *ftp://ftp.npac.syr.edu/pub/projects/pcrc/f90d/docs/ics94-out-of-core-hpf.ps.Z*

Harry, M., Miguel del Rosario, J., & Choudhary, A. (April 1995). VIP-FS: A VIrtual, Parallel File System for High Performance Parallel and Distributed Computing. *In Proc. of Ninth International Parallel Processing Symposium. ftp://ftp.npac.syr.edu/pub/docs/sccs/papers/ps/0650/sccs-0686.ps.Z*

Bordawekar, R., Miguel del Rosario, J., & Choudhary, A. (November 1993). Design and Implementation of Primitives for Parallel I/O. *In Proc. of Supercomputing'93*, Portland, OR.
http://www.npac.syr.edu/techreports/html/0550/abs-0564.html

Bordawekar, R., Miguel del Rosario, J., & Choudhary, A. (July 1993). An Experimental Evaluation of Touchstone Delta Concurrent File System. *In Proc. of International Conference on Supercomputing 1993.*
http://www.npac.syr.edu/techreports/html/0400/abs-0420.html

Bordawekar, R., Choudhary, A., & Thakur, R. (Sept. 1994). *Data Access Reorganizations in Compiling Out-of-core Data Parallel Programs on Distributed Memory Machines.* (NPAC Technical Report SCCS-622) *ftp://ftp.npac.syr.edu/pub/docs/sccs/papers/ps/0600/sccs-0622.ps.Z*

Bordawekar, R. (May 1993). Issues in Software Support for Parallel I/O. Master's Thesis, ECE Dept, Syracuse University. *http://www.npac.syr.edu/techreports/html/0450/abs-0487.html*

Miguel del Rosario, J., Bordawekar, R., & Choudhary, A. (April 1993). Improved Parallel I/O via a Two-Phase Run-time Access Strategy. *IPPS'93 Workshop on Input/Output in Parallel Computer Systems. http://www.npac.syr.edu/techreports/html/0400/abs-0406.html*

Jadav, D., Srinilta, C., Choudhary, A., & Berra, B. P. (May 1995). Design and Evaluation of Data Access Strategies in a High Performance Multimedia-on-Demand Server. *In Proc. of the Second Intl. Conf. on Multimedia Computing and Systems. ftp://ftp.npac.syr.edu/pub/projects/pcrc/f90d/docs/ICMCS95.ps.Z*

Jadav, D. & Choudhary, A. (Summer 1995) Design Issues in High Performance Media-on-Demand Servers. *IEEE Parallel and Distributed Technology Systems and Applications. ftp://ftp.npac.syr.edu/pub/projects/pcrc/f90d/docs/PDT.ps.Z*

Jadav, D., Srinilta, C., Choudhary, A. & Berra, B. P. (Fall 1995) Techniques for Scheduling I/O in a High Performance Multimedia-On-Demand Server. *The Journal of Parallel and Distributed Computing. ftp://ftp.npac.syr.edu/pub/projects/pcrc/f90d/docs/JPDC.ps.Z*

Jadav, D., Srinilta, C., Choudhary, A., & Berra, B. P. (January 1995) An Evaluation of Design Tradeoffs in a High Performance Media-on-Demand Server. *ACM Multimedia Systems Journal.*

Jadav, D., Srinilta, C. & Choudhary, A. (December 1995) I/O Scheduling Tradeoffs in a High Performance Media-on-Demand Server. *The 2nd Intl. Conference on High Performance Computing*, New Delhi, India.

Avalani, B., Choudhary, A., Foster, I., & Krishnaiyer, R. (Dec. 1994) Integrating Task and Data Parallelism Using Parallel I/O Techniques. *In Proc. of International Workshop on Parallel Processing. ftp://ftp.npac.syr.edu/pub/projects/pcrc/f90d/docs/task_data.ps.Z*

Ou, CW., Gunwani, M., & Ranka, S. (December 1994) An Architecture-Independent Locality-Improving Transformations of Computational Graphs Embedded in k-Dimensions. *ICS'95*, July 1995, pp. 289-298. *ftp://ftp.npac.syr.edu/pub/docs/sccs/papers/ps/0700/sccs-0728.ps.Z*

Ou, CW. & Ranka, S. (November 1994) Parallel Incremental Graph Partitioning. IEEE Transactions on Parallel and Distributed Systems. to appear. (A preliminary version appeared as "Parallel Incremental Graph Partitioning Using Linear Programming" in *Supercomputing '94*, pp. 458-467). *ftp://ftp.npac.syr.edu/pub/docs/sccs/papers/ps/0650/sccs-0653.ps.Z*

Ou, CW. & Ranka, S. (February 1995) Parallel Remapping Algorithms for Adaptive Problems. *Journal of Parallel and Distributed Computing, under revision.* (A preliminary version of the paper appeared in Frontiers' 95, pp. 367-374). *ftp://ftp.npac.syr.edu/pub/docs/sccs/papers/ps/0650/sccs-0652.ps.Z*

Ou, CW., Ranka, S., & Fox, G. (1996) Fast and Parallel Mapping Algorithms for Irregular Problems. *Journal of Supercomputing,10*, pp. 119-140. *ftp://ftp.npac.syr.edu/pub/docs/sccs/papers/ps/0700/sccs-0729.ps.Z*

UNIVERSITY OF FLORIDA

Distributed Indices for Distributed Data

Theodore Johnson
Department of Computer and Information Sciences
(ted@cis.ufl.edu)

1. Summary

Large-scale data systems need distributed indexing to manage and locate distributed data. This project is concerned with developing, implementing, and measuring and modeling the performance of distributed search structures. We focus on developing the dB-tree, a distributed B-tree. During the first year of funding, we developed some correctness theory results, developed an implementation, developed a simulator, and collected some simulation results. During the second, we completed correctness theory results, completed an implementation and collected results from it, completed the simulator and executed several studies, and developed an analytical performance model.

During the third year, we have concentrated on developing an extensible simulator for distributed index structures. The source code of the simulator, written in C++, is available via WWW at http://www.cis.ufl.edu/~ted/.

The project has also enabled us to perform research on related research issues of interest to NASA. Recent work on high performance distributed synchronization has developed distributed list algorithms. These algorithms can potentially be applied to distributed index maintenance. An initial investigation of this technique has resulted in a distributed reader/writer lock, that supports a higher throughput that previous mutual exclusion algorithms.

NASA's EOSDIS project will ingest and archive petabytes of satellite data per year, and distribute the data to a world-wide research community. To aid in planning and maintaining such a large archive we are developing performance models of mass storage archives. Part of this work is to analyze reference patterns to existing archives, and another part is to develop analytical queuing model of tertiary storage.

We discuss these results further in the following sections.

2. Distributed Index Simulator

We have developed a distributed simulator, which can be used to study the performance of distributed search structures. The design of the simulator is object-oriented and it could be adapted to develop a simulator for any distributed system. In this section, we describe the architecture of the simulator and discuss the implementation of the dB-tree with lazy updates.

3. System Model

There are five components in the model, three of which represent the distributed system itself, while the remaining two capture the applications utilizing the system services. The components, each of which is realized by a separate object in the simulator, are the following:

Source	which generates the system workload
Scheduler	which assigns the generated workload to different processors
Message Router	which mimics the inter-connection network between the processors
Processor	which models each node of the distributed system
Sink	which receives the completed requests

The interaction between these modules primarily consist of service requests nd completion replies. In the remainder of this section, we describe each of these components in detail.

3.1 Source

The source component represents the applications utilizing the system services. In our current implementation, it consists of an object which generates dB-tree insert and search requests.

The workload generator is modeled as an open system and requests are generated in a Poisson stream at a mean rate specified by the Arrival Rate parameter. The proportion of insert and search requests can be controlled by InsertProb and SearchProb, which are probabilities that define the proportion of insert and search requests.

3.2 Scheduler

All requests generated by the Source are delivered to the Scheduler. The Scheduler assigns the request to a particular processor, where the request is initiated. The execution of the request may generate subsequent relayed actions which are propagated to other processors through the Message router. In our current implementation, the Scheduler randomly assigns the requests to the processors. If we maintain the load profile of each processor, we can use some load balancing techniques to distribute the generated requests.

3.3 Message Router

The Message Router models the inter-connection network between the processors. Every request consists of the target processor and the action to be performed on the target processor. The request is delivered to the target processor after a time delay which is a function of the source and target processors of the request. This is done to model the delay associated with the communication channel. The time delay on a particular channel follows an *erlang* distribution with a mean and standard deviation, which are determined by the source and target processors of the channel.

3.4 Processor

The Processor models a node of the distributed system. Each processor is realized by an object of the Processor class and the number of processors is a parameter of the simulator. A processor consists of queue manager, node manager, resource manager and disk.

The queue manager maintains a queue of pending actions to be performed on the part of the search structure maintained by the processor (the message queue). The node manager repeatedly takes an action from the queue manager and performs the action on a node. The action execution typically generates a subsequent action on another node. If the next node to process is stored locally, then a new entry is put into the message queue. Otherwise, the node manager sends a request to the appropriate remote processor through the message router.

The resource manager handles the requests from the node manager to read(write) a node from(to) the disk. In our current implementation, we assume that all internal nodes are present in memory and the leaf nodes reside on the disk. This can be easily extended and we can incorporate a buffer manager between the node manager and the resource manager.

The disk models the physical resources at a processor.

3.5 Sink

The Sink module receives the completed requests from the node manager. A request is assumed to be complete if it does not generate subsequent actions. Each request is assigned a request-id and hence all the actions generated by a request can be identified. The sink module gathers statistics on these requests and measures the performance of the system from the perspective of the various applications using the system.

4. A Fair Fast Distributed Concurrent-Reader Exclusive-Writer Lock

Distributed synchronization is needed to arbitrate access to a shared resource in a message passing system. Reader/writer synchronization can improve efficiency and throughput if a large fraction of accesses to the shared resource are queries. In this paper, we present a highly efficient distributed algorithm that provides FCFS concurrent-reader exclusive-writer synchronization with an amortized O(log n) messages per critical section entry and O(log n) bits of storage per processor. We evaluate the new algorithm with a simulation study, comparing it to fast and low-overhead distributed mutual exclusion algorithms. We find that when the request load contains a large fraction of read locks, our algorithm provides higher throughput and a lower acquire time latency than is possible with the distributed mutual exclusion algorithms, with a small increase in the number of messages passed per critical section entry.

The low space and message passing overhead, and high efficiency make the algorithm scalable and practical for implementation. The algorithm we present can easily be extended to give preference to readers or writers. A paper describing this research was submitted to the Frontiers '96 conference. See also UF CISE technical report 96-018, available from http://www.cis.ufl.edu/cis/tech-reports/.

5. Performance Modeling of Mass Storage Archives

5.1 Mass Storage Archive Log File Analysis

The successful implementation of mass storage archives require careful attention to performance optimizations, to ensure that the system can handle the offered load. However, performance optimizations require an understanding of user access patterns. Since on-line archives and digital libraries are so new, little information is available.

The National Space Science Data Center (NSSDC) of NASA Goddard Space Flight Center has run an on-line mass storage archive of space data, the National Data Archive and Distribution Service (NDADS), since November 1991. A large world-wide space research community makes use of NSSDC, requesting more than 20,000 files per month. Since the initiation of their service, NSSDC has maintained log files which record all accesses the archive.

In the report, tr96-020, available from http://www.cis.ufl.edu/cis/tech-reports/ we present an analysis of the NDADS log files, spanning a four year period (1992 - 1995). We analyze the log files and discuss several issues, including caching, reference patterns, changes in user interest, user characterization, clustering, and system loading. A preliminary version of this report appeared in the 1995 NASA Goddard conference on Mass Storage Systems and Technologies. The full report has been submitted to the *International Journal on Digital Libraries.*

The Goddard Space Flight Center (GSFC) Distributed Active Archive Center (DAAC) has been operational for more than two years. Its mission is to support existing and pre-EOS Earth science datasets, facilitate the scientific research, and test Earth Observing System Data and Information System (EOSDIS) concepts. Over 550,000 files and documents have been archived and more than 6 Terabytes have been distributed. Information about user access patterns and their impact on system loading is needed to optimize current operations and to plan for future archives (i.e., EOS-AM1). To facilitate the management of the daily activities, the GSFC DAAC has developed a data base system to track all correspondence, requests, ingestion and distribution. In addition, several log files which record transactions on Unitree. This study identifies some of the users' requests pattern submitted at GSFC DAAC during 1995. The analysis is limited to a subset of all the orders which have their files under the control of the Unitree hierarchical storage management. Some of the results show that a large percentage of the volume of data requested came from two data types, were for high level (L3 and L4) products, were distributed mostly on 4mm and 8mm tapes, and that most requests came from North America, although there is significant world-wide use. We found a very wide range in the size of individual requests, and that most of the volume that is distributed is ordered by a few users. We evaluate some file caching algorithms and find that LRU/2-bin has the best performance, but that STbin also works well.

This report will appear in the 1996 NASA Goddard conference on Mass Storage Systems and Technologies.

5.2 Queuing Models of Tertiary Storage

Large scale scientific projects generate and use huge amounts of data. For example, the NASA EOSDIS project is expected to archive one petabyte per year of raw satellite data. This data is made automatically available for processing into higher level data products and for dissemination to the scientific community. Such large volumes of data can only be stored in robotic storage libraries (RSLs) for near-line access. A characteristic of RSLs is the use of a robot arm that transfers media between a storage rack and the read/ write drives, thus multiplying the capacity of the system.

The performance of the RSLs can be a critical limiting factor of the performance of the archive system. However, the many interacting components of a RSL make a performance analysis difficult. In addition, different RSL components can have widely varying performance characteristics. This paper describes our work to develop performance models of a RSL. We first develop a performance model of a RSL in isolation. Next, we show how the RSL model can be incorporated into a queuing network model. We use the models to make some example performance studies of archive systems.

The models described in this paper, developed for the NASA EOSIDS project, are implemented in C with a well-defined interface. The source code and accompanying documentation are available through WWW at: http://www.cis.ufl.edu/~ted/.

Papers describing these models will appear in the NASA Goddard conference on Mass Storage Systems and Technologies, and in Performance '96. See also UF CISE technical report 96-019, available from http://www.cis.ufl.edu/cis/tech-reports/.

UNIVERSITY OF ILLINOIS

High-Performance Input/Output Systems for Parallel Computers

Daniel A. Reed
Department of Computer Science
(reed@cs.uiuc.edu)

1. Introduction

During the past three and one half years, our research has focused on three areas:

(a) the performance analysis and optimization of input/output intensive NASA applications and parallel file systems,

(b) PPFS (portable parallel file system) development, ports and optimizations, and

(c) exploration of policies for adaptive, performance-directed file caching and migration.

These efforts were driven by data from performance characterizations of applications based on the NCSA/EOSDIS HDF library, code from the NASA Goddard SeaWiFS project, and code from the NOAA/NASA Pathfinder AVHRR (Advanced Very High Resolution Radiometer) project. As a prelude to input/output analysis, we first extended our Pablo performance instrumentation and analysis toolkit to support capture and processing of input/output data. We then instrumented and analyzed the input/output behavior of both the HDF library and the NASA codes. Using insights gained from this analysis, we were able substantially increase the performance of the HDF library and the Pathfinder and SeaWiFS codes by automatically classifying file access patterns. In addition, we have begun implementation of a complementary sensor/actuator model for performance-directed adaptive control of our portable parallel file system (PPFS).

The remainder of this report is organized as follows. First, in section 2 we describe our input/output characterization infrastructure and experiences characterizing file access patterns. Using the extended PPFS infrastructure described in section 3, section 4 describes our experiences with automatic classification of access patterns and the performance improvements that accrue when filepolicies are selected based on access pattern classification. In section 5, we also describe our ongoing implementation of an infrastructure for performance-directed adaptive control of file system policies. Finally, section 6 describes efforts in related input/output projects that buttress and augment the work supported by this contract.

2. Application Input/Output Characterization

As described in section 6, we are engaged in a broad-based effort to instrument and analyze the input/output behavior of large-scale parallel applications; NASA is funding this effort via the Scalable I/O (SIO) initiative. As a complement to this broader characterization work, this project focused specifically on the instrumentation, analysis and optimization of NASA satellite image processing codes.

As part of our NASA-funded effort, we have instrumented the NCSA HDF library, the storage access mechanism for NASA EOSDIS codes and analyzed a suite of NASA codes. Below, we describe our Pablo input/output instrumentation extensions, our instrumentation of the NCSA HDF library and analysis of a code from the NASA/NOAA Pathfinder project. (This work is also supported by NASA and Intel fellowships for Christopher Elford and Tara Madhyastha.)

2.1 Pablo Input/Output Extensions

Any instrumentation system must strike a balance between instrumentation detail and perturbation of application behavior. When characterizing input/output behavior, use of the input/output system to extract performance data can lead to particularly pernicious perturbations. To capture and analyze application input/output data while minimizing input/output perturbations, we exploited the Pablo [7, 8] data capture library's extension interfaces to develop a suite of input/output analysis routines [9, 12, 13, 14] that support both real-time calculation of statistical summaries and capture of detailed event traces. The former trades computation perturbation for input/output perturbation.

Detailed input/output event traces include the time, duration, size, and other parameters of each input/output operation. Statistical summaries can take any one of three forms: file lifetime, time window, or file region. File lifetime summaries include the number and total duration of file reads, writes, seeks, opens, and closes, as well as the number of bytes accessed for each file, and the total time each file was open. Time window summaries contain similar data, but allow one to specify a window of time for summarization. Finally, file region summaries are the spatial analog of time window summaries; they define a summary over the accesses to a file region.

2.2 NCSA HDF

Using the Pablo input/output characterization software [9], we developed an instrumented version of the Hierarchical Data Format (HDF) library [6] and internally distributed the code to NCSA. This instrumented version of the HDF library records the UNIX input/output calls generated in response to HDF requests and can record the time spent in classes of HDF procedures associated with accesses to particular higher-level HDF objects (i.e., 8 and 24-bit raster images, palettes, scientific data sets, annotations, and vdata).
In November 1995, NCSA released HDF version 4.0b2 with the Pablo-based embedded instrumentation. Because NCSA is working directly with NASA EOSDIS scientists, we expect this instrumented software release to provide greater insight into the behavior of EOSDIS codes. Locally, we have found the HDF instrumentation invaluable when studying the behavior of HDF-based applications.

2.3 Pathfinder Project Input/Output Analysis

As part of our characterization of NASA codes, we acquired and analyzed the behavior of a code from the NOAA/NASA Pathfinder AVHRR (Advanced Very High Resolution Radiometer) data processing project. (We are extremely grateful to Peter Smith of the NASA Goddard Space Flight Center for providing access to the Pathfinder code.) The goal of this project is to process existing data to create global, long-term, time series, remote-sensed data sets that can be used to study global climate change.

Although there are four types of Pathfinder AVHRR land data sets (daily, composite, climate, and browse images), our analysis [4] has focused on the creation of the daily data sets. Each day, fourteen files of AVHRR orbital data (approximately 42 megabytes each) in Pathfinder format are processed to produce an output data set that is approximately 228 megabytes in HDF (SDS) format. Based on our input/output analysis, over seventy percent of the Pathfinder code's time is spent in Unix input/output system calls – the code is heavily input/output intensive.

During program execution, ancillary data files and the orbital data file are opened, and an orbit is processed 120 scans at a time. Although the orbit file is accessed sequentially, the access patterns for other, ancillary files range from sequential to irregularly strided. The result of this processing is written to a temporary output file using a combination of sequential and two-dimensionally strided accesses. Finally, the temporary file is re-written in HDF format to create three 8-bit and nine 16-bit layers.

Tables 1-2 and Figure 1 show the performance of this Pathfinder code on a Sun SparcServer 670 running

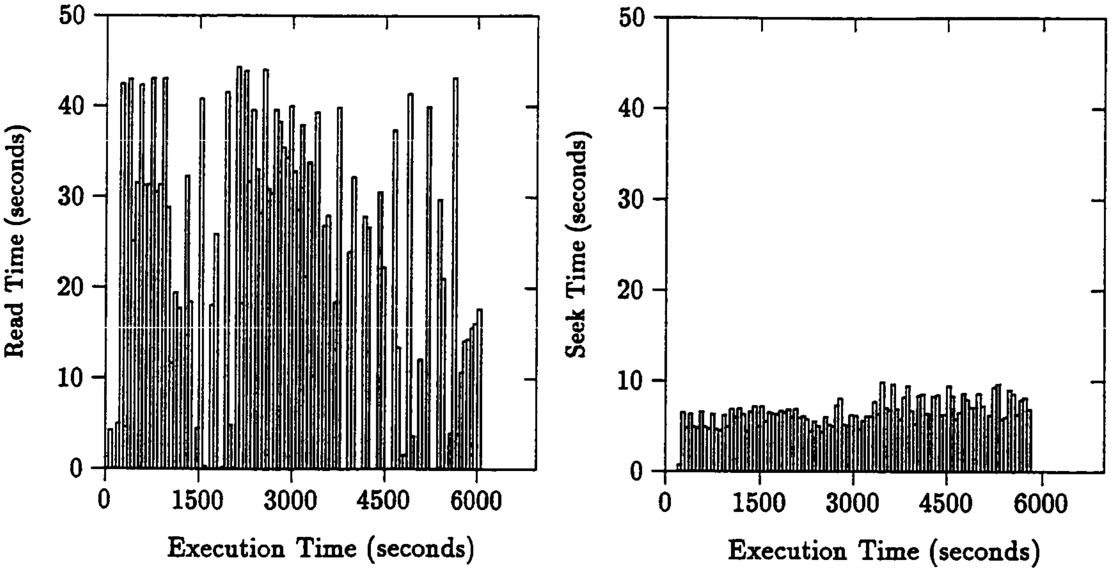
SunOS 4.1.3, with 64 megabytes of physical memory and a local SCSI disk; the results obtained with PPFS are the subject of section 4. This data was obtained by instrumenting the Pathfinder application using the Pablo instrumentation library. Because the number of input/output operations is large compared to the total execution time, and the overhead for the software performance instrumentation to compute input/output operation durations is high, performance instrumentation expands the execution time (The high cost of software performance instrumentation is due overwhelmingly to the overhead for clock accesses under UNIX.)

Experimental Environment	System Time	User Time	Total Time	Total Instrumented
UNIX	1578.2	1781.1	4299.3	6201.6
PPFS	400.4	1270.4	2300.8	4054.4

Table 1: Pathfinder Execution Times (seconds)

Experimental	Read		Write				
Environment	Count	Bytes	Count	Bytes	Seek	Open	Close
UNIX	3,030,382	2.48247e+10	4,077,265	625,698,239	10,961,293	41	41
PPFS	3,957,852	629,901,726	3505	766,433,994	3,897,144	42	42

Table 2: Pathfinder Input/Output Operations



(a) Read durations

(b) Seek durations

Figure 1: Pathfinder Input/Output Times (Instrumented Execution)

As we noted earlier, the Pathfinder code is dominated by irregular read and write accesses; Table 2 shows that many of these are small. The combination of small and non-sequential accesses is poorly matched to the standard UNIX file system policies, resulting in poor performance. As the PPFS data in Table 1 suggests, performance can be improved substantially by a judicious match of access patterns to file system policies.

3. PPFS: An Experimental Infrastructure

During the project, we developed a portable parallel file system (PPFS) [2] to study the interaction of application access patterns, file caching and prefetching algorithms, and application file data distributions. (This work is jointly supported by NSF grant 92-12369, "Multicomputer Resource Management Algorithms" and by NASA and Intel fellowships for Christopher Elford and Tara Madhyastha.) The PPFS distribution, which includes source code, documentation, and experimental data, is available via the WWW at http://www-pablo.cs.uiuc.edu/Projects/PPFS/.

3.1 PPFS Design and Performance

Quantitative assessment of file systems and policies necessarily presumes some underlying experimental infrastructure. To minimize implementation effort while focusing attention on the salient details, we have developed a user-level portable parallel file system (PPFS) [2,4,3,5,10]. PPFS is an input/output library, portable across parallel systems and workstation clusters, with a rich interface for application control of data placement and file system policies.

In the PPFS input/output model, files are accessed as either fixed or variable length records, and the PPFS library has an extensible set of interfaces for specifying file distributions, expressing input/output parallelism, and tuning file system policies. For example, one can specify how file records are distributed across input/output nodes, how and where they are cached, and when and where prefetch operations should be initiated.

As Figure 2 shows, the user-level PPFS library satisfies input/output requests via the interaction of three basic components: I/O servers, metadata servers and application clients. On a workstation cluster or

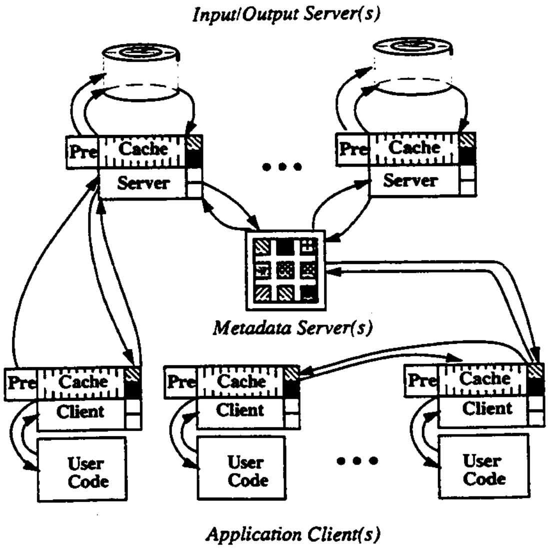


Figure 2: Basic PPFS Infrastructure

parallel system, PPFS servers are processes that mediate requests from application tasks, or clients, that are linked with PPFS interface functions. On the I/O servers, all "physical" input/output is performed through an underlying UNIX file system.

In this model, an application client opens a file by first contacting a metadata server that stores or creates information about the file storage order on the remote I/O servers. With this information, the application can specify caching and prefetching polices for local client caches and I/O servers. During application execution, the client caches either satisfy requests or forward them to I/O servers.

Experiments using large research codes on the Intel Paragon XP/S and IBM SP/2 have shown that tuning PPFS file system policies to application needs, rather than forcing the application to use inappropriate and inefficient file access modes, is the key to performance. In short, simple access pattern hints and cache policy controls can yield large performance benefits [2].

3.2 PPFS Ports and Distribution

The user-level implementation of PPFS allows one to quickly retarget the PPFS infrastructure to new hardware and operating system configurations. After developing PPFS on the Intel Paragon XP/S, we ported the software to several new configurations; Table 3 shows the platforms on which PPFS has been tested. As noted earlier, the PPFS software is available via the WWW at http://www-pablo.cs.uiuc.edu/Projects/PPFS.

Platform	MPI	PVM	NX
Intel Paragon XP/S			yes
IBM SP/2	yes		
SGI Indy/Challenge	yes		
Convex Exemplar	yes	yes	
Sun Solaris	yes	yes	
Sun SunOS	yes		yes
Linux PC	yes	yes	

Table 3: Currently Supported PPFS Platforms

The user-level implementation of PPFS makes possible software development and testing on workstations and workstation clusters. Because our local environment is expanding from predominately Sun workstations and file servers to include Silicon Graphics systems as well, we ported PPFS to SGI platforms. In addition, as our Sun systems were upgraded from SunOS to Solaris, we migrated PPFS development to Solaris.

Finally, we recently ported PPFS to PCs running the free Linux implementation of UNIX. In turn, NASA Goddard CESDIS researchers have used this port to migrate PPFS to the NASA Goddard Beowulf parallel PC testbed.

4. Automatic Access Pattern Classification

To achieve performance gains with PPFS, the application writer must understand both the application access pattern and the PPFS input/output cost model to specify appropriate policy controls and file data distributions. Because our characterization studies have shown that developers often do not know their file access patterns in sufficient detail to correctly choose file policies, we have investigated the benefits of automatic behavioral classification techniques for optimizing file system policies. We extended PPFS to include a trained neural net that accepts simple access pattern information. The neural net classifies the access pattern based on the predominance of operation types, sequentiality, and request size. This platform-independent classification is used by PPFS to select caching policies that are most efficient given the access pattern and input/output architecture. Most recently, we extended this single processor (local) classification scheme to accommodate global classification of access patterns on multiple processors.

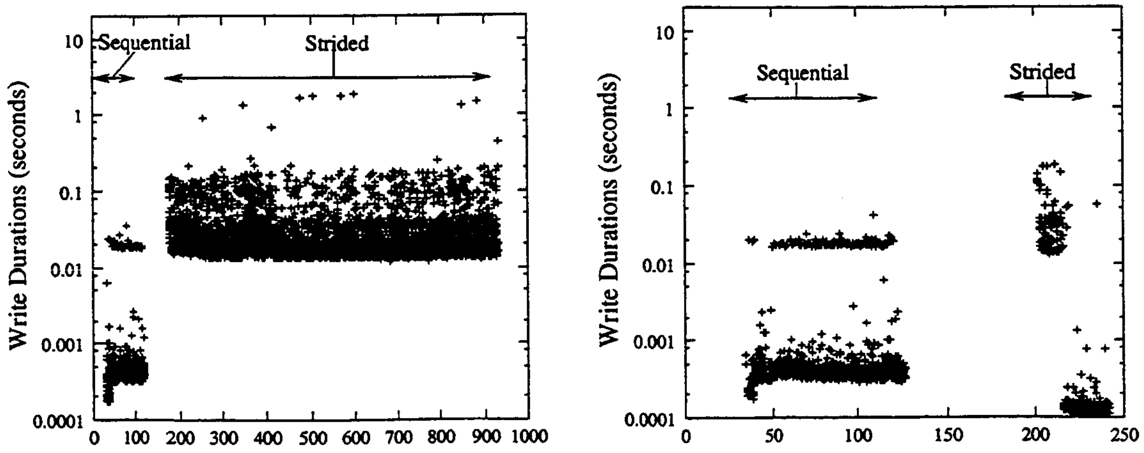
4.1 Local Classification Experiences

One of the main advantages of automatic access pattern classification is that PPFS policies can continually adapt to changing access patterns, yielding better performance than nonadaptive policy selection. Below, we describe recent experiences with a NOAA/NASA Pathfinder code.

In section 2, we described our input/output analysis of the Pathfinder code from the NOAA/NASA Pathfinder AVHRR (Advanced Very High Resolution Radiometer) data processing project. This analysis showed that the code is both input/output intensive and has a wide variety of changing access patterns. In consequence, the automatic behavioral classification and adaptive file system policies of PPFS potentially offer a significant performance improvement over that possible with UNIX buffered input/output alone.

To access the quantitative benefits of automatic classification and dynamic adaptation, we have used the Pathfinder code to conduct a series of comparative performance experiments. In a baseline set of experiments, the Pathfinder code relies on the standard UNIX file system for all input/output. Using PPFS, the neural net access pattern classifier monitors the Pathfinder access patterns and dynamically changes PPFS file management policies based on the detected access pattern. For the Pathfinder code, PPFS detects that the output file access pattern is initially write only and sequential, with large accesses; later, the access pattern changes to write only and strided, with very small accesses. PPFS chooses an MRU cache block replacement policy for the first phase. In the second phase, it enlarges the cache to retain the working set of file blocks.

Figure 3 shows the file write durations during the first phase of Pathfinder execution, both with and without the PPFS adaptive policies, when executed on a single processor Sun SPARC server. The first cluster of accesses at the left of Figures 3a and 3b is the sequential write phase. Performance for the first phase is roughly equivalent using either MRU or the default, non-adaptive LRU replacement policy. However, enlarging the cache in the second, strided access phase substantially decreases the average write duration and overall execution time.



(a) UNIX (Non-adaptive)

(b) PPFS (Adaptive)

Figure 3: File Write Durations (Pathfinder on Sun SPARC 670)

Read performance of one of the ancillary input data files also benefits from adaptive policies. First, the file is read sequentially with a large request size, and thereafter, accesses become small and variably strided. In the first phase, PPFS optimizes input/output by disabling caching and issuing the large reads directly. When the access pattern changes from sequential to variably strided, PPFS can either enlarge the cache or continue to read through to the underlying UNIX file system. Because the number of bytes requested with each read access is small, disabling caching outperforms UNIX buffered input/output due to the savings in unnecessary buffer copying.

4.2 Global Classification Experiences

As an extension of local access pattern classification schemes, we have developed a global classification infrastructure. Our global classification infrastructure is based on an access pattern temporal algebra. We combine local classifications and other local information to make global classifications. The number of processors contributing to the global access pattern is called the cardinality of the classification. Generally, we attempt to make global classifications with cardinality p, where p is the number of processors involved in the global input/output. However, a global classification involving a subset of the total number of processors is still useful for policy selection. A partial global classification may even be preferable, if it more accurately represents the temporal characteristics of the global access pattern.

A global access pattern classification is determined by a combination of local classifications. In addition, to identify global sequentiality, quantitative information about the input/output access stream is used to ``correlate" the local classifications within the global file context. For example, if every local access pattern is sequential, the beginning and end of every sequential stream is used to determine whether the global pattern is global sequential (every process reads the entire file sequentially) or partitioned sequential (the entire file is read in disjoint, sequential segments).

A second important consideration is that global classifications are valid only for a specific time interval. We define this time interval as the intersection of the local classification windows, ensuring that the local access patterns can and do overlap in time as well as within the file.

We have built a global classification framework that implements this temporal algebra. We have ported a version of this modified PPFS to the Intel PFS (i.e., input/output is performed through the native portable parallel file system) to demonstrate how global classification can accurately choose appropriate Intel PFS modes. Specification of the correct PFS mode, determined by the access pattern, is crucial for obtaining high performance. Preliminary experiences [5] shows major performance improvements are possible.

5. Performance-Directed Adaptive Control

Qualitative access pattern classification, used to select and tune resource management policies, is an important step toward creation of adaptive input/output controls. However, it must be coupled with quantitative data on resource utilization to provide true closed loop control. In a complementary effort to explore the efficacy of adaptive input/output control, we have coupled performance sensor data with user assertions that describe application input/output patterns and a prototype library of extensible, object-oriented resource policy actuators. Our hypothesis is that the PPFS sensor data can be processed in real time and used by the actuator library to identify and eliminate the current PPFS bottleneck.

5.1 Sensor Experiences

The PPFS actuator library contains three generic actuator types that can print actuator values to monitor system sensors, contain other actuators to form more complex combinations from simple components, or select an action based on sensor values that lie in specified ranges. We tested the PPFS sensors and this actuator library using both a parallel genome sequence matching code [2] and simple parallel input/output benchmarks. Our goal was to track changes in application access characteristics via their indirect effects on the performance sensors.

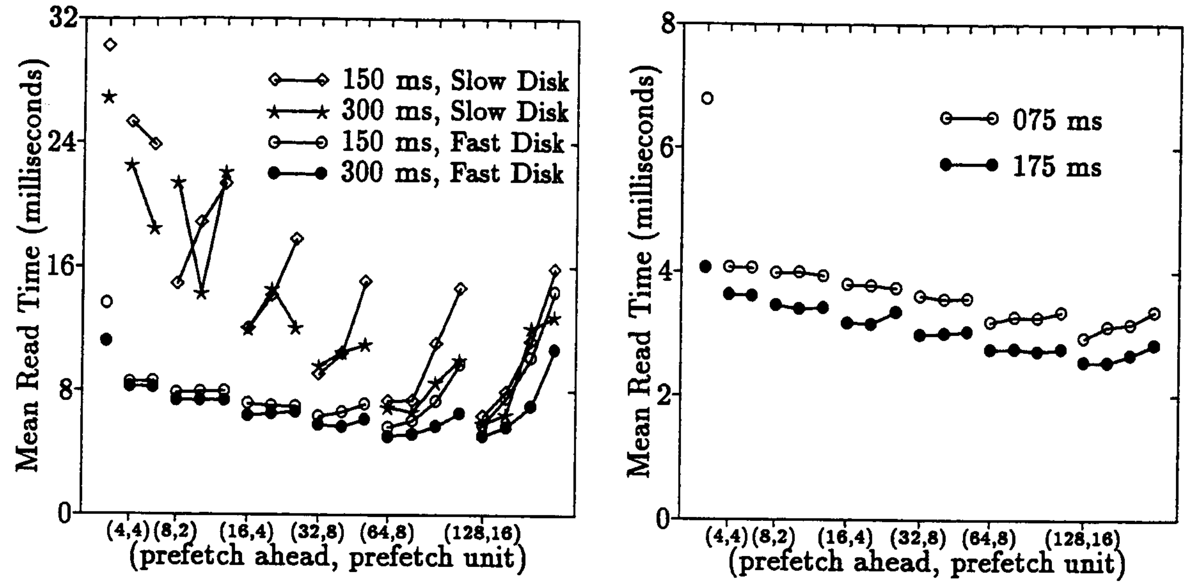
The results of these experiments showed that a richer set of performance sensors was needed to accurately monitor PPFS behavior and its responses to changing application access patterns. Hence, we have expanded number of distributed performance sensors contained in PPFS to include those shown in Table 4. Each of these sensors is currently computed for each processor by registering them with the Pablo sliding window instrumentation extension (SWAVE).

Performance Metric	Metric Description
Operation Count Operation Time	Total number of I/O requests Mean operation service time
Read Count Read Byte Count Read Time	Number of read Requests Number of bytes read Mean read service time
Write Count Write Time	Number of write Requests Mean write service time
Cache Hits Server Cache Hits Cache Check Time	Number of requests serviced by caches Number of requests serviced by off processor caches Time to check local cache
Off Node Time	Time for off processor demand accesses
Server Time Server Queue Time Server Queue Lengths	Time on I/O servers Time in disk queue Length of disk queue
Prefetch Byte Count Prefetch Cache Check Time Prefetch Off Node Time Hit Miss Time	Number of bytes prefetched Time to scan cache on prefetch initiation Time off processor for prefetch operations Time waiting for overlapped prefetch to complete

Table 4: PPFS Sensor Metrics

5.2 Performance Experiments

To verify that our sensors correctly tracked behavioral variations, we conducted a set of performance experiments using an input/output benchmark on the IBM SP/2 and on two different hardware configurations of the Intel Paragon XP/S. In this benchmark, all processors read disjoint, interleaved 4 KB blocks of the same file, forming a globally sequential access stream. Using the sensor-enhanced PPFS, we measured the average read access time as a function of file interaccess latency and file cache prefetch parameters (i.e., the number of file blocks prefetched ahead of the current access point and the number of blocks prefetched at a time). Figure 4 shows the results of these experiments.



(a) Intel Paragon XP/S

(b) IBM SP/2

Figure 4: PPFS Parallel Read Times (One File Server)

With large interaccess delays, moderate prefetching suffices to service all requests from the PPFS file cache. As the interaccess time declines and the aggregate request rate increases, PPFS must prefetch a larger number of blocks to maintain high cache hit ratios. Figure 4 shows that an efficient operating point depends on the request size, interaccess time, and disk performance. Thus, the slower disk system on the Paragon XP/S necessitates more aggressive prefetching than on the faster IBM SP/2. Intuitively, sensor data, together with access pattern information, would allow actuators to optimally configure prefetching policies for a given system. We are continuing to explore this approach with research funds from other sources.

6. Related Work

Several other projects have complemented the work supported by this contract. In particular, the Scalable I/O (SIO) initiative and other efforts supported capture and analysis of input/output access patterns and graphical display of the resulting data. Below, we briefly describe these research projects and how they are aided the work of this contract.

6.1 Application Characterization

Our application characterization efforts have been supported by a complementary, DARPA-sponsored project and an NSF Grand Challenge project that provide post-doctoral associates and staff software developers. Recent efforts have centered on analyzing the interactions of application access patterns, operating system versions, and disk hardware configurations. These characterizations [1, 12, 14, 9, 11] have shown that parallel applications exhibit a wide variety of input/output request patterns, with both very small and very large request sizes, sequential and non-sequential access, and a variety of temporal variations.

We have also tracked the performance of multiple codes across three releases of the Intel OSF/1 parallel file system (PFS) on the Caltech Paragon XP/S and across two hardware configurations [12]. These comparative operating system measurements show that Intel has reduced the cost of shared file opens for large numbers of nodes, a major bottleneck in earlier operating system versions, and has improved the performance of small input/output operations.

Finally, Caltech upgraded one of their Paragon XP/S systems to a configuration with a local disk on each of 80 nodes. This allowed us to conduct comparative measurements on two systems, one with remote I/O nodes and another with local disks. The results of these experiments showed that local disks not only increase input/output performance, they greatly reduce the need for application program input/output request aggregation (i.e., higher input/output performance can reduce program complexity).

6.2 Scalable Input/Output Initiative

Recall that the Scalable Input/Output (SIO) initiative is a broad-based, multi-agency (DARPA, DOE, NSF, and NASA) research program that includes application and system input/output characterization, networking, file systems and file system application programming interfaces (APIs), compiler and language support, and basic operating system services. The Illinois portion of the SIO project, supported by NASA through DARPA, is focused on unified characterization of application and system input/output data
We analyzed the Intel OSF/1 source code and built a new OSF/1 kernel for the Paragon XP/S that contains enabled disk input/output instrumentation [11]. The results show that disk hardware features profoundly affect the distribution of request delays and that current parallel file systems respond to parallel application input/output patterns in non-scalable ways.

6.3 Performance Data Immersion

We have exploited immersive visualization techniques to understand the temporal and spatial patterns of input/output present in parallel applications. At Supercomputing '95 and '96, we demonstrated two representations of input/output data: a time tunnel view of interprocessor interactions and a generalization of scatterplots that shows the temporal evolution of performance metrics.

We have tested these representations with several users and have found that virtual environment exploration can provide substantive new insights into software behavior and structure. In general, the ability to walk and fly through the data, to examine it from multiple perspectives, and to interactively change real-time display attributes has proven invaluable – we and others gained insights into behavioral and performance metric interactions that were not possible otherwise. For additional details, see the URL at http://www-pablo.cs.uiuc.edu/Projects/VR.

7. References

[1] Crandall, P. E., Aydt, R. A., Chien, A. A., and Reed, D. A. Input/Output Characteristics of Scalable Parallel Applications. In *Proceedings of Supercomputing '95* (Dec. 1995).

[2] Huber, J. V., Elford, C. L., Reed, D. A., Chien, A. A., and Blumenthal, D. S. PPFS: A High-Performance Portable Parallel File System. In *Proceedings of the 9th ACM International Conference on Supercomputing* (July 1995), pp. 385--394.

[3] Madhyastha, T. M., Elford, C. L., and Reed, D. A. Optimizing Input/Output Using Adaptive File System Policies. In *Proceedings of the Fifth Goddard Conference on Mass Storage Systems and Technologies* (Sept. 1996).

[4] Madhyastha, T. M., and Reed, D. A. Intelligent, Adaptive File System Policy Selection. In *Frontiers '96* (1996).

[5] Madhyastha, T. M., and Reed, D. A. Input/Output Access Pattern Classification Using HiddenMarkov Models. Submitted for publication (Mar. 1997).

[6] NCSA. NCSA HDF, Version 3.3. National Center for Supercomputing Applications, University of Illinois, Mar. 1993.

[7] Reed, D. A. Performance Instrumentation Techniques for Parallel Systems. *In Models and Techniques for Performance Evaluation of Computer and Communications Systems*, L. Donatiello and R. Nelson, Eds. Springer-Verlag Lecture Notes in Computer Science, 1993.

[8] Reed, D. A., Aydt, R. A., Noe, R. J., Roth, P. C., Shields, K. A., Schwartz, B. W., and Tavera, L. F. Scalable Performance Analysis: The Pablo Performance Analysis Environment. In *Proceedings of the Scalable Parallel Libraries Conference*, A. Skjellum, Ed. IEEE Computer Society, 1993.

[9] Reed, D. A., Elford, C. L., Madhyastha, T., Scullin, W. H., Aydt, R. A., and Smirni, E. I/O, Performance Analysis, and Performance Data Immersion. In *Proceedings of MASCOTS '96* (Feb. 1996), pp. 1-12.

[10] Reed, D. A., Elford, C. L., Madhyastha, T. M., Smirni, E., and Lamm, S. E. The Next Frontier: Interactive and Closed Loop Performance Steering. In *Proceedings of the 1996 ICPP Workshop on Challenges for Parallel Processing* (Aug. 1996).

[11] Simitci, H., and Reed, D. A. A Comparison of Logical and Physical Parallel I/O Patterns. Submitted for publication (Mar. 1997).

[12] Smirni, E., Aydt, R. A., Chien, A. A., and Reed, D. A. I/O Requirements of Scientific Applications: An Evolutionary View. In *Proceedings of the Fifth IEEE Symposium on High-Performance Distributed Computing* (1996).

[13] Smirni, E., Elford, C. L., and Reed, D. A. Performance Modeling of a Parallel I/O System: An Application Driven Approach. In *Proceedings of the Eighth SIAM Conference on Parallel Processing for Scientific Computing* (Mar. 1997).

[14] Smirni, E., and Reed, D. A. Workload Characterization of Input/Output Intensive Parallel Applications. In *Proceedings of the 9th International Conference on Modeling Techniques and Tools for Computer Performance Evaluation* (June 1997).

UNIVERSITY OF MINNESOTA

Fast I/O for Massively Parallel Applications

Matthew O'Keefe
Department of Electrical Engineering
(okeefe@ee.umn.edu)

Thomas Ruwart and Paul R. Woodward
Army High Performance Computing Research Center

1. Overview

The two primary goals for this research were the design, construction and modeling of parallel disk arrays for scientific visualization and animation, and a study of the IO requirements of highly parallel applications. In addition, we pursued further work in parallel display systems required to project and animate the very high-resolution frames resulting from our supercomputing simulations in ocean circulation and compressible gas dynamics.

2. Results and Transitions

With major additional support from the Army Research Office, NSF, and our corporate sponsors we constructed, modeled and measured several large parallel disk arrays. These arrays consisted of Ciprico 6700 RAID-3 devices (8 data + 1 parity drive) combined together in a variety of configurations, from a group of 8 RAID-3 from which we achieved nearly 100 MBytes/second transfer speed to a 31 array system that achieved a record 500 MBytes/second. These large bandwidths are necessary to support the high-resolution frame rates we require for the 2400x3200 pixel PowerWall parallel display system.

In addition to constructing these disk systems and measuring their performance, we developed performance models that capture many of the performance-limiting effects, such as start-up delays on RAID devices, fragmentation, and virtual memory page management overhead for very large transfers. We developed new techniques for instrumenting the kernel for taking filesystem performance data.

Other projects including performance measurements and experiments with D2 Helical Scan tapes from Ampex Corporation. We verified tape performance exceeding 15 MBytes/second for large transfers using the Ampex DST 310 tape device. In addition, Thomas Ruwart collaborated with storage vendor MTI on the construction of a 1-Terabyte filesystem using a collection of MTI RAID arrays.

Using the high speed disk subsystems to supply the bandwidth, we constructed a 4 panel PowerWall display system in our NSF-support Laboratory for Computational Science and Engineering following our successful (and partially NASA-sponsored) prototype at the Supercomputing '94 conference. A critical component of this system is the software that allows parallel rendering across the separate but seamlessly connected panels. Russell Catellan was partially supported by NASA to construct this software, which includes a version of XRaz used for scientific animation and also a modified version of VIZ, a 3D volume renderer developed in Norway. The PowerWall has inspired a host of imitations throughout the HPC community, including NASA Goddard. It is useful for a variety of high-resolution display applications, including our primary mission of visualizing and analyzing datasets generated by our simulation software on supercomputers.

Finally, we developed a package for performing parallel IO on the Cray T3D machine that is used by our regular grid applications such as the Miami Isopycnic Coordinate Ocean Model. This software is portable to other platforms, including the SGI Challenge class machines.

NASA support has helped produce two MS students and approximately 8 technical papers, as well as a variety of software and other research products, such as movies used by other researchers.

3. Graduate Theses Supported

Student's Name	Date	Degree	Thesis Title
Steve Soltis	June 1995	Masters	Instrumenting a UNIX Kernel for Event Tracing
Derek Lee	Feb 1995	Masters	Scientific Animation
Jeff Stromberg		pending Masters	Performance Effects of File Fragmentation

4. Research Products

4.1 Digital Movies

MPEG movies from the calculation described in journal reference [11] are available on the WWW at URL address: "http://www-mount.ee.umn.edu/~dereklee/micom_movies/micom_movies.html". These movies were recently reference by Semtner in his article on computer simulations of ocean circulation which appeared in the September issue of Science. As of November 16th, there have been 1557 accesses to this Web page. Actual data from our runs is also available at the Web site.

4.2 The PowerWall Project

In collaboration with Paul Woodward's team and several computer vendors, including Silicon Graphics Inc., Ciprico Inc., and IBM, my group helped to construct and demonstrate a high-resolution display system for datasets resulting from supercomputer simulations, medical imaging, and others. My group helped in the control software, data preparation and processing, and the actual physical construction. This system was demonstrated at the Supercomputing '94 conference and was described in conference publication [21]. A PowerWall, funded through an NSF CISE grant and with partial support from NASA and additional equipment grants from SGI and others, is now in operation in IT's Laboratory for Computational Science and Engineering. See our Web page on the PowerWall at URL: http://www-mount.ee.umn.edu/~okeefe.

5. Software Developed

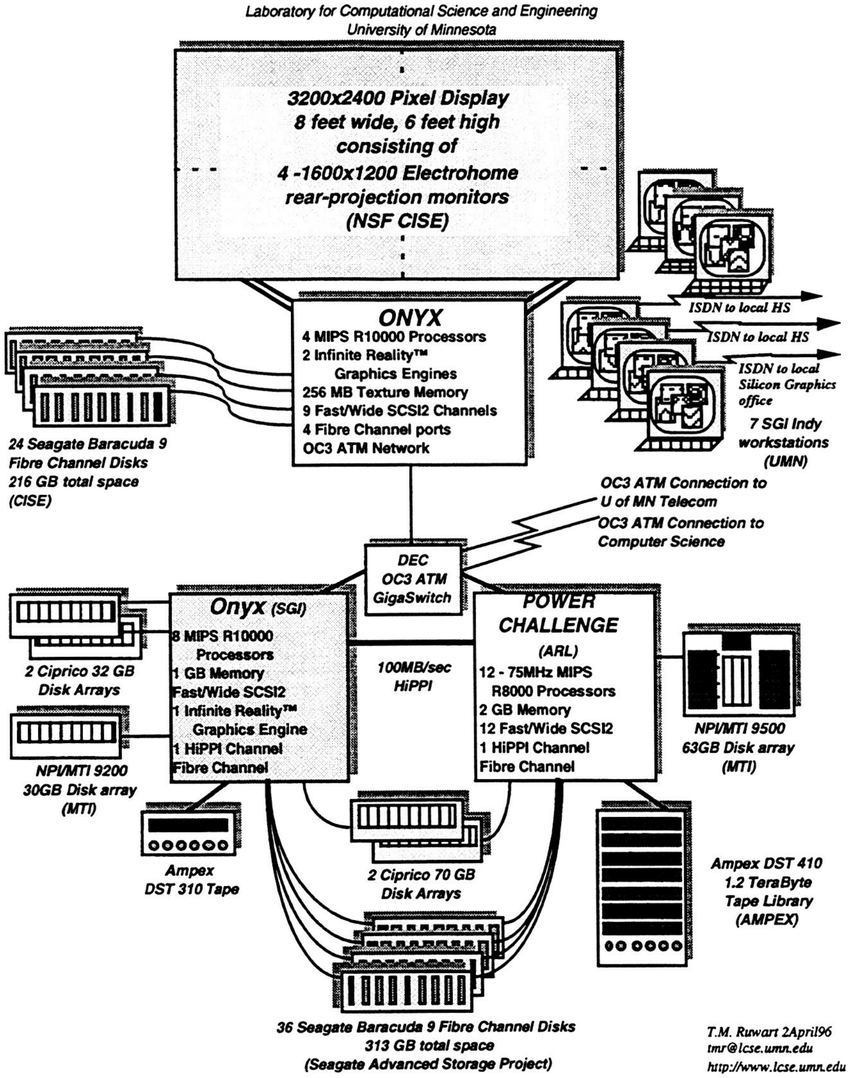
5.1 PowerWall Control Software

NASA support helped further the development of the control software for our parallel display system known as the PowerWall. This scalable display allows high-resolution supercomputer simulations to be shown in their totality to both small and large audiences. The disk array systems constructed partly with NASA support provided the more than 300 MegaBytes per second data throughput required by the PowerWall. First constructed at Supercomputing '94, we have constructed a PowerWall with NSF support in our own laboratory.

5.2 UNIX kernel trace and fragmentation measurement routines

These routines provide a means of measuring OS kernel performance and the effects of file fragmentation. Available on the Web at URL address: http://www-mount.ee.umn.edu/~soltis.

Current LCSE Equipment Configuration



6. Papers Published

[1] Thomas M. Ruwart and Matthew T. O'Keefe, Performance Characteristics of a 100 MegaByte/ Second Disk Array, *Storage and Interfaces '94*, Santa Clara, CA, January 1994.

[2] Aaron C. Sawdey, Matthew T. O'Keefe, Rainer Bleck, and Robert W. Numrich, The Design, Implementation, and Performance of a Parallel Ocean Circulation Model, *Proceedings of the Sixth ECMWF Workshop on the Use of Parallel Processors in Meteorology*, Reading, England, November 1994. Proceedings published by World Scientific Publishers (Singapore) in Coming of Age, edited by G-R. Hoffman and N. Kreitz, 1995.

[3] Paul R. Woodward, Interactive Scientific Visualization of Fluid Flow, *IEEE Computer*, Oct. 1993, vol. 26, no. 10, pp. 13-26.

[4] Thomas M. Ruwart and Matthew T. O'Keefe, A 500 MegaByte/Second Disk Array, *Proceedings of the Fourth NASA Goddard Conference on Mass Storage Systems and Technologies*, pp. 75-90, Greenbelt, MD, March 1995.

[5] Aaron Sawdey, Derek Lee, Thomas Ruwart, Paul Woodward and Matthew O'Keefe, and Rainer Bleck, Interactive Smooth-Motion Animation of High Resolution Ocean Circulation Calculations, *OCEANS '95 MTS/IEEE Conference*, San Diego, October 1995.

[6] Steve Soltis, Matthew O'Keefe, Thomas Ruwart and Ben Gribstad, The Global File System (GFS), to appear in the *Fifth NASA Goddard Conference on Mass Storage Systems and Technologies*, September 1996.

[7] Steven R. Soltis, Matthew T. O'Keefe and Thomas M. Ruwart, Instrumenting a UNIX Kernel for Event Tracing, submitted to *Software: Practice and Experience*, 1995, under revision.

[8] Aaron C. Sawdey and Matthew T. O'Keefe, A Software-level Cray T3D Emulation Package for SGI Shared-memory Multiprocessor Systems, submitted to *Software: Practice and Experience*, June 1995, under revision.

7. Technical Reports

[1] Aaron C. Sawdey, *Using the Parallel MICOM Code on the SGI Challenge Multiprocessor and the Cray T3D*, technical report, University of Minnesota, available on the WWW at http://www-mount.ee.umn.edu/~sawdey.

UNIVERSITY OF TEXAS AT ARLINGTON

Parallel Knowledge Discovery from Large Complex Databases

Diane J. Cook and Lawrence B. Holder
Department of Computer Science Engineering
(cook@cse.uta.edu, holder@cse.uta.edu)

1. Research Orientation

This report begins by restating the objectives, approach, and arguments from the original proposal "Parallel Knowledge Discovery from Large Complex Databases". The following section will describe progress made before this reporting period, progress made during the current reporting period, and planned activities for the next period. The last section will describe the educational benefits that have been derived from this effort.

1.2 Parallel Knowledge Discovery

NASA is focusing on grand challenge problems in Earth and space sciences. Within these areas of science, new instrumentation will be providing scientists with unprecedented amounts of unprocessed data. Our goal is to design and implement a system that takes raw data as input and efficiently discovers interesting concepts that can target areas for further investigation and can be used to compress the data. Our approach will provide an intelligent parallel data analysis system.

This effort will build upon two existing data discovery systems: the Subdue system used at the University of Texas at Arlington, and NASA's AutoClass system. AutoClass has been used to discover concepts in several large databases containing real or discrete valued data. Subdue, on the other hand, has been used to find interesting and repetitive structure in the data. Although both systems have been successful in a variety of domains, they are hampered by the computational complexity of the discovery task. The size and complexity of the databases expected from Earth and Space Science (ESS) program will demand processing capabilities found in parallel machines.

This effort focuses on combining the two approaches to concept discovery and speeding up the discovery process by developing a parallel implementation of the systems. The two discovery systems will be combined by using Subdue to compress the data fed to AutoClass, and letting Subdue evaluate the interesting structures in the classes generated by AutoClass. The parallel implementation of the resulting AutoClass/ Subdue system will be run on a massively-parallel machine (nCUBE), and will be tested for speedup on a number of large databases used by the ESS program.

1.3 Goals Stated in the Proposal

A general goal of the proposed research was to develop a combined discovery system (Subdue + AutoClass) to be applied to a variety of earth and space science databases. Our objective was to improve overall system speed, discovery power, and generality, by:

1. Improving the existing Subdue discovery system for application to structural Earth and space databases;

2. Using the Subdue system as a pre- and post-processor for NASA's AutoClass discovery system; and

3. Speeding up the combined discovery process through improved algorithms and parallel implementations on MIMD machines such as the nCUBE and Connection Machine 5.

2. Recent Work and Planned Work

This section organizes previous and future work around issues in machine discovery.

2.1 Parallel Subdue / AutoClass

The first step to parallelizing the Subdue / AutoClass discovery system is to ensure that both systems use the same programming language, and that a language be chosen that is supported by a majority of parallel systems. Joe Potts, one of our students who worked with the NASA Ames group in the summer of 1994, completed a port of AutoClass from Lisp to C. That port is now finished and is available from our anonymous ftp site (csr.uta.edu). The NASA Ames group has announced this distribution and is quite happy with the results and with the ability to distribute the code, as they cannot distribute the code from NASA. We have already received several requests for the C version of AutoClass from Lockheed, NASA Ames, and the Jet Propulsion Laboratory.

One benefit of Mr. Potts' stay at Ames was a redesign of the parallelization strategy for AutoClass. The initial strategy for parallelizing at a high level was abandoned after discussions with members of the Ames group indicated the size of typical data sets (e.g., Landsat images) would require too much data redundancy across processors. The high-level approach was set aside in favor of a low level approach that would spread the large number of observations out over the processors rather than distributing classes, classifications, or attributes.

The SIMD parallelization of AutoClass on the Connection Machine 5 is complete. Joe Potts will be defending his Masters thesis soon on the design and evaluation of AutoClass' parallelization. Through an NSF equipment grant awarded to Dr. Cook, the department has acquired a 128-processor nCUBE. Ports of both parallel AutoClass and Subdue will be performed for this machine as well.

A MIMD parallelization of Subdue is now complete, using a 128-processor nCUBE. Gehad Galal is working on MIMD-parallel and distributed version of the Subdue system. The parallel version of the system allocates distinct sets of substructures for each processor to evaluate and grow. Load balancing is performed as processors run out of substructures to consider. The parallel system is completed and we are currently performing speedup experiments. A distributed version of the system, in which the database itself is distributed over multiple workstation connected with a high-speed switch, will be implemented this summer.

In addition, a MIMD-parallel version of AutoClass is underway. In this version, distinct processors are given a distinct number of classes into which to fit the data. The MIMD-parallel version of AutoClass will be finished this spring, at which time we can compare performance between MIMD and SIMD versions of the system and combine parallel version of Subdue and AutoClass.

2.2 Improvements to the Subdue System

The two ongoing directions for the improvement of Subdue are the inclusion of facilities for accepting background knowledge to guide the discovery process and the evaluation and refinement of Subdue to work with structural features extracted from image data.

2.2.1 Domain-Dependent Discovery in Subdue

Subdue is designed as a domain-independent tool for discovering concepts in structured data. To make Subdue's discovery process more useful across a wide variety of domains, domain knowledge can also be used to guide the discovery process. We expect that compressing the graph using combinations of domain-independent and domain-dependent knowledge will increase the chance of finding interesting substructures and realize even greater compression.

Work continues on Subdue's ability to include both domain-dependent and domain-independent background knowledge. The domain knowledge can be input using a hierarchical collection of substructures known to be interested for the given domain, using a collection of domain-specific graph match rules, or using a set of feature extraction functions. All forms of domain-dependent search guidance have been incorporated into the Subdue system. Experiments have been performed in the domains of program analysis and CAD analysis to compare the results of discovery with domain-independent knowledge to discovery with domain-dependent knowledge and discovery with both types of knowledge.

Recently, a probabilistic version of Subdue in which background knowledge can be encoded as probabilistic models was completed. The results of this work have been submitted to the *Artificial Intelligence Journal.* Surnjani Djoko, the main student responsible for integrating domain knowledge into Subdue, successfully defended her Ph.D. dissertation on this topic in August 1995.

2.2.2 Processing Images with Subdue

With the eventual goal of processing NASA imagery using our proposed integration of Subdue and AutoClass, we are evaluating Subdue's ability to discover patterns from structural features extracted from images. The computer vision portion of the project involves processing images using standard image processing techniques (e.g., edge detection). Low level image data is extracted from the images, and then this data is transformed into higher level symbols which are input to the Subdue system. The main region features that will be extracted are lines and angles between lines.

The results of these experiments have been submitted to the Florida AI Research Symposium. Stephen Poe, the student mainly responsible for evaluating the use of Subdue on image data, successfully defended his Masters thesis on this topic in August 1995.

2.3 Integration of Subdue and AutoClass

The integration of the structural discovery process in Subdue and the attribute-value clustering process in AutoClass will yield a system capable of discovering previously-undiscoverable patterns using the combination of structural and non-structural data features of the data. Our initial design for the integration was to use Subdue as a pre-processor of the structural component of the data in order to construct new non-structural attributes for addition to the set of existing, non-structural attributes. The new data, augmented with this non-structural information about the structural regularities in the data, can now be passed to the AutoClass system. The structural information will bias the classifications preferred by AutoClass towards those consistent with the structural regularities in the data.

Discussions with the AutoClass group at NASA Ames have suggested a different approach to the integration in which Subdue is included as a parameterized model with AutoClass. AutoClass evaluates classes by tuning the parameters of statistical models of the classes (e.g., single normal distribution for each attribute). Expressing Subdue as a parameterized model whose parameters select the types of substructures preferred by Subdue would allow a more seamless integration of the two systems.

When both Subdue and AutoClass have been ported to the nCUBE machine, we plan to evaluate both approaches to the integration of Subdue and AutoClass by comparing the results of AutoClass alone on data with a weak, non-structural classification and a strong, structural classification. We also plan to compare previous AutoClass results with the classifications obtained with the integrated discovery system using the same data augmented with natural structural information such as temporal and spatial orientations.

2.4 Graphical User Interface

In order to make Subdue easier to use and the results easier to interpret, a Masters student will be developing a graphical user interface to Subdue this summer. This interface will display the input database (or selected portions of the database) as a graph, and visually display substructures as they are discovered.

3. Educational Benefits of This Work

One Ph.D. student and two masters students have complete theses based on research funded by this USRA subcontract. In addition, we are currently supporting one Ph.D. student and one Master's student to work on this project.

Surnjani Djoko finished her Ph.D. dissertation in August 1995 on investigating methods for incorporating domain-dependent knowledge into a discovery system, and is currently working at Bell Northern Research. Joe Potts, a masters student, used a portion of the NASA money to visit NASA Ames in the summer of 1994. Mr. Potts has completed the conversion of AutoClass from Lisp to C and the parallelization of AutoClass on the CM 5. Mr. Potts defended his masters thesis in December of 1996 and plans to continue investigating related issues for his doctoral work. Stephen Poe obtained his masters degree during this report period on the topic of substructure discovery in image data. Gehad Galal will finish his Masters thesis this May on the topic of designing parallel and distributed versions of Subdue.

The research sponsored by this USRA subcontract has contributed to the course Dr. Cook is teaching on Parallel Algorithms for Artificial Intelligence. The results of the parallel discovery algorithms has been incorporated into the new syllabus so that additional students can benefit from ongoing research and contribute to the state of the art in efficient machine discovery methods.

Many of the issues involved in discovery of patterns in NASA-related data sets have been included in Dr. Hoider's course on Machine Learning. These issues have been incorporated into the projects done by the students in the class. Knowledge of these issues has given the students a better appreciation for the difficulties and potential benefits of machine discovery in such domains containing both structural and non-structural information. Masters student Stephen Poe became interested in this project through discussions in Dr. Holder's class.

4. Publications

4.1 Publications supported by this NASA project.

Djoko, S., Cook, D. J., & Holder, L. B., An Empirical Study of Domain Knowledge and its Benefits to Substructure Discovery, to appear in *IEEE Transactions on Knowledge and Data Engineering.*

Galal, G., & Cook, D. J., "*Exploiting Parallelism in a Scientific Discovery System to Improve Scalability*", to appear in *Proceedings of the Tenth Annual Florida AI Research Symposium, 1997.*

Djoko, S., Cook, D. J., & Holder. L. B., Discovering Informative Structural Concepts Using Domain Knowledge. *IEEE Expert, 10*, pages 59--68, 1996.

Cook, D. J., Holder, L. B., & Djoko. S., *Knowledge Discovery from Structural Data, Journal of Intelligence and Information Sciences*, Volume 5, Number 3, pages 229--245, 1995.

Djoko, S., Cook, D. J., & Holder. L. B., Analyzing the Benefits of Domain Knowledge in Substructure Discovery. In the *Proceedings of the First International Conference on Knowledge Discovery and Data Mining,* pages 75--80, 1995.

Cook, D. J., & Holder. L. B., Substructure Discovery Using Minimum Description Length and Background Knowledge. In *Journal of Artificial Intelligence Research,* Volume 1, pages 231--255, 1994.

Djoko. S., Guiding Substructure Discovery with Minimum Description Length and Background Knowledge, Student Abstract in *Proceedings of the Twelfth National Conference on Artificial Intelligence, page 1442*, 1994.

Holder, L. B., Cook, D. J., & Djoko. S., Substructure Discovery in the Subdue System, in *Proceedings of the AAAI Workshop on Knowledge Discovery in Databases, pages 169--180*, 1994.

4.2 Related publications submitted prior to this report period.

Holder, L. B., & Cook. D. J., Discovery of Inexact Concepts from Structural Data. In *IEEE Transactions* on Knowledge and Data Engineering, Volume 5, Number 6, pages 992--994, 1993.

Holder, L. B., Cook, D. J., & Bunke, H., *Fuzzy Substructure Discovery,* Ninth International Machine Learning Conference, Aberdeen, Scotland, pages 218--223, 1992.

Holder. L. B., Empirical substructure discovery. In *Proceedings of the Sixth International Workshop on Machine Learning, pages 133--136*, 1989.

UNIVERSITY OF VIRGINIA

High Performance Databases for Scientific Applications

James C. French, Andrew S. Grimshaw
Department of Computer Science
(french@cs.virginia.edu)

1. Highlights

During the first two years of funding we completed our work on high performance parallel I/O support for multidimensional range searches. This work is described more fully in [KARP94a, KARP94b, KARP94c, KARP94d, KARP94e]. In addition we began shifting our work in a new direction, file system support in high performance metasystems [GRIM94a,GRIM94b,GRIM95]. These two topics are discussed separately in greater detail below. To summarize, our main achievements in multidimensional range searching are as follows:

- We developed a general approach for attacking the high-performance I/O problem.
- We developed a parallel file object based on PLOP files designed to provide high-performance range queries in multiple dimensions.
- We tested the performance of range retrievals on representative queries provided to us by the National Radio Astronomy Observatory.

During the final year of funding we directed our work on high performance parallel I/O to file system support in high performance metasystems [GRIM94a,GRIM94b,GRIM95]. This topic is discussed in greater detail below. Our main achievement over the period was the further development of the Campus Wide Virtual Computer (CWVC) deployed here at the University of Virginia. The campus-wide virtual computer is a prototype for the nationwide Legion system in that the computational resources at the University are operated by many different departments; sharing of resources is currently rare; resources are owned by the departments, and this equipment is used for "production" applications during the day.

Even though the CWVC is much smaller and the components much closer together than in the envisioned nationwide Legion, it still presents many of the same challenges. The processors are heterogeneous, the interconnection network is irregular with orders of magnitude differences in bandwidth and latency, and the machines are currently in use for on-site applications that must not be negatively impacted. Each department operates essentially as an island of service, with its own NFS mount structure, and trusting only machines in the island.

The next section describes our work with multidimensional range searching. It is followed by an overview of Legion, our ongoing metasystems project and a discussion of our I/O work.

2. High Performance Parallel I/O Support for Multidimensional Range Searches

We have developed a general approach for attacking the high performance I/O problem, namely the Extensible File Systems (ELFS) approach based on work in [GRIM91]. This report describes an implementation following the ELFS approach for a specific class of retrieval patterns, multidimensional range searches. Multidimensional range searches appear in a wide range of applications, including many scientific applications. Such applications view a data set as an n-dimensional data space, where each dimension represents the values along a key field present in the data. The coordinates of each data record are its values for each of the n dimensions. Using this view, subvolumes of the data space can be defined by specifying a range of values for each dimension. For example, a data set containing a set of time indexed two dimensional images can be viewed as a three-dimensional data space (time,x,y). Possible range searches for such a data set include retrieving a specified region of each image (a rectangle in (x,y)) for all time values, retrieving full images for a certain range of times, etc.

In the following sections we first present the current methods used for providing range search capabilities and then briefly describe the general ELFS approach and its benefits. This is followed by a discussion of an instance of this approach designed for high performance multidimensional range searches including details of the parallel structure of the implementation. We also describe the tests we executed using interferometry data sets from the National Radio Astronomy Observatory (NRAO). More details on these tests can be found in [KARP94a,KARP94c,KARP94d,KARP94e].

2.1. Current Methods

There are several approaches typically employed to provide range searching capabilities on a set of data, each with varying degrees of implementation effort and performance. Many implementations store the data sets as simple sorted sequential files and scan the file, filtering out unwanted data outside of the desired subvolume. This approach is easy to implement, but performs poorly, especially when small amounts of data are desired relative to the file size. An improvement to this scheme uses an indexed, sorted file, to reduce the number of accesses needed to the file, improving performance at a modest complexity cost. This scheme works well for a one-dimensional space (for the sorted, indexed key), but does not generally perform well for multidimensional accesses. Some implementations designed for parallel applications, improve the performance of a single file by either replicating the data sets or partitioning the data set into separate disjoint sets. Each of these approaches is designed to alleviate the contention for the single file resource among multiple concurrent processes, but does not improve upon the basic access methods for each distributed file.

Another common approach is to use a commercial database management system (DBMS) which allows for the specification of range queries. Relational DBMS are particularly popular where range searches are easily defined using the Structured Query Language (SQL). This approach is easy from the implementation standpoint, but may not achieve acceptable performance. DBMS are built to support a wide range of possible access patterns and types of data and are not tuned to range searching in multiple dimensions. In addition, DBMS incur overhead for the guarantee of consistency within the database, which may not be an issue for many applications.

A less often used approach implements file structures specifically tailored for range searching such as PLOP files[KRIE88a,KRIE88b], grid files [NIEV84], k-d & k-d-b trees [BENT79,ROBI81], or quadtrees [SAME84]. These file structures offer performance advantages by attempting to preserve physical data locality in all of the dimensions of the data space and by providing efficient methods for finding particular regions of the data space. The drawback is that these file structures can be difficult to implement properly, especially in a distributed manner. Even when these structures are implemented, the implementations are often highly application-specific and not reusable, so the common practice is to build them virtually from scratch.

2.2. The ELFS Approach

The ELFS approach is to create file objects that satisfy four criteria:

(1) Match the file structure to the access patterns of the application and the type of the data. As the examples in the previous section point out, the organization and structure of the underlying file can greatly influence performance by reducing the number of accesses required. For distributed file

structures, effective data placement can potentially improve performance by reducing latency. Therefore it is important to match file structures with their use.

(2) Use parallel and other advanced I/O techniques. In a file type-specific manner exploit parallelism to overlap application computation with I/O requests to reduce the effective latency of a request (i.e. the wait actually experienced after issuing a request). For distributed file structures, true parallel access can be used to better utilize the file system's bandwidth. Other I/O-related functions, such as data conversion and sorting, may be performed in parallel to speed the overall performance of using stored data. Prefetching and caching are two other well-known performance-oriented techniques that can be employed when applicable.

(3) Improve the I/O interface to application programs. There are two main reasons for improving the file interface. Most importantly from a performance standpoint, is to allow the user to convey useful information that can be exploited by the file object implementation. For example, knowing the stride of accesses in a matrix file can be exploited to effectively prefetch data, or knowing that the file will be used in a read only fashion can allow the implementation to avoid potentially costly consistency protocols. The second reason for improving the interface is to make file objects easier to use by application programmers, reducing their programming burden.

(4) Encapsulate the implementation details in file objects. This goal is aimed at increasing the maintainability and reusability of the file objects. By using the object-oriented paradigm for the file objects, application programmers can derive new file objects from existing base objects and can then extend them and tailor them without reimplementing much of the file object functionality.

A suite of extensible file objects can be developed using this methodology, each performing best for a particular class of data types and access patterns. Application designers can then choose the best file object for there purposes and extend or tailor the file definition as needed, hopefully requiring only a modest amount of effort. Our early design work in this area has been reported in [KARP94b].

2.3. Parallel File Objects for Multidimensional Range Searches

Using the ELFS approach we have created a parallel file object designed to provide high performance for range queries in multiple dimensions. Our implementation uses the PLOP file as the basic underlying file structure. Though other file structures could be used for multidimensional range searches, it is our opinion that none of these candidates is clearly superior to PLOP files, while PLOP files have a relatively straightforward implementation. For a more in depth analysis of the choice of file structure see [KARP94a]. A PLOP file views a data set as a multidimensional data space. The data space is partitioned by splitting each dimension into a series of ranges called slices. The intersection of a slice from each dimension defines one logical data bucket. Data points are stored in the bucket that has corresponding values in each dimension. Therefore, within a bucket, the data points exhibit spatial locality in all dimensions. A tree structure for each dimension tracks the physical location of each bucket within the file, so that each bucket can be accessed very efficiently. This structure allows retrievals to eliminate parts of the file that do not correspond to values within the range search based on all dimensions, while quickly accessing those parts that may contain valid data.

We first implemented a sequential version of the PLOP file based file object. Though unable to take advantage of parallel techniques, this version exploits the structure of PLOP files to achieve efficient accesses. In addition to the obvious benefits of using the tailor-made structure of the PLOP file, a subtle performance improving enhancement was implemented for sorting by a key that is one of the dimensions. Because the data points contained in each slice along a dimension are disjoint, the data in each slice along the sort key can be sorted separately, and with each slice returned in order. By sorting in smaller batches, the complexity of the sort is reduced from O(nlogn) to O(n/plogn/p) in the ideal case (p = number of slices spanned by the request).

The parallel version is being implemented using Mentat [GRIM93c], an object-oriented parallel processing system. The design of the parallel implementation includes several significant changes from the sequential version. First, PLOP files have been modified to accommodate distributed pieces. These pieces can be created and distributed in three patterns: segmented (or partitioned) along a dimension, striped along a dimension, or blocked by some set of dimensions (e.g. using two dimensions each piece would be a rectangular region). The distributed PLOP file allows not only parallel access to the data, but also allows an application program to map processes to nodes near the data they will require.

Second, parallel I/O workers have been added to access each distributed file piece and a manager has been added to coordinate their activities. The workers asynchronously handle all requests for data at their piece, including I/O device access, data conversion and, if possible, data sorting. Our initial design has only a single manager process for all worker processes and clearly does not scale well for increased numbers of application processes requiring data. We already plan to replace this design with a scheme that will scale for increases in the number of application processes by enabling the manager to replicate itself and assign different managers to different application processes.

Third, the interface has been improved. A major improvement to the interface to decouple the definition of a query request from the retrieval of the data. The idea is to allow the user to specify a query ahead of the time the data will be used whenever possible, and to submit the query to be performed. The file object can asynchronously begin buffering the request while the application continues to do useful work. When the application actually wants the retrieved values, a call is made to ask for the data.

2.4. Performance Tests

To test our implementation, we have converted two interferometry data sets from NRAO's Very Large Array (VLA) radio antenna installation. The first file, a line spectrum file, is ~50 megabytes and 126,092 records, while the second file, a continuum spectrum file, is ~270 megabytes and contains 8,440,092 records. Initial results for the sequential version have been very encouraging. The converted PLOP files utilize space fairly efficiently, 79% and 66% for the line and continuum files respectively (efficiency is calculated by comparing actual storage used for records versus the total storage allocated, including overhead and fragmentation).

To test the performance of range retrievals, a set of twelve representative queries for NRAO's applications has been developed. Both files have been tested using these queries for the sequential version, with impressive results. The parallel version will be tested with the same suite of queries for various file distribution patterns and numbers of file pieces. These results and a comparison of results across the various parameters can be found in [KARP94a,KARP94c,KARP94d,KARP94e].

3. Legion: File and Data Access

Our work in file systems for high-performance heterogeneous metasystems is being done within the Legion project. We describe that next.

3.1. Legion

Legion will consist of workstations, vector supercomputers, and parallel supercomputers connected by local area networks, enterprise-wide networks, and the NII. The total computation power of such an assembly of machines is enormous, approaching a petaflop; this massive potential is, as yet, unrealized. These machines are currently tied together in a loose confederation of shared communication resources used primarily to support electronic mail, file transfer, and remote login. However, these resources could be used to provide far more than just communication services; they have the potential to provide a single, seamless, computational environment in which processor cycles, communication, and data are all shared, and in which the workstation across the continent is no less a resource than the one down the hall.

A Legion user has the illusion of a single, very powerful computer on her desk, which is used to invoke an application on a data set. It is Legion's responsibility to transparently schedule application components on processors, manage data transfer and coercion, and provide communication and synchronization, while trying to minimize execution time via parallel execution of the application components. System boundaries will be invisible, as will the location of data and the existence of faults.

The potential benefits of Legion are enormous: (1) more effective collaboration by putting coworkers in the same virtual workplace; (2) higher application performance due to parallel execution and exploitation of off-site resources; (3) improved access to data and computational resources; (4) improved researcher and user productivity resulting from more effective collaboration and better application performance; (5) increased resource utilization; and (6) a considerably simpler programming environment for the applications programmers. Indeed, it seems probable to us that the NII can reach its full potential only with a Legion-like infrastructure.

Before the Legion vision can be realized, several technical challenges must be overcome. These are software problems; the hardware challenges are being addressed and are the enabling technologies that provide the opportunity. The software challenges revolve around eight central themes: achieving high performance via parallelism, managing and exploiting component heterogeneity, resource management, file and data access, fault-tolerance, ease-of-use and user interfaces, protection and authentication, and exploitation of high-performance communications protocols.

3.2. Objectives

From our Legion vision we have distilled six primary design objectives:

- **Easy-to-use, seamless computational environment:** Legion must mask the complexity of the hardware environment and the complexity of communication and synchronization of parallel processing. Machine boundaries should be invisible to users. Legion will provide both user and programmer with a uniform interface to service. As much as possible, compilers, acting in concert with run-time facilities, must manage the environment for the user.

- **High performance via parallelism:** Legion must support easy-to-use parallel processing with large degrees of parallelism. This includes task and data parallelism and their combinations. Because of the nature of the interconnection network, Legion must be latency tolerant. Further, Legion must be capable of managing hundreds or thousands of processors. This implies that the underlying computation model and programming paradigms must be scalable.

- **Single, persistent namespace:** One of the most significant obstacles to wide-area parallel processing is the lack of a single name space for file and data access. The existing multitude of disjoint name spaces makes writing applications that span sites extremely difficult. Therefore, Legion must provide a single name space for persistent objects (files).

- **Security for users and resource owners:** Because we cannot replace existing host operating systems, we cannot significantly strengthen existing operating system protection and security mechanisms. However, we must ensure that existing mechanisms are not weakened by Legion.

- **Manage and exploit resource heterogeneity:** Clearly Legion must accommodate heterogeneity, i.e., it must support interoperability between heterogeneous components. In addition, Legion will be able to exploit diverse hardware and data resources, executing subtasks of large applications on different heterogeneous processors, and using heterogeneous data sources. Some architectures are better than others at executing particular kinds of code, e.g., vectorizable codes. These affinities, and the costs of exploiting them, must be factored into scheduling decisions and policies.

- **Minimal impact on resource owner's local computation:** The noticeable impact of Legion on local resources must be small, particularly with regard to interactive sessions. If users notice a significant performance penalty when their site is attached to Legion, they will withdraw; an observed penalty must be more than offset by the benefits of Legionnaire status.

We have the additional objective of demonstrating the effectiveness of wide-area heterogeneous computing on serious applications drawn from a variety of application domains, including "grand challenges", such as global climate modeling and the human genome project, and from economically significant problem areas such as electrical engineering and medicine.

3.3. Approach

The principles of the object-oriented paradigm are the foundation for the construction of Legion; our goal will be exploitation of the paradigm's encapsulation and inheritance properties. Use of an object-oriented foundation will render a variety of benefits, including software reuse, fault containment, and reduction in complexity. The need for the paradigm is particularly acute in a system as large and complex as Legion. Other investigators have proposed constructing application libraries and applications for wide-area parallel processing using only low-level message passing services such as those provided by PVM [SUND90] and P4 [BOYL87]. Use of such tools requires the programmer to address the full complexity of the environment; the difficult problems of managing faults, scheduling, load balancing, etc., are likely to overwhelm all but the best programmers.

Objects, written in either an object-oriented language or other languages such as HPF Fortran, will encapsulate their implementation, data structures, and parallelism, and will interact with other objects via well-defined interfaces. In addition, they may also have inherited timing, fault, persistence, priority, and protection characteristics. Naturally these may be overloaded to provide different functionality on a class-by-class basis.

Our approach to constructing Legion is evolutionary rather than revolutionary. We have begun by first constructing a Legion testbed by extending Mentat, an existing object-oriented parallel processing system [GRIM93d]. Mentat attacks the problem of providing easy-to-use high performance parallelism to users. Mentat has been used to implement several real-world applications on hardware platforms spanning the bandwidth/latency space in a heterogeneous environment [GRIM93a,GRIM93b,GRIM93e]. Mentat's object-oriented structure, and its ability to achieve high-performance on platforms with very different communications characteristics are the key factors in our choice of Mentat as our implementation vehicle. The testbed provides us with an ideal platform to rapidly prototype ideas, forcing the details and hidden assumptions to be carefully examined, and exposing flaws in the ideas or in the system components.

3.4. File/data Access

File and data access is one of the most crucial issues for Legion, particularly with respect to providing a seamless environment. Today, distributed file systems such as NFS, Andrew, and Locus are commonplace in local area networks. The unified level of service and the naming scheme that they present to their users make them one of the most successful components of contemporary distributed systems. In Legion we intend to provide the same level of naming and access transparency provided in local area networks. This cannot be accomplished either by directly extending current systems onto a national scale, or by imposing a single file system for both local and Legion access. Instead we propose to adopt a federated file system approach. The Legion file system will provide naming, access, location, fault, and replication transparency. It will permit users (or library writers) to extend the basic services provided by the file system in a clean and consistent fashion via class derivation and file-object instantiation and manipulation. The extensions that we intend to design and implement ourselves include application-specific file objects designed to improve application performance by reducing observed I/O latency.

Issues such as naming, location transparency, fault transparency, replication transparency, and migration have been addressed both in the literature [LEVY90] and in several existing operational systems. Rather than duplicate those efforts we will build on them and extend them into a larger context. The difficulty that arises when borrowing from an existing system is that most of the systems do not have a scalable system architecture or flexible semantics, or they require the imposition of a unified file system model, contrary to our federated file system goal. Therefore we will borrow ideas but not implementations, looking more to combine the work of others with our own.

Although we will continue to refer to the Legion "file system," we intend to create a persistent object space as has been proposed for distributed object management systems [NICO93]. There are several other efforts in the distributed object literature with which we share many goals, e.g., SHORE [CARE94] and CORBA [MANO92,NICO93]. Legion is distinguished from these efforts by the emphasis we place on performance. Legion expects to provide a high performance computing environment and this goal is paramount. To this end we will focus more on file system support than database support.

The model that we will employ is simple and driven by the observation that the traditional distinction between files and other objects is somewhat of an anachronism. Files really are objects - they happen to live on disk, as a consequence they are slower to access, and they persist when the computer is turned off. We define a file-object as a typed object with an interface. The interface can also define object properties such as its persistence, fault, synchronization, and performance characteristics. Thus, not all files need be the same, eliminating, for example, the need to provide Unix synchronization semantics for all files even when many applications simply do not require those semantics. Instead, the right semantics along many dimensions can be selected on a file-by-file basis, and potentially changed at run-time.

3.5. Agenda

Our agenda consists of three stages:

(1) the construction of a campus-wide virtual computer at the University of Virginia,

(2) packaging the campus-wide system for preliminary experimentation and use by Legionnaires, and

(3) expansion to a nationwide demonstration system. Each of these three stages will build upon the previous.

Before any major project is undertaken, one must ask how to measure success. In parallel processing, success is measured by application performance (speedup, elapsed time) and the flexibility and ease of use of the tool. Other important metrics include acceptance by the user community, fault-tolerance, cost per used MIP/FLOP, and whether tasks can be performed that were not possible before (e.g., run an application in Virginia on data that resides at NASA-JPL, or collect and use data in real time from sensors in orbit, but have that data look like any other "file").

Application performance will be measured for a variety of real-world applications, as well as selected kernel codes and parallel processing benchmarks. The applications will be drawn from a diverse set of disciplines: biology, physics, electrical engineering, chemistry, economics, radio astronomy, and command and control. The applications will possess different granularity characteristics, as well as different latency tolerances. It is not our intent, however, to show that all applications will be capable of exploiting the nationwide resources of Legion. Some applications, those with inherently small granularity or that are latency intolerant, will remain best suited to local operation, e.g., on a single processor or on a single tightly-coupled parallel processor.

3.5.1. Construction of a Campus-Wide Virtual Computer (CWVC)

The campus-wide virtual computer is a direct extension of Mentat to a larger scale, and is a prototype for the nationwide system in that the computational resources at the University are operated by many different departments; sharing of resources is currently rare; resources are owned by the departments, and this equipment is used for "production" applications during the day.

Even though the CWVC is much smaller, and the components much closer together, than in the envisioned nationwide Legion, it still presents many of the same challenges. The processors are heterogeneous, the interconnection network is irregular with orders of magnitude differences in bandwidth and latency, and the machines are currently in use for on-site applications that must not be negatively impacted. Each department operates essentially as an island of service, with its own NFS mount structure, and trusting only machines in the island.

The CWVC is both a prototype and a demonstration project. The objectives are to: demonstrate the usefulness of network-based, heterogeneous, parallel processing to university computational science problems; provide a shared high-performance resource for university researchers; provide a given level of service (as measured by turn-around time) at reduced cost; and act as a testbed for the nationwide Legion.

The prototype consists of over sixty workstations and is now operational. In [GRIM95] we present the performance of two production applications that we have used to test the efficacy of our approach: complib, a biochemistry application that compares DNA and protein sequences, and ATPG, an electrical engineering application that generates test patterns for VLSI circuits. The performance results are encouraging.

4. I/O Status

We conclude this report with a summary of the I/O status in the Legion project. This section presents an overview of the current research initiatives related to file systems and I/O in the Legion Metasystem [GRIM94b] project. Given the intended nation-wide to world-wide scope of Legion, the system poses many new challenges in the area of scalable I/O applications, but at the same time holds the promise of exciting new tools for wide-area collaboration and large-scale information management and retrieval.

Legion is a distributed, object-oriented, virtual-machine based metasystem intended to present a single, seamless computational environment to users and application developers. A central part of the Legion environment is its single, persistent namespace. Current research in the area of persistent objects in Legion is focused on three of different levels. These are:

(1) Design and implementation of the basic system functionality to support persistent objects.

(2) Design and implementation of basic, useful, user-level persistent object classes.

(3) Development of applications to avail of the unique facilities supplied by Legion persistent object classes.

4.1. System Support for Persistent Objects

Current research at the low-level Legion system implementation level is addressing some of the basic problems in supporting distributed persistent objects. These include:

- **Naming and Binding:** Central to the ability to support persistent objects is the need for a well defined concept of the persistent object name-space. Legion objects have associated with them system-wide unique identifiers, LUIDs. A distributed scheme utilizing Binding Agent objects is employed to bind LUIDs to object addresses.

- **Persistent Object Creation and Scheduling:** Another active research area is related to the placement and instantiation of persistent objects in order to best utilize available resources.

- **Object States:** Legion objects can be in one of two basic states, active or inactive. Active objects have an associated thread of control and address space, while inactive objects are dormant and have all needed state saved to stable store. Currently, the basic mechanisms by which objects are moved between active and inactive status are being developed. All object classes will support "Save" and "Restore" asynchronous member functions which will be invoked by system scheduling and instantiation objects. Class authors will be responsible for utilizing Legion supplied mechanisms to save and restore user level object state, while system internal object state will be saved and restored automatically.

- **Advanced Features:** Features such as persistent object migration and object replication are also important goals of the Legion persistent object model. While such features will be under the control of persistent object class authors, the needed system support to elegantly utilize these mechanisms are currently being investigated and developed.

4.2. Basic User Level Persistent Object Classes

While the central core of the Legion system will provide the ability to develop persistent object classes, the system will be of little use unless an existing base of useful, user-level persistent object classes is also provided. Among the most important of these are:

- **File Objects:** The basic staple of information-based applications is the file. While the Legion system will not prescribe any limited set of file objects, it will provide a useful set of basic file classes. The most basic among these is the currently implemented "Byte Vector" object class – an unstructured vector of bytes supporting a set of member functions similar in functionality to Unix standard library file manipulation system calls. Along with the byte vector object class, utility programs such as "legion cat" have been developed to demonstrate the basic functionality of location independent, distributed files. Other more complex application utilizing basic Legion "Byte Vector" objects are described below.

- **Context Objects:** While the Legion LUID naming scheme addresses the basic issues of object naming and binding, it does not directly address the basic needs of name-space "navigation" mechanisms – the ability to explore the name space, create logical links between objects, and map human-user comprehensible string names to system readable LUIDS. This role will be played by "Context" object classes. These objects, at the highest level, will provide a mapping between user-level string names and system-level LUIDs. They will also provide mechanisms to create links between contexts, building a graph of object directories. Well constructed context objects will provide the basis for structuring the Legion namespace at the level of user comprehensible meaning.

4.3. Applications Utilizing Persistent Objects

In order to demonstrate the utility of Legion persistent objects, as well as to drive the further development and refinement of the Legion object model, a number of applications are currently being updated to utilize Legion file facilities. Some of these include:

- **Text Editing / Word Processing:** The ability to collaborate on documents conveniently in a wide-area distributed environment will be one of the basic benefits of Legion. Of the basic applications to support in this domain, text editing and word processing are among the most obvious candidates. Versions of the standard unix vi and emacs word processors have been updated to utilize Legion files. These have been demonstrated in truly wide-area (cross country) environments. Currently, an add-on module for the FrameMaker word processing system is being developed to allow Frame users to transparently manipulate files in Legion space.

- **Distributed Software Development:** Another potential area being investigated is distributed software project control. The ability to effectively collaborate on software development projects is a natural application of the Legion persistent object facility. The problem of developing a source code control / configuration management system utilizing Legion file objects is currently being investigated.

- **Distributed Simulation:** The use of distributed simulations in military and commercial applications is wide spread and growing consistently. The application of the Legion persistent object space to distributed simulation is an area of active research in the group. A recent paper examining this potential application of the system is [FERR95].

5. References:

[BENT79] Bentley, J. L. and Friedman, J. H., Data Structure for Range Searching, *ACM Computing Surveys* 11, 4 (Dec. 1979), 397-409.

[BOYL87] Boyle, J., et al., Portable Programs for Parallel Processors, Holt, Rinehart and Winston, New York, 1987.

[CARE94] Carey, M.J., et al., Shoring Up Persistent Applications, SIGMOD 1994, 1994, 383-394.

[FERR95] Ferrari, A. J., Distributed Interactive Simulation in the Legion System, ELECSIM 1995, Electronic Conference, 1995. <ftp://ftp.cs.virginia.edu/pub/ajf2j/legion_dis.ps.

[GRIM91] Grimshaw, A. S., and Loyot, E. C., Jr., ELFS: Object-Oriented Extensible File Systems, Tech. Rep. CS-91-14, Dept. of Computer Science, University of Virginia, July 1991.

[GRIM93a] Grimshaw, A. S., Strayer, W. T., and Narayan, P., Dynamic Object-Oriented Parallel Processing, IEEE Parallel & Distributed Technology: Systems & Applications, May 1993, 33-47.

[GRIM93b] Grimshaw, A. S., West, E. A., and Pearson, W. R., No Pain and Gain! - Experiences with Mentat on Biological Application, *Concurrency: Practice & Experience* 5, 4 (June 1993), 309-328.

[GRIM93c] Grimshaw, A. S., Easy to Use Object-Oriented Parallel Programming with Mentat, IEEE *Computer,* May 1993, 39-51.

[GRIM93d] Grimshaw, A. S., Easy to Use Object-Oriented Parallel Programming with Mentat, IEEE *Computer,* May 1993, 39-51.

[GRIM93e] Grimshaw, A. S., Weissman, J. B., and Strayer, W. T., Portable Run-Time Support for Dynamic Object-Oriented Parallel Processing, submitted to *ACM Transactions on Computer Systems*, July 1993.

[GRIM94a] Grimshaw, A. S., Wulf, W. A., French, J. C., Weaver, A. C. and P. F. R. Jr., A Synopsis of the Legion Project, Tech. Rep. CS-94-20, Dept. of Computer Science, University of Virginia, Charlottesville, VA, June 1994.

[GRIM94b] Grimshaw, A. S., Wulf, W. A., French, J. C., Weaver, A. C., and P. F. R. Jr., Legion: The Next Logical Step Toward a Nationwide Virtual Computer, Tech. Rep. CS-94-21, Dept. of Computer Science, University of Virginia, Charlottesville, VA, June 1994.

[GRIM95] Grimshaw, A. S., Nguyen-Tuong, A., and Wulf, W. A., Campus-Wide Computing: Early Results Using Legion at The University of Virginia, Tech. Rep. CS-95-19, Dept. of Computer Science, University of Virginia, Charlottesville, VA, March 1995.

[KARP94a] Karpovich, J. F., Grimshaw, A. S., and French, J. C., Breaking the I/O Bottleneck at the National Radio Astronomy Observatory (NRAO), Tech. Rep. CS-94-37, Dept. of Computer Science, University of Virginia, Charlottesville, VA, August 1994.

[KARP94b] Karpovich, J. F., Grimshaw, A. S., and French, J. C., Extensible File Systems (ELFS): An Object-Oriented Approach to High Performance File I/O, Proc. OOPSLA '94: Object-Oriented Programming Systems and Languages, 1994, 191-204.

[KARP94c] Karpovich, J. F., Grimshaw, A. S., and French, J. C., Extensible File Systems (ELFS): An Object-Oriented Approach to High Performance File I/O, Tech. Rep. CS-94-28, Dept. of Computer Science, University of Virginia, Charlottesville, VA, July 1994.

[KARP94d] Karpovich, J. F., French, J. C., and Grimshaw, A. S., High Performance Access to Radio Astronomy Data: A Case Study, Tech. Rep. CS-94-25, Dept. of Computer Science, University of Virginia, Charlottesville, VA, July 1994.

[KARP94e] Karpovich, J. F., French, J. C., and Grimshaw, A. S., High Performance Access to Radio Astronomy Data: A Case Study, *Proc. 7th Inter. Conference on Scientific and Statistical Database Management*, Oct. 1994, 240-249.

[KRIE88a] Kriegel, H. and Seeger, B., *Techniques for Design and Implementation of Efficient Spatial Access Methods, Proc. of the 14th VLDB*, 1988, 360-370.

[KRIE88b] Kriegel, H. and Seeger, B., PLOP-Hashing: A Grid File without a Directory, *Proc. of the Fourth Inter. Conf. on Data Engineering*, Feb. 1988, 369-376.

[LEVY90] Levy, E. and Silberschatz, A., Distributed File Systems: Concepts and Examples, *ACM Computing Surveys* 22, 4 (December 1990), 321-374.

[MANO92] Manola, F., Heiler, S., Georgakopoulos, D., Hornick, M., and Brodie, M., Distributed Object Management, *International Journal of Intelligent and Cooperative Information Systems* 1, 1 (June 1992).

[NICO93] Nicol, J. R., Wilkes, C. T., and Manola, F. A., Object-Orientation in Heterogeneous Distributed Systems, IEEE *Computer* 26, 6 (June 1993), 57-67.

[NIEV84] Nievergelt, J., and Hinterberger, H., The Grid File: An Adaptable, Symmetric Multikey File Structure, *ACM Transactions on Database Systems* 9, 1 (Mar. 1984), 38-71.

[ROBI81] Robinson, J. T., The K-D-B-Tree: A Search Structure for Large Multidimensional Dynamic Indexes, *Proc. of the Annual Meeting of the ACM Special Interest Group on Management of Data*, 1981, 10-18.

[SAME84] Samet, H., The Quadtree and Related Hierarchical Data Structures, *ACM Computing Surveys* 16, 2 (June 1984), 187-260.

[SUND90] Sunderam, V. S., PVM: A framework for parallel distributed computing, *Concurrency: Practice and Experience* 2, 4 (December 1990), 315-339.

UNIVERSITY OF WASHINGTON

A Visual Database System for Image Analysis on Parallel Computers and its Application to the EOS Amazon Project

Linda G. Shapiro, Steven L. Tanimoto, and James P. Ahrens
Department of Computer Science and Engineering
(shapiro@cs.washington.edu)

1. Introduction

1.1 Task Objective

The goal of this work was to create a design and prototype implementation of a database environment that is particular suited for handling the image, vision and scientific data associated with the NASA's EOS Amazon project. We are focusing on a data model and query facilities that are designed to execute efficiently on parallel computers. A key feature of the environment is an interface which allows a scientist to specify high-level directives about how query execution should occur. Using the interface does not require an understanding of the intricate details of parallel scheduling.

1.2 Introduction

This report summarizes research activities to date and serves as the final 3-year subcontract report. In the first year, we interviewed NASA scientists in order to understand their requirements and formulated an initial design for the database environment. In the second year, we refined the design and implemented a prototype. In the third year, we evaluated and documented the environment.

Our work was done in conjunction with the NASA Earth Observing System (EOS) Amazon Project at the University of Washington. The mission of the EOS Amazon project is to contribute to understanding the dynamics of the Amazon system in a natural state, and how it would evolve under possible change scenarios (from instantaneous deforestation to more subtle longer term climatic/chemical changes). The overall goal of the project is to determine how extensive land-use changes in the Amazon would modify the routing of water and its chemical load from precipitation, through the drainage system, and back to the atmosphere and ocean. The work is being undertaken by a number of groups here at the University of Washington including researchers in hydrology headed by Thomas Dunne, in biogeochemistry headed by Jeffrey Richey, and remote sensing headed by John Adams.

1.3 Scientists' Requirements

We interviewed the NASA scientists in order to understand their computing requirements. The scientists are working with data sizes on the order of hundreds of megabytes and processing algorithms whose completion time is on the order of minutes to hours. The scientists identified the following desirable properties for a computing environment to support scientific research:

Exploratory –The computing environment should facilitate the scientist's exploration of different algorithmic solutions.

Responsive – Algorithm results should be returned as quickly as possible, especially if the scientist is waiting for them.

Satisfies user requirements – The environment should schedule and execute algorithms based on the scientist's requirements for resource utilization and algorithm execution. For example, a scientist might like to specify which results are most important, what processing resources are available and how to utilize these resources.

High-level – The environment's interface should let the scientist specify a high-level description of his algorithms and requirements. The environment should provide support for scientists who are not computer experts.

Organized – The computing environment should record and organize the scientist's computer-based research work for later retrieval.

1.4 Approach

The scientific computing environment described in this report has these desirable properties. The approach we used to create this environment contains the following steps:

- An identification of how existing software tools fulfill the requirements described above.
- Creation of new algorithms and tools which fill the gap left by existing software tools.
- Integration of all these tools into a seamless whole.

In summary, we have identified two keys areas which are not well supported by existing software. These areas are:

1. Support for automated parallel program scheduling and execution.

   To achieve high-performance, programs are scheduled and executed on multiple processors. Parallel scheduling is a complex problem and automation is a welcome solution for scientists. One disadvantage of traditional tools is that they optimize for a fixed collection of preset scheduling goals. Another is that they do not fully automate the scheduling process. An automated scheduling system which is responsive to the scientists' scheduling needs would improve both scientists' satisfaction with their computer systems and their productivity.

2. Support for scientific experimentation.

   An environment needs to provide a computer-based framework for scientists' interactions. One typical interaction that scientists perform is parameterized experimentation with their programs. This experimentation helps the scientist to understand the effects of input parameter and coding changes. With automated support scientists could focus on analyzing their experimental results instead of the process required to generate the results.

1.5 Background

This section presents a high-level summary of existing software tools including programming languages, systems, and databases which scientists use to support their computer-based research work. This overview details how existing software tools fulfill the scientists' requirements and where they fall short. In addition, it provides a context for understanding how the computing environment described in this report builds upon and relates to existing tools.

1.5.1 Languages

Scientists have traditionally used sequential, imperative programming languages such as FORTRAN to express their scientific algorithms. Although FORTRAN is a low-level language, it is the language of choice for most scientists. One reason for this is that it is fairly straightforward to express efficient programs based on arithmetic expressions. It is one of the few programming languages which provides standardized support for complex arithmetic. Another reason is there is a legacy of FORTRAN programs that has been developed by scientists over the years. Scientists are very interested in reusing these programs, leveraging their work upon these existing successful programs.

An important advance in programming languages for scientists is visual programming languages. One of the most successful type of visual languages are data-flow-based visual programming languages. Examples include languages such as AVS [28] and Cantata/KHOROS [23]. Programs are expressed graphically as data-flow-based program graphs. Users can manipulate the program graph interactively, by adding and deleting tasks. Users have access to a library of existing tasks which are ready for use in their programs. These languages simplify program creation and the reuse of existing tasks. They support exploratory programming because changes can easily be made to programs without re-compiling.

One useful addition to a visual programming environment is support for parallel program scheduling and execution. Researchers at the Boeing Company created a data-flow based visual programming environment called Access Manager which allows distributed task execution [24,9]. The first version of Cantata/ KHOROS (version 1.0) also allows users to execute different tasks of their programs on different processors. Users are require to specify the details of this assignment. CM/AVS is an extension of AVS in which a parallel version of program tasks can be executed on the Thinking Machine CM-2 or CM-5 parallel computers[2]. Support for parallel program execution is a necessary first step in the process of providing support for automated parallel scheduling and execution.

1.5.2 Systems

Another way a scientist can improve his program's efficiency is to use distributed system software tools, such as Condor [19] or DQS. These tools execute a set of independent jobs on networks of workstations. The scientist formulates his program as a collection of independent jobs and submits them to a job queue. The tool then automatically schedules and executes the jobs on a set of available workstations. Work continues on the creation of efficient distributed systems support tools. Recent research focuses on methods of identifying and using idle workstations and avoiding scheduling conflicts [5].

There are many task scheduling algorithms that can be used to schedule the tasks of a data-flow program graph in parallel. Task scheduling algorithms attempt to maximize the number of tasks executing in parallel while minimizing inter-processor communication costs. A taxonomy of task scheduling algorithms can be found in [8]. Lewis et al [12] also provides a useful introduction to task scheduling. Since most types of task scheduling problems are NP-complete, solution algorithms are based on heuristics. Traditionally these heuristics optimize for a fixed preset collection of goals. This is a problem if the scheduling goals of the algorithms conflict with the scheduling goals of the user.

1.5.3 Databases

Databases provide support for storing, organizing and accessing scientific data. Key features of a database are its data model, which describes the stored data's relationships and semantics, and its query model, which describes how to retrieve the stored data. The relational data model represents data by tables of attributes. It is a simple model and popular for representing business data. Scientific data usually has more complex relationships than can be expressed using the relational model. Another concern is that scientific data, such as images and multi-level data structures, do not map well to relational tables. A second popular data model is the object-oriented model. Data is represented as a collection of data-structure-based objects. The object-oriented model usually lacks effective query models, because the structures it represents are so diverse that it is difficult to query them efficiently.

Recent research has focused on creating a data model and database system which supports scientists' needs. Examples include GAEA[15], MDBS [27] and DEVR [26]. These systems combine features of the relational and object-oriented data model, striving for the simplicity of the relational model with the expressiveness of the object-oriented model.

1.5.4 Constraints

A constraint expresses a relationship the user would like to hold in the solution of a particular problem. The environment described in this report uses constraints to express the user's task scheduling preferences. Related environments that use constraints include geometric layout systems [22,6], user interface builders [21], and machine vision systems [25].

An active area of research in constraint satisfaction is how to solve over-constrained systems (i.e., a set of constraints for which there is no solution that satisfies all constraints)[16]. Freuder and Wallace [14] adapt standard backtracking and consistency checking algorithms to satisfy a maximal subset of the constraints. Borning *et al* describe another solution to this problem: the user arranges his constraints in a hierarchy [7]. In the event that all constraints cannot be satisfied, the constraints at a higher level in the hierarchy are satisfied before constraints at a lower level in the hierarchy. The constraints at the top level are called requirements (or hard constraints) and must always hold. The constraints at the lower levels are called preferences (or soft constraints) and are satisfied based on their level in the hierarchy. Constraints within a level are solved based on a relative weighting provided by the user. The user-directed scheduler described in this report fits into this paradigm. The scheduler has two levels: the requirement level and one preference level. Preferences are satisfied based on their relative weights. Future work could consist of allowing the user to express a hierarchy of constraints to the scheduler, so that the user can control the order of constraint satisfaction.

1.5.5 Artificial-Intelligence-based Scheduling

Scheduling is the process of assigning a set of jobs to set of limited resources over time. The quality of a schedule is usually defined by a collection of user-defined criteria and constraints. Artificial Intelligence (AI) is the study and creation of theory, algorithms and computer systems that use knowledge and encoded intelligence to solve complex problems. Thus, scheduling is a natural area of interest for researchers in Artificial Intelligence.

AI researchers have built scheduling systems for a number of specific domains including systems for scheduling telescope usage [17], space shuttle maintenance [30], manufacturing [13] and defense logistics [10]. AI-scheduling solution methods are characterized by a number of features. *Constructive methods* build a complete schedule while *repair-based methods* incrementally update an existing but flawed schedule until a valid schedule is obtained. Fox's ISIS manufacturing scheduling system use a constructive solution method [13]. It iteratively builds a complete schedule by exploring a search space of partial schedules. It uses a beam-search which is guided by system and user constraints in order to find a schedule. Repair-based methods are useful for domains which change significantly over time. Repair-based methods only need to reschedule tasks affected by an external change to the problem. Zweben *et al* describe a repair-based scheduling system for space shuttle repair and maintenance [30]. It also uses a search-based solution method but explores a search space of complete schedules. A disadvantage of repair-based methods is that they usually use a local search-based solution method and therefore do not provide globally optimal schedules.

Many AI schedulers use constraints to express requirements and preferences on the problem domain. A characteristic of a scheduler is how it relaxes the problem constraints when they are in conflict in order to find a solution. Different methods include satisfying a maximal subset of constraints [17], using a fallback constraint if the original constraint cannot be satisfied [13], placing priorities on constraints and using a hierarchy of constraints [7].

The goal of this work is to create an automated task-scheduling environment. A critical component of the environment is a unique AI task-scheduler which allows the user to express task-scheduling constraints. The task-scheduling domain is different than other studied AI scheduling domains. For example, there are significant differences between the tasks in the task-scheduling domain and the jobs in the manufacturing domain. Tasks in a task-scheduling domain can usually be assigned to any processor whereas jobs in the manufacturing domain are assigned to specific machines. In the task-scheduling domain, if dependent tasks are scheduled on different processors a communication cost is incurred. There is no similar cost in the manufacturing domain. Furthermore, tasks in the task-scheduling domain usually do not have start and finish deadlines as jobs do in the manufacturing domain. Because of the many required manufacturing constraints, problems in the manufacturing domain are usually over-constrained. Therefore solution methods usually focus on finding an acceptable solution. Problems in the task-scheduling domain are usually significantly less constrained and therefore this work uses a constructive solution method which can often provide an optimized solution to its users.

1.5.6 A related environment

The members of the Intelligent Data Management Project led by Nicholas Short, Jr. at NASA Goddard are working on a prototype environment which can process the massive datasets generated by satellites that are part of NASA's Earth Observing System [18]. The environment supports the querying, real-time processing and storing of satellite image data. In order to cope with the changing volume of incoming satellite image data by a given deadline, the environment has access to different versions of processing algorithms, which offer varying trade-offs of result quality for shorter completion times.

The major subsystems of the environment are a set of processing request queues, a planning system, an execution engine/monitor and an object database.

- The processing request queues accept processing requests from users. Their requests are high-level and declarative allowing a user to express what processing should be done rather than how. For example, a user can specify that a satellite image be registered without specifying a specific algorithm to do the registration. A user can also specify a completion deadline for a processing request.

- The planning system inputs a processing request and selects and composes a set of tasks into a program graph which fulfills the user's request. The tasks are selected from a collection of Khoros tasks, LAS tasks[1] and user-defined tasks. Note that there are many tasks available to the planner, that perform the same type of operation but have different properties. For example, there may be multiple registration tasks: one which processes a specific image type, one which executes very quickly and one which produces registrations of very high quality. Task properties are formalized using conditions. Each task is annotated with a set of preconditions which must be true in order to execute the task and postconditions which are true after the task has executed. The planner unifies these conditions to create a program [3].

- The execution engine executes the program generated by the planner on a network of workstations. It uses a dynamic scheduling technique developed by Ma [20] to schedule based on network traffic, processor utilization and task dependencies.

[1] LAS is a geographic information system package used to process Landsat images.

In summary, Short *et al's* environment supports automatic program creation by allowing users to express requests for processing which are fulfilled by a planner. Scheduling requests are limited to completion deadlines. In contrast, our own work allows its users to express a full range of task-scheduling directives including the ordering of program results and the specification of task assignments and processor utilization levels. In addition, this work supports computer-based scientific experimentation.

1.6 Structure of the Environment

Having reviewed existing scientific software tools, we will now describe the components of the scientific computing environment presented in this report:

- Data-flow based visual programming environment – The scientist uses a visual programming environment to construct his programs.
- Scientific database – A database is used to organize and store information about program graphs and results.
- Distributed executor – The executor executes a program graph in parallel on a network of workstations in order to quickly generate the scientist's results. It handles inter-processor communication between distributed tasks in the program graph and records performance information for use by the performance prediction tool.
- Scheduler – The scheduler automatically schedules a program graph on a network of workstations based on the scientist's directives. The scientist's directives are specified declaratively as constraints.
- Performance prediction – Program performance prediction is necessary for efficient scheduling. The scheduling algorithm uses performance estimates to make scheduling decisions.

A diagram of the scientific computing environment is shown in Figure 1. The diagram shows the data-flow between the components of the environment. In this report, data-flow diagrams are represented visually with boxes representing operations and ovals representing data. Directed arrows define the flow of data through the data-flow diagram.

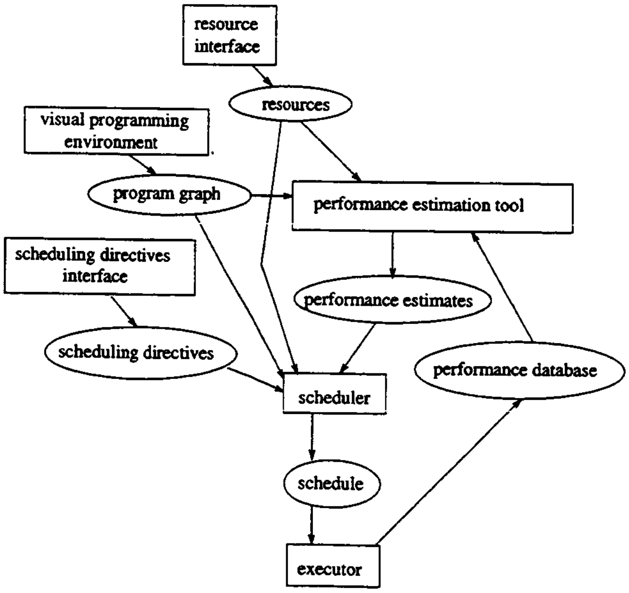


Figure 1: Structure of the Scientific Computing Environment

Data input to the environment includes resource information, a program graph and the user's scheduling directives. Available processors are specified initially by the system administrator. The program graph is specified using a visual programming environment. The user scheduling directives are specified using a constraint-based scheduling language. The program graph and resources are used by the automatic performance prediction tool to create a cost model of program execution and processor utilization. The scheduler inputs the resource information, the program graph, the user's scheduling directives and performance estimate information. The scheduler outputs a schedule which fulfills the user's scheduling directives. The program is then executed on a network of workstations using the distributed executor. During execution, performance data is collected and sent to the performance database for future use by the performance prediction tool.

1.7 Outline

Section 2 describes a problem space representation for task scheduling. This goal-oriented representation facilitates the specification of scheduling directives. It contains the definition of a language for specifying these directives and a number of examples, which show how to use the language to specify directives for task ordering, task placement, processor utilization and load balancing. In addition, it describes a search-based algorithm for fulfilling a user's scheduling directives. Section 3 describes the prototype and an algorithm for automatically creating parameterized scientific experiments. Section 4 reports the results of a study of the environment performance. Results are presented on the performance of the environment on a large number of realistic imaging graphs and on how well the environment fulfills the user's scheduling directives. Section 5 summarizes and describes future research directions.

2. User-directed scheduling

To achieve high performance, programs are scheduled and executed on multiple processors. Parallel scheduling is a complex problem and automation is a welcome solution for scientists. One disadvantage of traditional tools is that they optimize for a fixed set of preset scheduling goals such as simply minimizing completion time. Another is that they do not fully automate the scheduling process. A method for automatic scheduling which is responsive to their scheduling needs would improve both scientists' satisfaction with computer systems and their productivity.

This chapter describes an automatic scheduling method that was designed to meet these needs. First, a problem space representation for scheduling is described. This goal-oriented representation facilitates the specification of scheduling directives and is amenable to artificial-intelligence-based solution techniques including search and planning. Then a language for specifying scheduling directives is defined. Finally, a search-based algorithm for determining a schedule is described.

2.1 Preliminaries

A program graph consists of a set of functional tasks and set of input and output dependencies between these tasks. Figure 2 shows an example of a simple program graph with two tasks, one which inputs an image and another which displays an image. The output of the Input image task is used as input by the Display image task.

Task scheduling is the process of assigning and ordering the execution of tasks from a program graph onto a collection of processors. The parallel task execution model used by the environment assumes that each processor can run one task at a time. To execute a task on a processor:

1. All inputs that are the outputs of tasks executed on another processor in the distributed network are received in parallel. The processor blocks and waits until all inputs are received.

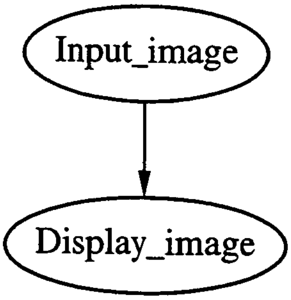


Figure 2: A Program Graph

2. The task is executed.
3. All outputs that are inputs of a task executed on other processors in the distributed network are sent to these processors in parallel.

Blocking communications assure the correct parallel execution of the task graph by guaranteeing a task is not executed until all its inputs are available.

2.2 A Problem Space Representation for Task Scheduling

This section describes a problem space representation for task scheduling. A problem space is defined as a set of states and operators that moves between these states. A particular problem to be solved in a problem space is known as a problem instance and is defined by an initial state and a set of goal states.

2.2.1 States

A state represents an empty, partial or complete schedule of tasks to processors. It must represent task and processor scheduling information as well as other related information such as estimates of scheduled task start and finish times. A state consists of a collection of tasks, a collection of task dependencies and a collection of processors. Elements of these collections are entities. Each entity consists of a set of attributes, each of which consists of a name and type. Attributes are detailed below using the following syntax: <attribute name>:<attribute type>; <descriptive comment>. The tasks, processors and task dependency entities are as follows:

- . Task Entity

id:integer	; a unique task id
name:string	; the task's name
exec-time:integer	; the task's execution time
	; Note: all timings are expressed in seconds
start-time:integer	; the task's start time
	; Note: the start of the schedule is time 0
finish-time:integer	; the task's finish time
assigned-proc-id:integer	; the id of the processor this
	; task is assigned to

- Processor Entity

id:integer	; a unique processor id
name:string	; the processor's name

```
finish-time:integer               ; the total running time of
                                  ; the tasks scheduled on this processor
                                  ; assuming no gaps or idle periods
assigned-task-ids                 ; an ordered list of tasks scheduled on
  :list                           ; this processor
util:integer                      ; the processor's CPU utilization
```

- Dependency Entity

```
task-id:integer                   ; a task id
dep-task-id:integer               ; the id of the task that depends on
                                  ; the output of the task with task-id
                                  ; as input
comm-time:integer                 ; the time to required to communicate this
                                  ; dependency data
                                  ; Note: if no communication is required than
                                  ; comm-time is 0.
non-local-comm-time:integer       ; the time to communicate this
                                  ; dependency data to another processor in
                                  ; in the network
```

2.2.2 Initial State

The initial state has the following values initialized:

- There is a task entity for each task in the input data-flow program graph.
- There is a processor entity for each identified available processor.
- There is a dependency entity for each dependency in the program graph.

All other attributes of these entities are assigned to a special symbol which represents unknown values.

2.2.3 Operators

An operator makes a transition from one state to another state. There is one operator in the problem space representation for task scheduling. Its name and type is: *schedule-task-to-processor (integer, integer, state) → state.*[2] The result of executing the call, *schedule-task-to-processor (task-id, proc-id, original-state) → new-state* is that the task identified by *task-id* is scheduled on the processor identified by *proc-id.*

2.3 Goal State

The conditions required of a goal state are:

- Each task is scheduled to a processor.
- The task dependencies are respected by the schedule. That is, if a task is dependent upon another task for input, it runs after that task has completed.

This completes the specification of a problem space representation for task scheduling.

[2] We will use the following syntax to describe function types in this document: <function name>(<param type 1>, <param type 2> etc.) → <return param type>.

2.4 A Language for Expressing Scheduling Directives

The problem space representation for task scheduling defines any complete valid schedule of tasks to processors as a goal state. Traditional task scheduling algorithms add another condition to these criteria. They optimize performance by working to minimize a particular performance variable, such as processor completion time, or task finish times. These optimizations are always hard-encoded into the scheduling algorithm, and these algorithms do not allow other optimization criteria to be used. In this section, I describe a language in which a user can specify a variety of optimization criteria, by describing relationships he would like to hold between values in the goal state and values he would like to be minimized or maximized in the goal state. These *scheduling directives* allow the user to optimize for performance as well as specify other desirable properties of a schedule including the ordering of task outputs, specific task to processor assignments and specific processor utilization levels.

2.4.1 Preliminaries

The scheduling language is an extension of SQL [1, 11] a relational database query language. SQL is the pre-eminent database language in use today, enjoying wide acceptance among non-computer experts because of its ease of use.

In SQL, a relation is a collection of entities with the same sets of attributes. A state in the task scheduling problem space representation is composed of three relations: tasks (task), processors (proc) and dependencies (dep).

A basic SQL expression has three clauses: select, from and where. The from clause specifies the relations to be operated on. The where clause specifies a boolean predicate on entity attributes which are used to select entities from the relations. The select clause specifies the resulting relation in terms of the attributes of the selected entities. The syntax is:

```
select <attributes from the selected entities>
from <relations>
where <boolean predicate on the entity attributes of the relations>
```

The scheduling language defines importance and type constraints. Importance constraints are either requirements or preferences. Requirements must always hold, preferences are fulfilled based upon user-defined priorities. Constraint types include relationship-based constraints that express a desired relationship between attributes of relations, value-based constraints that express a desire for a value to be minimized or maximized, and ordering-based constraints that express a desire for a particular ordering on a relation. The basic syntax for constraints is:

```
assert {relationship | value | ordering} {requirement | preference}
(
<specific assertion constructs>
)
```

The bracket and slash notation used above (i.e {A|B|C}) means that one of the elements in the collection of choices is utilized. For example, valid constraints include: assert value requirement and assert ordering preference.

Selecting elements from a collection: An SQL expression can be used to select entities which pass a given test. Using the * symbol in the select clause returns all the attributes of an entity. Note that the attributes of a relation are referred to by appending the attribute name to the entity type name. For example, the *id* attribute of the *task* entity is *task-id.*

Aggregating the elements of a relation: SQL also provides a way to compute a single summary value from a collection of attribute values. In the select clause the user identifies a specific attribute to aggregate. Possible aggregate functions include: average, minimum, maximum, sum and count.

2.4.2 Requirements

The first type of scheduling directive is a requirement. A requirement guarantees that a user-specified constraint will hold in a goal state. Requirements are specified and tested with a requirement function.

Relationship requirements: A relationship requirement guarantees that a user-specified relationship will hold in a given state. The name and type of the relationship requirement function is:

assert relationship requirement (expression, test, expression) → boolean.

It returns TRUE when applied to a valid state. For the call, *assert relationship requirement (expression-1, test-1, expression-2)*:

- expression-1 and expression-2 are SQL expressions. The function applies the SQL expressions to the given state. The returned values are used to create *relation-1* and *relation-2.*

- *test-1* is run on each element of the cross product of the previously created resulting relations (i.e. all possible pairs of an input value from the first relation and an input value from the second relation). If any test returns FALSE the requirement is FALSE.

Example 1 - Ordering task output generation time: To assert that the task with id 1 finishes before the task with id 2, the following requirement is defined:

```
assert relationship requirement
(select task-finish-time from task where task-id = 1) <
(select task-finish-time from task where task-id = 2) )
```

Example 2 - Deadlines on task output generation time: To assert that all tasks finish before a 30 second deadline, the following requirement is defined:

```
assert relationship requirement
(select task-finish-time from task) < 30 )
```

Example 3 - Controlling task/processor assignments: To assert all FFT tasks are run on lillith, the following requirement is defined:

```
assert relationship requirement
(select task-assigned-proc-id from task where task-name = “FFT”) =
(select proc-id from proc where proc-name = “lillith”))
```

Ordering requirements: An ordering requirement function provides a means for asserting relationships which hold on an ordered sequence of values. Thus, the relationship test holds between each element of the sequence and any subsequent elements. Its name and type are:

assert ordering requirement (sequence, ordering-test) → boolean.

For the call, *assert ordering requirement (sequence-1, ordering-test-1)*:

- The *ordering-test-1* is applied to *sequence-1*. The order-test clause is an extension to standard SQL, allowing the user to specify a sort order to test. The order-test clause identifies the attributes to test and whether to test if the sequence is sorted in ascending or descending order. If any entity of the sequence is out of order the ordering test returns FALSE.

Example - Ordering task output generation time: To force the tasks to be scheduled in order of id number the following requirements is made[3,4]:

```
assert ordering requirement
(select * from task where task-assigned-proc-id <> UNKNOWN)
(order-test task-id asc))
```

Ordering-based requirement functions are useful for scheduling tasks to processors in a particular order. Many traditional task algorithms define an order in which to schedule tasks. With ordering-based requirement functions this behavior can easily be mimicked.

Additional goal state condition: Requirements add an additional condition to the problem state representation of a goal state: when applied to a goal state all defined relationship and ordering-based requirements must be TRUE.

2.4.3 Preferences

Relationship and Ordering Preferences: The second type of scheduling directive is a preference. A preference specifies a relationship the user would like to hold in a goal state or a value the user would like to minimize or maximize in the goal state. There are relationship and ordering based preference functions and they are very similar to relation and ordering requirement functions. The only difference between these types of preference and requirement functions is their return values. Requirement functions return TRUE if all tests are passed and FALSE otherwise. Preference functions return the number of failed tests. The name and type of the relationship and ordering preference functions are:

assert relationship preference (expression, test, expression) → integer and
assert preference order (expression, ordering-test) → integer.[5]

The ordering preference function computes for each element in the sequence the number of subsequent elements that should precede it in the specified ordering. The sum of these values is returned by the function. This calculation places decreasing emphasis on the correct ordering of entities as their distance from the beginning of the
sequence increases.

Example 1 - Balancing the task load on processors: To specify a preference for a balanced task load among the processors the following function is specified:

```
assert relationship preference
all (select task-start-time from task) <=
all (select proc-finish-time from proc)
```

[3] Note that the order-by clause creates a sequence from the unordered relation using one key and the order-test clause tests if the sequence is ordered based on a different key.

[4] The order-by clause considers entities out of order if the *task-assigned-proc-id* value of the task earlier in the sequence is UNKNOWN and the *task-assigned-proc-id* value of the task later in the sequence is known.

[5] Requirements can be implemented with preferences as follows: *assert relationship requirement* calls *assert relationship preference* with the same parameters. If *assert relationship preference* returns 0 (tests failed) then return TRUE else return FALSE.

This expression states that there is a preference that all task start times be less than the total running time of each processor. The intuition for why this balances workload is that in an unbalanced workload, tasks start on some processor after other processors have finished. Note that this relationship should not be expressed as a requirement because when communication costs are excessive, optimal schedules are not balanced.

Example 2 - Controlling processor utilization: To specify a preference for the processor *calvin* to be assigned at least twice as much task load as the processor *lillith,* the following function is specified:

```
assert relationship preference
2 * (select proc-finish-time from proc where proc-name = "lillith") <=
   (select proc-finish-time from proc where proc-name = "calvin")
```

Value-based preferences: Value-based preferences allow the user to specify values they would like minimized or maximized in the goal state. The name and type of the value-based preference function is:

assert value preference (optimization-type, integer, function, integer, integer) → integer.

For the call *assert value preference (opt-type, priority, value-function, min, max)*:

- *opt-type* states whether to minimize or maximize the value function.

- *priority* is a measure of the importance of fulfilling this preference. Specifically, priority values have the following semantics: The relative importance of a particular preference is equal to its priority value over the total of all priority values. For example, if three preferences have priorities, 1, 2, 1, the relative importance of the preferences is 0.25, 0.50, 0.25. For example, when choosing between two goal states, the environment will prefer a state which fulfills the second preference but not the first or third over a state which fulfills the first preference but not the second or third because the second preference is twice as important to the user as the first.

- *value-function* is a SQL expression which when applied to a given state returns an integer value.

- *min, max* are estimates of lower and upper bounds on the result of the *value-function*. These values are used by the environment to scale the result of the value-function so that comparisons with other value-function results make sense.

Example 1 - Minimizing processor run times: To specify a preference for minimizing processor run times the following function is specified[6]:

```
assert value preference
opt-type = minimize, priority = 1,
function = (select max (task-finish-time) from task)
min = 0, max = (select sum (task-exec-time) from task) +
   (select sum (non-local-comm-time) from dep))
```

All relationship and ordering-based preferences are expressed using value-based preferences because the environment can use value-based preferences to create a numeric measure of how much a state is preferred.

Additional goal state condition: Preferences add an additional condition to the problem state

[6] The maximum finish time value is bounded by the serial execution of all tasks plus the serial non-local communication of all dependency data.

representation of a goal state: goal states which fulfill preferences based on their priority values are preferred. A formal description of how this condition may be met is described in the next section.

2.5 A Search-based Scheduling Algorithm

In this section, I describe a search algorithm for user-directed scheduling. Best-first search is used to find optimized goal states in the problem-space representation. A best-first search algorithm requires three functions: a successor function, which defines how to create the successor states of a state, an evaluation function, which gives each state a score, and a goal function, which identifies goal states.

Best-first search selects from the set of states generated so far the state with the minimum score. It checks if the selected state is the goal state, if it is then the state is returned. Otherwise the successors of the selected state are created and evaluated and the process continues.

2.5.1 Successor Function

The name and type of the successor function is: *successor(state)* → *set of states.*

For the call *successor(state1)* the function creates:

- *set1* – a set of all tasks that could be executed. This set is composed of each non-scheduled task whose dependent tasks are already scheduled.
- *set2* – a set of all processors on which the tasks could be executed. This set is a list of all the available processors.

For all pairs of elements, *ele1* ∈ set1 and *ele2* ∈ set2, *scheduled-task-to-processor(ele1, ele2, state1)* is executed. These executions create a set of new states.

All defined requirements are applied to each new state. If any requirement fails when applied to a new state, the state is removed from the set of new states. After this is complete, the remaining set of new states are returned as successors.

2.5.2 Evaluation Function

Semantics of priorities: Preferences provide a mechanism for comparing states. For a call, *assert value preference (opt-type, priority, value-function, min, max)* the *opt-type, priority, min* and *max* values allow the environment to scale the results of value functions so that comparisons make sense. The following variables are used to calculate a global preference comparison value, g_{total} for a state from a set of *1... vp* value-based preferences:

- p_i is the priority of preference i where $i = 1 \ldots vp$.
- p_{total} is the sum of all the preferences priority values, that is, $p_{total} = \sum_{i=1}^{vp} p_i$
- v_i is the result of the value function of preference i.
- min_i, max_i is the lower and upper bound values of preference i.
- s_i is the scaled preference value of preference i (s_i values are between 0 and 1 with 0 preferred), that is, *if (type = minimize) then* $s_i = \dfrac{v_i - min_i}{max_i - min_i}$ *else* $s_i = \dfrac{min_i - v_i}{max_i - min_i}$
- g_i is the scaled prioritized value of preference i, that is, $g_i = s_i \ * \dfrac{p_i}{p_{total}}$

- g_{total} is the sum of all the preferences scaled prioritized values, that is, $g_{total} = \sum_{i=1}^{vp} g_i$

The name and type of the evaluation function is: *evaluation(state) →* integer. The evaluation function returns the global preference value, g_{total} defined in the previous section. Best-first search find an optimized goal state but not necessary the optimal goal state because it stops as soon as it finds a goal state. Branch and bound search could be used to find the optimal goal state but the extra time it requires to search through the problem space is prohibitive.

2.5.3 Goal Function

The name and type of the goal function is: *goal(state) → boolean*. The goal function returns TRUE if all the tasks are scheduled and FALSE otherwise.

2.5.4 Soundness and Completeness

- Soundness is the property that if a goal state is returned by the search, it is valid. Informally, this is true because:

  1. Only valid states are identified as goal states since the goal function only returns TRUE if all tasks are scheduled.

  2. Only valid states are generated because the successor function only schedules tasks whose dependent tasks have already been scheduled.

  3. Only valid states are generated because the successor function eliminates states which do not satisfy the user's requirements.

- Completeness is the property that if a goal state exists, it can be found by the search.

  1. The successor function lists all valid task-to-processor assignments. Thus, all possible valid schedules can be generated.

2.5.5 Computational Complexity

The computational complexity of a search algorithm is the branching factor raised to the depth of the search tree (i.e. $O(b^n)$ where *b* is the branching factor and *n* is the depth). Let *tasks* be the number of tasks and *procs* be the number of processors. The worst case computational complexity is $O((tasks \times procs)^{tasks})$. The average computational complexity is usually better than this, because the branching factor is usually significantly less than the total number of tasks. The removal of states that do not meet the user's requirements further reduces the branching factor. A study of the performance of this algorithm on a large number of imaging graphs is presented in Section 4. The study reports on the number of states the algorithm generates.

3 SCE: The prototype

This section describes a prototype of the scientific computing environment SCE developed in this research. The first subsection describes how the prototype supports computer-based scientific experimentation. The second subsection describes how the user interacts with the prototype and the outputs that are generated. The last subsection presents an overview of the implementation of the prototype.

3.1 Computer-based Scientific Experimentation

Scientists are interested in experimenting with their programs. They make parameter and coding changes to their programs and then analyze their results in order to understand the effects of these changes. With automated support, scientists can focus more on analyzing their experimental results than on how to generate these results. This section describes how SCE supports computer-based scientific experimentation. An efficient algorithm for automatically creating a computer-based experiment is presented. This is followed by a discussion of another environment which provides support for experimentation and the specific advantages of the prototype's implementation.

An experiment specifies the controlled substitution of tasks, data or parameters in the program graph.[7] All possible combinations of substitutions may need to be tested. For example, a geologist working on a remote sensing problem might be interested in testing the quality of a set of edge detection tasks on a collection of satellite images. Using the prototype's visual programming environment, a program graph is created which consists of nodes for an input image task, edge detection task and display-image task connected as a sequence. The created program graph is shown in Figure 3.

In the experiment, the first task, Input_image_region_a, which contains data for the northern region of the Amazon river basin, is to be replaced with Input_image_region_b, which contains data for the southern region of the basin. The second task, the Sobel edge detector is to be replaced with two different edge detection tasks: the Prewitt edge detector and the Canny edge detector as shown in Figure 4. All possible combinations of substitutions of input images and edge detection tasks are instantiated and executed as shown in Figure 5. The output images are labeled and stored in the database for later analysis.

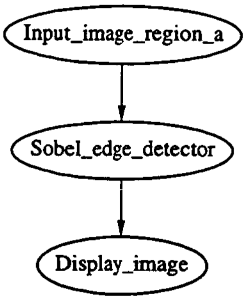


Figure 3: A Sequence of Tasks in a Program Graph

[7] In most data-flow based visual programming environments, parameters and data are not represent explicitly in a program graph. Instead they are considered part of each task. For example, parameters and data in Cantata/Khoros are specified as input values. Thus, to specify parameter and data substitutions a corresponding task is specified with modified input values.

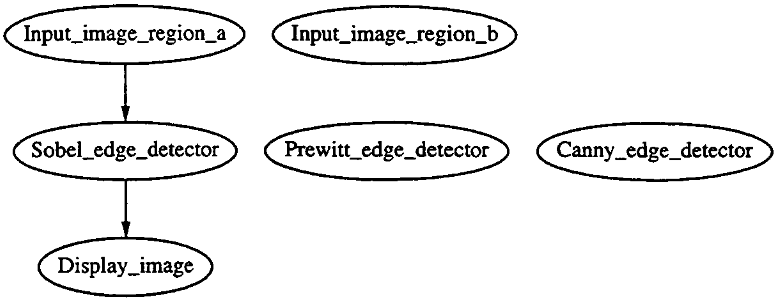


Figure 4: Substitutions For the Experiment

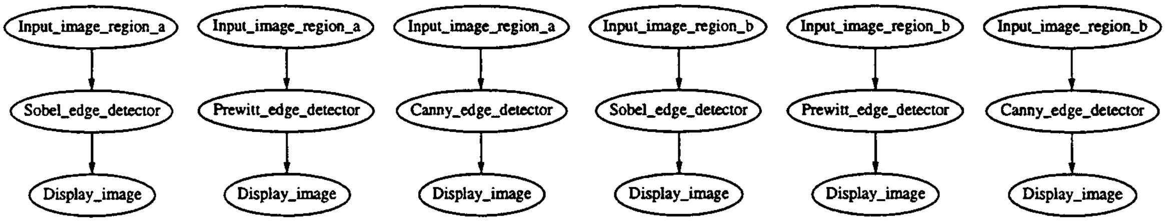


Figure 5: An Instantiated Experiment

A simple way to create an experiment is to replicate the original program graph for each possible combination of substitutions and then make one set of substitutions to each replicated graph. This method was used to create the experiment shown in Figure 3. This simple method requires more task executions than are necessary. For example, in Figure 3 notice that the Input_image_region_a task is executed three times although it is only necessary to execute it once. SCE uses a new experiment creation algorithm that avoids this problem by reusing the results of executed tasks. Reusing task results helps to minimize experiment execution time.

3.1.1 Discussion

A related environment which executes experiments on a collection of distributed workstations in parallel was created by D. Abramson *et al* [4]. The environment, Nimrod, allows a user to express a set of input parameters and data changes for a program. Nimrod creates experiments in a similar manner to the example shown in Figure 3.3. The cross-product of user parameter changes is generated and elements from this set are input to copies of the original programs. These copies are scheduled and executed on a collection of distributed workstations.

Nimrod and SCE both provide a concise and useful interaction model. Experiments provide a concise method for scientists to express a set of controlled changes to a program graph. With this support, scientists can express what changes they want to experiment with, but not how to implement these changes. Nimrod and SCE also both provide efficient experiment executions. Experiments execute efficiently because their program graph representations contain many independent execution paths which can be scheduled and executed in parallel.

In addition, SCE simplifies experimentation with task substitutions in a program. Nimrod allows its users to experiment with different data and parameter inputs to their programs. Nimrod has no knowledge about the inner workings of the program on which it is running experiments. Thus, in order to make a task substitution in Nimrod, a scientist must modify his program by hand, removing the code to be substituted for, replacing it with new code and recompiling their program. After this process is complete he can use Nimrod to run experiments. In SCE, programs are represented as a collection of communicating tasks. SCE allows its users to experiment with program tasks. Thus, it is a simple matter to have the user identify which task to replace and to automatically substitute the user's new task in its place. Specifying task substitution is useful when the user wants to experiment with a collection of different tasks which perform the same function, such as edge detection.

Furthermore, SCE reduces the total amount of work required to execute an experiment. SCE uses an experiment creation algorithm which reuses task results whenever possible during an experiment. This algorithm allows scientists to obtain their experimental results faster than Nimrod's experiment creation algorithm. Nimrod's algorithm replicates the entire program for each substitution. It cannot optimize the experiment creation process because it does not have any knowledge of the inner workings of the program on which it is running experiments.

3.2 A Sample Session with SCE

A sample user session with SCE is now presented. This includes a description of the components of SCE the user interacts with and the results of this interaction. This presentation helps the reader become familiar with the interface provided by SCE.

3.2.1 User Inputs

Visual program environment: The scientist uses the visual programming environment, Cantata [23], to construct his programs. Figure 6 shows a Cantata workspace. The boxes represent tasks and lines connecting the tasks represent dependencies. The user selects tasks from the pull-down libraries at the top of the screen and connects the tasks together using dependencies to form a program graph.

Scheduling directives interface: The scientist uses a text editor to express his scheduling directives. A set of default directives are supplied by the environment. Note that these directives do not have to be utilized, they are provided as a suggestion. The goal of these directives is to minimize program completion time. The default directives are described in more detail in Section 4.

Resource interface: The scientist uses a text editor to create a list of available processors.

Experiment interface: The scientist currently defines an experiment using a text editor to specify locations for task substitutions and sets of the tasks to substitute into the program graph. Future work on the environment could consist of modifying Cantata's interactive graphical interface to allow the user to express experiments graphically. Another useful feature would be to extend the experiment creation interface to allow the user to express experiments which do not create the full cross product of task substitutions. This is useful when the user is not interested in all experimental results. For example, in the example shown in Figures 3-5 the user may only be interested in testing the Sobel edge detector on Input_image_region_a and in testing the other edge detectors on both images.

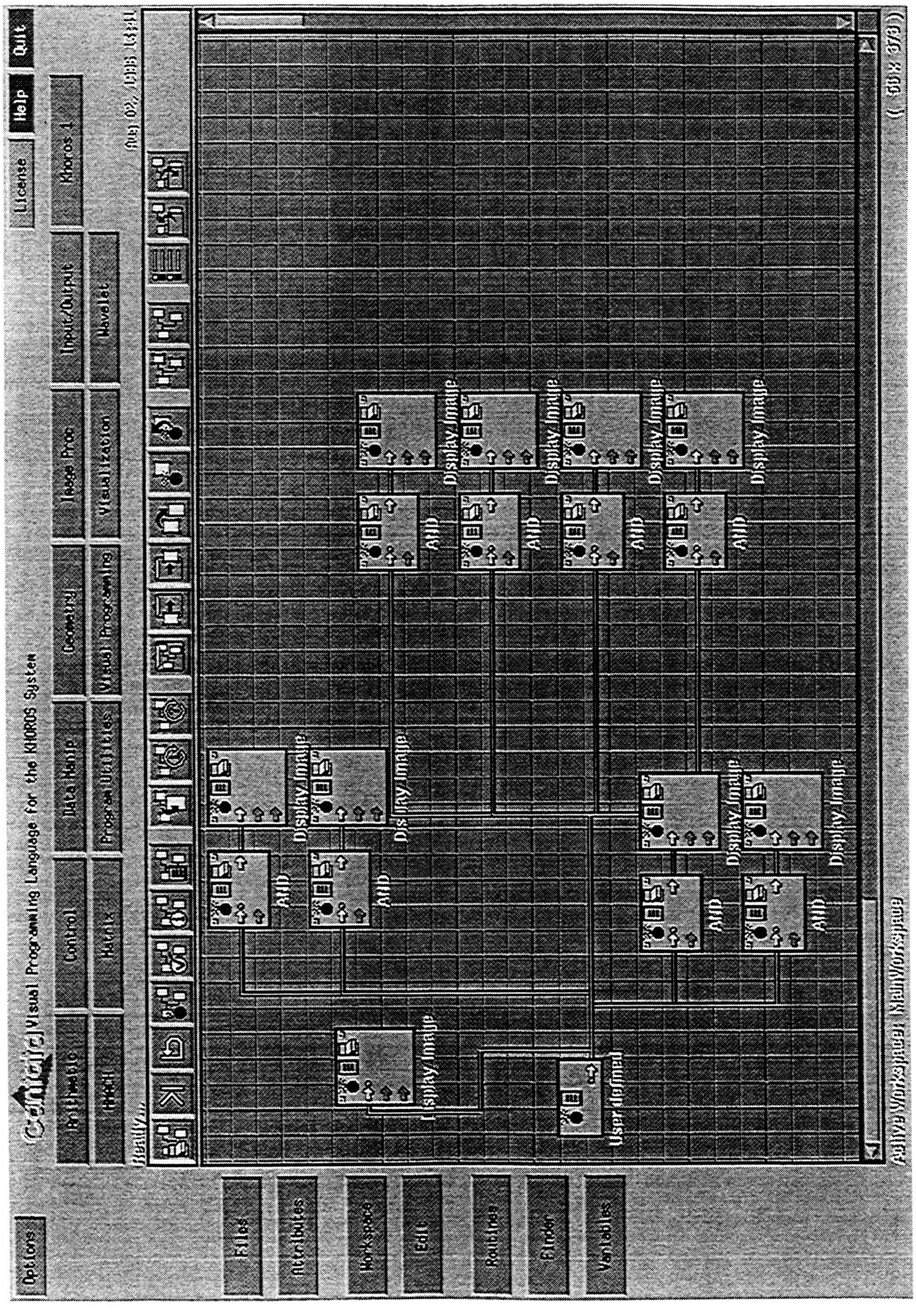
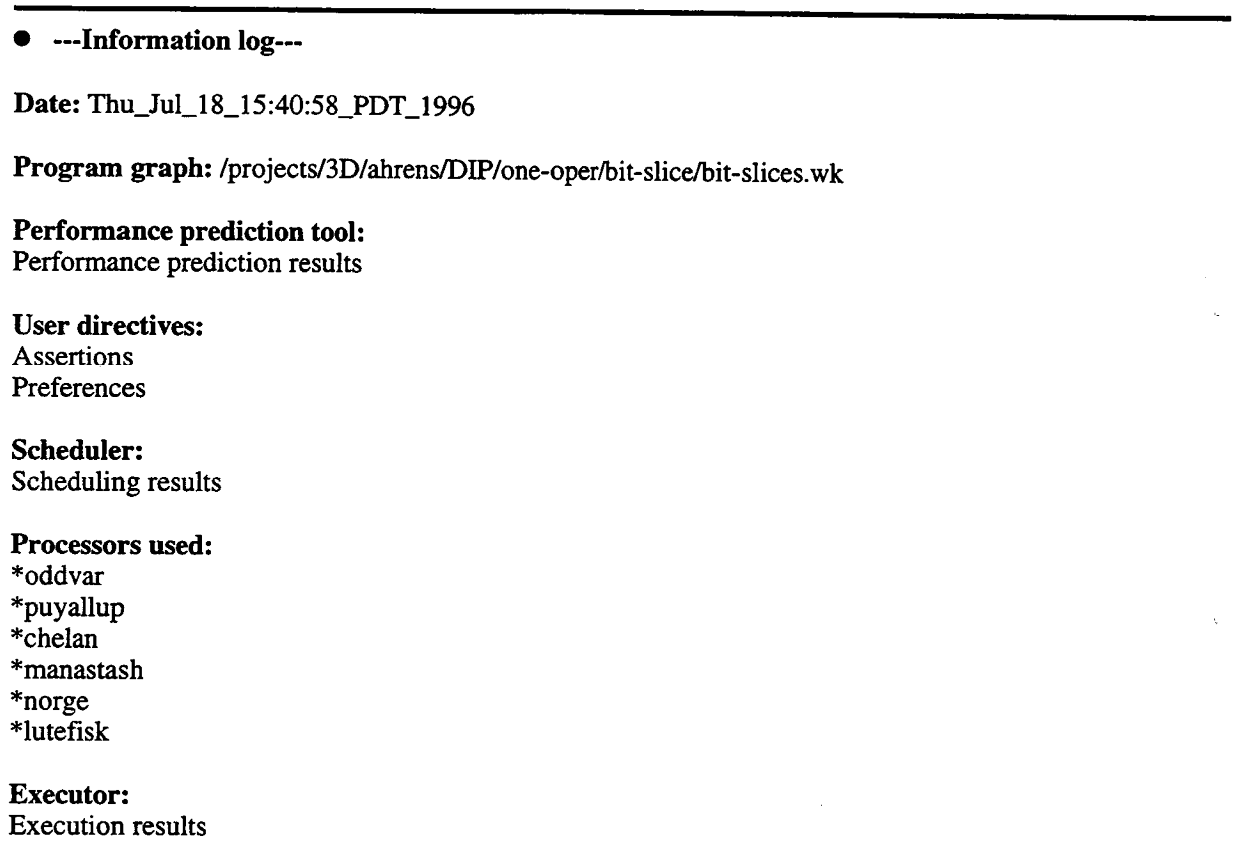


Figure 6: The Cantata Visual Programming Environment

3.2.2 Prototype Results

After the user creates his program graph, scheduling directives, and resources, SCE uses this data to create performance prediction information. This information is then passed to the scheduler which creates a schedule. Once the schedule is created, the tasks are executed on the workstations and the program outputs are generated. SCE creates an information log which records the details of each run. Details include the program graph and directives used, the generated performance prediction information and schedule and the program execution statistics. The log is written in HTML and the user can browse the information with a browser such as Netscape Navigator. Figure 7 shows an example of the information log. Scheduling information is stored graphically as part of the information log.



- **---Information log---**

Date: Thu_Jul_18_15:40:58_PDT_1996

Program graph: /projects/3D/ahrens/DIP/one-oper/bit-slice/bit-slices.wk

Performance prediction tool:
Performance prediction results

User directives:
Assertions
Preferences

Scheduler:
Scheduling results

Processors used:
*oddvar
*puyallup
*chelan
*manastash
*norge
*lutefisk

Executor:
Execution results

Figure 7: Information Log

This completes Section 3. Section 4 presents results of a performance evaluation of the environment which include results on the efficiency of the experiment creation algorithm and the performance of the environment when executing experiments.

4 Results

This section reports the results of a performance evaluation of the environment and survey of usefulness of the environment to scientists to support their computer-based scientific research work. The performance evaluation consists of three different studies. The first study explores the performance of the environment using the default scheduling directives on a diverse collection of image processing program graphs. Results are presented on the performance of the prediction tool, the scheduler and the executor. The second study investigates how well the environment responds to the user's scheduling directives. The third study examines the performance of the environment on computer-based scientific experiments.

4.1 Testing Method

The environment was tested by scheduling and executing a collection of program graphs which are a part of the Digital Image Processing (DIP) course for the cantata/Khoros visual programming environment. The course presents lessons on topics in image processing and provides forty-seven example program graphs for students to modify and execute. Topics include image representation, image manipulation, linear and non-linear operators and pattern classification.[8] The average number of tasks in the program graph is 18 and the average number of dependencies is 18. This data shows that the program graphs have a significant number of tasks and dependencies. All tests were executed on a collection of nine ethernet-connected Sun SPARCstation-IPXs.

4.2 Performance Study 1 - Default Scheduling Directives

The first study explores the performance of the environment using the default scheduling directives. The goal of these directives is to minimize program completion time. The default directives and their purpose are now described.

The first default directive requires the scheduler to only use processors with utilizations of less than or equal to three percent. This allows a program graph to execute efficiently without interference from other user's programs. The directive works by requiring that all processors with utilizations greater than three percent have their task assignment list be empty (i.e. equal to UNKNOWN). The second default directive directs the scheduler to prefer states with more scheduled tasks. This directive allows the search algorithm to make efficient progress. The next three directives emulate Wu and Gajski's task scheduling algorithm [29]. The goal of their algorithm is to minimize program completion time. The algorithm first determines an order in which to schedule the tasks. Then, as each task is scheduled, the algorithm chooses the processor that allows its earliest start time. The ordering is computed as follows: for each task, the length of the longest path between the task and any output task is calculated. The path length is the sum of the execution times and non-local communication times of the tasks and dependencies on the path. The tasks are arranged in non-increasing order based on their calculated path lengths.

Using these scheduling directives, the environment executes the DIP course program graphs. Results are presented on the performance of the scheduler, performance prediction tool, and executor.

[8] The Digital Image Processing course can be found on the World Wide Web at http://www.eece.unm.edu/dipcourse/.

Figure 8 shows the number of states explored by the scheduler for the program graphs. Notice that for most graphs the environment explores less than five hundred states.[9] Thus, the scheduler, when using the default directives, only needs to explore a small portion of the search space.

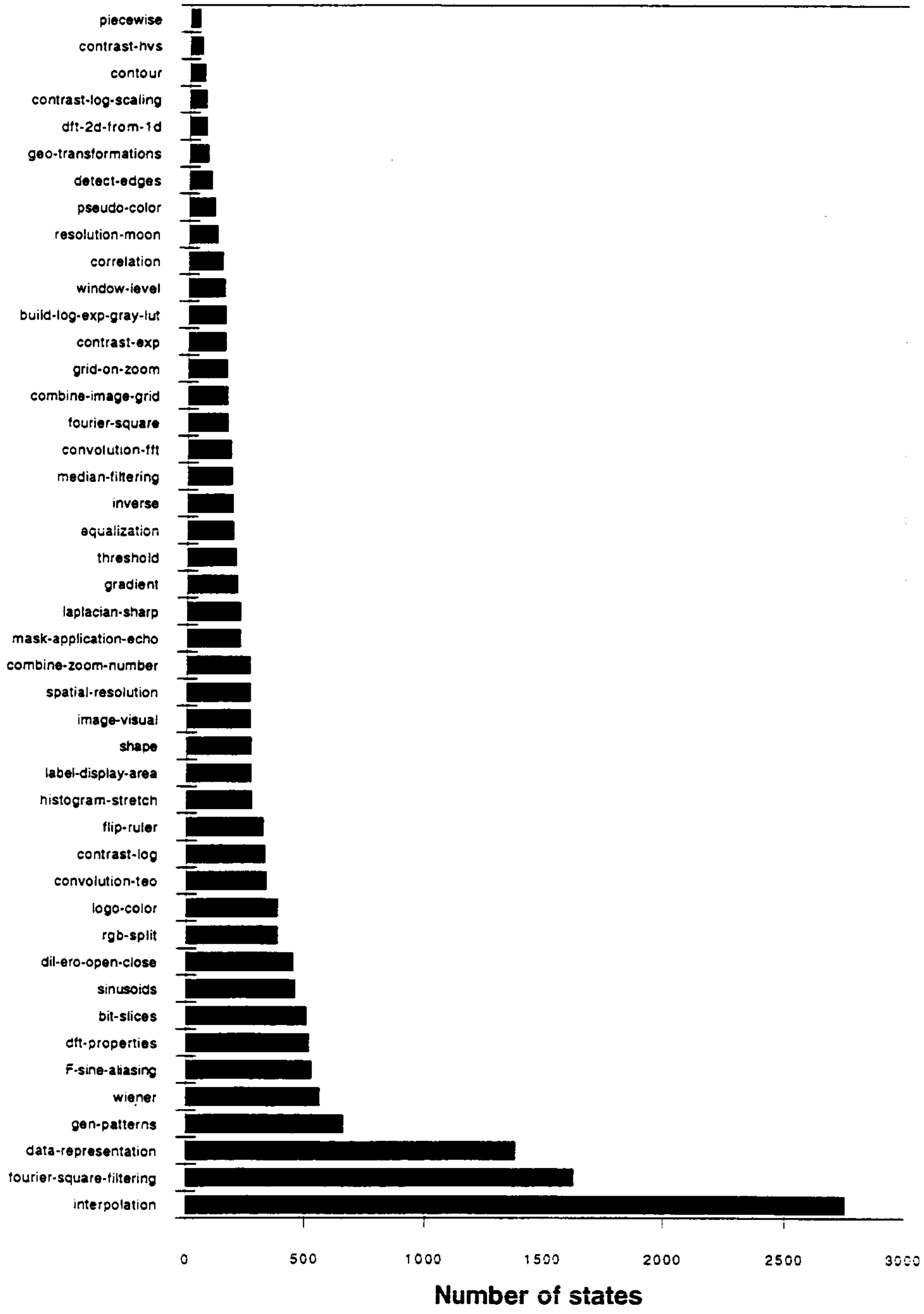


Figure 8: Number of States Explored by the Scheduler

[9] A worst case estimate on the average number of states in the search state is 18 $9^{18} = 162^{18}$.

Figure 9 presents the speedup achieved by the environment using the default directives for the collection of program graphs.[10] It is important to study the speedup achieved by the environment to assess the

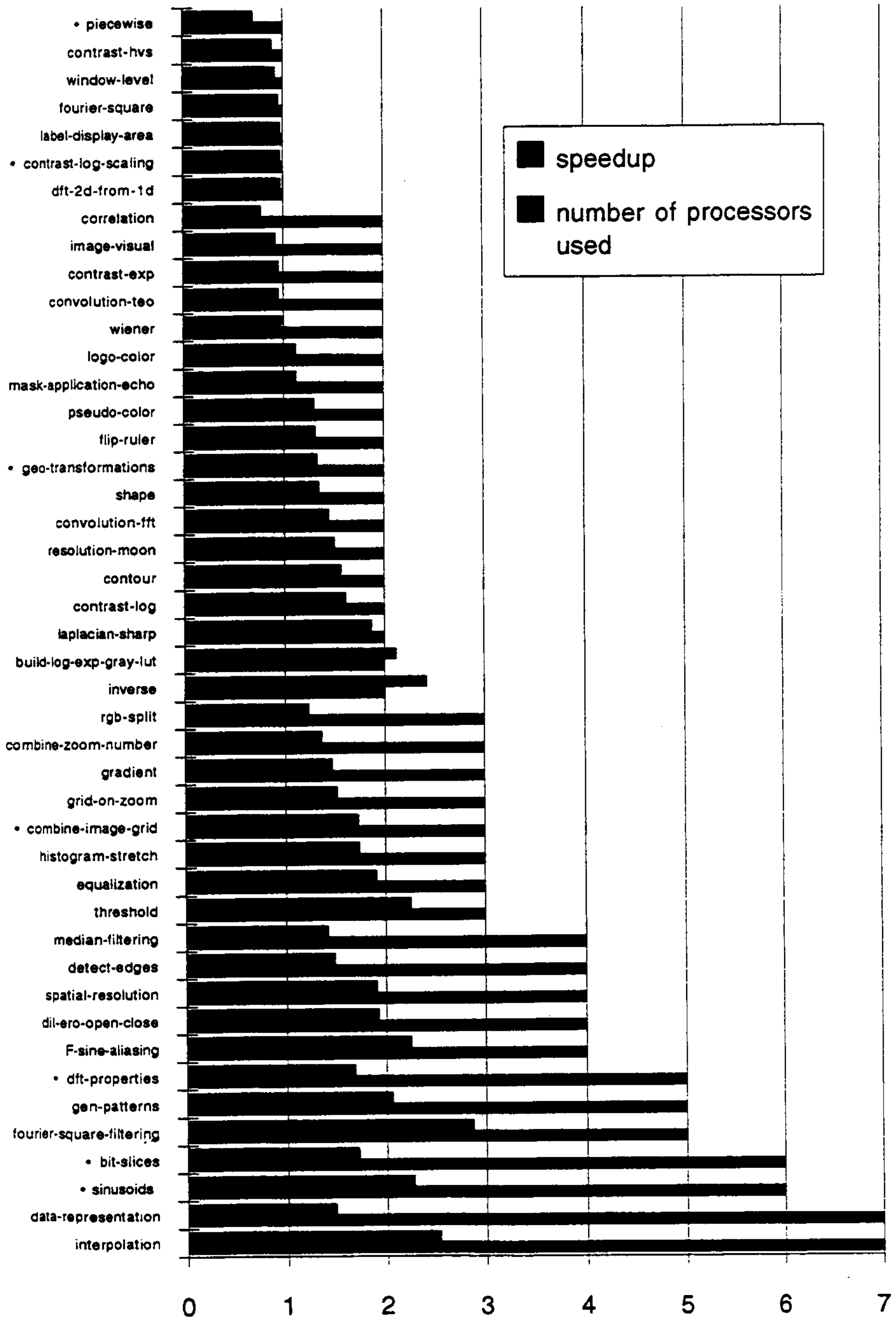


Figure 9: Speedup of the Program Graphs Using Default Directives

[10] Note that the input data used by the program graphs in this test was expanded to be 36 times larger (i.e. a factor of six expansion on the row and columns of the input images) in order to simulate the massive data sizes used by scientists such as geologists working on remote sensing problems.

performance of the default directives. During the scheduling process, the utilization assertion selects the number of processors that have a utilization of 3 percent or less. From this set of selected processors, the scheduler then schedules tasks on a subset of these processors. This subset is called the scheduled processors. When the number of scheduled processors is equal to the number of selected processors, it is possible that the scheduler could have used more processors to obtain better speedup. These instances are identified in Figure 9 by a dot in front of the program graph name.

The speedup data is grouped according to the number of scheduled processors (i.e. all program graphs scheduled on one processor, all program graphs scheduled on two processors, etc.). Within each group, the data is sorted from worst speedup to best speedup. The average speedup achieved was 1.4 on an average of 2.8 scheduled processors.[11] Note that the speedup the scheduler can obtain is limited by the existing data-flow parallelism in the program graphs. It is also important to note that this speedup was achieved without user intervention. The user provides a program graph to the environment, and it is automatically scheduled and executed.

4.3 Performance Study 2 - User Directed Scheduling

The second study investigates how well the environment responds to the user's scheduling directives. Multiple tests were executed as part of this study:

1. A program completion time preference test
2. A processor finish time preference test
3. A task ordering preference test
4. A task-to-processor assignment preference test.

4.3.1 A Program Completion Time Preference Test

The goal of this test is to minimize program completion time. The default directives fulfill this goal. This is evidenced by the speedup of 1.4 obtained on the program graphs in performance study 1. Additional speedups were also obtained using the default directives on a set of computer-based scientific experiments. This data is presented in Section 4.4.

4.3.2 A Processor Finish Time Preference Test

The goal of this test is to prefer the finish time of one processor be at least twice the finish time of another processor. The processor that finishes early can be used for other computing tasks the user has in mind. For the test, two directives are used in addition to the second through fifth default directives. The first new directive requires that the environment only schedule tasks on the processors oddvar and norge. The second new directive requests that the finish time of the processor oddvar be at least twice that of the processor norge. Figure 10 shows the results of the test. Notice that the finish time of processor oddvar is always at least twice the finish time of processor norge as the user requested.[12]

[11] The geometric mean is used to average normalized values such as speedups.

[12] Note that the reported results are execution times. Therefore they show the accuracy of the performance prediction tool as well as the quality of the scheduler. That is, the scheduler might fulfill the user's directives, but if its performance prediction information was incorrect, the execution results would most likely not fulfill the user's directives.

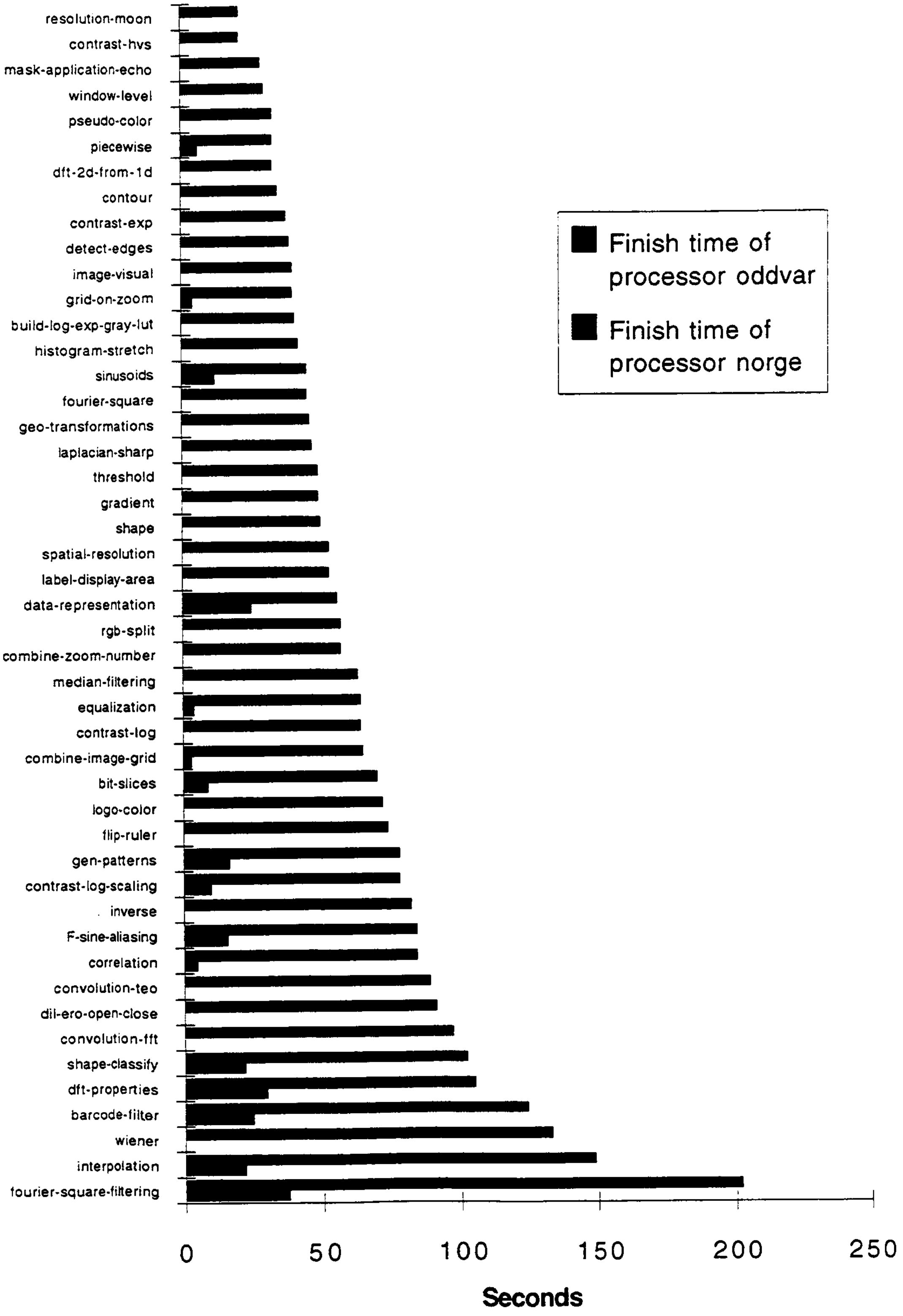


Figure 10: Processor Finish Time Directive Results

4.3.3 A Task Ordering Preference Test

The goal of this test is to prefer a particular ordering of task outputs. For the task ordering preference test, multiple directives are added in addition to the default directives. Each directive adds a dependency between a pair of output tasks to achieve this goal.

For the test, the output tasks of each DIP course program graph were identified, a random ordering of the tasks was generated and this ordering was preferred. The environment ordered the output tasks of all tested program graphs as requested. Table 1 presents a sample of the results of the test. The first column of the table lists the name of the tested program graph. The remaining columns lists the finish time in seconds of the tasks the user preferred to be output first, second, third, etc. Notice that the tasks are output in the order the user requested.

Program Graph	Finish Time 1st Task	Finish Time 2nd Task	Finish Time 3rd Task	Finish Time 4th Task	Finish Time 5th Task
combine-zoom-number	20 sec.	29 sec.	29 sec.	42 sec.	43 sec.
detect-edges	19 sec.	19 sec.	20 sec.	28 sec.	29 sec.
label-display-area	9 sec.	16 sec.	34 sec.	34 sec.	51 sec.
spatial-resolution	15 sec.	17 sec.	18 sec.	19 sec.	27 sec.

Table 1: Task Ordering Directive Results

4.3.4 A Task-to-Processor Assignment Preference Test

The goal of this test is to prefer a particular task-to-processor assignment. For the test, a single directive is added in addition to the default directives. The new directive prefers that all "Display Image" tasks execute on the processor willow.

The DIP course program graphs were scheduled and executed using these directives and all "Display Image" tasks of each program graph were scheduled on the processor willow. Figure 11 shows an example result schedule. Notice that all "Display Image" tasks are scheduled on willow. Notice also that the default directives work in concert with the task-to-processor assignment directive to cause the tasks to be scheduled on multiple processors in parallel, reducing program completion time.

4.4 Performance study 3 - Computer-Based Scientific Experimentation

The third study explores the performance of the environment on a set of computer-based scientific experiments. Experiments are created using the program graphs of the DIP course. For each experiment, an input task and non-input task are randomly chosen. In the experiment, four different versions of both the input and non-input tasks were tested. Figure 12 presents the speedup of computer-based scientific experiments. The result data is presented in the same manner as the speedup data in Figure 9. The average speedup is 3.4 on an average of 5.5 scheduler processors. Notice the significant increase in speedup of these graphs. This is because experimentation creates many independent execution paths.

Figure 13 presents a comparison of the experiment creation algorithm described in this report to the simple method of replicating the entire program graph for each experimental substitution used by Nimrod [4]. This graph shows the finish time of the experiment created with the experiment creation algorithm described in this report along with an estimate of the finish time of an experiment created with the simple method. The estimated finish time for the simple method is calculated by multiplying the time to run the original program graph on a single processor by the number of replications (i.e. in this case, 4 x 4 = 16 replications). This is the time required to execute the experiment on one processor. This time is divided by the number of scheduled processors used when scheduling the experiment created by the experiment creation algorithm described in this report. This provides the optimal finish time possible for the simple method. Notice that because the experiment creation algorithm described in this report reduces the workload required to create experimental results, its finish time is usually less than the optimal finish time of the simple method.

Time	Processor lutefisk	Processor manastash	Processor norge	Processor oddvar	Processor willow
1	User defined 83	User defined 31	User defined 117	User defined 35	User defined 3
2					
3			User defined 43	User defined 39	
4	2D Plot 87	Data Object Info 23			Display Image 7
5	Data Object Info 95				Data Object Info 11
6					
7			Animate 47	Data Object Info 55	
8			Data Object Info 125		
9					
10					
11					
12					File Viewer 15
13	File Viewer 91				
14			File Viewer 129	File Viewer 59	
15			Data Object Info 79	Data Object Info 67	Display Image 27
16		File Viewer 19			Display Image 51
17					Display Image 121
18					
19					Display Image 63
20					
21					
22			File Viewer 75		
23				File Viewer 71	

Figure 11: A Schedule Created Using a Directive Which Prefers all "Display Image" Tasks be Scheduled on the Processor Willow.

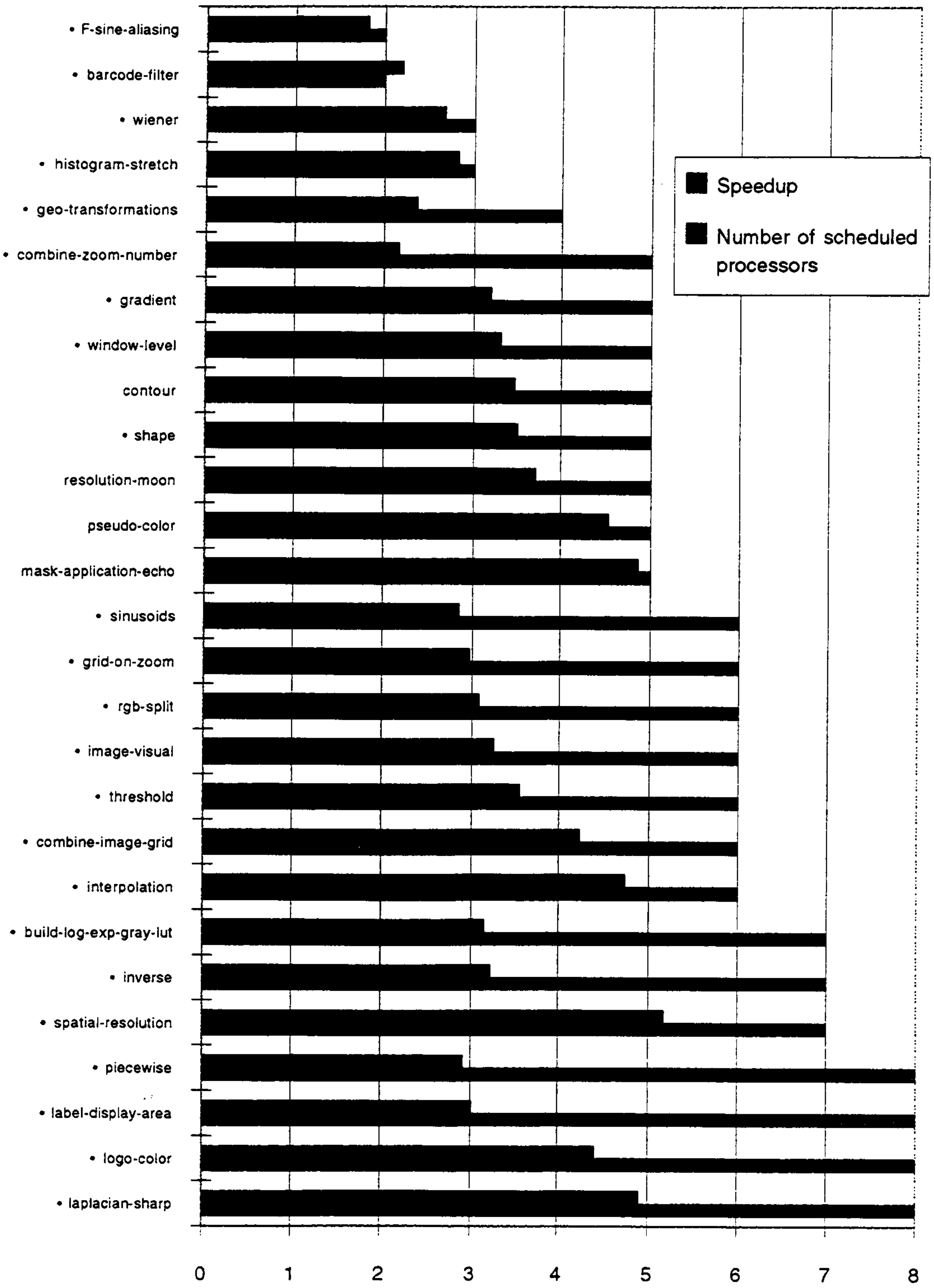


Figure 12: Speedup of Experiments

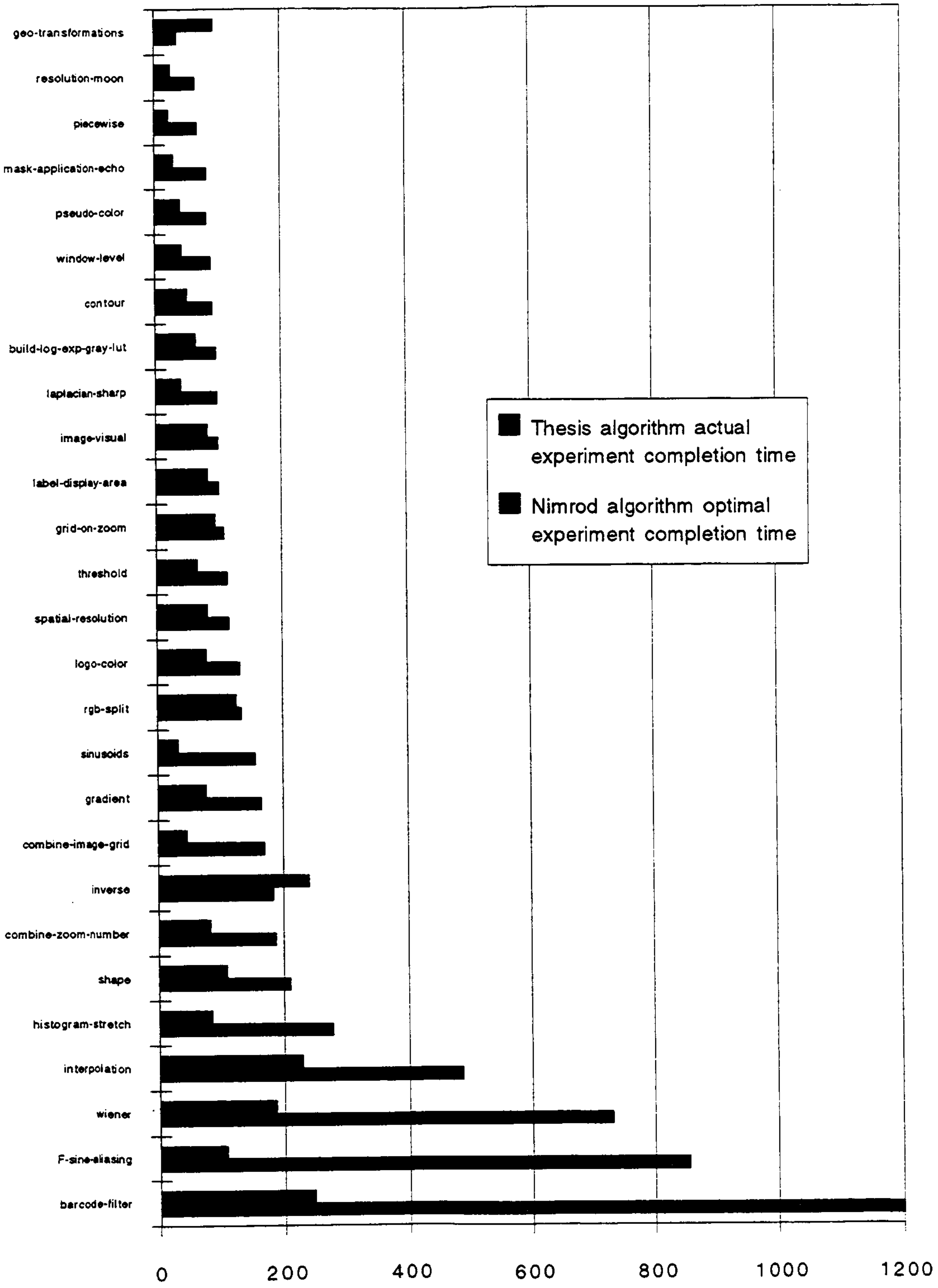


Figure 13: Comparison of Experiment Creation Techniques

4.5 User surveys

A survey was given to potential users of the scientific computing environment in order to assess its usefulness. Three vision researchers and a geologist who works on remote sensing applications saw a demonstration of the environment and completed a survey. In summary, the users felt the environment would be useful for their computer-based scientific research work. Specifically, in response to the question, "If you were running programs on a shared distributed network of workstations, is the scheduler a tool you would find useful for your scientific research work?" The geologist responded, "Yes, this would be useful now in the remote sensing lab as many users attempt to share a network of workstations". The survey also tried to assess how familiar the scientists were with the tools used in the environment. Most had used a visual programming environment but not the relational database language SQL. They did not think that this would be a hindrance to learning the scheduling language, however. In fact, in response to the question, "Is the scheduling language easy to learn and use?", all responded affirmatively. The users were also asked to order the usefulness of a collection of specific directives. The following list summarizes the user's choices:

- Minimizing program completion time: **1**
- Controlling task/processor assignments: **2**
- Output ordering: **3**
- Controlling processor utilization: **4**
- Time-related directives (after 3:00, before 6:00): **5**

Finally, the users were asked: "Is the support for computer-based scientific experimentation, a feature you would find useful for your scientific research work?" and most scientists responded positively with specific examples of research problems which would benefit from automatic experiment creation and execution. The full results of the geologist's survey are presented in section 6.

4.6 Summary

This completes the performance study of the environment. In summary, the study has shown that:

- Using the default directives, an average speedup of 1.4 on an average of 2.8 scheduled processors is achieved on the DIP course program graphs.
- The environment is responsive to the user's scheduling directives. In a variety of tests including a processor finish time test, task ordering test, and task-to-processor assignment test, the environment fulfilled the user's directives for all program graphs.
- The environment achieves very good performance on scientific experiments. An average speedup of 3.4 on an average of 5.5 scheduled processors is achieved. In addition, the experiment creation method presented in this report creates more efficient experiments than the simple method used by the Nimrod environment. On average, the experiments execute 2.1 times faster than an optimal execution of the experiments generated by the simple method.
- A survey was given to potential users of the scientific computing environment in order to assess its usefulness. In summary, the users felt the environment would be useful for their computer-based scientific research work.

5. Conclusions and Future Work

This report describes a computing environment which supports computer-based scientific research work. Key features include support for automatic distributed scheduling and execution and computer-based scientific experimentation. A new flexible and extensible scheduling technique that is responsive to a user's scheduling directives, such as the ordering of program results and the specification of task assignments and processor utilization levels, is presented. An easy-to-use constraint language for specifying scheduling directives, based on the relational database query language SQL, is described along with a search-based algorithm for fulfilling these directives. A set of performance studies show that the environment can schedule and execute program graphs on a network of workstations as the user requests. An algorithm for automatically generating scientific experiments is presented. Experiments provide a concise method of specifying a large collection of parameterized program executions. The environment achieved significant speedups when executing experiments; for a large collection of scientific experiments an average speedup of 3.4 on an average of 5.5 scheduled processors was obtained.

Future work on the environment could consist of a high-performance implementation of the scheduler and extensions to support other types of parallelism. A more efficient implementation of the scheduler would allow the environment to quickly find solutions to very complex directives. Ideas for a more efficient implementation include using an imperative programming language, parallelism and incremental user directive calculations. Also, in addition to data-flow parallelism, the environment could be extended to support operator, pipeline and loop parallelism. The performance prediction tool would be extended to predict the performance of parallel and pipelined tasks. The scheduler and executor would need to be extended to handle these type of tasks as well.

6. A Survey of Users of the Scientific Computing Environment conducted by Milton Smith, geologist, member of the University of Washington EOS Amazon project team

- Requirements

  Will a computing environment which fulfills the stated requirements (i.e. exploratory program creation, high-performance program execution, responsive to scheduling directives) be useful to you in your scientific research work?

  Yes, it will assist in utilizing the computing resources available in the Remote Sensing Lab.

- Test programs

  Are the Digital Image Processing course program graphs representative of the types of programs you use in your scientific research work?

  They are representative but not comprehensive. Research involves the continuous ingestion, evolution and development of new algorithms.

- Visual programming environment

  Is the visual program environment a tool you would find useful for expressing programs for your scientific research work?

  Visualization is very important to communicating research results.

- Scheduler

  If you were running programs on a shared distributed network of workstations, is the scheduler a tool you would find useful for your scientific research work?

  Yes - this would be useful now in the remote sensing lab as many users attempt to share a network of workstations.

  If you were running programs on a shared distributed network of workstations, which of the following directives do you think would be of useful to you?

  1. Output ordering:
  2. Controlling task/processor assignments: **3**
  3. Controlling processor utilization: **4**
  4. Minimizing program completion time: **1**
  5. Time-related directives (after 3:00, before 6:00): **2**
  6. Other directives you create: Are you familiar with the database query language SQL?

  Yes, to a limited extent.

  Do you feel that the scheduling directive language would be easy to learn and use?

  Yes, no problem.

  Any other comments you have about the scheduler?

  None.

- Distributed program executor

  Is the distributed program executor a tool you would find useful for your scientific research work?

  Yes, it makes sense in terms of our distributed computing resources.

- Computer-based scientific experimentation

  Is the support for computer-based scientific experimentation, a feature you would find useful for your scientific research work?

  This is definitely the wave of the future. We are interested.

  Any other comments about the environment's support for computer-based scientific experimentation?

  Make it easy for the user community to take responsibility for its evolution. Simple modular interfaces that allow expansion of capabilities.

- Improvements

Do you have any suggestions for improving any component of the environment so that it would be useful to you for your scientific research work?

Actually use it.

7. References

[1] American National Standard for Information Systems. (1986) *Database Language SQL.* ANSIX3(135-1986). American National Standards Institute, New York.

[2] *CM/AVS User's Guide.* (1992) Thinking Machines Corporation, Cambridge, Massachusetts.

[3] Lansky, A. and Philpot, A. (February 1994) AI-Based Planning for Data Analysis. *IEEE Expert, 9* (1), 21-7.

[4] Abramson, D., Sosic, R., Giddy, J. and Hall, B. (1995, August) Nimrod: A Tool for Performing Parameterized Simulations Using Distributed Workstation. In *Proceedings of the Fourth IEEE International Symposium on High Performance Distributed Computing,* (pages 112-121).

[5] Arpaci, R., Dusseau, A. Vahdat, A., Liu, L., Anderson, T. and D. Patterson. (1995, May) The Interaction of Parallel and Sequential Workloads on a Network of Workstations. In *Proceedings of the ACM SIGMETRICS Conference,* (pages 267-78).

[6] Borning, A. and Duisberg, R. (October 1986) Constraint-Based Tools for Building User Interfaces. *ACM Transactions on Graphics, 5* (4), 21-70.

[7] Borning, A., Freeman-Benson, B. and Wilson, M. (1992) Constraint Hierarchies. *Lisp and Symbolic Computation, 5,* 223-270.

[8] Casavant, T. and Kuhl, J. (1988) A Taxonomy of Scheduling in General Purpose Distributed Computing Systems. *IEEE Transactions on Software Engineering, 14* (2).

[9] Cigel, R., Carlson, D. and Maloney, J. (1992) Graphical Executive Language for Engineering Applications. In *MacNeal Schwendler World User's Conference.*

[10] Cross, S. and Walker, E. (1994) Applying Knowledge Based Planning and Scheduling to Crisis Action Planning. In M. Zweben and M. Fox (Eds.), *Intelligent Scheduling.* Morgan Kaufmann Publishers.

[11] Date, C. J. *(1989) A Guide to the SQL Standard* (2nd ed.). Addison-Wesley.

[12] El-Rewini, H., Lewis, T. and Ali, H. (1994) *Task Scheduling in Parallel and Distributed Systems.* Prentice Hall.

[13] Fox, M. (1987) *Constraint-Directed Search: A Case Study of Job-Shop Scheduling.* Morgan-Kaufmann.

[14] Freuder, E. and Wallace, R. (1992) Partial Constraint Satisfaction. *Artificial Intelligence, 58,* 21-70.

[15] Hachem, Nabil I., Qiu, Ke, Gennert, Michael and Ward, Matthew. (1993) Managing Derived Data in the Gaea Scientific DBMS. In *Proceedings of the Nineteenth Very Large Data Base Conference.*

[16] Jampel, M., Freuder, E. and Maher, M. (editors). (1996) *Over-Constrained Systems*. Springer.

[17] Johnston, M. (1990) SPIKE: AI Scheduling for NASA's Hubble Space Telescope. In *Proceedings 6th IEEE Conference on AI Applications*, (pages 184-190).

[18] Short Jr., N. M. and Dickens, L. (Jan.-Feb. 1995) Automatic Generation of Products from Terabyte-Size Geographical Information Systems Using Planning and Scheduling. *International Journal of Geographical Information Systems*, *9* (1), 47-65.

[19] Litzkow, M., Livny, M. and Mutka, M. (1988, June) Condor - A Hunter of Idle Workstations. In *Proceedings of the 8th IEEE International Conference on Distributed Computing Systems*, (pages 104-111).

[20] Ma, P., Lee, E. and Tsuchiya, M. (January 1982) A Task Allocation Model for Distributed Computing Systems. *C-31* (1), 41-7.

[21] Myers, B., Guise, D., Dannenberg, R., Vander Zanden, B., Kosbie, D., Marchal, P. and Pervin, E. (November 1990) Comprehensive Support for Graphical, Highly Interactive User-Interfaces: The Garnet User Interface Development Environment. *IEEE Computer*, *23* (11), 71-85.

[22] Nelson, G. (1985, July) Juno, A Constraint-Based Graphics System. In *SIGGRAPH 1985 Conference Proceedings*, (pages 235-243).

[23] Rasure, J. R. and Williams, C. S. (1991) An Integrated Data-Flow Visual Language and Software Development Environment. *Journal of Visual Languages and Computing*, *2* (3), 217-246.

[24] Ridlon, S. (1996, September) A Software Framework for Enabling Multidisciplinary Analysis and Optimization. In *6th AIAA/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization.*

[25] Shapiro, L. and Haralick, R. (1981) Structural Descriptions and Inexact Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *3*, 504-519.

[26] Shapiro, Linda G., Tanimoto, Steven L., Brinkley, James F., Ahrens, James P., Jakobovits, Rex M. and Lewis, Lara M. (1994, February) A Visual Database System for Data and Experiment Management in Model-Based Computer Vision. In *Proceedings of the Second CAD-Based Vision Workshop*, (pages 64-72).

[27] Smith, Terence R., Su, Jianwen, Agrawal, Divyakant and El Abbadi, Amr. (1993, February) MDBS: A Modeling and Database System to Support Research in the Earth Sciences. In *Proceedings of the Workshop on Advances in Data Management for Scientist and Engineer*, (pages 90-99).

[28] Upson, C., Faulhaber Jr., T., Kamins, D., Laidlaw, D., Schlegel, D., Vroom, J., Gurwitz, R. and van Dam, A. (1989) The Application Visualization System: A Computational Environment for Scientific Visualization. *IEEE Computer Graphics and Applications*, *9* (4), 30-42.

[29] Wu, M. and Gajski, D. (November 1988) A Programming Aid for Hypercube Architectures. *Journal of Supercomputing*, *2* (3), 349-372.

[30] Zweben, M., Davis, E., Daun, B. and Deale, M. (Nov.-Dec. 1993) Iterative Repair for Scheduling and Rescheduling. *IEEE Systems, Man and Cybernetics*, *23* (6), 1588-96.

UNIVERSITY OF WISCONSIN AT MADISON

Paradise - A Parallel Information System for EOSDIS

David J. DeWitt
Department of Computer Science
(dewitt@cs.wisc.edu)
(http://www.cs.wisc.edu/paradise/)

1. Project Overview

The goal of the Paradise project is to prototype a scalable database system for storing, browsing, and reprocessing EOSDIS data sets. Paradise is taking a DB-centric point of view in which both data and metadata is stored in the database system. We think that this approach is superior for a couple of reasons. First, commercial parallel database management systems have been shown to provide excellent scalability. Second, database systems already automatically deal with two levels of the storage hierarchy (primary and secondary storage). When a query is submitted for execution, the database system optimizes it to minimize execution time by selecting an execution plan that minimizes both CPU usage and the movement of data pages between disk and primary memory. Extending the optimizer and execution algorithms to handle a three-level storage hierarchy provides a number of opportunities. For example, consider a scientist who wants to process a years worth of AVHRR images corresponding to a particular region (described by a polygon). With a non-integrated approach, the scientist first must query the database system for the names of the files containing the images of interest. Then he/she must submit a request to the hierarchical storage system to move the appropriate files to disk from tertiary storage. Finally, the user can then execute his program to process the images. With a database-centric approach, the user can simply issue a query for the data of interest and let the database system deal with migrating the data from tertiary to secondary storage. Furthermore, since the user's request specifies retrieval of only a subset of each AVHRR image (the portion clipped by the polygon of interest), the database system may be able to move only a subset of each AVHRR image from tertiary storage to secondary storage and/or primary memory. The primary goal of the Paradise project is to demonstrate that a database centric approach, when combined with integrated support for tertiary storage and scalable parallelism, can provide a superior solution to many of the problems facing EOSDIS.

2. Summary of Year 3 Activities

During year 3 we focused on five major activities:

- Implementation of the parallel version of Paradise
- Completing HDF support in Paradise
- Integrating support for tertiary storage
- Implementation of a new spatial join algorithm in Paradise
- Ports and bug fixes

3. Implementation of the Parallel Version of Paradise

Over the past year we completed the initial parallel implementation Paradise and got it working both on a cluster of Sparc workstations and a 16 node SP2 that IBM donated to the project. This effort involved a major rewrite of the client-server version of the system. First, each operator process was redesigned (and re-implemented) to take each of its inputs from an input stream and send its output to an output stream.

This pipelined approach to query processing makes it simple to connect operators running on the same or different processors in a fashion that is transparent to the operator. Next, the overall software structure of the Paradise was rearchitected. Instead of a single, multithreaded process that performs all functions associated with query execution, the new architecture consists of a master process that is responsible for optimizing and compiling queries plus a slave process on each processor of the cluster or multiprocessor. After a query has been optimized and compiled, the master process walks the execution plan, initiating operators on each of the slave processors. Third, we implemented a communications infrastructure that enables operators executing on different processors to communicate with one another.

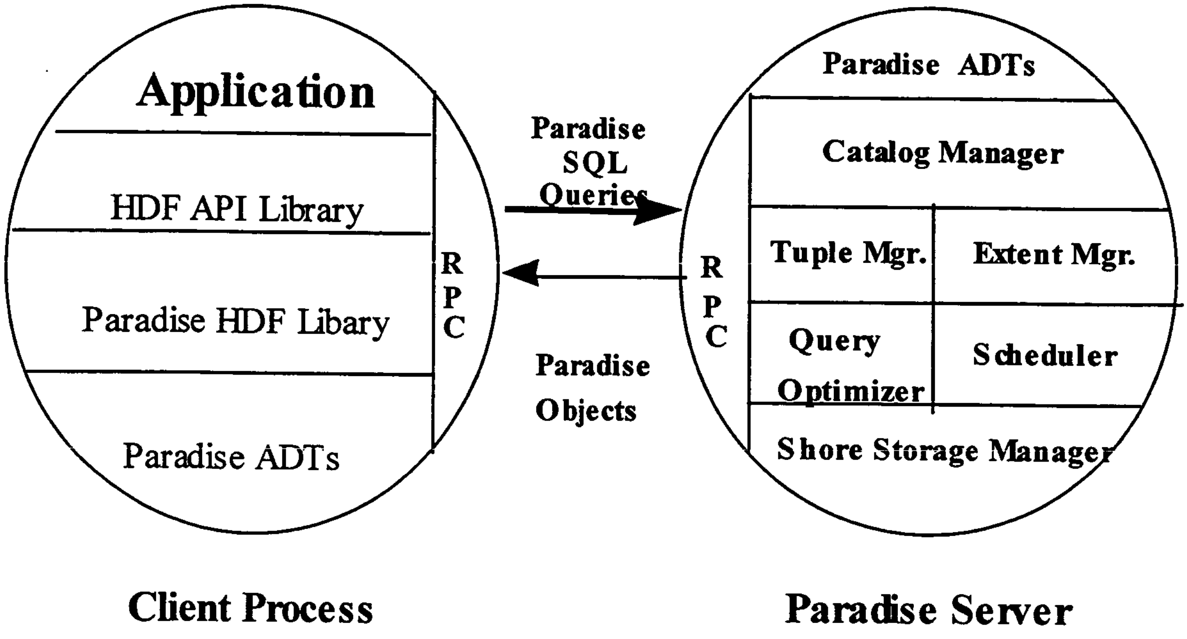
We are currently adding support for the parallel manipulation of rasters/arrays, all of Paradise's geospatial types including polygons, polylines, and points, plus video.

In February 1996, Intel donated 20 dual processor Pentium boxes to the project. This cluster is interconnected using 100 Mbit/second Ethernet and a CISCO Catalyst 5000 switch.

4. HDF Support

To simplify the task of using Paradise for those who do not "speak" SQL, we implemented an HDF-compatible, call-level interface to Paradise. Two major extensions were needed. First, we extended Paradise's type system by adding support for 8 and 24 bit raster images as well as multidimensional arrays. Each of these three types are a standard Paradise base type and thus can be used like any other type (int, float, string, polygon, etc.) when defining a relation. Since HDF's concept of V-data is analogous to the concept of a relation in a relational database, nothing new was necessary to support this construct.

The second extension involved modifying the HDF library to replace that portion that deals with reading and writing HDF files with calls to Paradise instead. To use the Paradise version of HDF, all one has to do is to relink an HDF application with the Paradise version of the HDF library. At run-time when the application makes an HDF call, the call is converted to a database query that gets shipped to the Paradise server for execution. As tuples are returned by the Paradise server to the application process, they are converted by the Paradise HDF Library to the format expected by the HDF API Library. This process is illustrated by the following figure.



The compatibility of this mechanism was tested by taking a copy of NCSA Collage and relinking it with the Paradise HDF library. As we had hoped, this was done without recompiling or modifying Collage. At the February 1996 NASA meeting for MTPE grant awardees, we demonstrated Collage running on top of Paradise. We are in the process of conducting a through benchmark of HDF on top of Paradise, comparing it with the standard version of HDF as distributed by NCSA.

5. Integrated Support for Tertiary Storage

A key benefit of taking a DB-centric approach to EOSDIS is that the database system can manage migration of data from tertiary storage to secondary storage as an extension of its existing mechanisms for migrating data from secondary storage to primary storage. To support tape-based tertiary storage we extended the Shore Storage Manager to support the DLT 4700 drive. Since DLT tapes cannot be updated in place, SHORE builds a log-structured file system on the tape. When a block on tape is first referenced it is initially cached in a memory resident buffer pool. This buffer pool is managed on an LRU basis. Tape blocks that are removed from the buffer pool by the LRU policy are buffered in a disk cache in case they are subsequently re-referenced. When a block is updated, the updated block is appended to the logical 'end' of the tape followed by a new directory block.

While modern tape technology such as the Quantum DLT 4700 is dense and relatively fast, a typical tape seek still takes almost a minute! Our solution is two pronged. First, we employ a novel query execution paradigm that we term query pre-execution. The idea of pre-execution grew from the experimental observation that queries which accessed data on tape were so slow that we could actually afford to execute the query twice! During the pre-execution phase, Paradise executes the query normally except when a reference is made to a block of data residing on tape. When such a reference occurs, Paradise simply collects the reference without fetching the data and proceeds with the execution of the query. Once the entire query has been "pre-executed", Paradise has a very accurate reference string of the tape blocks that the query needs. Then, using a cache-conscious tape scheduling algorithm, which reorders the tape references to minimize the number of seeks performed, the query is executed normally. While the idea of query pre-execution sounds impractical for a disk-based system, we demonstrate that it actually works very effectively when dealing with large raster images on tape.

The second major technique that we employ to make query processing on tape efficient is termed query batching. Query batching is a variant of traditional tape-based batch processing from the 1970s and what Gray refers to as a data pump. The idea of query batching is simple: dynamically collect a set of queries from users, group them into batches such that each batch uses the same set of tapes, pre-execute each query in the batch to obtain its reference string, merge the reference strings, and then execute the queries in the batch together. The processing of a batch is done essentially in a "multiple instruction stream, single data stream" (MISD) mode. The ultimate goal is to scan each tape once sequentially, "pumping" tape blocks through the queries that constitute the batch as the blocks are read from tape.

We have completed an implementation of these mechanisms in Paradise and have conducted a a detailed performance evaluation. A copy of the paper is available from the Paradise web site. It will be submitted to the 1997 SIGMOD Conference.

6. PBSM Join Algorithm

Users of a spatial database system frequently need to combine two inputs based on some spatial relationship - for example, a user might want to find all rivers that overlap with some landuse polygons. This operation, called a spatial join, can be very expensive and efficient algorithms for evaluating it are required. As part of handling spatial data in Paradise, we have developed a new spatial join algorithm called PBSM (Partition Based Spatial-Merge), which partitions large inputs into manageable chunks, and joins them using a computational geometry based plane-sweeping technique. A novel spatial partitioning function has been developed as part of this algorithm. A performance comparison of PBSM with existing spatial join algorithms demonstrates the advantages of PBSM, especially in cases when neither of the inputs to the join has an index on the joining attribute. A paper describing the PBSM was presented at the 1996 SIGMOD conference. A copy of the paper can be found on the Paradise web site as well as in the proceedings of the conference.

We are currently implementing a parallel version of PBSM.

7. Ports

In addition to the activities above we have also been actively porting the client-server version of Paradise to a number of other platforms including Solaris on both Pentium and Sparc processors, SGI, HP, and NT.

Over the past year, ARPA "discovered" Paradise and is planning on using the system for a number of projects including the JTF Metoc Anchor Desk and as part of a new program to disseminate geo-spatial and satellite data sets via direct-broadcast satellite into battlefield environments. In the future, ARPA will be providing support for the Paradise project.

8. Publications

We completed the following papers this year:

"Partition Based Spatial Merge Join", (Jignesh Patel and D. DeWitt), to appear, *Proceedings of the 1996 SIGMOD Conference*, Montreal, CA, June, 1996.

"Query Pre-Execution and Batching in Paradise: A Two-Pronged Approach to the Efficient Processing of Queries in Tape-Resident Data Sets", (JieBing Yu and D. DeWitt), to be submitted to the 1997 SIGMOD Conference, Tucson, Arizona.

"Processing Raster Images on Tertiary Storage: A Study of the Impact of Tile Size on Performance," (JieBing Yu and D. DeWitt), to appear at the *1996 NASA Mass Storage Conference*.

Copies of all Paradise publications can be found on the Paradise web site http://www.cs.wisc.edu/paradise/

9. Presentations

Talks on Paradise were given this past year at ARPA (August 1995), Georgia Tech (October 1995), NASA AISRP (February 1996), Intel (January 1996), IBM (March 1996), and Oracle (March 1996) and NASA CESDIS (May 1996).

10. Technology Transfer

JieBing Yu a graduate student working on the Paradise project spent the summer at HAIS working on the EOSDIS prototype.

HPCC Earth and Space Science Project Scientist

George Lake
University of Washington
Department of Astronomy
(lake@hermes.astro.washington.edu)

Statement of Work

Dr. Lake serves as the HPCC/ESS Project Scientist through a CESDIS consulting agreement. As Project Scientist, Dr. Lake is responsible for the following:

- Chairs the Science Working Group which consists of the PIs of the nine funded Grand Challenge Teams;
- Represents the interests of this group to the Project Manager and advises on the allocation of project resources;
- Provides scientific/technical oversight to the in-house computational science team;
- Designs and participates in the implementation of forums for communication between the various constituents of the project; and
- Works to advance the scientific goals of the HPCC/ESS project.

Report

1. Service as HPCC/ESS Project Scientist:

This was the first year that I served as the HPCC/ESS Project Scientist. There has been modest progress on several fronts:

- Reinvigoration of the In-house Team
- Greater Connection of HPCC with other GSFC Projects
- Finding common themes to enable the HPCC/ESS Science Team to function as a Team.

Toward these goals:

1. The In-house computational scientist positions were advertised and are in the process of being filled.

2. The concept that all In-house Science Team members should be co-funded by other sources has been approved by 930 management. This will insure a wide impact of HPCC technology at GSFC. It slightly enlarges the community of computational scientists. With time, we expect the co-funded Team members to migrate to other codes within GSFC, leading to a steady enlargement of the computational science community as this technology becomes better integrated into other NASA missions.

3. Key projects have been defined for the In-house Team that will enhance the communication between the current Grand Challenge Teams and speed the technology transfer of tools developed by the HPCC/ESS Project to the broader scientific community. These include building libraries of modules derived from Grand Challenge Team codes and leading an effort to build a flexible AMR code.

4. The first round of allocations on the T3D/E to scientists outside the project has been made.

2. Scientific Results:

Work has continued on Galaxy Harassment – the evolution of galaxies in clusters of galaxies that results from rapid fly-by collisions and global cluster tides. Papers have been submitted that elucidate the mechanism that can lead to "quasar feeding" and another that details the evolution of galaxies from the distorted spirals seen by HST to the spheroidal galaxies observed in local clusters.

Simulations were performed that showed the evolution of the largest scales of the Universe and renormalized volumes were studied to follow the evolution of clusters of galaxies and environments that look like our own local group. What is emerging from this work is that NONE of the current models can explain the structure that we observe in the Universe. However, they have also clarified the reason for that failure, enabling us to start building better models that will be explored in the coming year. The cosmological N-body code is now being modified to simulate the origin of the solar system. Richardson describes this work in his contribution to the annual report.

Other scientists involved in the projects described here: T. Quinn, B. Moore, F. Governato, D. Richardson, J. Stadel, and R. Cen.

HPCC Earth and Space Science Project PR

Adam Frank
University of Rochester
(afrank@alethea.pas.rochester.edu)

Statement of Work

Dr. Frank works with George Lake, Jarrett Cohen (Hughes STX), and members of the HPCC/ESS Project Science and Management Team to broaden the communication and impact of their scientific results and methods.

Report

I was asked by George Lake to join the program and help with the outreach effort. Since I am both a computational scientist and a science writer, he felt I might be able help place stories in national media outlets. My overall goal for the consulting is to get as many HPCC-related feature articles as possible in magazines and newspapers.

For the period January to June my activities fell into two areas. First, since I am new to the program, I have worked to learn what each of the groups is doing and look for those angles of the various research projects which can be turned into stories for the popular press. This included a 2-day trip to the 1997 simulation multi-conference in April to observe the science team meeting.

The second part of my activity has focused on placing stories in magazines and other media outlets. In Feb and March I got *Astronomy* magazine to agree to a full length story on chaos in the solar system which among other things, focused on the work of George Lake and his group. The story stresses the role of high performance computing in understanding the long term evolution of the solar system and the need for pushing the state of the art forward via hardware and software techniques. The story will published in early 1998.

In May, I wrote and submitted a proposal to *Earth* magazine concerning Dr. Olson's work on the geodynamo. After some negotiations they agreed to take the piece. This is the first time I have written for *Earth* magazine, and it is likely I will be able to place other HPCC stories in the pages. In particular I have already spoken with them about doing a story on the magnetosphere which would focus on Dr. Gomobosi's team. In June I traveled to Washington/Baltimore to meet with Dr. Lyster in preparation for a story proposal based on his work. I also traveled to Johns Hopkins to meet with Dr. Olson and interview him for the *Earth* magazine story.

In addition to this work I have also made contact with K. C. Cole, the science editor of the *L.A. Times*. K. C. is formerly the editor of *Discover* magazine where she edited some of my pieces. In the future I will stay in close contact with her and alert her to HPCC-related stories which she might use in the *L.A. Times*

For the next year I hope to convince *Discover* magazine to take a story on dynamos which would focus on both the HPCC solar and terrestrial science teams' work. In addition, I will try and place more stories in *Astronomy* and *Earth* magazines as well as branch out into other venues such as *Air & Space*, *Smithsonian*, and *Natural History*.

Formation and Stability of Planetary Systems

Derek C. Richardson
University of Washington
Department of Astronomy
(dcr@hermes.astro.washington.edu)

1. Introduction and Scientific Goals

I started consultation work for CESDIS in October 1996 on the problem of developing a massively parallel processor (MPP) application for simulating planet formation with a suitable machine such as the Cray T3E at GSFC. In what follows I will outline the goals of the project, the work performed to date, and the research plan for the next year.

The ultimate goal of this project is to model Solar System formation via the Planetesimal Hypothesis, the supposition that planets formed by the pairwise accretion of small chunks of material (planetesimals) that coalesced out of the nebula. Previous attempts to do this have been forced to rely on analytical approximations, statistical techniques, or direct N-body methods with comparatively few particles and severe spatial restrictions. The chief advantage of direct numerical simulation (where the gravity of all particles is included) over analytical and statistical techniques is that many assumptions to do with the nature of the planetesimal interactions can be eliminated, replaced by the exact laws of Newtonian mechanics. However, hardware and algorithmic constraints have limited direct methods to 10^4 particles for 10^4 dynamical times (yr), or 10^2 particles for 10^8 yr, which is either too few particles to resolve the interesting dynamics (it also makes implicit assumptions about initial conditions), or too short to follow the evolution to its natural conclusion.

The work I report on here promises to yield simulations of at least 10^6 particles for 10^6 years, or permutations thereof, in a large spatial domain (and fully 3D), giving insight into disk dynamics and planet formation that have previously been out of reach of direct methods. The work will quantitatively show for the first time the transition from runaway growth to the final accumulation of protoplanets into planets. It will also address other fundamental questions in cosmogony, such as: the nature of the primordial mass distribution, the likely extent of radial mixing, the cause of mass depletion in the asteroid belt, the origin of planetary spins/obliquities, the likelihood that the Moon was formed by a late massive impact, and the role of giant planets in terrestrial planet formation.

2. Method

To accomplish these goals, I have begun to modify a spatially adaptive cosmology code called "pkdgrav" developed at the University of Washington and designed to run on MPPs. Spatial adaptivity is achieved by using a tree-code to achieve "N log N" scaling of the force calculations rather than N^2. A balanced k-D tree is constructed by recursively bisecting the longest axis of the particle distribution. The leaf nodes are buckets that contain several particles (usually 8 to 32) whose force calculations are collectively optimized. At each level of the tree, multipoles are calculated to speed distant force evaluations. While this code was designed for parallel implementation, the serial version is three times faster than any previous force solver using trees. The complete parallel code achieves a sustained performance of 28 Gigaflops on the 512-node Cray T3E; about a hundred times the speed of a Cray C90.

The modifications I am making (with help from Tom Quinn, Joachim Stadel, and George Lake, all at UW) include: collision detection and resolution, sensitive hierarchical time-steps, double precision data storage and manipulation, and external potentials for gas drag and possibly the Sun and giant planets. These ideas will be expanded upon below, along with comments of what has been accomplished so far to implement them and what remains to be done.

2.1 Collision Detection and Resolution

This is the most fundamental addition to the code and takes the place of the traditional force softening used in cosmological codes. Collision detection is important because this is how planetesimals grow: by colliding and merging with other planetesimals. This means it is imperative to detect all collisions between bodies and decide what the outcome should be in each case. The idea for now is that collision outcomes will depend on the relative impact energies, the lowest energies leading to mergers and the highest energies leading to fragmentation. The energy thresholds can be determined from laboratory experiments that other researchers have already undertaken. Currently collision resolution is still in the testing phase and I haven't implemented any realistic energy thresholds yet. Also, only merging and bouncing have been coded. Since the code must allow for a dramatic change in the total number of particles during the simulation, the balanced parallel code requires considerable bookkeeping. These changes are currently being tested.

Collisions are detected by examining the closest N_n neighbors of each particle. The leapfrog integrator used in pkdgrav detects collisions using linear trajectories; there are no complicated parabolas to follow. It also allows the detection of ALL collisions in the correct sequence even if a single particle suffers more than one collision during the interval. N_n is parameterized by the local number density, velocity dispersion, and time-step, and is kept large enough to ensure the chance of missing a collision is exceedingly small (this is still in the testing phase). Note that N_n << N until most particles are incorporated into protoplanets.

2.2 Sensitive Hierarchical Time-steps

These are critical in order to follow the gravitational scattering of the planetesimals. This cross section is naturally larger than that for collision, so a particle's time-step must take into account the gravitational influence of its neighbors. After some experimentation I have chosen the following time-step formula for now: dt = eta (a / a_dot), where a is the net acceleration on the particle, a_dot is the first derivative, and eta is a dimensionless constant. In the absence of nearby perturbers, this expression reduces to dt = eta t_orb, the local orbital timescale in a Keplerian disk. The value of eta can be chosen to ensure a minimum number of steps per orbit. The ideal choice for eta still needs to be determined. Note that pkdgrav uses hierarchical time-stepping where the steps differ by powers of 2.

2.3 Double Precision Data Storage and Manipulation

The cosmology code pkdgrav was originally written in single precision (accuracy of 1 part in 10^8) since this was sufficient for the expected dynamic range. But for the planetesimal problem with N = 10^6 or more particles spread out in a disk between 0.5 and 2.0 AU and integrated for 10^6 or more years, double precision (1 part in 10^{16}) is required to have more than 2 digit accuracy (which is the absolute minimum for accurate collision detection). This is because the planetesimals are between 10 and 100 km in size initially and range over distances of order 10^9 km.

Similarly, the time-steps can be as small as 1 part in 10^5 yr to ensure careful stepping through encounters (see discussion above). The switch to double precision is actually a substantial change for pkdgrav and has yet to be implemented. Initial tests (see below) used fewer, larger particles over relatively short intervals, so double precision has not yet been a big concern.

2.4 External Potentials

These are needed to include gas drag, and possibly the giant planets as well. Gas drag is relatively straightforward to implement since the amount and direction of the drag is characteristic only of the particle in question. As for the giant planets, in initial tests (discussed below) they (and the Sun) were treated like any other particle. Preliminary results from these tests indicate that tree cells may be overwhelmed by the large masses involved and it would be better to keep them separate from the tree. It may be that they should be included as external potentials, but that would make it difficult to account for back reaction from the disk. More tests need to be performed in this area.

2.5 Auxiliary Development

In order to supply pkdgrav with sensible initial conditions and in order to analyze the results in a way suitable to the planetesimal problem, I have developed the auxiliary code and scripts described below. I have been able to exploit data structures used in my previous work by designing filters to convert back and forth between these and pkdgrav's data structures.

- Initial Conditions:

  I have developed a code called "ssic" (for Solar System Initial Conditions) to accept the following parameters and generate appropriate initial conditions using a random number generator:

  - Giant planets to include (Jupiter, Jupiter and Saturn, or none)
  - Number of planetesimals
  - Total mass (typically 2 Earth masses)
  - Planetesimal density (typically 2 g/cc)
  - Planetesimal radius scaling (to exaggerate particle size if desired)
  - Inner and outer orbital radius (currently using 0.5 to 2.0 AU)
  - Projected surface density exponent (use -1.5; true value unknown)
  - Eccentricity and inclination dispersion (can be zero for cold disk)

- Output Analysis:

  In addition to providing simple graphical animation, I have written an analysis package "ssa" and scripts to plot the following key indicators:

  - Time evolution of N
  - Time evolution of mean and maximum planetesimal mass
  - Time evolution of the velocity dispersion
  - Snapshot of surface mass distribution
  - Histograms of eccentricity, inclination, and mass
  - Snapshots of eccentricity and inclination as a function of semimajor axis and planetesimal mass
  - Various diagnostic quantities, such as the motion of the barycenter

  * Note there is no provision for analysis of particle spin distribution yet.

- Sample Test Results

I have run a number of cases to test various aspect of the code development. For this summary, I will discuss only a few of the more recent and meaningful cases that used N = 1000 in a cold disk and lasted 1000 years (note even these test cases are competitive with previous research!). These tests used fixed uniform steps of 0.01 yr (hierarchical stepping still needs work) and assumed perfect accretion (i.e., all collisions lead to mergers). Varied parameters were: the tree opening angle theta (which determines the force accuracy); the number of giant planets; and the value N_n used in collision searching. For the initial conditions, the nominal values given in the "Initial Conditions" section above were used, except where indicated.

The following conclusions were derived from these preliminary tests:

1. The final value of N after 1000 yr of accretional evolution is largely determined by how fast the e & i of the planetesimals increase. The larger the values of e & i, the fewer the collisions. Hence in cases without giants, N dropped the fastest as there was no i pumping (essentially the calculation was 2D). The growth of e & i is also slower when only one planet (i.e., Jupiter) is included.

2. The magnitude of the barycentric drift (i.e., the total linear momentum conservation error) can be controlled by theta. Only two values were tested, 0.8 and 0.5, but the latter resulted in a considerable reduction in barycentric drift (factor of 10) and a slight reduction in e & i growth as well, indicating that force errors are contributing to disk heating.

3. The barycentric velocity oscillation is correlated with the orbital periods of the giant planets. The magnitude of the oscillation is always much less than the corresponding position drift and is close to single precision noise.

4. There is no difference in outcome for different values of N_n ranging from 16 to 128. There IS a difference in computation time however, with N_n = 16 being 2.3 times as fast to compute as N_n = 128. This indicates that even for N_n as low as 16, no collisions are being missed.

I have included three figures illustrating key aspects of these tests. The first shows a plot of N vs t for a "nominal" run with two giants; there were 920 planetesimals left at the end of the run. The second plot shows the barycentric position and velocity drift for a run that included only Jupiter; note the 12 yr oscillation in the velocity that corresponds to Jupiter's orbital period. The third plot shows e & i vs t for a nominal run with theta = 0.5; the fairly rapid increase in e & i causes the merger rate to slow down.

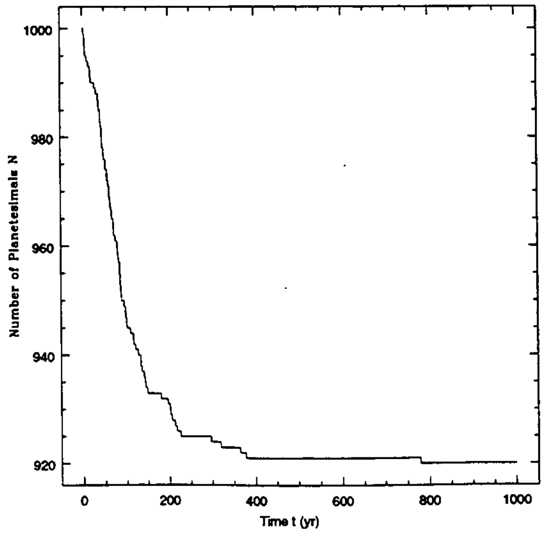


Figure 1: Plot of N vs. t for a Nominal Run With Two Giants.

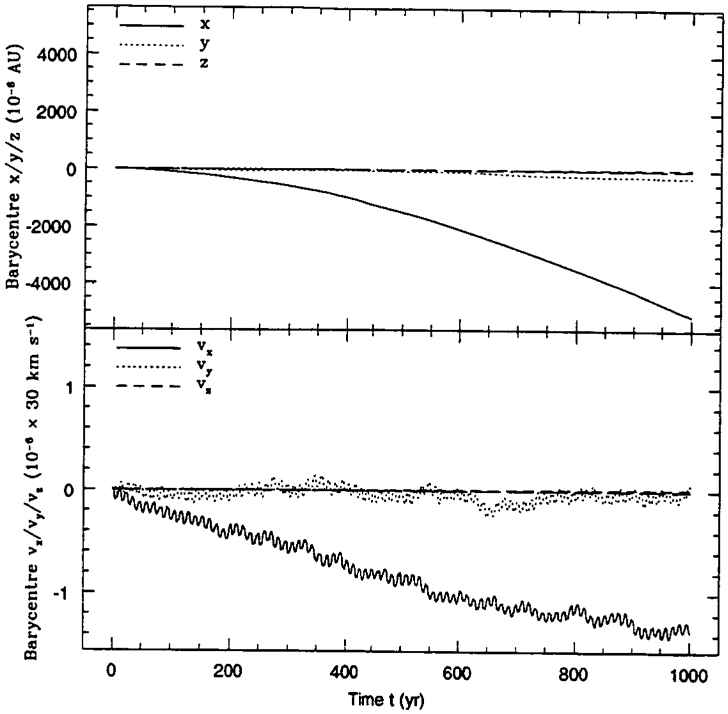


Figure 2: Barycentric Position and Velocity Drift for Run Including Only Jupiter.

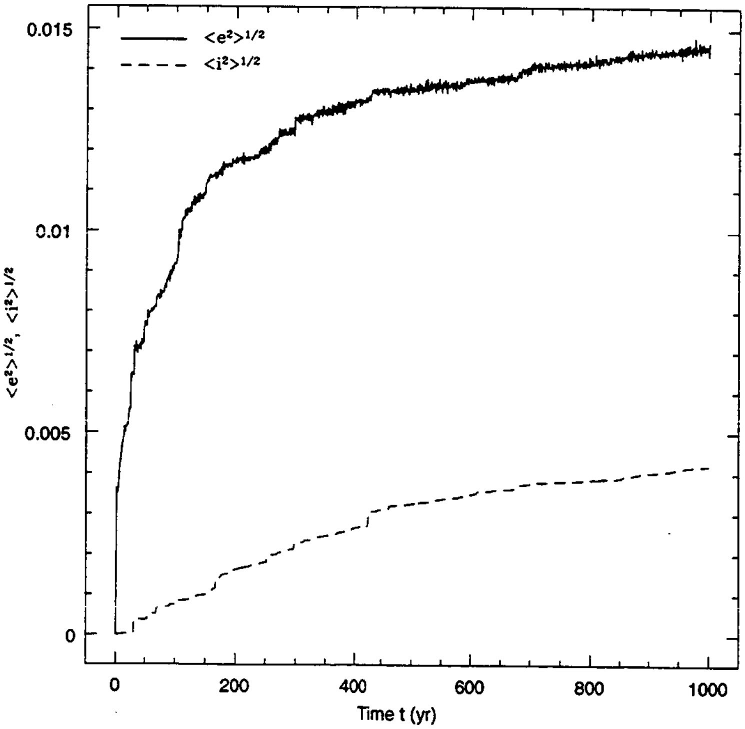


Figure 3: Shows e & + i vs. t for a Nominal Run With Theta = 0.5.

3. Other Work Done to Date

Beyond the preliminary assessment performed in October I also performed tests with my own non-symplectic integrator (which was quickly shown to be unsuitable as it led to unacceptable radial drift over long intervals). I also performed limited tests of pkdgrav in parallel mode. I am currently in the process of running tests to establish the contributions to disk heating in the sample runs discussed above, whether it be from dynamics intrinsic to the disk, perturbations by the planets, or the step size. This is important since it directly affects the merger rate.

4. Future Work

There is not much more coding to do, but there is still a lot of testing and tweaking of parameters to be done. The energy thresholds for collision outcomes need to be established, and a model for fragmentation (i.e., number of fragments, their mass and velocity distribution, etc.) needs to be implemented. Parameters that need to be tweaked include N_n, eta in the time-step formula, and theta. How best to handle the Sun and giant planets needs to be determined, and a gas drag model should be incorporated eventually. It is also necessary to finish the conversion to double precision and do more testing on parallel platforms. Once these items and smaller tests are complete, a full-scale run on the T3E will be performed.

The following estimated timeline gives a more detailed research plan for the new year:

Month 1-3: Finish initial coding (collision handling, hierarchical steps, double precision, and external potentials). Refine choice of simulation parameters (N_n, eta, theta, maximum time-step to ensure energy stability, etc.)

Month 4-6: Test performance on local cluster (8 Alpha workstations with a fast switch). This test will use $N = 10^4$ and $t = 10000$ yr (better than any previous study) under the assumption of perfect accretion. Jupiter and Saturn will be included in some form to be determined.

Month 7-9: Resolve any problems arising from test run. Compare test results with previous work. Publish short paper, if warranted. Start small scale tests on T3E.

Month 10-12: Implement and test fragmentation model. Add modifications to handle I/O with $N = 10^6$ particles in preparation for full scale runs on T3E. Test with short runs.

It is anticipated that the first full-scale run on the T3E will be underway within the year.

APPLIED INFORMATION TECHNOLOGY BRANCH

Yelena Yesha, Acting Branch Head

Digital Libraries Technology
Nabil Adam, Rutgers University
Yair Amir, Johns Hopkins University
Susan Hoban, University of Maryland Baltimore County
Konstantinos Kalpakis, University of Maryland Baltimore County
Aya Soffer, University of Maryland Baltimore County

Global Legal Information Network (GLIN)
Nabil Adam, Rutgers University
Tarek El-Ghazawi, George Washington University
Konstantinos Kalpakis, University of Maryland Baltimore County
Russell Turner, University of Maryland Baltimore County

Digital Libraries Consultants
Nabil Adam, Rutgers University
Hans Mark, University of Texas at Austin

Executive Secretariat to the U. S. Global Change Research Program's Data and Information Management Working Group
Les Meredith, Senior Scientist

Executive Secretariat to the Committee on Environmental and Natural Resources (CENR) Task Force on Observations and Data
Sushel Unninayar, University of Maryland Baltimore County

Direct Readout Image Processing Ground System
Fran Stetina, Fran Stetina and Associates

High Data Rate Satellite Communications
Burt Edelson, Neil Helm, George Washington University

NASA and the Private Sector
Murray Felsher, Associated Technical Consultants

3-D Unstructured-Grid Adaptive H-Refinement Module
Rainald Lohner, George Mason University

Linearized Riemann Solver for Numerical Magnetohydrodynamics (MHD)
Dinshaw Balsara, National Center for Supercomputing Applications, University of Illinois

Scalability Analysis of ECS, Data Server
Daniel Menascé, George Mason University
Mukesh Singhal, Ohio State University

DIGITAL LIBRARIES TECHNOLOGY

CESDIS has been tasked with conducting research in areas related to digital library technology, specifically in areas which will complement the work proposed by the investigator teams funded by NASA Cooperative Agreement Notice CAN-OA-94-01. Additional areas of research suggested include:

- Advanced query capabilities that provide location independent access, anticipate users' needs, assist the user in performing precise requests, and support approximate queries.
- Adaptive user profiles, information filtering, and time-constrained delivery.
- Translation methodologies, content languages, and development of ontologies that facilitate integration and interoperation of multiple information resources.
- Repository management, replication, and caching of objects in a distributed, heterogeneous environment.

This work is performed by Dr. Nabil Adam, Rutgers University, and Drs. Aya Soffer, Susan Hoban, and Konstantinos Kalpakis of the University of Maryland Baltimore County.

CESDIS has also been tasked to provide consultants to develop plans and coordinate interagency, university, and industry collaboration in the area of digital libraries. This work has been undertaken by Dr. Nabil Adam of Rutgers University and Dr. Hans Mark of the University of Texas at Austin. Reports on these activities follow.

Geodata Modeling and Query in Geographic Information Systems

Nabil Adam
Rutgers University
Center for Information Management, Integration, and Connectivity (CIMIC)
(adam@adam.rutgers.edu)

Statement of Work

Dr. Adam and graduate students at Rutgers University were funded to conduct research in the broad area of digital libraries, specifically geodata interoperability specifications. They proposed to develop a better understanding of computational modeling systems (CMS) and open geodata interoperability specifications (OGIS) and identify their commonalities, limitations, and strengths. They proposed to develop a full implementation of a realistic and meaningful example in both CMS and OGIS by developing a prototype implementation. They also proposed to investigate the hypothesis that CMS and OGIS are complementary and study ways by which the two components could be integrated into a cohesive system for supporting model development and query by domain experts in the area of GIS. The following report is the introduction from a longer technical report which may be obtained through the CESDIS administrative office.

Report

Geographic information systems (GIS) deal with collecting, modeling, managing, analyzing, and integrating spatial (locational) and non-spatial (attribute) data required for geographic applications. Examples of spatial data are digital maps, administrative boundaries, road networks, and those of non-spatial data are census counts, land elevations and soil characteristics.

GIS shares common areas with a number of other disciplines such as computer-aided design, computer cartography, database management, and remote sensing. None of these disciplines however, can by themselves fully meet the requirements of a GIS application. Examples of such requirements include: the ability to use locational data to produce high quality plots, perform complex operations such as network analysis, enable spatial searching and overlay operations, support spatial analysis and modeling, and provide data management functions such as efficient storage, retrieval, and modification of large datasets; independence, integrity, and security of data; and concurrent access to multiple users. It is on the data management issues that we devote our discussions in this monograph.

Traditionally, database management technology have been developed for business applications. Such applications require, among other things, capturing the data requirements of high-level business functions and developing machine-level implementations; supporting multiple views of data and yet providing integration that would minimize redundancy and maintain data integrity and security; providing a high-level language for data definition and manipulation; allowing concurrent access to multiple users; and processing user transactions in an efficient manner. The demands on database management systems have been for speed, reliability, efficiency, cost effectiveness, and user-friendliness. Significant progress has been made in all of these areas over the last two decades to the point that many generalized database platforms are now available for developing data intensive applications that run in real-time. While continuous improvement is still being made at a very fast-paced and competitive rate, new application areas such as computer aided design, image processing, VLSI design, and GIS have been identified by many as the next generation of database applications.

These new application areas pose serious challenges to the currently available database technology. At the core of these challenges is the nature of data that is manipulated. In traditional database applications, the database objects do not have any spatial dimension, and as such, can be thought of as point data in a multi-dimensional space. For example, each instance of an entity EMPLOYEE will have a unique value corresponding to every attribute such as employee_id, employee_name, employee_address and so on. Thus, every Employee instance can be thought of as a point in a multi-dimensional space where each dimension is represented by an attribute. Furthermore, all operations on such data are one-dimensional. Thus, users may retrieve all entities satisfying one or more constraints. Examples of such constraints include employees with addresses in a certain area code, or salaries within a certain range. Even though constraints can be specified on multiple attributes (dimensions), the search for such data is essentially orthogonal across these dimensions.

In contrast with the traditional database applications, GIS applications require both spatial and non-spatial objects as data. Unlike a non-spatial object, a spatial object may have dimensions such as length (for lines), area (for surfaces), and volume (for solids). Furthermore, spatial objects have locations identified by a coordinate system. The locational property of spatial objects gives rise to spatial search even for zero-dimensional objects such as points. For instance, a user may request to retrieve all points within a specified radius from a point given as the center. Such a query will require a two dimensional search. If the information on spatial proximity among all point objects is not preserved, the above query will require an exhaustive comparison of the coordinates of the center with that of all other point objects in storage, resulting in a prohibitively expensive query execution plan. The information on spatial proximity will be lost if the point objects are represented in the same way as traditional database objects (for instance, as rows in a relation).

Additional requirements arise with higher dimensional objects. Examples include spatial relationships such as intersection between linear objects, overlap, containment, and shared boundaries between aerial objects, incidence relationship between a point and a line, and containment and distance relationship between a point and an area. Simple geometric objects such as line and area can be combined into larger objects. Such objects can be further classified into compound, where the constituent objects are similar, or complex, where the constituent objects are dissimilar. Examples of compound objects are dyad (pair of point domains), network (collection of curves), lattice (collection of points), and tessellations (collection of areas). Complex objects can be decomposed into a finite number of constituent domains of different types. An example of a complex object is a spatial unit consisting of a land parcel, a house, and utility network.

In the above discussion we took the object-based approach of dealing with geographic data. Another approach used frequently in developing a GIS, called the field-based approach, uses a complementary view of spatial information. Instead of associating attributes with individual spatial objects, it addresses the variation of the individual attributes across a spatial domain. This gives rise to the layer-based organization of data, where each layer represents the spatial variation of a specific attribute. Such representations impose additional functional requirements such as polygon overlay and reclassification, which cannot be supported using existing database technology.

Since both non-spatial and spatial data are used in a GIS, it requires a seamless integration between spatial data such as extent, location, and orientation, and attribute data such as ownership and valuation of land parcels. Users must be able to retrieve spatial data given set of attribute values and retrieve a set of attribute values for a specified spatial object. Modifications of spatial data such as polygons representing land parcels as well as their attribute values must be supported. This would require maintaining integrity of data if an update is performed. For example, in a GIS for land information, if the size of a land parcel is changed, it would affect that of its neighboring land parcels, and the attribute data must be changed for all land parcels that are affected by the changes made. Integration of spatial and non-spatial data, and maintaining data integrity for both types poses a significant challenge to existing database technology that has been primarily designed for non-spatial data.

Other challenges include dealing with spatio-temporal data, providing multi-valued logic capabilities, and handling extremely long transactions.

The rest of the book is organized as follows: Chapter 2 contains an analysis of the data requirements in various GIS applications. GIS applications are categorized into object-based and field-based applications, depending on the nature of their spatial information requirements. Field-based applications deal with the spatial distribution of data, whereas object-based applications deal with objects with geospatial references. We show how the data requirements of the object-based applications differ from those of the field-based applications. Within each group, further classification is made by functional areas and the spatial data modeling techniques that are predominantly used. We thus develop a taxonomy of GIS applications based on the nature of their information requirements.

Chapter 3 starts with the operations required for various field and object-based applications. We discuss the differences in these operations and illustrate how each operation is supported more naturally by either representing the application domain as a spatial distribution of certain attributes (fields) or as a number of discrete objects plotted in an Euclidean space. There are fundamental differences between the organizations of spatial data in the field and object based representations. For instance, in field-based applications, the spatial distribution of an attribute is captured by dividing the space into tessellations and studying the spatial variation of attribute values across tessellations. Object-based applications, on the other hand, need to manipulate the spatial extents of geographic objects. Different representations of spatial data give rise to interoperability issues: how to convert data between different methods. We describe spatial organization methods based on the discretization of space using regular and irregular tessellations, and those for representing the topologies of spatial objects. We describe the methods developed for intra- and inter-format conversion methods. We then discuss the issue of integrating the two methods in a way that would facilitate spatial data organization in applications requiring both types of data organization.

In chapter 4 we discuss the usefulness of using database technology for GIS applications. We discuss the various strengths and weaknesses of database technology in dealing with spatial data. We also discuss the various database architectures that have been proposed to deal with applications such as GIS. We perform a requirements analysis by evaluating the needs of GIS applications and then identify the functionalities that can be provided by a database platform. This exercise is geared towards determining the feasibility of a generalized database platform for supporting GIS applications across multiple domains.

Next (chapter 5) we take a top-down view of developing a spatial database for GIS applications. We start with a discussion on the various attempts made in spatial data modeling. The data models are divided into two categories: application-dependent, and application-independent. We further categorize existing data models into underlying paradigms on which they are based. Two such paradigms are extensions of the entity-relationship and object-oriented models. The data models are evaluated by their ability to support the spatial operations required by both field and object-based applications. In addition, we suggest ways of combining existing data models by drawing upon the strengths and eliminating the weaknesses of each.

Chapter 6 addresses the topic of spatial query processing. We categorize spatial queries into the type of data manipulated, the type of operations performed, and the language in which the queries can be expressed. For each of these, we discuss the studies done in the literature, their shortcomings, and ways of improving them. We also address the issue of optimization of spatial queries, by studying the effect of processing strategy on performance, and possible ways of restructuring to improve operational efficiency.

We next (chapter 7) describe the physical database design issues for storing and retrieving spatial data. Efficient storage and retrieval of data is accomplished by indexing. We discuss the various spatial indexing methods that have been developed for different data types such as point, line segment, rectangle, and volume data. We then discuss what extensions are required to incorporate spatial indexing capabilities in relational databases.

In conclusion, we discuss the open research issues in each of the above areas, and provide future research ideas for possible improvements of the existing methodologies.

Combining Satellite Communication in Commedia

Yair Amir
Johns Hopkins University
Department of Computer Science
(yairamir@cs.jhu.edu http://www.cs.jhu.edu/~yairamir)

Statement of Work

Using the Internet currently, it is possible to pass a message between almost any two points within the U.S. with a latency of about 80 milli-seconds (turn around time), with a relatively high probability of success. Preliminary measurements for satellite communication show that latency of about half a second (turn around time) will be experienced for each satellite hop. This drawback creates an interesting problem for protocols that are designed to achieve interactivity. Satellite communication may provide high bandwidth with access to almost any point on the Earth, including places where Internet connection is not yet supported or which lacks the necessary bandwidth for systems such as Commedia, a crossplatform infrastructure for multimedia conferencing. A subcontract was recently put in place with Johns Hopkins University for research by Dr. Amir on the possibility of utilizing satellite communication within Commedia. A report on early work follows.

Report

- We have evaluated ways to build Mpeg viewers on Unix and Windows platforms. Based on our evaluation, we are going to use mainly software player. In the windows (95 and NT) environment, our choice is the Active Movie architecture from Microsoft. In the Unix environment, our choice is the MpegTV player. We have hosted Tristan Savetier, President of MpegTv Inc. and discussed several enhancement to the player that will help control the player from an outside program. We have provided one of our BSDI machines to MpegTV so that the company can port the mtv player to BSDI Unix. This was done and now allows us to use the player in the following Unix environments: BSDI, Linux, SGI, Sun/Solaris.

- We have created an Mpeg client for both the Unix and Windows environment. We also are able now to simulate a live Mpeg source from any Unix machine. We can also use our two live Mpeg hardware sources on Windows (NT and 95).

- We have created a client for the Connectix camera, both for regular Unix and for Java. The Connectix camera is capable of providing about 8 frames per seconds (64 gray scale). We have a source running on both BSDI and Linux. We get very good performance running the Java client and using the Internet Explorer browser. Netscape 3 does not provide adequate performance (for rates of 1Mbits/sec).

- I have adapted the group communication protocol for the wide area network we are using. This is not the perfect solution, but it is working and gives us lots of valuable information. We can push around 160Kbits/sec reliable multicasting between all of the machines in this network. The network configuration, detailed below, contains 19 machines from Hopkins, UMBC, GSFC, Rutgers, and DIMACS. It is fully operational as of July 1st.

- Existing wide area network layout (5 sites):

```
# cnds.jhu.edu domain       # cs.umbc.edu domain        # gsfc.nasa.gov domain
5 128.220.221.255           3 130.85.100.255            3 128.183.0.0
commedia  128.220.221.1     topdog    130.85.100.62     cesdis     3128.183.38.27
com1  128.220.221.11        stavro    130.85.100.121    cesdis7    128.183.38.31
com2  128.220.221.12        retriever 130.85.100.32     what       128.183.38.63
com3  128.220.221.13
com5  128.220.221.15

# rutgers.edu domain        # dimacs.rutgers.edu domain
3 128.6.42.255              5 128.6.75.255
cimic   128.6.42.134        dimacs       128.6.75.16
cimic1  128.6.42.127        lunar        128.6.75.43
adam    128.6.42.5          iyar         128.6.75.51
                            av           128.6.75.54
                            browning     128.6.75.22
```

- Although apparently we cannot use cesdis1(CESDIS web server) in our experiment, we have three other web-servers (at Hopkins, UMBC, and DIMACS) that are taking part in the experiment.

- We have reached the point where we have a limited working version of Commedia, with multicast protocols, group communication services, very simple media sensitive protocols, and representative applications. We also have a wide area network test-bed constantly available in the East coast.

- The goal for the next few months will be to use this test-bed to investigate the required properties for multicast and group communication protocols that will work despite high latency and possible omission and partition failures, such as in the wide area network, or when using satellite communication. Based on these properties, we plan to design and build a new set of low level protocols that will be able to work efficiently in these environments.

NASA Digital Library Technology Project Support

Susan Hoban
University of Maryland Baltimore County
Department of Computer Science and Electrical Engineering
(shoban@pop900.gsfc.nasa.gov)

Profile

Dr. Hoban received a B.S. in astronomy, an M.S. in physics, and a Ph.D. in astronomy, all from the University of Maryland. Prior to joining CESDIS through a subcontract with the University of Maryland Baltimore County, Dr. Hoban was a Principal Scientist with Hughes STX, providing support to the GSFC Digital Library Technology Project as the Assistant Manager. She served as a guest lecturer for the Maryland Space Grant Consortium, teaching a course entitled *Introduction to the Internet for K-12 Educators.* As the Principal Investigator on a project funded through NASA's Innovative Developments in Education in Astronomy Science, Dr. Hoban became Dr. Sue in *Astronomy On Line: Ask Dr. Sue.* In this capacity she developed science education curricula as well as its World Wide Web implementation. Dr. Hoban was also instrumental in developing a homepage for NASA's Chief Scientist, Dr. France Cordova.

Dr. Hoban's association with NASA began when she was selected to participate in the NASA Graduate Student Researchers Program for work in charged coupled device imaging (astronomical observations and analysis). As a National Academy of Sciences Research Associate, she performed research on infrared spectroscopy of astronomical sources and served as the IRAF Data Reduction Package Manager for installation, maintenance, and user assistance. This work was continued as a research scientist with USRA's Goddard Visiting Scientist Program. Dr. Hoban's research interests include image processing of remotely sensed data, two-dimensional data analysis (spectral and spatial), and multi-wavelength studies of comets and young planetary systems.

Report

Direct support is provided to Dr. Nand Lal (Code 935), Manager of the HPCC/IITA Digital Library Technology (DLT) project. The DLT project consists of a group of seven Cooperative Agreement Teams funded by NASA and six university-led consortia funded jointly by NSF, DARPA, and NASA. Support to Dr. Lal consists of preparing the DLT Monthly report in HTML, and preparing summaries and presentations as required, such as a briefing for HPCC/IITA Center Review held at Goddard on September 6, 1996; a presentation and demonstration delivered by Dr. Lal to Dr. Robert J. Hansen, Director of the NASA Center of Excellence for Information Technology, on March 20, 1997; input for a proposal regarding GSFC National Earth and Space Science Library; the development of a proposal for a follow-on program; and input for the HPCC Independent Annual Review, held in June at NASA Ames Research Center. Also, attendance at weekly management telecons as well as other management meetings, such as the HPCC Learning Technologies Project Strategic Planning Meeting, held at NASA Ames Research Center, May 22-23, is required.

Other DLT-related events which require support included the IITA Principal Investigators Meetings, which were held at NASA HQ in Washington, DC on Sept. 16 - 18, 1996 and at Lockheed-Martin in Sunnyvale, CA, on May 19 - 21, 1997.

Also, the DLT Team (Lal, Maurer and Williams/933, Soffer/CESDIS, Burrows and Harberts/HSTX, Rosati/Adnet) organized and held an Information Technology Workshop on March 11 - 13, 1997, at Goddard. Over sixty attendees were present, including investigator teams from the IITA Digital Library Technology project and the OSS Applied Information Systems Research program, as well as members of the space and Earth science communities. The investigators presented products resulting from their funded projects during oral presentations and also at the Goddard Atrium Teas and Posters on March 11 and 13. On March 12, the investigators held technical discussions.

Advances in Digital Libraries Forum

Support for the coordination of the IEEE Advances in Digital Libraries Forum (ADL97) was provided.

Global Legal Information Network

The Global Legal Information Network (GLIN) is a joint effort between the Library of Congress and NASA. CESDIS is providing technical support to the NASA component of GLIN. As Project Coordinator for the CESDIS/NASA GLIN effort, Hoban coordinated the Goddard component of the GLIN Director's meeting held in September 1996, drafted the GLIN Project Plan, and wrote an article for the NASA Information Systems newsletter with Judy Laue (HSTX) about GLIN. Hoban also worked with Kalpakis/CESDIS on a proposal to NASA to assimilate remotely sensed data into the GLIN system (see Proposal Support, below).

Regional Validation Centers

Hoban has worked with GSFC Code 935 in their Regional Validation Centers projects. Hoban organized the Training Workshop July 16 - 19, 1996 and assembled the RVC Training Manual. Hoban also serves as the liaison between Code 935 and the UMBC RVC.

Proposal Support

Hoban was a co-author of two proposals in response to the NASA NRA-96-OSS-10 (Applied Information Systems Research Program):

1. with K. Kalpakis/CESDIS-UMBC, Yaacov Yesha (UMBC) and F. P. Schloerb (U. Mass): FCRAO Telescope Archive and Retrieval System, and

2. with J.M. Hollis/GSFC and S. Chettri (GST): Cometary Science Information System
   Neither proposal was funded.

Hoban was a co-author of a proposal in response to the NASA CAN-97-MTPE-02 (Earth Science Information Partnerships) with K. Kalpakis/CESDIS-UMBC, D. Zaelke and D. Hunter/Center for International Environmental Law, R. Medina and N. Kozura/Library of Congress and J.P. Gary and W. Campbell/GSFC: "Integrating Environmental and Legal Systems." Selections have not yet been announced for this solicitation.

Hoban is collaborating with CESDIS and GSFC personnel to prepare a white paper outlining possible CESDIS participation in the development of a data system for NASA's Stratospheric Observatory For Infrared Astronomy (SOFIA).

Other

Hoban prepared an HTML version for "NASA's Science Policy Guide" and "NASA's Science Communications Strategy" at the request of Dr. France Cordova, who was Chief Scientist of NASA at the time.

During the summer of 1996, Hoban supervised Arika Anderson, a Master's student in mathematics who worked on developing a science library on the Web. Anderson's project involved developing a set of evaluation criteria for science-related Web sites to enable users to narrow the search space for items of interest. Anderson evaluated 200 Web sites and prepared a report of her findings.

Hoban prepared and delivered with George Lake (University of Washington) a presentation for the Petaflops workshop in Annapolis, Maryland (10/28/96) entitled "The Digital Sky" which discussed issues pertaining to the data volume expected to be produced by the next generation of astronomical sky surveys.

Hoban is currently working with Bill Campbell/935 to coordinate a symposium in honor of remote sensing scientist, Nick Short, Sr., to be held at Goddard Space Flight Center in the fall of 1997.

Awards

Hoban received a NASA Headquarters Special Service Award for her efforts on the "Alliance for Science" project. Hoban also received a NASA Goddard Space Flight Center Group Achievement Award for her participation in the Regional Validation Center project.

Digital Research Technology

Konstantinos Kalpakis
University of Maryland Baltimore County
Department of Computer Science and Electrical Engineering
(kalpakis@cs.umbc.edu)

I was member of the Working Group on Digital Libraries and Electronic Commerce at the ACM Workshop on Strategic Directions in Computing Research held at MIT in July 1996. I coauthored a paper, based on the conclusions of that Working Group, entitled "Electronic Commerce and Digital Libraries: Towards a Digital Agora", which appeared in *ACM Computing Surveys*, Vol. 28, No. 4, pp. 818--835, December 1996.

I completed and submitted to *Algorithmica*, a technical paper coauthored with Yaacov Yesha on the problem of finding explicit tight upper and lower bounds on the makespan of schedules of tree dags on linear arrays, and the problem of polynomial time algorithms to find schedules that are optimal within a small constant. I refer the reader to the paper itself for the exact statements of my results since I find them too complex to be included here.

I completed a technical paper, coauthored with Bella Bellagradek (Ph.D. student) and Yelena Yesha, on "Strategies for Maximizing Seller's Profit under Unknown Buyer's Utility Values" which I submitted to the CASCON'97 Conference organized by IBM and the NRC, Canada. Suppose there is a seller that has an unlimited number of units of a single product for sale. The seller at each moment of time posts a price for his/her product. Based of the posted price, at each moment of time, a buyer decides whether or not to buy a unit of that product from the seller. The only information about the buyer to the seller is the seller's sales history. Further, I assume that the maximal unit price the buyer is willing to pay does not change over time. The question then is how should the seller price his/her product to maximize profits? To address this question, I use the notion of loss functions. Intuitively, a loss function is a measure, at each moment of time, of the lost opportunity to make a profit. In particular, I provided a polynomial–time algorithm that finds a pricing algorithm (strategy) for the seller that minimizes the cumulative (total) losses over time. Further, I presented preliminary results on pricing strategies that minimize the maximum possible loss at every moment of time. I also showed that there is no strategy minimizing both the total loss and the maximum loss at the same time.

George Durham, an M.S. student of mine, and I, developed a multi-level security model for object-oriented databases. A technical paper based on this work will be presented at the 20th National Information Systems Security Conference in October 1997. Our model is based on and extends the requirements of the Department of Defense 5200.28-STD, DOD Trusted Computer System Evaluation Criteria (TCSEC) dated December 1985, commonly known as the Orange Book. Currently, there exists no database model in any technology which meets the requirements of the Orange Book. There has been little interest outside of the U.S. Government and the academic community because the Orange Book is believed to focus on military needs rather than commercial interests. This is an unfortunate belief because, in fact, commercial espionage is growing daily, and without proper protection, commercial information will be pilfered both nationally and internationally. Previous work has focused on Discretionary Access Controls (DAC), Mandatory Access Controls (MAC), or other security requirements not included in the Orange Book, but no work includes all three. We developed policies for access controls, inference controls, and an implementation strategy based on the MAC, DAC, and other security requirements. The access authorization mechanism is based on a combination of DAC and MAC requirements, and the proposed model is easily extended to include other access requirements. We also described a system implementation.

Digital Research Technology

Aya Soffer
University of Maryland Baltimore County
Department of Computer Science and Electrical Engineering
(aya@cesdis.edu http://www.cs.umbc.edu/~soffer/)

Profile

Dr. Soffer received a B.S. degree in computer science from the Hebrew University of Jerusalem in 1986, and M.S. and Ph.D. degrees in computer science from the University of Maryland College Park in 1992 and 1995 respectively. She is currently a research assistant professor in the Department of Computer Science and Electrical Engineering at the University of Maryland Baltimore County, has an appointment as a research scientist at the Center for Automation Research (CFAR) at the University of Maryland College Park, and has an appointment as a research scientist at CESDIS through a subcontract with the University of Maryland Baltimore Country. Prior to coming to the U.S., Dr. Soffer worked as a software engineer for Elscint Ltd. in Haifa where she designed and developed a distributed database system for managing clinical information and images for the Nuclear Medicine Division and a software package for clinical evaluation of cardiac performance based on echocardiograms for the Ultrasound Division. She also developed educational software for use in Israel's public schools. Dr. Soffer's research interests include pictorial information systems, document analysis and recognition, digital libraries, spatial databases, geographic information systems, and non-traditional database systems.

Report

This year I have concentrated on four projects that are related to digital libraries (DL), geographic informations systems (GIS), and image databases.

Web-Based Geospatial Metadata Input and Extraction

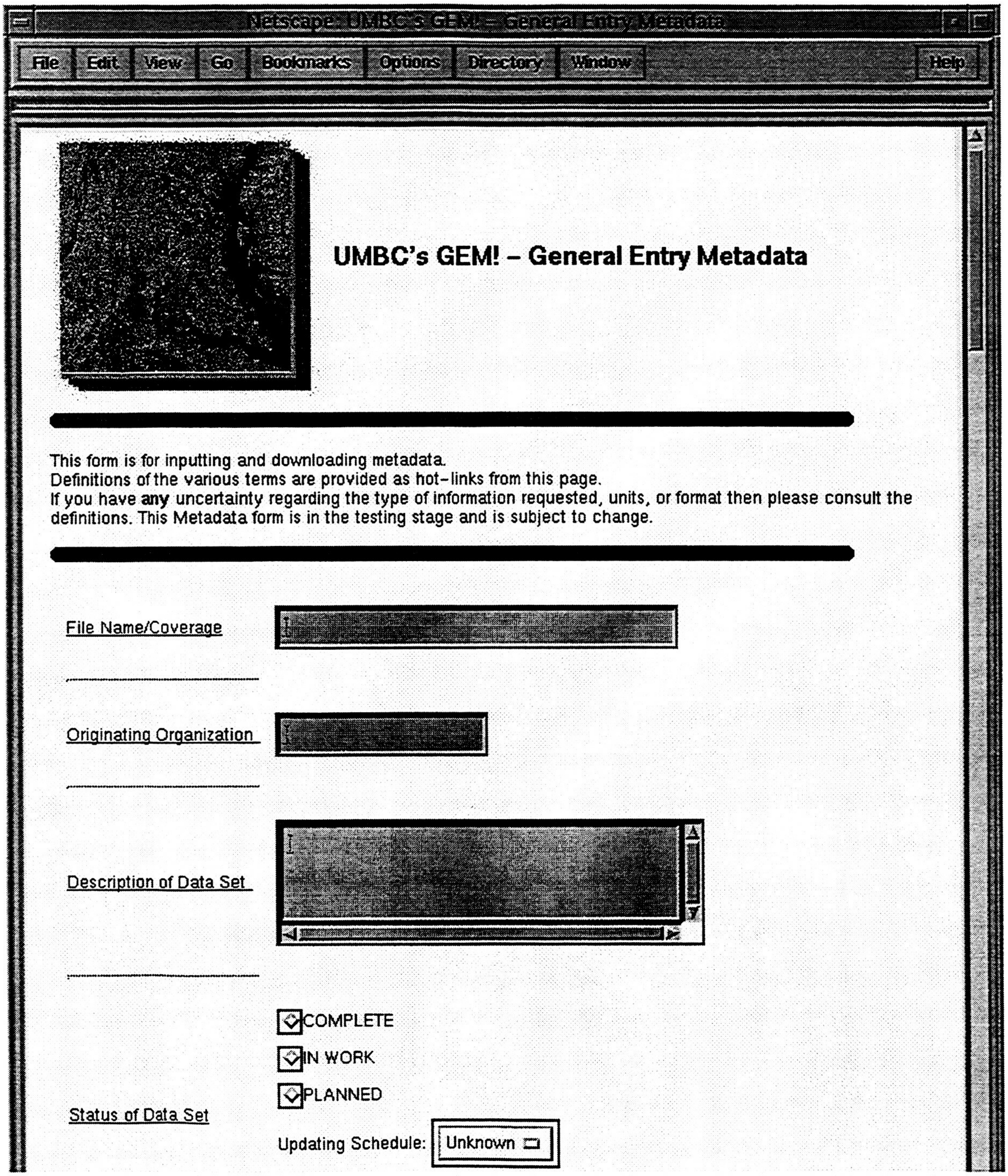
The University of Maryland at Baltimore County (UMBC) and Collaboratory partners are developing a three-tier approach to document metadata for geospatial data. As a multi-entity Baltimore Washington Regional Collaboratory using multiple data types, sources, and platforms, the task of documenting our data sets is onerous. Therefore, we are proposing an improved metadata standard and an efficient methodology to implement this standard. Levels I-III represent subsets of the Federal Geographic Data Committee's (FGDC) metadata standards. Level I represents the lowest common denominator fields for data browsers and general users. Level II is an extension of Level I and contains more detailed metadata fields that are institution specific. Level III conforms to the original FGDC metadata standards. We are in the process of interviewing federal, state, and local GIS users about the usability of this metadata standard and identifying case study participants for a more in-depth study of this subject.

In order to support the use of this metadata standard, we are developing a Web-based Level I metadata input tool. Members of the Collaboratory can use this tool to document their own datasets as well as to include their data in the Baltimore Washington Regional Validation Center (BWRVC). The tool is accessed via the World Wide Web and provides a form-based interface to the Level I metadata fields. After filling out this form, a metadata document is sent back to the user for his own documentation purposes. Furthermore, this data is immediately recorded in the BWRDC database.

We are also developing a Web-based metadata search and extraction tool that can be used by members of the Collaboratory as well as by the general public to access the BWRVC database. The search tool is tightly coupled with our metadata standard. Users can search for data based on a subset of the Level I metadata fields that have been identified as useful search criteria. The search engine is adaptable so it can be easily modified to accommodate changes in the metadata that may be mandated based on our case studies. Furthermore, the search engine is dynamic and reflects the current holdings of the repository and can also be utilized as a browser.

Maryland National Capitol Parks and Planning Commission (Montgomery County), Howard County, National Park Service, and Maryland Department of the Environment are currently testing the tools. The Maryland State Government Geographic Information Coordinating Committee is currently considering adopting our metadata standard and using these tools for input and extraction of metadata statewide.

I have played a major role in defining the metadata input tool, the database, and the search engine. Together with a student, Zhiguang Han, we have set up a database for storing Level I metadata. We are currently using POSTGRESS for this purpose. In addition, we have developed a Web-based interface that accesses this database. This has been implemented in Java. The input tool can be accessed at http://cgi.umbc.edu/~bwrdc/cgi-bin/test.pl (a username and password is required). The search engine can be accessed at http: www.cs.umbc.edu/~zhan/demo. Figure 1 shows the first page of the input tool. Figure 2 shows the Web-based interface of the search engine. We demonstrated this tool at the Forum on Research and Technology Advances in Digital Libraries (ADL97). A paper and demo have also been accepted to the Second IEEE Metadata Conference to be held September 1997.



UMBC's GEM! – General Entry Metadata

This form is for inputting and downloading metadata.
Definitions of the various terms are provided as hot-links from this page.
If you have **any** uncertainty regarding the type of information requested, units, or format then please consult the definitions. This Metadata form is in the testing stage and is subject to change.

File Name/Coverage

Originating Organization

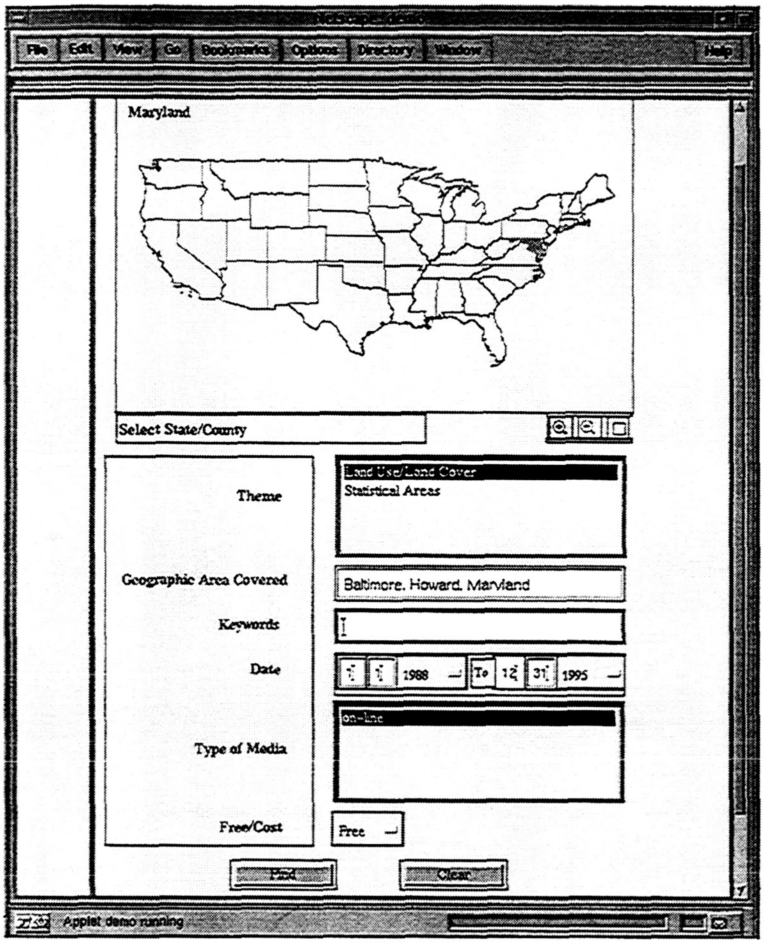
Description of Data Set

Status of Data Set

◇ COMPLETE
◇ IN WORK
◇ PLANNED

Updating Schedule: Unknown

Figure 1: Input Tool



Theme: Land Use/Land Cover (selected); Statistical Areas

Geographic Area Covered: Baltimore, Howard, Maryland

Keywords:

Date: 1 1 1988 To 12 31 1995

Type of Media: on-line

Free/Cost: Free

Figure 2: Search Engine Interface

Image Categorization

I have continued my work on image categorization. In particular I have improved and extended a method for image categorization based on NxM-grams. The goal is to find other images from the same category as the query image rather than the more general goal of finding all images that are "similar" to a given query image as is the case in most similarity-based image retrieval systems. Some example categories are handwritten documents, printed documents, satellite images, and fingerprints. With this goal in mind, it is easier to assess how well the system performs. While categorization may not be relevant for every image archive, we believe that for large image archives such as the World Wide Web, having the ability to find all satellite images, for example, is very useful. The method that we suggest is able to implicitly extract structural information from the images, and we hypothesize that images from similar categories have similar structures.

In this work we explore whether images can be categorized based on texture features. We hypothesize that images from similar categories have similar texture features. In particular, we present a method for categorizing images using a new texture feature termed NxM-gram. This method is based on the N-gram technique that is used for determining similarity of text documents. Our approach for categorizing images is based on extending the definition of N-grams to 2D. Intuitively an NxM-gram is a small patch or pattern in an image. The hypothesis that we examine is that two images that have the same recurring patterns are likely to belong to the same category. We defined the notion of NxM-grams and developed a procedure to compute an image profile in terms of its NxM-grams termed an NxM-gram vector. We proposed three similarity measures to compare images based on their NxM-gram vectors. We conducted an experiment with images from various categories. We compared the results using NxM-grams with these similarity measures to each other as well as to results of categorization using other well known texture features such as co-occurrence matrices, local standard deviation, Laws texture features, and to a method based on color distribution features.

Our results show that for our test images, NxM-gram-based methods were more successful in finding images from the same category as a given test image than other texture features or greyscale distribution-based methods. Figure 3 shows the first 16 ranked images for a query image from the musical notes category. The three parts (a), (b), and (c) report the results using 3x3-grams with histogram intersection similarity measure, local standard deviation texture feature, and greyscale distribution, respectively. The first image is the query image. The names of all images that are from the same category as the query image are displayed in reverse video. In this case, the 3x3-gram texture feature was best ranking the nine musical scores in our database perfectly, while local standard deviation found six, and greyscale distribution only found three.

NxM-grams work very well on simple documents such as floorplans, music notes, comics, and tables. The NxM-gram method is, however, weak on richer images such as aircraft, aerial photos, and road maps. As a first attempt to improve the results for such images, we have experimented with using three binarization methods. In the first case a global threshold operation was applied to the original image. In the second case, we first applied a gradient operator (sobel) to the original images and followed this with a global threshold operator. In the third case, we first computed the local standard deviation images and then applied the global threshold operator to these images. This modification did in fact improve the results for the more complex images. However, for simple documents such as musical notes and text, the accuracy

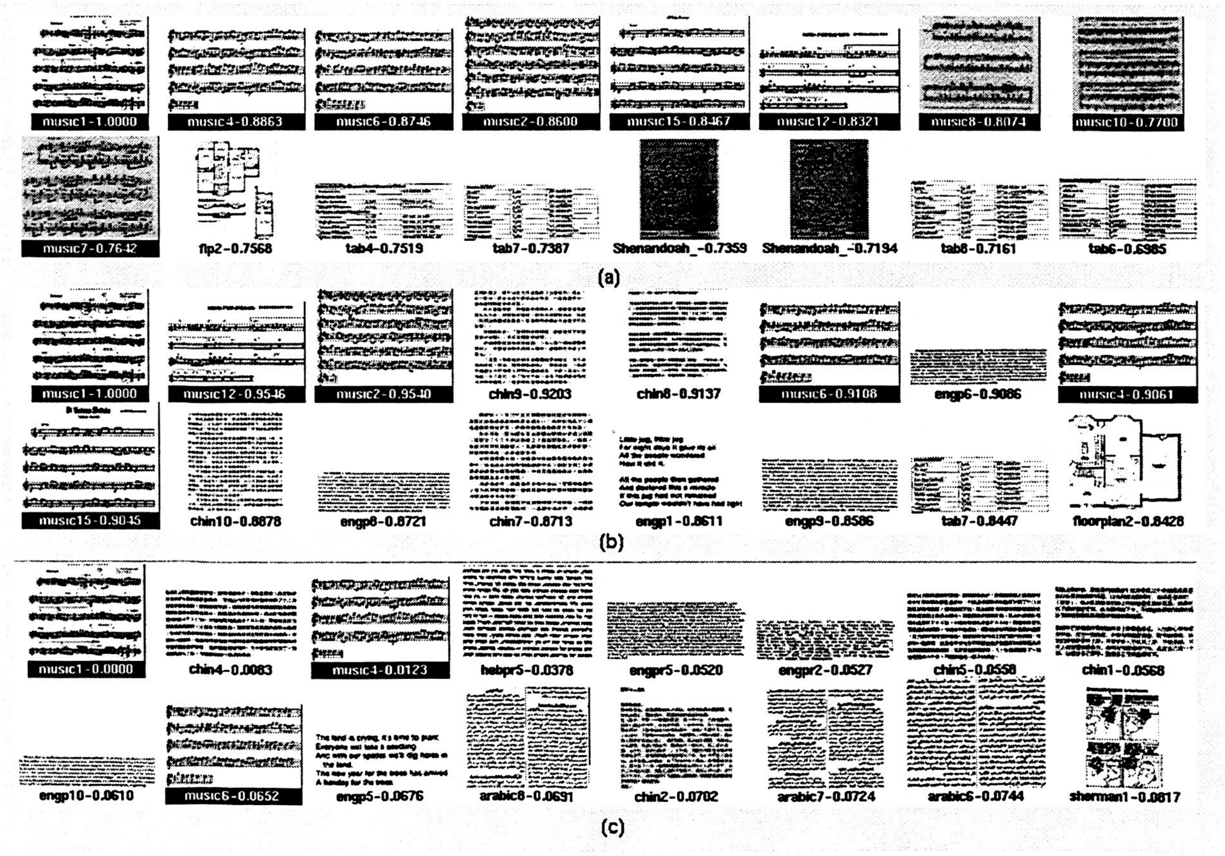


Figure 3: Ranked Images for Query Image

deteriorated. Thus, some hybrid approach is necessary. We are currently looking into using multiple window sizes by applying the same size window to various resolutions of the image using the wavelet transform. By computing NxM-grams on both the low pass and the high pass components of the wavelet decomposition of the images, we will, in effect, get both edge information and multi-resolution images.

Map Image Database

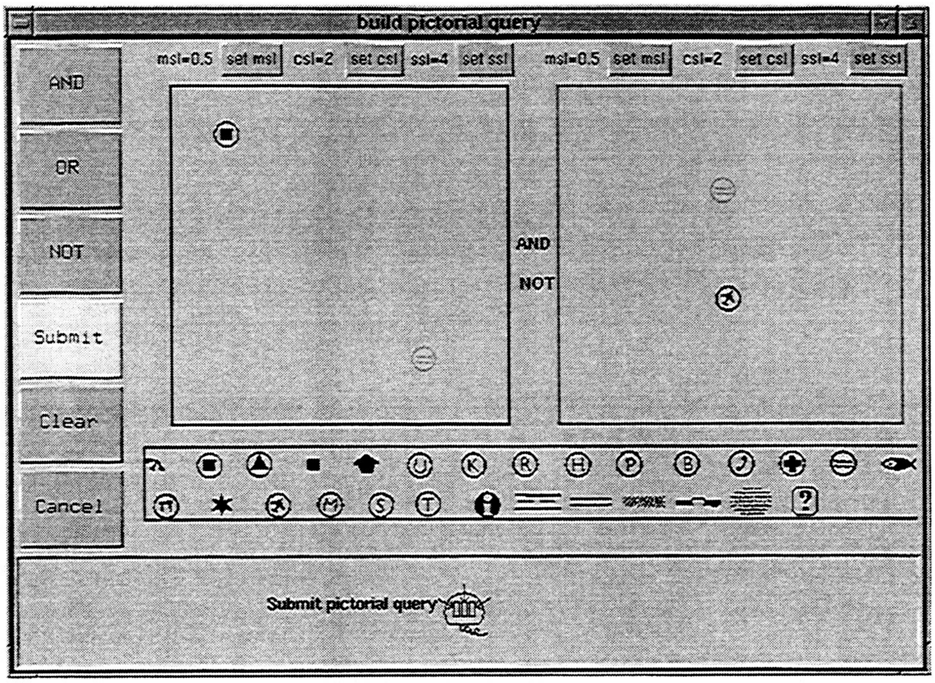
I have also continued my work on map (or symbolic) image databases. In particular, I have extended a method for querying a symbolic image database pictorially that I have developed.

The goal of this tool is to enable finding all of the images in the database that contain a number of specific objects in a particular spatial configuration with respect to each other. One method to achieve this is by an extension of a standard database query language (e.g., SQL) that provides additional predicates corresponding to spatial relationships. One problem with this method is that the objects in the images must be pre-classified so that the user can specify them by some alphanumeric tag. In addition, if we wish to find more complex images that are composed of several objects that must satisfy a particular spatial configuration, or if we wish to specify a choice among objects that satisfy some spatial configuration, then the corresponding SQL query would be very complex.

An alternative method, and the one we have chosen to investigate, is to specify the queries graphically. This is a more "natural" method that facilitates the use of more complex constraints based on the implicit characteristics of the graphical query (i.e., the particular objects in the graphical query and their spatial arrangement). There are, however, several difficulties associated with graphical query specifications. First of all, graphical queries are inherently ambiguous which gives rise to several questions. In particular, what criteria should be used in order to determine that an object in a database image is the same as a particular object in the query image (*matching ambiguity)*? In addition, when query images are composed of several objects, are we looking for images that contain all of these objects, or would we be satisfied with any subset of these objects (*contextual ambiguity)*? Finally, is the spatial arrangement of the query objects of significance? For example, if one object in the query image is placed above and within 30 units of another object, what database images satisfy this query? One possibility is that only database images with exactly the same spatial configuration satisfy the query. However, the intent may be that only the distance must be the same, or maybe that any configuration may suffice (*spatial ambiguity*).

Another difficulty with graphical queries is that they are not always as expressive as textual queries in terms of specifying combinations of conditions and negative conditions. For example, how do we specify graphically images that contain beaches but do not contain camping sites within three miles of these beaches? What we want is a graphical query specification method that leverages on the expressiveness of graphical queries in terms of describing what objects the target images should contain and their desired spatial configuration, while simultaneously resolving the matching, contextual, and spatial ambiguities as well as the limited expressiveness of graphical query specifications.

The graphical query specification technique that we have developed addresses the issues of matching, contextual, and spatial ambiguity in a comprehensive manner. This method enables the formulation of complex graphical queries that describe the target images in terms of their required contextual and spatial properties as well as the degree of matching that is needed. The desired objects can be specified as well as how many occurrences of each object are required in the target images. Moreover, spatial constraints can be imposed that specify bounds on the distance between objects, as well as the relative direction between objects. Figure 4 shows the graphical query builder. The user has constructed a query to retrieve all database images that contain a hotel within six miles of a beach and do not have an airport within one mile of the beach. Furthermore, the certainty that the database-image symbols are in fact a hotel, beach, and airport is ≥ 0.5. The symbols are "dragged and dropped" from the menu of symbols displayed in the bottom of the window. The query builder constructs this menu of symbols directly from the database which stores one example of each symbol relevant for the application at hand. These example symbols are taken from the legend of the map in our example application.



msl=0.5 set msl csl=2 set csl ssl=4 set ssl

msl=0.5 set msl csl=2 set csl ssl=4 set ssl

Figure 4: Graphical Query Builder

NASA Digital Library Initiative

As part of my involvement in the digital library task, I have assisted Dr. Nand Lal in tracking the progress of some of the projects that are sponsored under NASA's digital library initiative. In particular, I have been involved this year with the two of the projects (Bellcore and IBM). I received brief tutorials on how to run the servers for these two projects that are now installed on machines in the digital libraries studio and on how to operate the clients.

The digital library group organized and held an Information Technology Workshop on March 11-13, 1997. This was a 3-day workshop where digital library technologies researchers presented brief highlights of their activities to the Earth and space science community. In addition, technical discussions of various areas of interest were held among the information science researchers. I assisted Nand Lal and Susan Hoban in planning, organizing, and running the workshop. This included surveying the projects that were presented at the workshop in terms of the technologies that they are studying and suggesting topics for the technical discussions held on the second day of the workshop that would be of interest to the visiting scientists. I also prepared an html document with abstracts of the projects that were presented.

Publications

Samet, H.& Soffer, A. (Aug. 1996) MARCO: MAp Retrieval by COntent. IEEE transactions on Pattern Analysis and Machine Intelligence. *Special Issue on Digital libraries: Representation and Retrieval, 18(8)*:783-798.

Soffer, A. & Samet, H. (August 1996) Pictorial Queries by Image Similarity. *Proceedings of the 13th International Conference on Pattern Recognition, volume III*, pages 114-119, Vienna, Austria.

Soffer, A. & Samet, H. (August 1996) Handling Multiple Instances of Symbols in Pictorial Queries by Image Similarity. *Proceedings of the First International Workshop on Image Databases and Multi Media Search*, pages 51-58, Amsterdam, The Netherlands.

Soffer, A. (February, 1997). Image Categorization using NxM-grams. *Proceeding of the SPIE, Storage and Retrieval of Still Image and Video Databases V*, pages 121-132, San-Jose, California.

Soffer, A. & Samet, H. (May 1997) Negative Shape Features for Image Databases Consisting of Geographic Symbols. *The 3rd International Workshop on Visual Form*, Capri, Italy.

Soffer, A. (August 1997) Image Categorization using Texture Features. To appear: *Fourth International Conference on Document Analysis and Recognition*. Ulm, Germany.

GLOBAL LEGAL INFORMATION NETWORK

The Global Legal Information Network (GLIN) is an international, non-commercial, cooperative network of government agencies working in conjunction with the Law Library of the U. S. Library of Congress to create a database of international law documents which will be available to member countries throughout the world and which will facilitate international cooperation and joint ventures. The Library of Congress and NASA have signed a Memorandum of Understanding to establish a framework for coordinating cooperative efforts on updating and enhancing the technological infrastructure of GLIN. The intent is for the work to be conducted through collaborative and cooperative research by the Law Library, NASA GSFC, industry, academia, participating GLIN members, and other relevant international bodies.

In order to more efficiently collect and disseminate current legal information, a prototype system has been established to acquire, process, and retrieve digitized legal texts. The application of advanced digital technology is necessary to maintain the GLIN database and to increase the speed and flexibility of the system as the volume and complexity of the data expands. Upgrades and enhancements to GLIN are desired in order to share the benefits and burdens of obtaining, processing, and retrieving legal texts among cooperative partners throughout the world. Nabil Adam (Rutgers University), Tarek El-Ghazawi (George Washington University), and Konstantinos Kalpakis and Russell Turner (University of Maryland Baltimore County) have contributed to the CESDIS portion of this effort. Their reports follow.

Information Extraction Applications for GLIN

Dr. Nabil R. Adam
Rutgers University
Center for Information Management, Integration and Connectivity (CIMIC)
(adam@adam.rutgers.edu)

This report discusses the work undertaken at Rutgers CIMIC for the CESDIS and Library of Congress work on the GLobal Legal Information Network (GLIN). The specific work focuses on applying information extraction (IE), a form of natural language processing, to the problem of law summary classification and retrieval. For this work, we developed an incremental modeling methodology that reuses an existing hierarchical classification scheme followed by a systematic method for identifying common concepts within and between classes of documents. Concepts are words or phrases that appear in specific linguistic contexts (e.g., as specific parts of speech or sentence fragments) and are expressed by a set of semantic and syntactic constraints. We then train a series of IE "extractors" based on this model that are capable of identifying these sets of concepts. The concept identification process is used to determine if a novel document should be assigned to a class. Novel documents are passed down the hierarchy and are processed by extractors at each node. Successful extraction of key concepts leads to a class assignment. In.the case of GLIN, this results in index terms being assigned to the GLIN summary.

Information retrieval (IR) is achieved by gathering the instantiated concept definitions (sets of constraints) and using them to form an index of the document. This index can then be queried using a standard user interface. However, rather than return all of the documents that contain a specific query word, our IR system first returns a set of class/concept pairs that are indicative of the query terms supplied. The user may then filter the query further by choosing some or all of the class/concept pairs. This results in much smaller and more precise result sets.

In August 1996, we received access to a collection of approximately 50,000 GLIN summaries that were classified/indexed using terms from the GLIN thesaurus. Using the modeling methodology just described, we modeled a subset of the GLIN documents as a hierarchy of classes. The 18 classes modeled represent about 2% of the total index terms found in GLIN (18/700), while the documents associated with these index terms represent over 10% of the total documents found in our test collection. Within the classes, 32 unique concepts were found.

Based on this hierarchy, we then trained a series of IE systems to recognize these concepts. Following the standard IR practices, we used 80% of the documents in a class for training and tested using the remaining 20%. The IE software was licensed from the University of Massachusetts at Amherst and adapted for use on this project. The modeling, training, and software development effort took just over four months to complete. Much of this initial time was spent in developing the methodology and software required to automate much of the work. Using the incremental modeling methodology, a new class can be modeled and added to the overall system in approximately four person hours.

A classifying system was built using the hierarchy and trained extractors. A web interface was designed to allow a user to type or paste in a new GLIN summary and have that summary classified with up to 18 different GLIN thesaurus terms.

A Web interface to the GLIN classifier can be found at the URL: http://cimic.rutgers.edu/~holowcza/glin/ling/

Experimental Results

After creating the extractors, each one was run on a test set of summaries. 100 of the summaries came from within the sub-domain (relevant texts) and 100 were randomly chosen from outside of the sub-domain (irrelevant texts). Recall and Precision measures were recorded. The system favors recall with results ranging from 70 to 100%. Precision measured ranged from 63 to 100%. In most cases, a classic recall/precision trade-off was identified. Future work may focus on adjusting the training tolerance error values (a parameter to the training function) to try and improve recall and precision for some of the classes.

Information Retrieval Application

We also created an IR application as described previously. We classified 5,000 GLIN summaries using our classifier and then formed document indexes from the instantiated concept definitions (constraints). The index attributes include the sentence number, segment number (within the sentence), class, concept, phrase type (noun phrase, verb phrase, prepositional phrase), the actual words used, and a document id (pointer to the actual document). 5,000 documents created 12,129 records (note that more than one CN definition can apply to the same document). A WWW forms interface and several CGI scripts were also written for the IR application. The URL for the GLIN query application is: http://cimic.rutgers.edu/~holowcza/glin/ling/query.html.

Other Events

Our work has led to the following publications and presentations:

1) Holowczak, R. D. Extractors for Digital Library Objects. Ph.D. Dissertation, Rutgers University. May, 1997.

2) Holowczak, R. D. and Adam, N. R. Information Extraction-based Multiple-Category Document Classification for the Global Legal Information Network. *Proceedings of the Ninth Annual Conference on Innovative Applications of Artificial Intelligence* (IAAI-97). July, 1997. Providence, Rhode Island.

3) Holowczak, R. D. Extractors for Digital Library Objects. Presentation given to Columbia University Department of Computer Science. February, 1997.

The Global Legal Information System (GLIN)

Tarek El-Ghazawi
The George Washington University
Department of Electrical Engineering and Computer Science
(tarek@seas.gwu.edu)

In this work, I have contributed to the creation and refinement of the GLIN development plan for Phase I (system upgrade) and Phase II (System Enhancement). To this effect, I have worked with other team members and NASCOM identifying communications resources that can be allocated for demonstrations and experiments needed by the GLIN project. Further, I developed new versions of the communications tasks in Phase I and in Phase II based on discussions with the pertinent members of the group. The Phase I communications task, as a result, was to be restricted to demonstrations and experiments with GLIN country members. Mexico and Romania, and possibly Israel, were identified as some of the countries with which to start these demonstrations with due to their technical readiness. In phase 2, a thorough effort was to be conducted to assess the GLIN communications requirements and provide canned solutions (at the system level). Such solutions will take into account countries that have adequate connections to the Internet global backbone as well as those that do not have adequate national communication infrastructures. The power of satellite communication systems will be of particular interest in the latter situations and will be fully exploited to support GLIN in these regards.

I also conducted a survey of compression techniques which can be used in GLIN and which can affect the required communication bandwidth and storage requirements. The survey sought the most popular schemes for compressing text, still pictures, moving pictures, and sound. Evaluations were conducted qualitatively with the following identified criteria in mind: (1) whether the method is free (no patent or royalty for use), (2) high-level of compressibility, and (3) whether it allows progressive transmission. In addition, I participated in the GLIN Directors' meeting which was held at the Library of Congress in order to gauge the progress of the project in the member countries and exchange experiences and learned lessons.

Adam, N., Edelson, B., El-Ghazawi, T., Halem, M., Kalpakis, K., Kozura, N., Medina, R., & Yesha, Y. (December 1996) The Global Legal Information Network (GLIN). *American University Law Review,* Vol. 46, NO. 2.

Architectural Design for GLIN

Konstantinos Kalpakis
University of Maryland Baltimore County
Department of Computer Science and Electrical Engineering
(kalpakis@cs.umbc.edu)

GLIN Related Activities

CESDIS has been collaborating with NASA and the U. S. Library of Congress on the development of the Global Legal Information Network (GLIN) (http://glin.gsfc.nasa.gov). Since June 1996, I have been the technical leader of the GLIN project at CESDIS. Further, except for the summer months of 1996, in which I had 2 summer students (a high school student and a sophomore college student), I was the only programmer available to this project.

GLIN is an on-line repository of legal instruments, providing global access to the legal information of participating nations. Currently there are 22 nations involved in the project. The primary goal of the GLIN system is to provide efficient, flexible, and reliable access to authentic, accurate, and current legal information. The GLIN member nations have committed to providing the appropriate content to its legal digital library.

Efforts on GLIN are on two fronts/phases running in parallel. Phase 1, the upgrade phase, calls for upgrading the Law Library's prototype with additional functionality required by the Law Library's staff. Phase 1 has three tasks. Task 1 is the upgrading of the prototype, Task 2 is surveying infrastructure of member countries, and Task 3 analyzing the communications system. Phase 2, the enhancement phase, calls for designing and developing the next generation system.

I developed two prototypes for Phase 1. The first one was using the INQUERY text-retrieval engine available at the Law Library. This prototype was eventually abandoned, primarily because of various limitations imposed by using INQUERY. The second one is using a WAIS text-retrieval engine based on the vector-space model for text retrieval. This prototype is the current GLIN prototype available at the http://glin.gsfc.nasa.gov URL. This prototype was first released at the end of January of 1997. The current version of this prototype was released in March 1997.

I completed all the sub-tasks of Task 1 in Phase 1 that were requested except for handling multi-lingual legal instruments in their native character set/language. This subtask of Task 1, Phase 1, had been moved to Phase 2, since completing it would have required too many drastic changes to the existing system. Furthermore, I incorporated certain additional features in this prototype that are envisioned in Phase 2. The rationale is to demonstrate preliminary versions of them to the GLIN user community at an early stage in order to capture their requirements for delivering a successful system.

The current version of the GLIN prototype serves as the bridge between Phases 1 and 2. The current GLIN prototype consists of a database (Postgres) server, a WAIS server, a Web server, together with application software built using the functionality provided by the database and WAIS servers, and interacting with the users primarily via the Web server. Currently, legal documents are submitted to the data servers in SGML-format, and accessed either via SQL or Z39.50-type queries. Documents are indexed using the Legal Thesaurus developed by the U.S. Library of Congress, and their full-text summary (currently mostly English). In addition, digitized images of the legal documents are also stored in the system. At the same time, I have been working on the architecture of the GLIN system in Phase 2. The architecture for GLIN is based on the agent–oriented programming approach and is inspired by ARPA's reference architecture for the intelligent integration of information. I also collaborated with Dr. Susan Hoban on refining the GLIN Project Plan.

I demonstrated the use of ACTS communications capabilities for GLIN at the ADL'96 Conference. This demonstration was made possible by the assistance of technical staff members of Code 930, especially Mr. Pat Gary.

I will present my agent-based architecture for the next generation GLIN system at the 3rd Annual GLIN Directors meeting to be held at the Library of Congress in September 1997.

I also gave a tutorial during the February 1997 GLIN Training Session at the Library of Congress on CGI and Javascript scripts.

I have written the GLIN/Digital Libraries and Electronic Commerce attachments to the proposal submitted by SCDC (Code 930) and CESDIS to the US-Israel technology commission in the Fall of 1996.

I coauthored a paper with title "The Global Legal Information Network (GLIN)" which appeared in *The American University Law Review*, Vol. 46, No. 2, pp. 477–491, December 1996.

GLIN Project Report

Russell Turner
University of Maryland Baltimore County
Department of Computer Science and Electrical Engineering
(turner@cs.umbc.edu)

Goals

The purpose of this project was to explore possible graphical user-interface approaches for performing interactive Web-based searches of the Global Legal Information Network database in order to provide input, editing, and querying functionality for the Phase 1 GLIN tasks.

Currently, the GLIN database can be accessed via the World Wide Web using a standard form-based user-interface similar to what can be found on most Internet search engines or on-line databases. Unlike many of these databases, for which a simple key-word search is sufficient, the GLIN database can be searched using more complex queries based on several search criteria. For this reason, it was felt that a more sophisticated user-interface was needed which would provide a more natural and intuitive means for users to query the GLIN database.

Java-Based Approach

One of the most powerful features of the World Wide Web is the ability of users to query on-line databases directly from their Web browsers. Normally, this can be implemented using the forms mechanism of the HTML file format which provides a small set of standard graphical user-interface mechanisms such as text fields and selection boxes directly in the Web browser. These can be used to send query data directly back to the Web server using CGI scripts. From a user-interface design point of view, this technique is extremely limited since the interface designer is restricted to the small set of user-interface techniques provided by the HTML syntax.

To implement more sophisticated user interfaces that are accessible using a standard Web browser, the only practical alternative is to use Java. Java is an interpreted computer language that can be downloaded from a Web server and run directly in any Java-enabled browser. Since it is a general-purpose language with a graphics API, and not simply a file format, Java provides much more flexibility in implementing the visual appearance and interactive behavior of a Web-based graphical user-interface.

GLIN Database Search Criteria

Queries to the GLIN database can be specified using the following criteria:

- Geographical region, specified by a list of countries
- Index terms, specified by a list of words
- Publication or enactment dates, specified using a range of dates
- Search keywords
- Maximum number of returned items

While it is possible for all of these parameters to be entered by the user through a standard HTML form-based interface, there are a number of ways this can be made more intuitive and efficient through more advanced graphical user interface techniques. For example, geographical regions can be specified directly using maps, and dates can be selected with interactive calendars.

User-Interface Prototype

To demonstrate the feasibility of such a Java-based user interface for querying the GLIN database, we have constructed a prototype which can be viewed at the following URL: http://www.cs.umbc.edu/~turner/glin.

This prototype has the following features:

- World map used for selecting search by country.
- Pop-up menus for country selection.
- Specialized check-boxes for selecting search query constraints.
- General-purpose interactive calendar for selecting query date ranges.
- Specialized help icons for links to help information.

The world map provides a natural starting point for users selecting geographical regions to limit their search. The non-standard pop-up menus, which are implemented completely in Java, allow the user to select a series of countries which are then displayed in a search list prior to submitting the query. The time period of the search may be specified using the general purpose interactive calendars which display the current date by default and can be interactively set to display any desired month and year. Specific search restrictions may be specified using the non-standard check-boxes, and help pages may be accessed via the interactive" light bulb" icons.

Conclusion

While standard HTML form-based user interfaces for querying on-line databases provide a minimal capability, it is now possible to construct sophisticated graphical user interfaces using Java which are more natural and intuitive. The feasibility of such a user interface for accessing the GLIN database has been demonstrated by building a prototype demonstration Web page.

DIGITAL LIBRARIES CONSULTANTS

Nabil Adam
Rutgers University
Center for Information Management, Integration, and Connectivity (CIMIC)
(adam@adam.rutgers.edu)

- I served as the General Chair of the Forum on Research and Technology Advances in Digital Libraries (ADL,97) sponsored by the IEEE Computer Society that was held May 7-9, 1997 at the Library of Congress in Washington, D. C.

- I spent a good amount of time on the U. S.-Israel proposal.

- I worked on the following paper and presentations:

Adam, N., & Naqvi, S. (1996). Universal access in digital libraries. *ACM Computing Surveys*, 28(4), December.

International Conference on Digital Libraries and Information Services for the 21st Century (KOLISS DL,96), Seoul, Korea. September 1996.

Matsushita Information Technology Laboratory, Panasonic Technologies, Inc., Princeton, NJ. September 1996.

Hans Mark
University of Texas at Austin
John J. McKetta Centennial Energy Chair in Engineering
(betty_richardson@asemailgate.ae.utexas.edu)

I have been working with Dr. Milton Halem [Chief, NASA Goddard Earth and Space Data Computing Division (Code 930)] on the prospect of developing a three-dimensional optical memory that could be employed for high density information storage as well as very high speed readout and writing. There are two things which give the advantage to optical memories:

1. The fact that optical memories do not require connectors to anything since the reading and writing is done by light beams.
2. The fact that other memory devices, such as CD-ROMs and magnetic tapes, are two dimensional whereas optical memories can be three dimensional. Thus, higher information storage densities can be achieved.

When these factors are considered, three-dimensional optical storage devices can hold information more efficiently than conventional ones by about a factor of 50 when compared to the same volume of silicon. In addition, optical storage devices can be read and written on very rapidly because specially designed beams of light are used for that purpose. These can be moved very quickly using fast-moving mirrors and, of course, the information, which is carried by the phase relationship between two coherent beams of light, travels at the speed of light.

About 60 years ago, the Dutch physicist Frits Zernike invented the phase contrast microscope. In an ordinary microscope, contrast is achieved because the object being examined absorbs the light going through it strongly enough so that the object is visible to the observer. However, some objects are not absorbent enough even when stained with appropriate dyes and, therefore, no direct image results. Zernike recognized that even though the light going through the object might not have been absorbed, the object always had a different refractive index from the surrounding medium.

Zernike realized that he could use the phase difference induced by the region having a different refractive index to make the region visible through special optical means. He knew that a region with a different refractive index actually diffracted some of the light passing through it. He, therefore, put a phase-shifting filter in the path of the diffracted beam which changed its phase by 180° with respect to the beam passing directly through the sample. When the beams recombined in the microscope, there was destructive interference and the region with the different refractive index showed up by what Zernike named "phase contrast".

It is exactly the same physical principle (diffraction by a region with a different index of refraction) that is used to write and read the information contained in three-dimensional optical memory devices. In order to do this, there must be a "write" system that creates a region of different refractive index in a suitable material and then a "read" system somewhat like Zernike,s phase contrast microscope that retrieves the information. I will return to the optics after describing how a material can be created in which exposure to light can induce small regions with differing refractive indices.

The class of materials we are considering are called "photo-refractive". In such materials, exposure to light will create a region of charge separation that leads to localized internal electric fields in small regions of space. These regions will have a slightly different refractive index from their surroundings and it is therefore possible to produce a pattern of such regions from which light can be diffracted. There are several inorganic materials (crystals) which exhibit this photo-refractive effect, the most prominent being Lithium Niobate (LiNbO3). These crystals tend to be expensive and hard to make. Recently it has been discovered that certain blends of organic molecules embedded in a polymer matrix can also be made to exhibit photo-refractive properties. These have the great advantage of being both cheap and easy to produce.

The technology of polymer blends is one of the new and really exciting areas in polymer chemistry. One of the leaders in the field is Professor Donald R. Paul who is a chemical engineer at the University of Texas at Austin and who heads the Institute for Polymer Research. The essential discovery involving polymer blends is that within very broad limits, it is possible to engineer desired material properties by blending polymers with different properties. For example, if one polymer is hard and brittle and another is tough and flexible, through blending them a material can be created that is both hard and tough. What is interesting is that this works over an astonishingly large range of material properties and mixing ratios. Photo-refractive polymers are complex blends of organic materials that are specifically engineered to have the desired properties.

In order to produce an efficient photo-refractive polymer, the material must have four properties:

1. It must absorb light at the correct wave length that corresponds to the lasers that will be used to "write" and "read" the information.
2. It must be a photo conductor so that charge separation can occur in order to create the internal electric field which causes the refractive index difference.
3. It must be able to make this refractive index difference as large as possible.
4. It must be able to maintain the refractive index difference long enough to be useful as a high density data storage device.

The development of the desired material starts with a photo conducting polymer called poly N-vinyl carbazole (PVK). This material can be produced in thin films, and the final data storage device will consist of many layers of thin films of this polymer. PVK is about 33% by weight of the final blended material.

The second component is the photo absorbing material. The desired photo absorption is achieved by adding a small amount (1% by weight) of 2, 4, 7 trinitro-9-fluorenone (TNF). The combination of PVK and TNF creates the charge separation complexes that cause the internal electric fields.

In order to enhance the internal electric fields, a material is added that has molecules with a very large dipole moment. These can be aligned by applying an external electric field that polarizes the whole film and enhances the internal electric fields created by the PVK/TNF charge separation complexes. This material is called 2,5-dimethyl-4-(p-nitrophenylazo) anisole (DMNPAA), and it is present in the polymer blend at about 50% by weight. It is also very important that the "enhancing" material be transparent so that optical writing and reading methods can be used.

Finally, a material must be added that prevents the polymer from becoming a rigid glass-like material in which it would not be possible to align the DMNPAA molecules to enhance the refractive index differences. This is done by adding N-ethylcarbazole (ECZ) to the blend in the amount of 16% by weight.

The polymer-blend films created in this manner have the desired photo-refractive properties. In order to build a memory device, a great many of these polymer films are stacked together separated by inactive layers. The optical system uses laser diodes as the light source, and there is a lens system that focuses the light in the proper photo-refractive polymer film. The film stack can be thought of as a book with each photo-refractive film being a page in the book. The information is carried by modulating the intensity of one of two coherent light beams created by using a beam splitting optical system. The two coherent beams are then focused on one of the "pages" in the photo-refractive stack where they form an interference pattern.

As a result, there will be regions where the light is more intense (constructive interference) and those where the light is only at the intensity level of the reference (unmodulated) beam. Regions of the "page" where the light intensity is high in which many charge separations have been produced, will have an intense internal electric field. In these regions, the refraction induced by the two interfering beams will be much higher than in the surrounding regions where the light intensity is lower. It can be seen that the smallest region of high refractive index difference that can be produced by this method will have linear dimensions of the order of the wave length of the light used (about 4 micros) to "write" the "page". The high electric field (or refractive index difference) is produced by the collective action of a great many PVK/TNF/DMNPAA/ECZ molecular complexes that are located in the region.

The "page" in the photo-refractive memory stack can be read out by illuminating it with an unmodulated light beam having the same wave length and also the same propagation direction as the original modulated beam. The pattern of refractive index differences on the "page" will cause diffraction of the "read" beam by the same mechanism as in Frits Zernike,s phase contrast microscope. The diffraction pattern created in this way is essentially the Fourier Transform of the original message. This transform can be inverted by the appropriate optical system and thus the information originally "written" on the "page" is retrieved.

As a memory, the photo-refractive system described has a principle drawback in that the regions of high refractive index have a finite lifetime. Even with an applied electric field, the charges separated by the incident light eventually want to recombine. At this writing, storage times of up to an hour have been achieved by Professor Ray Chen (University of Maryland Baltimore County) using photo-refractive polymers of the kind I have described. Memories with such short time constants might be useful for some applications, but it would clearly be advantageous to somehow stabilize the images on the "pages" so that information storage for much longer periods would be possible. There is another drawback that is particularly important for applications in which the memory is used in a spacecraft. The photo-refractive polymers are radiation sensitive. Thus the information stored would be erased quickly unless heavy shielding were carried along.

I have no doubt that these devices have great potential. Much research must be done, however, before these memories can be applied in practice.

EXECUTIVE SECRETARIAT TO THE DATA AND INFORMATION MANAGEMENT WORKING GROUP OF THE U.S. GLOBAL CHANGE RESEARCH PROGRAM

The Data and Information Management Working Group (DIMWG) acts as the data management arm of the U.S. Global Change Research Program (USGCRP) and provides an informal mechanism for interagency coordination and cooperation. Working Group agencies are the Department of Commerce, the Department of Defense, the Department of Energy, the Department of the Interior, the Environmental Protection Agency, NASA, the National Science Foundation, and the U.S. Department of Agriculture. The Department of State and the National Academy of Sciences serve as liaison members. The Data and Information Management Working Group has six subgroups and more than 50 active participants. The DIMWG supports collaboration between computer and Earth scientists involved in database, data management, and data distribution research by facilitating access to global change-related data and information in useful forms.

This task was assigned to CESDIS through the Global Change Data Center (GCDC) in the NASA Goddard Earth Sciences Directorate (Code 900). It requires the provision of Executive Secretariat support to the Data and Information Management Working Group including the guidance and coordination necessary to ensure future accomplishments which can be endorsed by the National Academy of Sciences and which enhance the level of general cooperation and participation of the DIMWG agencies. Les Meredith and is responsible for providing the support required by this task.

Les Meredith, Senior Scientist
(les@usra.edu)

Profile

Dr. Meredith holds Bachelors, Masters, and Ph.D. degrees from the State University of Iowa. He is a Fellow of the American Association for the Advancement of Science, a Fellow of the Royal Astronomical Society, and a member of the American Geophysical Union, the American Physical Society, Phi Beta Kappa, and Sigma Xi.

Dr. Meredith's contributions to space science span more than 40 years and include employment as Head of Rocket Sonde Branch and Meteor and Aurora Section of the Naval Research Laboratory and a variety of positions at NASA Goddard Space Flight Center including Space Science Division Chief, Deputy Director of Space and Earth Sciences, Assistant Director, Acting Director, Director of Applications, and Associate Director. He spent a year as Liaison Scientist for Space Science in Europe with the Office of Naval Research in London, four years as the General Secretary of the American Geophysical Union, and more than five years as its Group Director.

Dr. Meredith is the recipient of the NASA Exceptional Scientific Achievement Medal (1965), the NASA Outstanding Leadership Medal (1975), the Senior Executive Service Presidential Meritorious Award (1981), and the NASA Distinguished Service Medal (1987).

Report

1. Organized, briefed the chair, wrote the minutes, and followed up on the action items of CENR's SGCR and TFODM Data Management Working Group, DMWG, meetings. These meetings were held about monthly. Between meetings, worked with DMWG members on special issues, responded to requests

for help, and performed the multiple actions needed to keep the DMWG interagency coordination process productive.

2. In my role as Program Associate for Data Management to the USGCRP, I fully participated in all their planning meetings and responded to action items and questions on about a weekly basis. I also drafted the data management section of the USGCRP's "Our Changing Planet - FY1998".

3. Drafted recommended DMWG positions on the proposed WIPO database treaty. They were subsequently adopted and provided the first real alert to the agencies of the potential negative impacts of the treaty if signed. The issues raised subsequently went to the President's Science Advisor and the U. S. position on the treaty has been reversed.

4. Was invited by the Deputy Director of EPA's Office of Research and Development to be part of a four-person group from outside the agency to review their plans for establishing a data management program spanning their organization. The seven recommendations I gave formed the basis for the review group's recommendations.

5. Was invited to participate in the April 1997 three-day meeting of a special EOSDIS Review Group that Harriss said would be the most important review EOSDIS has had in terms of making changes. I made twelve specific recommendation to the EOSDIS Project Scientist which he subsequently further distributed.

6. Proposed that a specific definition of "full and open" data access be adopted by the U. S. Subsequently drafted a letter requesting this action that was forwarded by the DMWG basically unchanged for CENR approval.

7. Was a member of the White House-sponsored National Environmental Monitoring and Research (NEMR) Workshop and subsequently drafted a proposal for the role the DMWG should play in this program. This proposal was endorsed by the DMWG and I was a member of its OSTP briefing team. OSTP's response is pending their NEMR definition.

8. Made an evaluation of the DMWG's GCDIS home page that's serving as the primary basis for its present restructuring.

EXECUTIVE SECRETARIAT TO THE COMMITTEE ON ENVIRONMENTAL AND NATURAL RESOURCES (CENR) TASK FORCE ON OBSERVATIONS AND DATA

The function of the Secretariat is to act on behalf of the CENR Task Force as the primary CENR interface for international consultations on scientific planning and implementation of the Global Observing System and its related data management system. This includes coordination with the international efforts underway be the Global Terrestrial Observing System (GTOS), the Global Climate Observing System (GCOS), the Global Ocean Observing System (GOOS), the Committee on Earth Observation Satellites (CEOS), the World Climate Research Programme (WCRP), and the International Geosphere-Biosphere Programme (IGBP).

This task was assigned to CESDIS through the Global Change Data Center (GCDC) in the NASA Goddard Earth Sciences Directorate (Code 900). It requires the provision of all the necessary technical and administrative support to assist the CENR Executive Director in implementing the responsibilities of the Secretariat. This includes coordinating the activities of the Task Force and its working groups, planning and coordinating U.S. participation in the International Global Observing System in accordance with the strategy outlined in the OSTP concept paper on the GOS, coordinating relevant observations and data management budget justification and advocacy material among the CENR subcommittees for submission to the Task Force, and coordinating with the Task Force's Data Management Working Group to promote effective access data management systems for CENR relevant global, regional, state, and local environmental and natural resources data.

Sushel Unninayar is responsible for providing the support required by this task. He works with CESDIS through a subcontract with the University of Maryland Baltimore County.

Sushel Unninayar
University of Maryland Baltimore County
Department of Computer Science and Electrical Engineering
(sushel@cesdis.usra.edu)

Profile

Dr. Unninayar holds a B. Tech. degree from the A.M.I.E.E. in London, an M.S. in electrical engineering from the University of Hawaii, and a Ph.D. in meteorology from the University of Hawaii Institute of Geophysics. His career to date has involved scientific planning and management of national and international programs dealing with climate change, global change, observing systems and data management, impact assessments, sustainable development, and environmental issues.

Specific positions have included: Director of Research, GLOBE-UCAR in Boulder, Colorado where he designed and developed the framework plan for GLOBE; Director of International Projects for the United Nations Institute for Training and Research in Geneva, Switzerland; Senior Scientist/manager at NASA HQ where he developed plans for the Greenhouse Effect Detection Experiment and its strategic implementation; Physical Science Administrator at the National Science Foundation; Senior Scientific Advisor to the United Nations Environment Programme GRID Center in Geneva; and Division Head of the World Climate Programme, World Meteorological Organization also in Geneva.

Dr. Unninayar is a member of the American Meteorological Society, the American Geophysical Union, the American Association for the Advancement of Science, the Institute of Electrical and Electronics Engineering (London), and the Royal Geographic Society (London). His research interests include the atmosphere, oceans, biospheres and Earth sciences, Earth system modeling and prediction, greenhouse gases and the detection of climate change, integrated impact assessments, global observing systems (surface- and space- based), environmental monitoring, data and information analysis, synthesis, and integration, data management and data exchange systems, instrumentation and engineering design, science education, and public, national, and international environmental policy.

Report

Activities in 1997 included the further development of an Integrated Global Observing Systems Strategy (IGOS) under the auspices of the CENR/TFODM, and scientific input to several programs and projects of the National Academy of Sciences related to atmosphere, climate system, and Earth science thematic areas.

The initial part of 1996 was spent in planning the first international conference on global in-situ observations. The conference/workshop was held in September 1996, in Geneva, Switzerland and was co-sponsored by the World Meteorological Organization (WMO), the United Nations Environment Programme (UNEP), the World Food and Agricultural Organization, the International Oceanographic Commission (IOC), and the International Council of Scientific Unions (ICSU) among other agencies. This was the first conference held to explore the needs for comprehensively monitoring all key components of the Earth system in an integrated manner. The major international observing programs included the Global Climate Observing System (GCOS), the Global Ocean Observing System (GOOS), and the Global Terrestrial Observing System (GTOS) spearheaded by WMO, IOC and FAO (and UNEP) respectively. The primary emphasis was on is-situ and surface-based observation systems for variables and parameters which could not be monitored otherwise (e.g., by remote sensing) and/or were needed for the calibration and validation of space-based measurements. I prepared for this meeting the first consolidated review/summary of scientific requirements. The draft review was widely distributed and discussed at the meeting.

The summary of observing system requirements evolved over the next several months into a Compendium of Requirements and Systems covering In-Situ Observations for Global Observation Systems. The compendium included a scientific rationale for each observational variable/parameter, requirements in terms of space/time resolutions and accuracy, as well as an assessment of the existing state of observing networks and data exchange systems. Cross references were provided to satellite observations. Deficiencies in existing systems were highlighted as a backdrop to recommendations for improvement. The compendium (286 pages), co-authored by Robert Schiffer (NASA HQ YSM), is available in hard copy (from CESDIS) and on Internet (under NASA's Mission to Planet Earth Program). Prior to finalization, the draft version was reviewed extensively by both national and international scientists. We intend to maintain the report as a "live" document which will be revised as scientific understanding improves and as observing and modeling technology (and requirements) change.

In addition to the above, I was involved in several activities of the Task Force on Observations and Data Management (TFODM), and in particular the User Needs Working Group. In this context work was continued on the identification of a key or "core" set of monitoring variables and parameters after reviewing needs arising from the various programs under the auspices of the OSTP Committee on Environment and Natural Resources (CENR). CENR subcommittees and working groups include those covering the Global Change Research Program (GCRP), Ecosystems and Biodiversity, Natural Disasters, and Environmental Monitoring and Research among others. All these working groups are coordinated interagency activities. The TFODM is considered a cross-cutting group to look into the requirements for observations stemming from the broad cross-section of disciplines represented by the CENR. From an international or global perspective, the climate system was chosen as an initial focus for the development of a global observations strategy or "system." The above activity included substantial interagency coordination as well as linking with various scientific groups and panels of the National Research Council.

Other activities included providing scientific advice to the National Academy of Sciences Panel on the Global Ocean-Atmosphere-Land Systems (GOALS) program – in particular the strategy for the U. S. participation in the international GOALS program, considered to be one of the major component programs of CLIVAR (Climate Variability) program. CLIVAR looks at both seasonal-to-interannual climate prediction (the objective of GOALS) as well as Decadel-to Centennial (DecCen) changes. I participated in several discussions on the subject at NAS and contributed to the development of the strategic plan with the chairman of the panel and academy staff.

Besides various coordination and interagency working group meetings, I attended the following conferences/workshops: (1) The 21st Climate Diagnostic and Prediction Workshop, Huntsville, Alabama, October 1996; (2) The International CLIVAR Scientific Steering Group meeting, April 1997, Washington DC.

Pilot EOS Direct Readout Ground Systems Support

Fran Stetina
Fran Stetina and Associates
(stetina@gsti.com)

Objective

This task provides technical support to develop the system design and implementation plans for NASA/GSFC code 935 (Applied Information Science Branch) Regional Validation Centers to become pilot EOS Direct Readout Ground Receiving Stations and for these Centers to become regional MTPE product validation centers.

Background

To effectively conduct research into global change problems and issues, it is necessary to solicit the cooperation of the broadest user community. To meet this long term outreach objective, Code 935 has developed the concept of Regional Validation Centers. This concept has been accepted as an effective approach to support the long term objectives of NASA's Mission to Planet Earth and to effectively transfer NASA's information technology to the broadest user community and to solicit the help of a broader community to both use and evaluate MTPE data products.

A number of these regional centers are being implemented as prototype centers to test the effectiveness of new information system technologies under real operating conditions. Emphasis will be concentrated on two such centers:

> In Hawaii a consortium of state, government, university, and private sector organizations are developing a concept called the Pacific Disaster Center(PDC). A regional validation center has been colocated with the PDC. The RVC is expected to provide valuable information regarding natural hazards. Efforts undertaken in support of this activity are to define the relationship between the RVC and PDC.
>
> In Lafayette, Louisiana, at the University of Southwestern Louisiana, a regional validation center has been established to concentrate on providing value-added weather products to the oil and gas industry. Emphasis will be on fusing all available weather products and providing new value-added products to minimize the impact of severe storms on the operations of the oil and gas industry and to determine the impacts of severe weather on the fragile coastal wetland areas.

These centers would contribute to the efficient and effective utilization of human and natural resources and the development of an information infrastructure to support knowledgeable decision making. Such an infrastructure must not only gather and store data, but it must contain sufficient processing power and intelligence to produce useful output products. The system must facilitate rapid retrieval and distribution of information so that decision making can be made based on objective criteria using expert knowledge and simple visualization techniques. This philosophy requires a systems design approach which emphasizes integration, automation, user friendly interfaces, and thorough understanding of the user's requirements.

Implementation of systems with these features is based on 10 years of project management experience for NASA/GSFC in implementation of satellite weather receiving systems, ground processing; specifically it includes the development of a modular system concept called SAMS, Spatial Analysis & Modeling System. The SAMS system has been defined as a potential model for the development of the MTPE Regional Validation and Calibration Center.

One of the key components of such a system is a real-time direct readout capability. Thus, the design and development of the Regional Environmental and Technology Center concept (Regional Validation Center), has been defined as an important objective of NASA's Mission to Planet Earth.

As Co-Pi for the development of SAMS, I am uniquely qualified to apply this information to the design and development of the Regional Validation Center Prototype System.

Scope

The activities to be undertaken under this task include hardware and software system design which are required to develop a Prototype Regional Validation Center to support MTPE. Included in this concept is the need to develop a core EOS Direct Readout capability and general support of end-to-end system software to provide EOS core instrument algorithms and basic mission products. The system concept should include an archiving and distribution capability. In addition, strategies should be developed to test EOS Direct Readout System components and concepts in an operational environment. This includes the use of aircraft high spatial and spectral resolution instruments to support algorithm development and evaluation, integration of insitu measurements to validate remote sensing measurements, the integration of a Geographic Information System. In addition, the system should include the design of a local user analysis system to interface with the Regional Validation Center.

Task Elements

Provide expert advise to determine user requirements for EOS Direct Readout core instrument algorithms and products.

- Develop project plans to utilize hyperspectral instruments to facilitate the development of Regional EOS MODIS algorithms.

- Determine weather product requirements for various applications which will be implemented at Regional Validation Centers for both operational and research users.

- Assist in defining MTPE core algorithm processing capability for a direct readout facility.

- Define the relationship and operating scenarios between the Pacific Disaster Center and the NASA Hawaii Regional Validation Center.

- Provide expert advice in defining EOS Direct Readout system concepts, define end-to-end system components and functions. Utilize SAMS concept to determine the requirements of Regional Validation Center.
- Develop operational scenarios for Direct Readout System and its interfaces with the GSFC EOSDIS.

- Provide expert advice in the development of strategies to develop and test various components of the Regional Validation Center System using existing operational facilities at University of Southern Louisiana and University of Hawaii.

- Assist in the development of an implementation plan for the use of unmanned aerial vehicles to support MTPE regional algorithm development and MTPE product validation.

- Represent the NASA/GSFC Regional Validation Center manager in meetings and conferences as required.

The Earth Alert personal warning system has been defined as a potential important technology which has significant value to the Hawaii Pacific Disaster Center. A number of activities relating to bringing this technology to a successful commercial product line and introduction of this capability to the Hawaii Civil Defense and to FEMA have also been undertaken as part of this task.

RESEARCH IN SATELLITE-FIBER NETWORK INTEROPERABILITY

This task requires CESDIS to perform research with high data rate satellite communications to effect its seamless integration with terrestrial fiber optic-based networks. Researchers are to collaborate with Goddard scientists to produce a specific proposal involving the hybrid use of satellite communications and fiber optic networks to advance a jointly defined set of NASA-related projects. The research will be performed through a subcontract with the George Washington University personnel listed below.

TEST PLAN FOR ACTS SPACE SCIENCE EXPERIMENTS

Burt Edelson, Neil Helm
George Washington University
Institute for Applied Space Research
(edelson@seas.gwu.edu, helm@seas.gwu.edu)

Introduction

The Advanced Communications Technology Satellite (ACTS) represents a large investment by NASA to design, develop and demonstrate new satellite technologies to support and promote the nation's civil, commercial and military satellite communications programs.

To test and demonstrate high data rate satellite communications, NASA and ARPA developed ground terminal equipment that can operate at SONET data rates up to 622 Mb/s (OC-12). A series of high data rate experiments have been approved. These experiments will characterize the use of the ACTS satellite with new digital protocols, algorithms, and architectures and will demonstrate space science applications that require the high data rate of supercomputer networks.

George Washington University has prepared the following test plan to evaluate the performance of high data rate transmission links using the ACTS satellite, and to provide a preparatory test framework for two of the space science applications that have been approved for tests and demonstrations as part of the overall ACTS program. This test plan will provide guidance and information necessary to find the optimal values of the transmission parameters and then apply these parameters to specific applications. The test plan is comprised of four parts. The first part will focus on the satellite-to-Earth link. The second part is a set of tests to study the performance of ATM on the ACTS channel. The third and fourth parts of the test plan will cover the space science applications, Global Climate Modeling and Keck Telescope Acquisition Modeling and Control.

The ACTS high bit rate terminals were delivered to the Jet Propulsion Laboratory, the Goddard Space Flight Center, and the State of Hawaii in October 1995. This test plan will begin by establishing the initial pointing accuracy of the antennas and continue through the demonstration of the applications. The test plan is prepared as a roadmap and introduces an evolutionary approach to thoroughly test and demonstrate the merits of high data rate networking via the ACTS satellite. Finally, the test plan provides an easy framework for evaluating the results of the experiments and assists in preparing technical papers and final reports that will validate the level of accomplishment of the experiments.

Part I. Satellite–to–Earth Link Tests

1. Introduction

The ACTS high data rate Earth stations operates at 30/20 GHz (Ka–Band). It is known that Ka–Band frequencies are subject to disruptions of service due to precipitation attenuation. Further, the beamwidths of the antenna for both transmit and receive functions are quite small, with the transmit 1/2–power beamwidth being only 0.3 degrees. That means that a pointing error of 0.15 degrees from the correct direction will reduce the power transmitted by a factor of two. Thus, antenna pointing must be done quite accurately.

The ACTS satellite moves with a period of 24 hours within a small 'box' in the pointing space, azimuth, and elevation. It is necessary, therefore, to optimize the pointing of the antenna when the satellite is in the center of its box.

To optimize the transmission parameters, it is prudent to make some measurements on the satellite link as part of the high data rate digital experiments. These measurements also will include effects of rain, so that we can obtain some operational familiarity with the service reliability of the link rather than simply measure propagation using the beacon frequency.

2. Pointing Accuracy

To obtain an estimate of the pointing accuracy, it will be necessary to measure the beacon power received over several 24–hour periods. Unfortunately, the antenna does not pass the 30 GHz beacon frequency to the receive side on the communications equipment. With the ability to obtain the 20 GHz beacon frequency, we intend to plot received power vs. time. From this we can extract the pointing accuracy, both in angle and in time. If it is not possible to view the beacon, then we will ask to have a signal put up through the satellite to serve as a beacon. The latter is somewhat less desirable, since the transmitted signal includes the effects of the uplink. But, since the satellite is hard limiting on the uplink, these effects are minimized.

From the orbital elements for the satellite, it is possible to plot the daily motion of the satellite within its nominal box as a function of time. Using that information and the observed beacon power, we should be able to determine the actual pointing angles of the Earth station antenna and, if necessary, make required corrections in the pointing angles.

Figure 1, shows the satellite motion for a 24–hour period as a change in azimuth and elevation for the GSFC position of the Earth station. To obtain a current plot of figure 1, we require the orbital elements of the ACTS satellite. This information is available from the Lockheed Martin Astro Space satellite orbit manager, Mr. Kent Mitchell. The satellite is being kept within a very tight box, about +/– 0.05 degrees in latitude and longitude, centered on the nominal satellite position of 100 degrees west longitude.

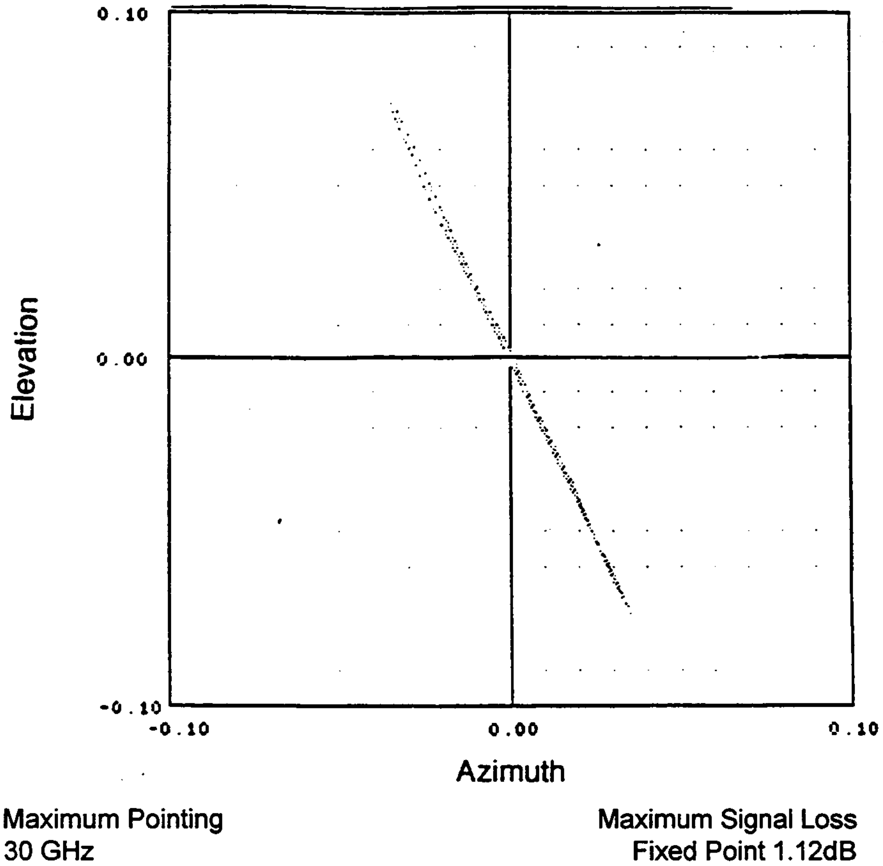


Figure 1: Satellite Position - Earth Station Pointing

3. Propagation Effects

Propagation effects at 30/20 GHz can be fairly severe. Heavy rain can cause path losses of ten dB or more. Studies have been made by others, e.g., COMSAT, of the propagation effects by measuring the power received from the satellite beacons and correlating these measurements with observed rainfall, etc.

We do not intend to duplicate these measurements, but it will be instructive to determine the effects of rain on the actual performance of the link. For that purpose we need to install a simple rain gauge near the antenna and make BER measurements during periods of rain. In addition to that, we must make some visual observations of the actual weather conditions. Severe rain tends to be fairly local, e.g., a thundershower, but to cause rain fading, the rainfall must be within the antenna beam, and it can often be observed visually and/or by making intelligent estimates. Since we do not have access to the link continuously, we need to keep a log of observed weather conditions and BER values during our allocation of satellite availability.

The results obtained should yield an estimate of the link in–service reliability. The link reliability obtained from the actual operational experience will then be compared to the predicted link reliability obtained from link budget calculations.

Since the link operation is affected by both the up and down links, it will be necessary to know the prevailing weather conditions at the other end of the link. Thus, a similar station log needs to be maintained at the other end. In view of the limiting conditions of the satellite transponder on the up link, mild degradations on the transmitting end should not show up on the receiving end. Therefore, we need not obtain very accurate rainfall information from the transmit station.

4. Link Calculation Verification

4.1 Theoretical Link Performance

We will make theoretical link performance calculations based on the performance data for the Earth station and the satellite. We will try to include any known effects or degradations that pertain to the particular Earth station.

4.2 Observed Link Performance

We will use a spectrum analyzer at the 3 GHz intermediate frequency to make Signal–to–Noise measurements. We will at the same time measure the error performance of the modem. The two measurements should show a fairly good correlation. We will make an analysis of the results to verify the operational parameters of the system.

5. Analysis of Performance Results

The areas listed below will be noted in the operation log books and then analyzed in more detail. These analyses will then be described in the final report on the transmission characteristics associated with the ACTS high data rate space science experiments.

Weather
Pointing
Antenna Performance (Tx, Rx, other)
Up Link
Down Link
Performance at GSFC vs. Performance at LeRC, JPL, and Hawaii

The initial results obtained at the GSFC site will be compared with the transmission parameters obtained at NASA Lewis Research Center (LeRC) during its experiment with the Boeing Corporation. This comparison will be made as soon as possible to determine that the re–installation of the Earth terminal at GSFC has not changed the overall Earth station performance measured when the HDR terminal was located at LeRC.

6. Conclusions

In addition to a comprehensive final report on Part 1, the Satellite– to–Earth Link Tests, an initial report will be prepared on the effects of moving the terminal from LeRC, if any, as well as its performance at the GSFC site. This initial report will assist the space science team in the conduct of their experiments.

Part 2. ATM over an ACTS Channel

1. Introduction

The Part 2 tests will characterize the performance of the basic protocols which are used in the Global Climate Modeling applications [11] and in the Keck Acquisition Visualization and Control over ACTS [13] experiments. Previous work has been done to characterize the application level protocol, namely PVM, performance for the Global Climate Modeling application [3] on a high speed satellite link. Additional work has been done over a low speed (T1) ACTS link that addresses the limitations of the TCP protocol. It suggests an elegant solution that can be applied to the file transfer protocol (FTP) and other applications [12].

In this part of the test plan we are proposing a set of tests which will evaluate the limitations and characterize the performance of each one of the protocols (IP, TCP and UDP) by itself, before addressing the application layer protocol performance. These tests will provide a solid background to the science teams in terms of what the performance ceiling of the ACTS ATM link would be. In addition, we are proposing to use the application layer solution to the file transfer protocol (XFTP) [12] and compare the performance results with those of the standard FTP using the TCP extended window option. In summary, we will provide the throughput, latency, and packet loss rates of the above protocols as functions of the packet and window sizes, the link speed, and the number of nodes.

2. Preliminary Tests

2.1 Test Configuration and Methodology

In order to characterize the ACTS channel with regard to ATM traffic, we propose an end-to-end test. ATM cells will be fed to the GSFC ATM (Fore Systems) switch by an ATM analyzer (generator function). These will be received by the other ATM analyzer (analyzer function) connected to the ATM switch at JPL.

During the preliminary phase, loop back tests (figure 2) will be undertaken at GSFC. These tests will characterize the overall communication link and system, from the ATM switch to the satellite. This will include the fiber optic OC-3 link, the HDR terminal and the satellite up and down links. This test will require the use of only one analyzer. This analyzer will perform both traffic generation and measurement functions. The ATM switch will be configured so that the forward and backward traffic are always switched between the same ports. These ports will be identified as port1 and port2 (OC-3 155.52 Mbit/s fiber link interfaces).

In both configurations, the switch management function will be turned off to avoid the influence of band traffic on the performance. Transfer latency, cell loss ratio, cell delay variation, and cell error ratio will be measured for different physical Bit Error Rate values. (This will be simulated by a noise generator.)

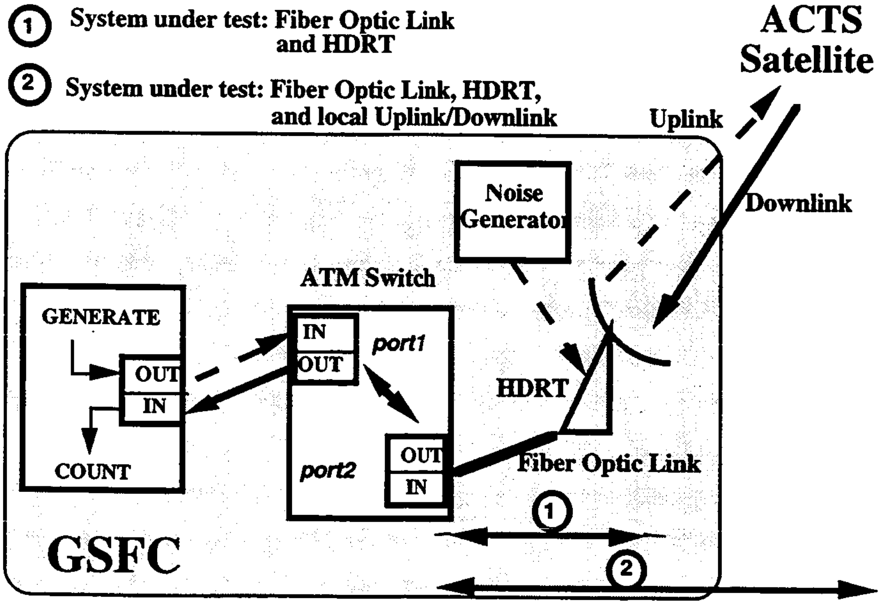


Figure 2: Loop Back Test Configuration

2.2 Latency - Cell Transfer Delay

This parameter measures the amount of time required to establish a virtual connection and to transmit cells between two end-stations. The end-to-end configuration will allow measurement of the latency between cell transmission and reception, for cells flowing within a single virtual circuit. A constant bit rate stream (for example, one cell every 20 cell times) will be used for the test.

2.3 Cell Loss Ratio

By comparing the cells sequence number on an end-to-end basis, we can estimate the amount of cells that eventually have been lost over the ACTS channel. Then, cell loss ratio will be derived by dividing the number of cells lost by the number of cells transmitted.

2.4 Cell Delay Variation

This cell delay variation parameter refers to the differences in end-to-end transit times for a given series of cells. It allows the investigator to estimate the consistency of the transmission of cells on the ACTS channel. In fact, when latency varies, cell inter-arrival times fluctuate. Cell delay variation will be measured by taking the standard deviation of cell delay over a certain number of cells, transmitted via a constant bit rate stream.

2.5 Cell Error Ratio

This parameter measures the accuracy of the ATM connection between GSFC and JPL. It will indicate how the ACTS channel latency and characteristics affect end-to-end ATM cells flow. The cell error ratio is computed by dividing the number of error cells by the number of cells transmitted. ATM level Bit Error Rate is given by analyzing the integrity of data within cells payload.

PART 3. CBR/VBR Traffic

This CBR/VBR traffic test will verify the ability of the ACTS channel to protect delay-sensitive CBR (constant bit rate: simulation of a voice traffic) flow when such traffic is sent to a port already fed with bursty VBR (variable bit rate: simulation of a data traffic) cells flow. As bursty traffic, we will consider "IP-over-ATM" like traffic generated by the analyzer. The frame size will be set to 9-kbyte, which is the maximum size defined by the IETF (Internet Engineering Task Force). The data flow burst will be a certain percentage of the available bandwidth on the OC-3 ATM interface, once the CBR cells were accounted for. This test will indicate whether or not the latency introduced by the ACTS channel will disrupt delay-sensitive traffic as voice and video.

PART 4. TCP/IP Tests

The following TCP/IP measurements should be performed on incremental loopback, point to point, multiple hop, and point to multipoint (multicasting) configurations between the available HDR terminals. All these different architectures will illustrate the peculiarities of an IP/ATM ACTS network.

4.1. IP Tests

The purpose of the IP tests in the test plan is to characterize the performance of the IP/ATM link. This set of measurements will evaluate the performance of the satellite link over a wide range of rates from OC-3 to

OC-12. Prior to the space science experiments over ACTS, it is important to evaluate the throughput of the IP/ATM link parameterized by the link speed and the message size.

In addition to these parameters, IP tests will discuss the impact of the number of hops between origin and destination. This test is important because the obtained results can be compared to similar test results from the NSF terrestrial IP/ATM backbone (vBNS). Such a comparison will clarify the role of the IP as the network protocol in satellite and hybrid high speed networks.

In this experiment, we will use a diagnostic tool, namely "windowed ping" [1], which provides direct measurement of IP performance, including queue dynamics. It uses a transport style sliding window algorithm combined with either ping or traceroute to sustain packet queues in the network. It can directly measure such parameters as throughput, packet loss rates and queue size as functions of packet and window sizes. Other parameters, such as switching time per byte or per packet can also be derived.

The need to understand the performance limitations of IP equipment and technologies vis-a-vis a high rate, high delay link drives this set of experiments. As the author of windowed ping discusses[1], rate-based IP performance tools can measure performance at rates below queue formation but only at ill defined points beyond congestion. They cannot measure switch performance under conditions of sustained partial queues. Clearly, in a megabit networking environment, the behavior of IP formed queues is of great importance. Windowed ping measures throughput (delivered data) and packet loss as functions of packet size and window size. During the IP link evaluation, we will measure the above quantities and in addition we will introduce as test parameters, the amount of hops in the IP path between the source and the destination, and the line speed. The amount of hops can either be two or four, depending on whether we use a loopback configuration. Figure 3 illustrates the experiment architecture.

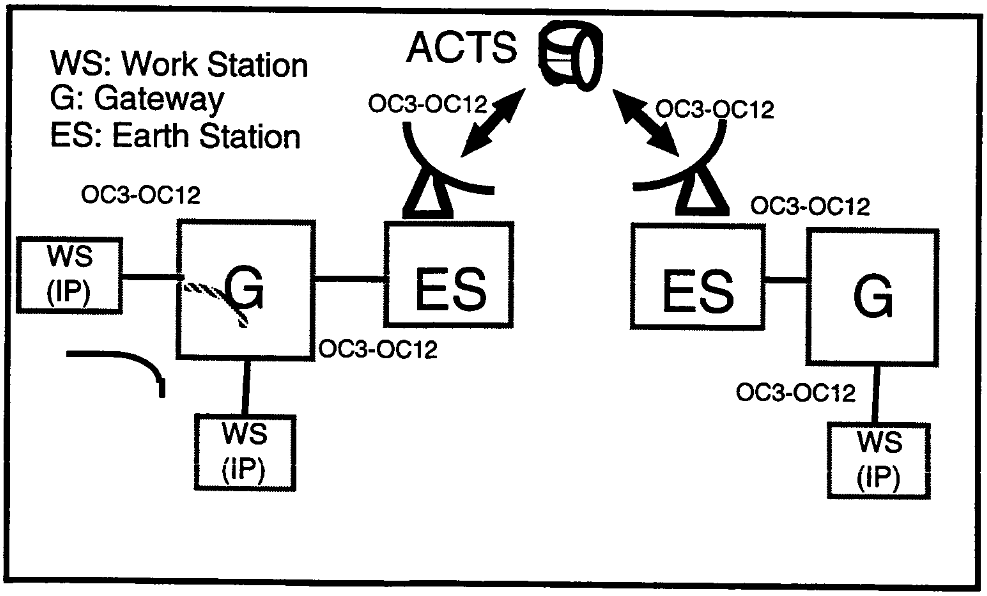


Figure 3: IP Testing

In addition to the IP performance, IP multicast performance will be tested. IP multicast uses a substantially smaller number of packets than the IP when it delivers data to a group of users. Therefore, it is expected that the protocol switch performance should be better due to the smaller size of the partial IP queues throughout the network.

4.2 TCP-UDP Tests

The purpose of the TCP tests is to characterize the performance of the TCP protocol over the ACTS IP/ATM link. We will perform a series of measurements to evaluate the throughput and the latency of the TCP/IP suite, as a function of the link speed, the message size, and the window size. Similar network architecture as in the IP tests will be used (figure 1). This architecture will enable us to run point-to-point tests as well as tests between two or four nodes (hops) and measure the TCP performance in a single and a multiple link satellite network.

Previous work has been done on evaluating TCP and UDP on a local testbed [2], [4] and by using PVM both locally and over the ACTS [3], [12]. The proposed tests will take into consideration all the previous results on the TCP performance over the ACTS T1 link, and they will investigate TCP on the IP/ATM ACTS backbone. The main tool to evaluate the latency and the throughput parametrized by the window size, the message size, the number of hops, and the line speed is TTCP [8]. Other analysis tools that may be used include tcpdump, netstat, and/or atmstat [7].

In addition to the investigation of the performance of the TCP Extended Window option, the proposed solution (for T1 satellite links) at the application layer [12] will be investigated as well.

The application layer protocol used at the Global Climate Modeling experiment is PVM, which uses mainly UDP for remote process communication. Since UDP is not a reliable protocol, PVM uses its own error recovery mechanism. The proposed techniques [4] to overcome the PVM error recovery protocol limitations will be used. The UDP protocol also is used in the proposed multimedia multicasting demonstration to carry IP packets to multiple groups of users.

Since UDP is the protocol used in applications over ACTS, such as PVM and MBONE, it is crucial to analyze its behavior in a high rate, high delay environment. Because UDP operates at the same level as TCP, i.e., on the top of IP, the same TCP tests will be performed to evaluate the throughput and the latency of UDP as a function of packet and speed link over the IP/ATM ACTS backbone.

After performing the above measurements for the individual protocols of the link and understanding the expected performance of the TCP-UDP/IP suite in correlation with the high data rate terminal characteristics and the ATM level tests, we will investigate the performance of a network application. The proposed experiment follows.

PART 5. Applications Tests

The GWU test plan now merges with the test plan activities of the two space science applications. A preparatory level of tests may be conducted to obtain transmission link data. But the conduct of the majority of the applications tests will be done to further the experiment behavior of the science applications.

5.1 HIPPI/ATM/SONET Test

The HIPPI/SONET architectures provide this test plan with the potential of using the full OC-12 (622 Mb/s) bandwidth of the ACTS channel, with the lowest host or network overhead. This test will use the HIPPI/SONET Gateway device made at the Los Alamos National Laboratory. It is likely that this HIPPI/SONET transmission will provide the highest ACTS channel throughputs and thus it will be coordinated closely with the space science applications

The SONET/ATM tests will operate in the OC-3 (155 Mb/s) speed, and a complete end-to-end ATM test using multi-node hybrid networks is proposed. For example, an ATM signal would originate at a supercomputer at one of the Advanced Technology Development (ATD) network nodes, and route the terrestrial traffic to the Earth terminal at GSFC and the via the ACTS satellite to the JPL and Hawaii Earth terminals, where it would be routed through terrestrial fiber cable networks to locations such as the San Diego Supercomputer Center or the Maui Supercomputer Center.

5.2 Multicasting Multimedia Test

The main goal of this applications test is to implement and evaluate a multimedia multicasting session between the three Earth stations (GFSC, JPL, Hawaii) where the number of participants changes dynamically. During the sessions we will apply all the previous results to evaluate how user level performance correlates to the network utilization for different link speeds and different sources of traffic. Specifically, we will try to evaluate user level performance in the presence of data only, voice and data, video, voice and data, etc. The multicasting architecture will be both, one to many configuration (classroom environment) and many to many (teleconference environment).

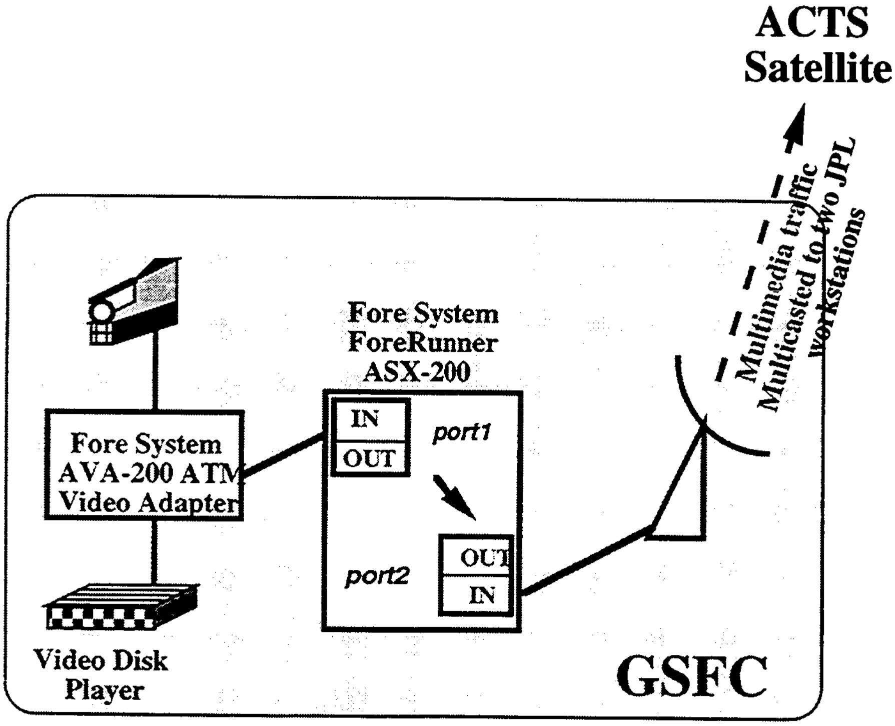

ACTS Satellite
Multimedia traffic Multicasted to two JPL workstations
Fore System ForeRunner ASX-200
IN
OUT
port1
port2
OUT
IN
Fore System AVA-200 ATM Video Adapter
Video Disk Player
GSFC

Figure 4: Multimedia Communications Over the ACTS Channel

5.2.1 Option 1: Fore System Platform.- Multimedia traffic will be multicasted from GSFC via the Fore System AVA-200 ATM Video Adapter, to one or many Workstations at JPL (figure 4).

5.2.2 Option 2: MBONE Framework.- The main protocol that supports network multicasting is IP multicast. IP multicast provides for the delivery of IP datagrams to multiple hosts across an IP-based backbone. First defined in 1988 by Steve Deering [9], IP multicast is used today across the global Internet. IP multicast is composed of three parts: extensions to the IP protocol, which is little more than a group addressing standard; a group management protocol called IGMP to allow hosts to join and leave groups dynamically; and, a multicast protocol to enable routers to correctly forward multicast datagrams.

IP multicast traffic is carried over a virtual network called the MBONE [10]. Traffic on the MBONE is audio, video, and imagery all carried over UDP. A set of publicly available tools allows a user connected to the MBONE to sit in on teleconferences, video conferences of special events, and audio programs. A user can scan the list of active sessions at any time and join one or more of them. As a result of this degree of freedom and the lack of resource reservation, user level performance can degrade sharply during periods of heavy use. These are the phenomena in conjunction with the IP partial queue sizes that we would like to investigate over the ATM ACTS link parameterized by the link speed (OC3-OC12).

The advantage of multicasting versus unicast transmission is illustrated in figures 5 and 6. A single source sends the same message to five destinations. In figure 5, five data units (one for each destination) are sent through the network traveling the same route for much of their journey. We would like to take advantage of the common paths and send only one copy where the route is shared. This idea is the one behind multicasting, as shown in figure 3, where one data unit is injected into the network and travels along the shared path. The benefits to be gained by multicasting in terms of reduced bandwidth and other network resources should be clear in this example scenario.

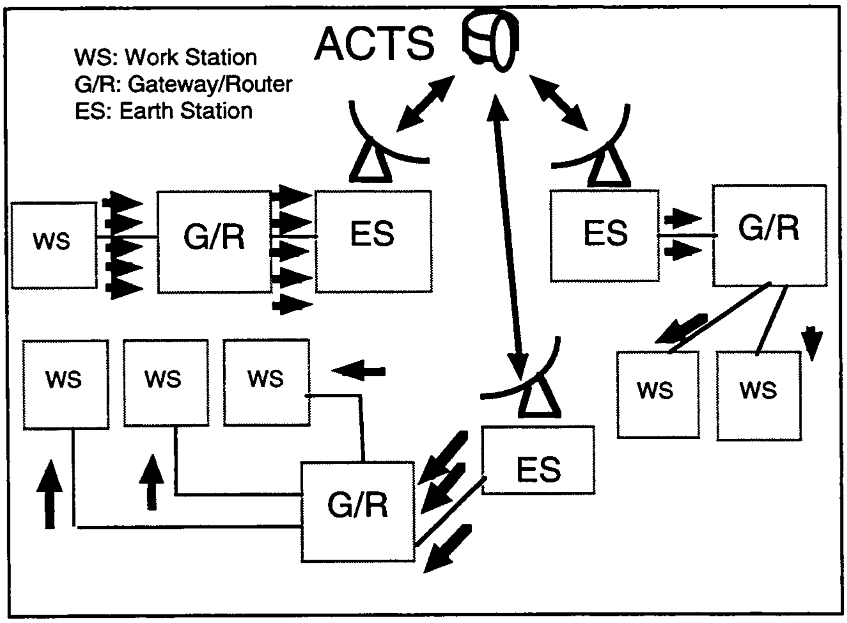


Figure 5: Unicast Transmission

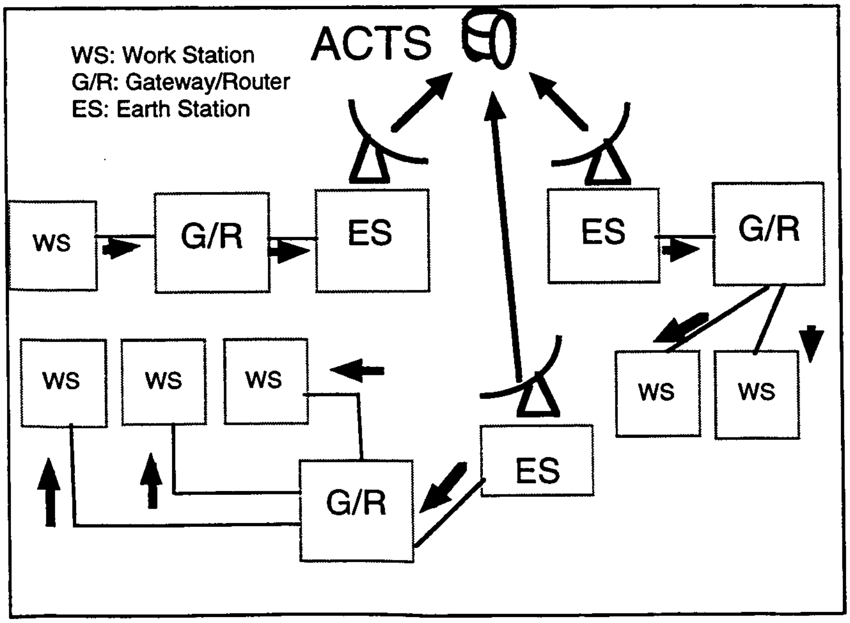


Figure 6: Multicast Transmission

The vast majority of traffic over the MBONE is audio and video streams running over UDP. Applications have been built by the Internet research community to send and receive audio and video. Additional hardware (a camera and microphone) is needed only if a user wishes to send. In our experiment, we would like to explore both cases and see how the performance of the application degrades when the traffic increases. All the previous measurements, in specific IP and UDP/IP, should be taken into account in order to understand the limitations of ACTS ATM multicasting sessions. Finally, empirical results of session performance as a function of the number of participants, the speed of the link, and the nature of the traffic will be furnished.

Part 6. Test Equipment

m = Mandatory; o = Optional

1. Two ATM analyzers for the end-to-end test (Wandel & Goltermann ATM-100 analyzer or any ATM analyzer). (m)
2. Two Fore Systems Fore Runner ASX-200 ATM switches. (m)
3. Three workstations with a Fore Systems Network Adapter Card on each of them (for the multimedia test). (m)
4. A Fore Systems ATM Video Adapter model AVA-200. (o)
5. MBONE software. (o)
6. A camera and a video disk player. (o)
7. Noise generator. (m)
8. Spectrum analyzer with a frequency coverage including 3.5 GHz. (m)

Part 7. References

Mathis, M. (1994). Windowed Ping: An IP layer performance diagnostic. (DRAFT) *Proc. INET '94/JENC5.*

Dowd, P. W., Srinidhi, S. M., Blade, E., Claus, R. *Issues in ATM support of high performance geographically distributed computing.* NASA-Lewis Research Center.

Brooks, D. E., Carrozi, T. M., Dowd, P. W., Lopez, I., Pellegrino, P. A., Srinidhi, S. m. *ATM based geographically distributed computing over ACTS.* NASA-Lewis Research Center.

Dowd, P. W., Srinidhi, S. M., Pellegrino, P. A., Carrozi, T. M., Claus, R., Guglielmi, D. *Impact of transport protocols and message passing libraries on cluster-based computing performance.* NASA-Lewis Research Center.

Wehner, M. F. et all. Toward a high performance distributed memory climate model. *Proc. 2nd International Symposium on High Performance Distributed Computing,* pp.102-113.

Mechose, C., Ferrara, J., Spahn, J. (1994) Achieving superlinear speedup on a heterogeneous distributed system. *IEEE Parallel and Distributed Technology,* Summer, pp 57-61.

Gary, J. P. (1995, September) Distributed global climate model experiment via ACTS, slide presentation. ACTS results conference.

Gary, J. P., Srinidhi, S. M. E-mail conversation and early performance data on the SONET/ATM satellite link.

Deering, S. (1988, August) Host extensions for IP multicasting. RFC 1112, SRI Network Information Center.

Macedonia, M., Brutzman, D. (1994, April) MBONE provides audio and video across the Internet. *Computer magazine,* IEEE Computer Society.

Helm, N. R., Edelson B. E. (1994, February) ACTS supercomputer network of global science applications. *AIAA 15th International Communications Satellite Conference*, San Diego, CA.

Allman, M., Kruse, H., Ostermann, S. (1995, August) *Data transfer efficiency over satellite circuits using a multi socket extension to the File Transfer Protocol (FTP).* ACTS Results Conference.

Bergman, L. A., Cohen, J. G., Shopbell, P. *KECK acquisition, visualization, and control over ACTS.*

NASA SCIENCE AND THE PRIVATE SECTOR

This task requires CESDIS to design and implement activities intended to facilitate private sector awareness and utilization of existing and expected NASA remote sensing data sets and technologies, incorporate early private sector requirements input to NASA programs, and assure continuing private sector participation in NASA undertakings. The work will be performed by Dr. Murray Felsher of Associated Technical Consultants.

Murray Felsher
Associated Technical Consultants (ATC)
(felsher@tmn.com)

Tasks assigned to Associated Technical Consultants (ATC) by the Director, Science Division, NASA Headquarters Office of Mission to Planet Earth (OMTPE) under this subcontract have been associated with strengthening the relationship between the science-oriented program/project management activities of NASA headquarters and the private sector commercial remote sensing industry. The objectives of these activities include (1) providing industry with the means of establishing an agglomerated, coherent input to NASA of private sector pursuits, concerns, and plans related to remote sensing, and (2) providing NASA with the means of transferring information on its science-related remote sensing activities/technologies into the private sector remote sensing community.

Examples of specific accomplishments this past year include:

1. Serving on the steering committee of the *NASA/Industry Workshop on OMTPE's Commercial Strategy*, a conference held on 22-23 July 1996 in Greenbelt MD. The Workshop, attended by NASA Administrator Dan Goldin, brought together senior NASA management and key industry

personnel to discuss, critique, and fine-tune OMTPE's overall commercial strategy. As part of the subject subcontract, ATC was responsible for contacting, inviting, and working with high-level private sector participants of two of the four workshop panels: (1) Data provider companies, and (2) Value-added companies. Follow-on activities related to the workshop carried through to December 1996.

2.. Setting up ongoing meetings between Science Division management and industry through a continuing program of on-site visits by OMTPE to local (Washington DC-area) commercial remote sensing firms. These have included value-added firms and primary satellite data acquisition companies. Such meetings will continue throughout the term of the subcontract.

3. Initiation of a major task to have the remote sensing private sector, through the North American Remote Sensing Industries Association (NARSIA) provide "...advice and recommendations on how to best shape a scientific initiative in the NASA Office of Mission to Planet Earth that will address top priority needs of the commercial remote sensing industry." This activity will allow NASA to receive, for the first time, direct input from the remote sensing industry, per NASA's request that they (NASA) "...appreciate that the review and advisory process would be run by industry, address industry concerns, and establish industry priorities."

Further tasks assigned have been related to providing outreach for NASA science activities to other federal agencies, so as to spin-off NASA science and applications into goals and missions of these other agencies.

A specific example was:

1. Serving on the steering committee of an EPA/NASA Workshop on Water Monitoring, Remote Sensing, and Advanced Technologies. The workshop took place on December 11-12, 1996 at the Holiday Inn in Washington DC. The purpose of the Workshop was to expose technical and management personnel of both agencies to (1) NASA's remote sensing science and technology, and (2) EPA's water resources monitoring requirements and data bases. The goal of the Workshop was mutual education and the opportunity to explore future collaboration in water monitoring/remote sensing research and applications. ATC, as having worked with both NASA and EPA headquarters, was responsible for outlining the structure of the workshop and coordinating the pre-workshop activities, as well as providing coordination of all early drafts of the workshop results.

Additional activities under the subject subcontract have included:

1. Serving on various NASA panels to evaluate proposals dealing with remote sensing applications activities.

2. Providing ongoing advice and recommendations to NASA personnel on programs and plans related to remote sensing applications.

3. Keeping Science Division personnel apprised of current plans of commercial remote sensing licensees as to status of upcoming launches.

3-D Unstructured-Grid Adaptive H-Refinement Modules

Rainald Lohner
George Mason University
Institute for Computational Sciences and Informatics
(lohner@rossini.gmu.edu)

The task assignment was to combine a GSFC 3-D High-order Godunov MHD solver based on unstructured grids with an Adaptive Mesh Refinement (AMR) module developed by the author. This task was successfully completed. This report details some of the difficulties encountered, and the solutions found for them.

1. Different Solvers

There are two versions of the MHD field solver. One stores the relevant quantities (e.g., pressure, magnetic fields, etc.) at the nodes. The second one stores these quantities at the elements. This implies that different data structures (i.e., arrays) will have to be used for both versions. The AMR module, as originally developed, operated only on node-based quantities. A new version of the AMR module was written that allows operation on element and node quantities at the same time. This solved most of the compatibility problems posed by the two different MHD field solvers.

2. Different Versions of Surface Definition Software

When refining the mesh, new surface points may have to be introduced. The new points will have to be placed on the proper surface, which is non-trivial for curved boundaries, such as the Earth. The AMR module and MHD field solvers had different types of surface defining software. These two versions were brought up to date, making them compatible.

3. Different Geometrical Arrays

The AMR module and MHD field solver need a number of so-called geometry-information arrays which are used to calculate derivatives, integral averages, etc. The arrays that contain this information were differently named, and in some cases differently structured, between the AMR and MHD field solver. Whenever an array of the MHD field solver could be used by the AMR module, it was passed. In some instances, new arrays had to be incorporated to the MHD field solver.

After solving all of the problems reported above, a working version was obtained. This version was tested on the CRAY-J90 at GSFC, as well as the Indigo-4Ks at George Mason University.

Additionally, a new criterion for limiting the region of refined elements was incorporated into the AMR module. It allows one to specify, based on the local physics, a lower limit for the "physically sensible element length". This could be based on the ion giroradius or any similarly relevant quantity.

Linearized Riemann Solver For Numerical Magnetohydrodynamics (MHD)

Dinshaw Balsara
National Center for Supercomputing Applications
University of Illinois
(u10956@ncsa.uiuc.edu)

This task required CESDIS to provide consulting support to develop a linearized Riemann solver for numerical MHD for work with NASA scientists which was to be based on developing code that concretized ideas contained in a paper by Dr. Dinshaw Balsara entitled The Linearized Formulation of the Riemann Problem for Adiabatic and Isothermal MHD. Dr. Balsara was retained on a consulting basis to perform this work.

- Wrote a subroutine for calculating eigenvectors for numerical MHD. This entailed being sensitive to the exception cases that arise when evaluating the eigenvectors for numerical MHD. This was then tested for several cases where the solution was already known in order to validate the sanctity of the eigenvector subroutine that was developed.
- Incorporated the subroutine into a linearized strategy for solving the MHD Riemann problem. This entailed projecting the difference in the conserved quantities onto the right eigenvectors. These projected quantities are called weights. The weights were then used to make intermediate states. The values in the intermediate states were then used to make entropy fixes.
- Tested the above subroutine to verify good performance on problems where the solution was known. Then started writing another subroutine to make Roe-averaged eigenvectors for the linearized problem for numerical MHD. Again the exception cases had to be handled. Also the calculation of weights and intermediate states had to be done.
- Integrated the two subroutines into one driver subroutine. Tested the driver subroutine. Verified that it sent easy situations to the easy version of the linearized Riemann solver and hard cases to the hard version of the linearized Riemann solver. Verified good functionality of the integrated package of codes.

Scalability Analysis of ECS' Data Server

Daniel A. Menascé
George Mason University
Department of Computer Science, Center for the New Engineer
(menasce@cs.gmu.edu)

Mukesh Singhal
Ohio State University
Department of Computer and Information Science
(singhal@cis.ohio-state.edu)

Statement of Work

The objective of this study is to carry out an analysis to determine if the current ECS Data Server design is scalable to the near and far term data volume requirements of EOSDIS.

Reports

Daniel A. Menascé

During this reporting period several meetings were held with Ben Kobler, Ted Willard, Jeanne Benke, and Bearnie Peavy with the purpose of clarifying the understanding of the ECS Data Server Architecture. We received a substantial number of documents that were analyzed.

From these documents and from the various technical discussions with Kobler, Willard, Benke, and Peavy, we derived a description of the operation of ECS's Data Server. This description is contained in the document "An Overview of the Ingest and Retrieval Systems of ECS's Data Server" included as part of this report below.

Also, a presentation was prepared and delivered on June 13, 1997 to a group of about a dozen NASA people including Karen Moe, Gail McConaughly, and Ben Kobler.

The scalability model development phase started and is scheduled to extend throughout the summer.

Mukesh Singhal

The goal of this study was to carry out a scalability analysis of NASA's mass storage server, the Data Server subsystem of ECS. The analysis took into account the proposed hardware and software architecture as well as the expected workload and would generate a report containing an analysis of how well the current design would handle the expected workload as well as workloads of higher intensity. The report would point out components of the system that will become bottlenecks as the workload intensity increases.

The consultant made several trips to CESDIS and participated in discussions with the researchers at CESDIS, GSFC, and George Mason University to understand the operations and architecture of the mass data archive so that it can be mathematically modeled and its scalability properties can be studied.

The consultant studied several documents, reports, and published articles that describe the operations and the architecture of the mass data archive. The consultant in coordination with the Co-PI of the project developed an understanding of the "Ingest Data" and the "Retrieval and Processing" operations for a performance characterization and analytic model of the mass storage. Special emphasis was placed on the completeness and correctness of the understanding of the operations, hardware mapping, and assumptions that will be made during the modeling of the system. In a meeting with GSFC personnel on March 24, the operation was fine tuned for a realistic performance characterization and analytic modeling of the mass storage server. The details of the operations are below.

An Overview of the Ingest and Retrieval Systems of ECS's Data Server

Mukesh Singhal and Daniel A. Menascé

1. Introduction

The purpose of this document is to provide a high level description of the Ingest and Retrieval processes as they relate to ECS's Data Server. This description will be used to support the development of a model to analyze the scalability of the Data Server. In the context of the current study, scalability analysis means the determination of whether the current architecture of the ECS Data Server supports an increase in the workload intensity with possibly more processing and data storage elements of possibly higher performance. This study will not address the types of architectural changes that may be needed to make the architecture scalable.

2. Subsystems of the Data Server

The following subsystems of the Data Server will be considered for the purpose of the scalability analysis considered in this study:

- **Software Configuration Items:**

  a. Science Data Server (SDSRV): responsible for managing and providing access to non-document Earth science data.

  b. Storage Management (STMGT): stores, manages, and retrieves files on behalf of other SDPS components.

- **Hardware Configuration Items:**

  a. Access Control and Management (ACMHW): supports the Ingest and Data Server subsystems that interact directly with users. Of particular interest here is the SDSRV.

  b. Working Storage (WKSHW): provides high performance storage for caching large volumes of data on a temporary basis.

  c. Data Repository (DRPHW): provides high capacity storage for long-term storage of files.

  d. Distribution and Ingest Peripherals (DIPHW): supports ingest and distribution via physical media.

3. Flow Diagram of Ingest Data

Figure 1 shows a diagram that depicts the flow of control and data for the Ingest process. We have not included Document Repository nor the Document Data Server due to their small impact on scalability, if compared with ingest of L0 data. Circles in the diagram represent processes. The labels in square brackets beside each process indicate the hardware configuration item they execute on. Bolded labels indicate hardware configuration items that belong to the Data Server.

The main aspects of the diagram in figure 1 are discussed below:

- Incoming L0 data is first stored into the system on the Staging Disk. The metadata are extracted and entered into a Metadata database managed by COTS DBMS (e.g., Sybase) and the actual data are archived. The archival process is essentially composed of the steps depicted in figure 2. The actual data is transferred into AMASS' disk cache. From the cache, the data migrates to robotically mounted tapes managed by AMASS. The metadata extracted from the data is stored into a Metadata Database managed by Sybase.

- Data products resulting from processing are also inserted into the Data Server. These products are stored directly into the AMASS cache.

- The SDSRV (Science Data Server) gives the metadata templates to the Ingest Request Manager for it to extract metadata.

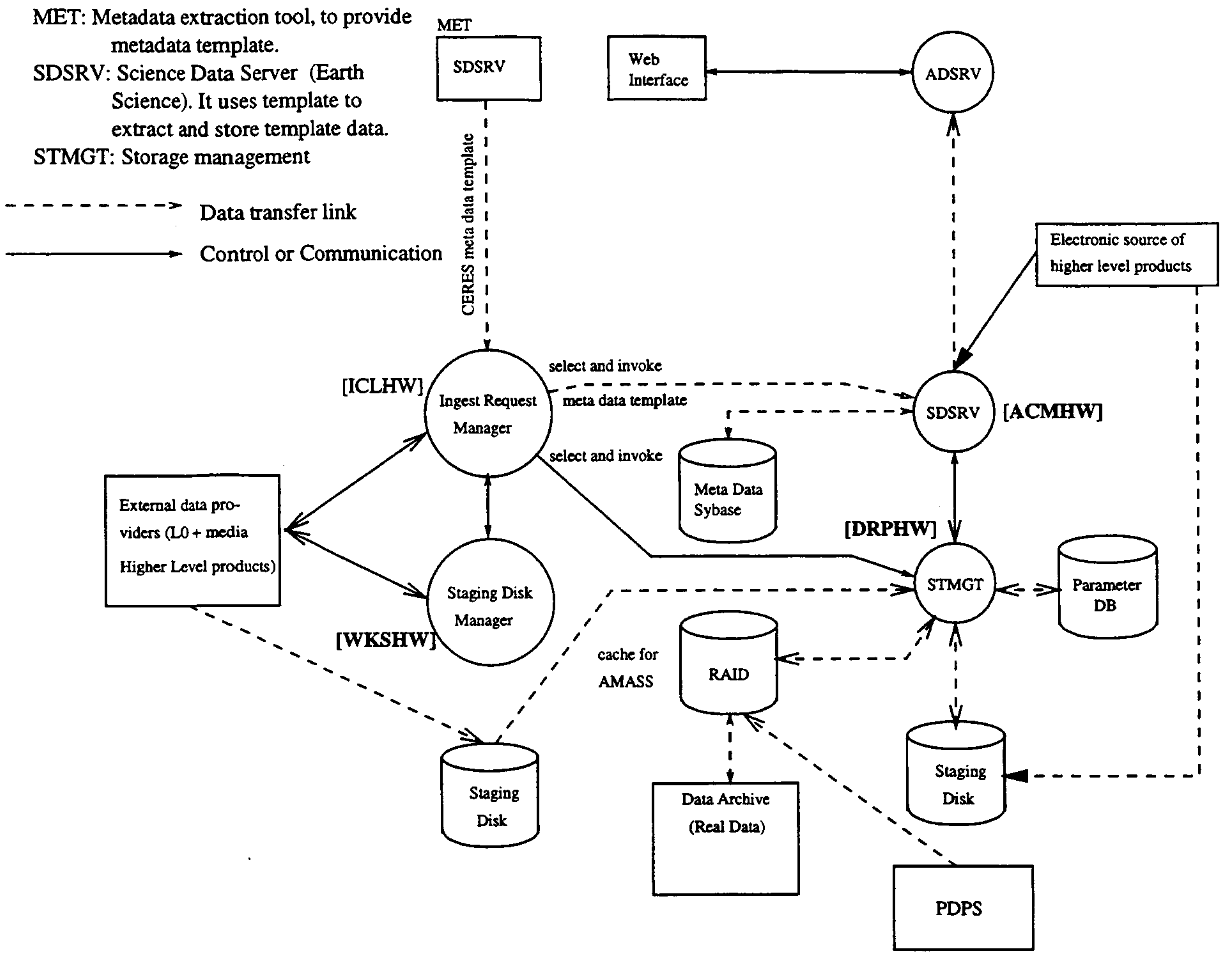


Figure 1: L0 Ingest Control and Data Flow

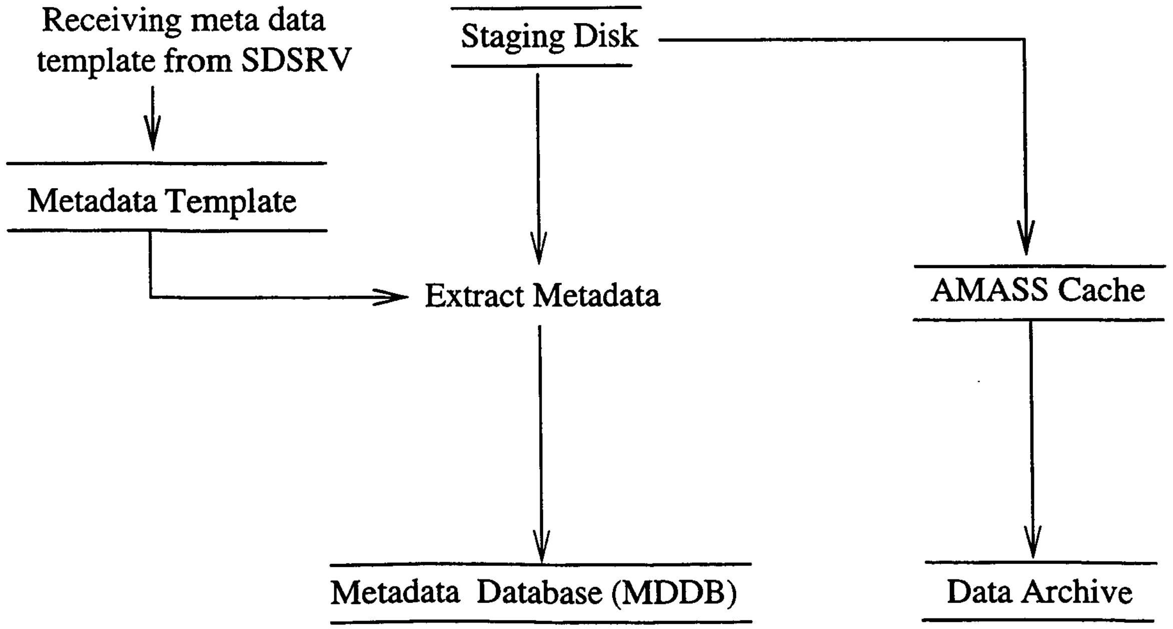


Figure 2: Data Flow Diagram for Ingest Data

- There could be several SDSRV's and one STMGT (Storage Management) processes. The Ingest Request Manager process selects which SDSRV to use.

The scalability analysis will among other things determine possible performance bottlenecks. The staging disk, the AMASS disk cache, and the metadata extraction process are likely candidates for bottlenecks.

4. Flow Diagram for Retrieval

This section examines the retrieval and processing operation on L1+ data as depicted in figure 3. Circles in the diagram represent processes. The labels in square brackets beside each process indicate the hardware configuration item they execute on. Bolded labels indicate hardware configuration items that belong to the Data Server.

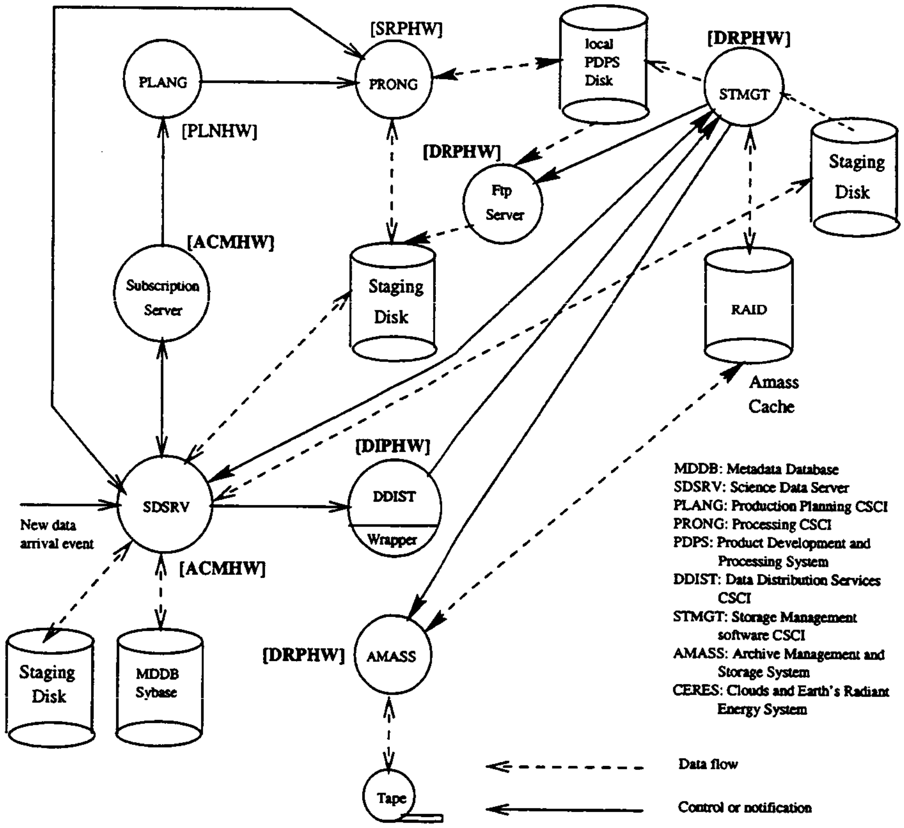


Figure 3: A Flow Diagram of Data Retrieval and Processing

The retrieval and processing operation proceeds in the following three stages:

Stage 1: Checking data and deciding what processing is required:

- SDSRV initiates the retrieval process by notifying the Subscription Server of the new data arrival.

- The Subscription Server performs a subscription check (given a UR) for this data and performs an appropriate notification, e.g., email notification, etc.

- The Subscription Server notifies PDPS PLANG of new data arrival.
- PLANG figures out (e.g., retrieves) a processing plan and based on the processing plan, passes the processing request to PRONG.
- PDPS PRONG connects to the appropriate SDSRV (may not be the SDSRV which initiated the retrieval and processing operations).

Stage 2: Retrieving data:

- The SDSRV requests that the Data Distribution Services CSCI (DDIST) retrieves the data files.
- SDSRV $\rightarrow^{requests}$DDIST $\rightarrow^{requests}$STMGT. The STMGT retrieves the files from AMASS archive into the AMASS cache if it is already not present in the cache.
- SDSRV notifies PRONG of data (identified by UR) availability.

Stage 3: Processing data and archiving, both data and metadata:

- Data is first moved into the staging disk from the AMASS cache prior to being transferred to the PDPS disk.
- PRONG processes the retrieved data to produce a higher level product.
- PRONG processes the data to a higher-level product and extracts metadata from the higher-level data using the Metadata Extraction Tool and populates the target metadata template and writes a metadata file (on MDDB Sybase).
- PDPS PRONG sends an insert request to SDSRV.
- SDSRV $\rightarrow^{requests}$STMGT $\rightarrow^{requests}$AMASS. The AMASS file manager archives the files. Archiving is done in two steps:
  - STMGT copies data from PDPS (local disk) to Working Storage via an ftp command.
  - Data are copied from the Working Storage to AMASS cache (and then to AMASS archive).
- SDSRV inserts metadata in the Metadata Database (MDDB) and then notifies PRONG that the archival operation has been completed.

5. Software Hardware Mapping

This section presents a mapping from the relevant software processes in the Data Server to the hardware configuration items for the GSFC, EDC, and LaRC DAACs. This mapping assumes the fourth procurement for Release B dated December 30, 1996.

Table 1 indicates the hardware configuration item needed to run each of the software processes of the Data Server at the three DAACs mentioned above. A description of each of the hardware items mentioned in table 1 is given in table 2.

Process: HWCI	GSFC	EDC	LaRC
Staging Disk Manager: WKSHW	G059	E064	00001210
STMGT Archive Manager: DRPHW	00001188	E046	00001219
SDSRV/Wrapper: ACMHW	00001185	E040	L090
Subscription Server: ACMHW	00001185	E040	L090
DDIST/Wrapper: DIPHW	G021	E030	L023

Table 1: Software Hardware Mapping per DAAC

Component Id	Description
E030	SUN 4000/2; 512MB; 8GB; D
E040	SGI PC XL; 10 CPUs; 1GB; 6GB; H/D/S
E046	SGI PC XL; 6 CPUs; 512 MB; 6GB; H/D
E064	SGI C XL; 6 CPUs; 512MB; 8GB; D/S
G021	SUN 4000/2; 512MB; 8GB; D
G059	SGI C L; 4 CPUs; 512 MB; 6GB; DAT; H/D
L023	SUN 4000/2; 512 MB; 6GB; D
L090	SGI C XL; 12 CPUs; 1 GB; 6 GB; H/D/S
00001188	SGI C XL; 6 CPUs; 512 MB; 6GB; D/H
00001185	SGI C XL; 14 CPUs; 1GB ; 6GB; D/S/H
00001210	SGI C XL; 6 CPUs; 512 MB; 6GB; DAT; D/H
00001219	SGI C XL; 6 CPUs; 512 MB; 6GB; DAT; D/H

Table 2: Hardware Component Descriptions

The symbols H, D, and S in table 2 mean the following: H (HIPPI), D (Dual attached FDDI), and S (single-attached FDDI).

6. Assumptions

- Processing of "Ingest data" and "Data retrieval and processing" constitute the main load on the storage server. Thus, we will model only these two operations.
- We will not model users' requests for data to be subsetted or subsampled. We will also not model compressed data.
- In data retrieval operation, PLANG retrieves a processing plan from a database (say Sybase).
- AMASS cache and working storage may be implemented on the same disk.
- We will drop the servers that are not potential bottlenecks from the model. For examples, "subscription server" and PDPS.
- We assume that mean arrival rate of both types of requests (ingest data and data retrieval) and service demands of these requests at various service stations are available or can be easily estimated.

- The only DAACs to be modeled are EDC, GSFC, and LaRC due to their higher data volumes as compared with other DAACs.

7. Data Volumes

Data Volumes for Release B are shown in table 3. The tables shows the amount of L0 data received per DAAC and the amount of data generated per DAAC per day.

DAAC	L0 (GB/day)	Products (GB/day)	Total (GB/day)
EDC	236.53	383.12	619.65
GSFC	70.21	395.74	465.95
LaRC	47.78	188.50	236.28

Table 3: Data Volumes for Release B

Data volumes for EDC reflect L0 data volume plus ASTER L1A data volume. Product volumes were obtained from "Volume Timelines v3.01" of Attachment C of "Ad Hoc Working Group on Production (AHWGP) Information", February 1996 (technical baseline available at http://edhs1.gsfc.nasa.gov/wais-data/toc tp2100106toc.html). TRMM products (LIS, VIRS, PR, TMI, CERES(TRMM)) were not included in the total for product volumes, but products that were based on CERES data from multiple platforms including TRMM where included (i.e., products identified as CERES(TRMM-AM) were included.) For each quarter, products were summed by archiving DAAC. The maximum volume per DAAC was taken over the quarters through 4Q99. (According to Attachment K of the technical baseline, Release C becomes operational towards the end of January, 2000.)

L0 volumes were obtained from Appendix E (SDPS Performance Parameter Synopsis) of DID 304 (SDPS Requirements Specification) available at http://edhs1.gsfc.nasa.gov/waisdata/toc/cd3040502toc.html. L0 data comes from Landsat-7, AM-1, RADAR ALT, SAGE III, ADEOS II, and ACRIM.

INFORMATION TECHNOLOGY RESEARCH ISSUES

Oktay Dogramaci, Technical Specialist
(oxd@cesdis.usra.edu)

Profile

Mr. Dogramaci received a B.A. in political science from Johns Hopkins University in May 1996. While a student he worked as a research assistant in the Political Science Department performing literature reviews on current international immigration policies and researching and writing article summaries. As an intern at the Southeast Asia Resource Action Center (SEARAC), he provided assistance with corporate outreach and program development, assisted with the development of a skills data bank, and investigated resources for expanding the organization onto the Internet. During the summer of 1996, Mr. Dogramaci participated in the USRA Goddard Visiting Scientist Program's Visiting Student Enrichment Program where he researched legal and regulatory impediments to electronic commerce.

Report

After working with Dr. Yelena Yesha from June to August of 1996 on the G-7 pilot project "Global Marketplace for SMEs," I began working full-time at CESDIS to further collaborate with Dr. Yesha on the interdisciplinary research issues which affect the underlying technology involved in that project as well as related disciplines. The research issues involved are broadly categorized within the field of information technology. In order to support various initiatives undertaken by Dr. Yesha, I have been performing research within the fields of electronic commerce, telemedicine, distance learning, and digital libraries, which all share similar technical requirements and a common infrastructure. CESDIS has been involved in research in all these areas for some time now.

My work within these areas has resulted in various reports on specific aspects of these information technologies, several presentations given by Dr. Yesha on these subjects in the past year, and recommendations to Dr. Yesha on research areas which may be beneficial for CESDIS to pursue. Additionally, I was a contributing author for the paper "Electronic Commerce and Digital Libraries: Towards a Digital Agora" [*ACM Computing Surveys* 28(4), December 1996. Copyright 1996 by the Association for Computing Machinery, Inc.], edited by Dr. Yesha and Dr. Nabil Adam, and am currently co-authoring a monograph to be published this fall by the MIT Press, titled *Electronic Commerce: An Interdisciplinary Focus*, with Dr. Yesha and Dr. Adam as well.

In addition to working in these research areas, I have been involved in determining various other areas of research within NASA which CESDIS may be able to provide expertise for or desire to participate in. I wrote a report for Dr. Adam detailing the proposed restructuring of the Office of Space Sciences at NASA Headquarters. Also included in the report were other activities being funded by Joe Bredekamp at OSS that I learned of. The goal of this work was to provide Dr. Adam, Dr. Yesha, and others with necessary information for future proposals to involve CESDIS in applied information technology programs that Mr. Bredekamp and the OSS sponsor.

Other aspects of my work this past year include a NASA agency-wide formal search I conducted to find collaborative research projects that CESDIS would benefit from, and for which CESDIS could provide expertise. The search specifically was focused on on-going work at Ames Research Center and The Office of Space Science at NASA Headquarters, but was not restricted to it. The most fruitful result of this project was making contact with Mr. Bredekamp at NASA Headquarters. Following this search, I assisted in preparation of a presentation to be given to Mr. Bredekamp by Dr. Yesha and Dr. Adam. The presentation highlighted CESDIS' recent achievements and projects. The goal was to stimulate discussion for future funding for CESDIS in information technology-related projects supported by the Office of Space Science and Mr. Bredekamp's programs.

After Dr. Yesha's and Dr. Adam's meeting with Mr. Bredekamp from NASA HQ, OSS, they met with me to discuss CESDIS' desire to be involved in the NASA/USRA SOFIA project. I prepared a report for them on what the SOFIA project is, what partners are involved, and which components of the project are furthest advanced or in need of support. It was determined that CESDIS would try to begin a formal collaborative project with Ames Research Center to handle the Data Management tasks for SOFIA. I opened a line of communication with the curator of SOFIA documents at Ames, and acquired the technical papers regarding the SOFIA data system which have been written in the past. This project is now being led by Susan Hoban, and a formal proposal to begin CESDIS' involvement is expected in the near future.

ADMINISTRATION BRANCH

Nancy Campbell, Senior Administrator – Branch Head

Annemarie Murphy, Administrative Assistant 3 (Financial)
Georgia Flanagan, Administrative Assistant 3 (Conference Management)
Jillian Lusaka, Administrative Assistant 3 (Web Site Administration, Database Management, Presentation Graphics, Desktop Publishing)

Joyce Hines, Administrative Assistant 2 (Financial/Subcontract Support)

Michele Meyett, Administrative Assistant 1 (General Clerical)

This branch is responsible for supporting the CESDIS Director, Senior and Staff Scientists, Technical Specialists, funded project personnel and graduate students, and USRA's corporate office. Branch personnel:

- Serve as the liaison among funded research personnel, NASA scientific and administrative personnel, and USRA accounting and procurement personnel,
- Monitor subcontracts and consulting agreements,
- Monitor the contract's Small and Small/Disadvantaged Business Plan,
- Prepare and monitor task budgets,
- Prepare contract reports,
- Obtain Contracting Officer permission for foreign travel by staff and university scientists,
- Obtain Contracting Officer permission for equipment purchases with contract funds and report purchases to Goddard's property personnel,
- Assist with conference planning and provide on-site support at conference, workshop, and seminar locations,
- Assist foreign national visitors in gaining access to Goddard,
- Provide peer review support to NASA program personnel for proposals submitted in response to NASA Research Announcements and Cooperative Agreement Notices,
- Maintain CESDIS Web site.
- Provide desktop publishing assistance for paper preparation, the CESDIS newsletter, and presentation graphics,
- Make travel arrangements and provide assistance with travel voucher completion,
- Perform functions of remote site data entry for USRA's centralized accounting system including pay roll, purchasing, and accounts payable.

BRANCH ACTIVITIES

Seminar Series

CESDIS sponsors seminars by visiting scientists from universities, government laboratories, and the public sector. These presentations are open to everyone at Goddard as well as interested off-site attendees. Announcements of speakers and dates are posted on the CESDIS homepage. Seminar presentations during this reporting year are listed below. Abstracts appear in Appendix B.

- Alfred **Aho**, Columbia University. *How Reliable Can We Make Software?*

- Eduardo **Alvarez-Cordero**, Harpy Eagle Program, The Peregrine Fund. *Eagles on the GIS: Satellite Tracking Harpy Eagles and Mapping Their Rain Forests in Venezuela and Panama.*

- Amnon **Barak**, The Hebrew University (Israel). *Measures for Performance Scalability in Networks of Workstations and Servers (NOWS).*

- Alex **Brodsky**, George Mason University. *The CCUBE Constraint Object-Oriented Database System: An Overview.*

- William **Carlson** and Jesse **Draper**, Supercomputing Research Center. *AC and the T3D.*

- Robert **Chervin**, National Center for Atmospheric Research (NCAR). *A Global Ocean Model for Climate Change Applications on Massively Parallel Processors.*

- Mel **Ciment**, National Science Foundation. *Research and Development Opportunities in Information Services.*

- Farouk **El-Baz**, Boston University. *Origin and Evaluation of Desert Landforms: the Case of the Western Desert of Egypt.*

- Simon **Julier**, IDAK Industries. *A Revolutionary General Extension for Nonlinear Kalman Filtering.*

- Ashfaq **Khokhar**, University of Delaware. *A Poly-algorithmic Approach for Achieving Scalable Performance on MPPs.*

- Norman **Kopeika**, Ben-Gurion University of the Negev (Israel). *Image Restoration From Atmospheric Blur and Its Application to Satellite Imagery.*

- David **Landgrebe**, Purdue University. *On Information Extraction Principles for Hyperspectral Data.*

- Miron **Livny**, University of Wisconsin at Madison. *High Throughput Computing on Clusters of Workstations.*

- Michael **Mascagni**, University of Southern Louisiana. *A Scalable Library for Pseudorandom Number Generation: Theory and Practice.*

- Tova **Milo**, Tel-Aviv University (Israel). *Correspondence and Translation for Heterogeneous Data.*

- Gagan **Mirchandani**, University of Vermont, Burlington. *Wreath Product Group-based Correlation Applications to Problems in Signal Processing.*

- Joel **Morris**, University of Maryland Baltimore County. *On Discrete and Discrete-time Wavelets With*

Optimum Time-frequency Resolution.

- Timothy **Murphy**, GN Nettest. *Signal Processing in Optical Time Domain Reflectometer.*
- Anthony **Norcio**, University of Maryland Baltimore County. *People, Interfaces, and Systems.*
- Arthur **Secunda**. *A GRAPHIC VOYAGE Through COLOR AND FORM: Unraveling the Creative Process by Giving Technology a Soul.*
- Aya **Soffer**, University of Maryland Baltimore County. *The Retrieval by Content in Symbolic-image Databases.*
- Willy **Zwaenepoel**, Rice University. *Shared Memory Computing on Networks of Workstations.*

Workshop on Data Mapping and Matching

A one-day workshop to present the results of the research project entitled *A CESDIS-University Collaboration on Data Mapping and Matching: Languages for Scientific Datasets* was held at GSFC on November 7, 1996. Individual presentations were made by the following:

- Nabil **Adam**, Rutgers University. *Maximizing the Value of Large Heterogeneous Knowledge Bases.*
- Mariano **Consens**, University of Waterloo. *Discovering Resources Across Internet Databases.*
- Susan **Davidson**, University of Pennsylvania. *Data Mapping for Scientific Databases using Morphase.*
- James **French**, University of Virginia. *Database Support for Correlation and Fusion Algorithms.*
- Robert **Grossman**, University of Illinois. *Data Mapping Issues for Data Mining Applications.*
- Yannis **Ioannidis**, University of Wisconsin. *Querying External Scientific Systems.*
- V. **Jagadish**, AT&T Laboratories. *Data Mapping for Business Data, and Why the Problems Are Not As Different From Mapping Scientific Data As We Might Imagine.*
- George **Lake**, University of Washington. *The Digital Sky.*
- Miron **Livny**, University of Wisconsin. *Data Staging in a High Throughput Computing Environment.*
- Marc **Postman**, Space Telescope Science Institute. *The Hubble Space Telescope Data Archive.*
- Raghu **Ramakrishnan**, University of Wisconsin. *Content-based Queries in Image Databases.*
- Peter **Wegner**, Brown University. *Multiple Interface Models.*

A panel discussion on *Data Mapping and Matching: Ad Hoc Art or Well-grounded Science* was presented by Serge **Abiteboul** (Stanford University), Peter **Buneman** (University of Pennsylvania), David **Maier** (Oregon Graduate Institute), and Stanley **Zdonik** (Brown University).

CESDIS Science Council

The CESDIS Science Council met on September 9, 1996 at NASA Goddard Space Flight Center. Presentations on work-in-progress were given by Yelena Yesha, Kostas Kalpakis (University of Maryland Baltimore County), Daniel Menasce¢ (George Mason University), Aya Soffer (UMBC), Yair Amir (Johns Hopkins), George Lake (University of Washington), Jacqueline Le Moigne, Phillip Merkey, Don Becker, Rick Lyon (UMBC), and Nathan Netanyahu (University of Maryland College Park). Joe Rothenberg (Goddard Director) and Al Diaz (Goddard Deputy Director) spoke about new directions for NASA GSFC.

The next meeting of the Science Council will be in August 1997.

APPENDIX A

In Review

The CESDIS Newsletter

CESDIS

In Review

Universities Space Research Association

Issue 3 Winter 1997

Featuring: Computational Sciences Branch

REMOTE SENSING IMAGE PROCESSING AT CESDIS

Based on collaborative research with several NASA/Goddard groups, the CESDIS Image Processing Group, part of the Computational Sciences Branch, focuses on the development of new computational techniques which can be utilized to process and analyze Earth and space science image data. This research is illustrated below, and includes image classification, image registration, and image restoration, all new methods being developed on a Massively Parallel Processor, the MasPar MP2. ■

Landsat-TM Image

Metadata Extraction from Remotely Sensed Images
Nathan Netanyahu

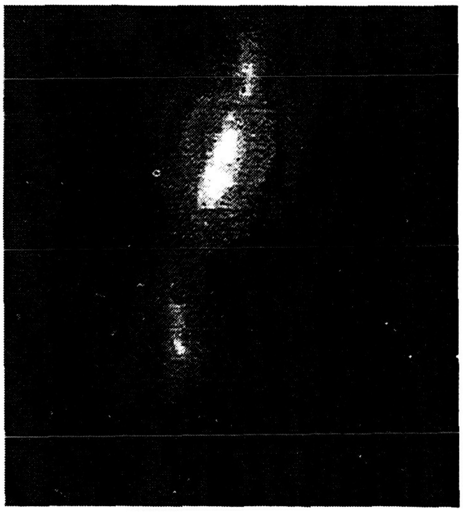
To prepare for the challenge of handling fficiently the archiving and querying of ra-byte sized scientific spatial databases, we have pursued, as part of the verall development of the Intelligent formation Fusion System (IIFS) at the pplied Information Science Branch, ode 935, NASA Goddard, a number of

continued on page 4)

Wavelet-Based Image Registration
Jacqueline Le Moigne

In the near future, new satellite remote sensing systems will provide large amounts of global coverage and repetitive measurements representing multiple-time or simultaneous observations of the same features by different sensors. Also, with the new trend of smaller missions, most sensors will be carried on separate platforms, resulting in a tremendous amount of data that must be combined. In meeting some of the Mission To Planet Earth objectives, the combination of all this data at various resolutions - spatial, radiometric and temporal - will allow a better understanding of Earth and space science phenomena. Accomplishing this will require fast, accurate and automatic image registration, which will help develop "ready to use" global datasets from multi-instrument/multi-platform/multi-temporal observations, and new image products summarizing some basic understanding of the original data may be created.

(continued on page 7)



Hubble Space Telescope Image

Model-Based Image Processing and Restoration
Richard Lyon

Classical image processing typically consists of a toolbox of algorithms which operate on an image, transforming it in some manner. Some examples would be

(continued on page 6)

FROM THE CESDIS DIRECTOR

Following a recent reorganization of the research activities of CESDIS from three to two branches, the Scalable Systems Branch was combined with the Computational Sciences Branch. This merging of research areas which include image processing, development and evaluation of parallel algorithms as well as scalable architectures is proving to be very fruitful and has already given rise to new projects combining several of these research aspects. An example is the implementation of the wavelet-based registration technique (described in this newsletter) onto the Commodity-Off-The-Shelf architecture known as Beowulf (see *In Review*, Fall 1995). Through these activities and those of the Applied Information Technology Branch, CESDIS is providing the NASA community with high-quality research which is applicable to many Earth and space problems. ■

CESDIS SUMMER '96 VISITORS

Serge Abiteboul, Stanford University
- to investigate the high-level description of source behavior based primarily on an integration of well-understood paradigms: object databases (and high-level query languages) and active databases.

Nabil Adam, Rutgers University
- to coordinate, provide guidance to, and assess the progress of the research projects relating to the digital libraries initiative, including the Global Legal Information Network (GLIN) and the US-Israel Information Technology Development Program.

Alfred Aho, Columbia University
- to develop efficient approaches for determining and keeping track of time - varying information about complex objects such as those found in geographic databases.

Ian Akyildiz, Georgia Institute of Technology
- to work on ATM over Satellite research problems and to assess current long-term and short-term research directions.

Neil Helm, George Washington University
-The Institute for Applied Space Research is working on a development plan for a "Testbed for Satellite and Terrestrial Interoperability." This testbed is planned to allow government, industry and academic users to test and demonstrate the next generation of interoperable computing and communications systems and networks.

Walter Ligon, Clemson University
- to port one or more parallel file systems to a Beowulf machine at NASA Goddard Space Flight Center and evaluate both performance and suitability for the T-Racks/Beowulf mass storage system.

Alberto Mendelzon, University of Toronto
- to conduct a preliminary study of database issues in internet-based electronic commerce, including: data integration, virtual catalogs, data warehousing, and declarative query languages.

Mukesh Singhal, Ohio State University
-to conduct research on performance modeling of digital libraries, of massive video data and images, and fast parallel/distributed techniques to retrieve, analyze, store, and display of data.

Brooke Stephens, University of Maryland Baltimore County
-to investigate optimal resource allocation for distributed computing problems.

Russell Turner, University of Maryland Baltimore County
- to provide expertise in creating interactive graphical user interfaces (GUI) for providing input, editing and querying functionality for Phase I of the Library of Congress Global Legal Information Network task.

Jeffrey Ullman, Stanford University
- to consult on GLIN and look into tele-education.

Peter Wegner, Brown University
- to develop a conceptual framework for distributed and embedded information systems in terms of interactive computing, and a common interactive architecture for the NASA earth observation system and for digital libraries based on this conceptual framework.

Ouri Wolfson, University of Illinois at Chicago
- to develop and analyze a set of algorithms for dynamic allocation and replication of objects with the purpose of adapting the allocation scheme to the current read-write pattern in distributed databases.

CESDIS
NASA GSFC
Code 930.5
Greenbelt, MD 20771
301-286-4403
cas@cesdis.gsfc.nasa.gov
http://cesdis.gsfc.nasa.gov/

Yelena Yesha
Director

Jillian Lusaka
Editor/Designer

In Review is published quarterly by CESDIS and NASA Goddard Space Flight Center.

Advances in Digital Libraries (ADL '96) May 13 - 15, 1996 -- Washington, DC

On May 13 -15, 1996 the second Annual Forum on Research and Technology Advances in Digital Libraries (ADL '96) was co-sponsored by NASA Goddard Space Flight Center, The National Library of Medicine, The IEEE Computer Society, and The Library of Congress. The forum was held at the Library of Congress and featured research papers, panels, exhibits, and prototypes and applications of digital libraries in science and industry. This technology has attracted widespread interest from many sectors and brought together universities, private industry and government to discuss constantly changing research ideas and advances in the digital library community. Approximately 175 attendees discussed state-of-the art technologies for global networked libraries for electronic commerce, environmental monitoring, law and medicine. Among the keynote speakers were Dr. Donald Lindberg (National Library of Medicine) and Dr. Harry McDonald (NASA/Ames Research Center). Senator Robert Kerrey (D-NE) was the banquet speaker. A student exhibition highlighted innovative research projects, such as a remote interface for the National Library of Medicine's Visible Human Digital Library. ■

(For information on ADL '97, visit ADL's web page: http://cesdis.gsfc.nasa.gov/admin/adl97/adlcall.html)

The IEEE Frontiers'96 Conference October 27 - 31, 1996 - Annapolis, Maryland

The IEEE Frontiers '96 Conference provided a major forum for exploring the technical issues that define the outer boundaries of effective high performance computing. This decade-long series of symposia was one of the principal meetings for scientists to present new and original research results extending the threshold of computational capability through advances in hardware, software, methods, and technology.

The spectrum of fields addressed by the Frontiers' sessions included applications and algorithms, system software and languages, component technologies, and system architectures. A central theme of Frontiers'96 was research related to the exploitation of massive parallelism, and any aspects of the design, analysis, development, and/or use of massively parallel computers. The realm of computing considered included general purpose, domain specific, and special purpose systems and techniques.

Topics illustrating both near-term practical results and those having long-term implications were addressed. This dynamic forum provided a stimulating and exciting environment for scientists, industry representatives, and government policy planners to present ideas, findings, product capabilities, and future directions through a series of sessions, panels, and workshops. The Conference sessions were held Tuesday through Friday; the Workshops were conducted Sunday afternoon and all day Monday. ■

(For further information, visit the Frontiers '96 web page: http://cesdis.gsfc.nasa.gov/front96.html)

Global Legal Information Network (GLIN)

The Global Legal Information Network (GLIN) is an international, non-commercial cooperative network of government agencies which seeks to create a database of international law documents which would be available to member countries throughout the world and would facilitate international cooperation and joint ventures. The Library of Congress and NASA Goddard Space Flight Center have joined forces in helping to realize the goals of GLIN.

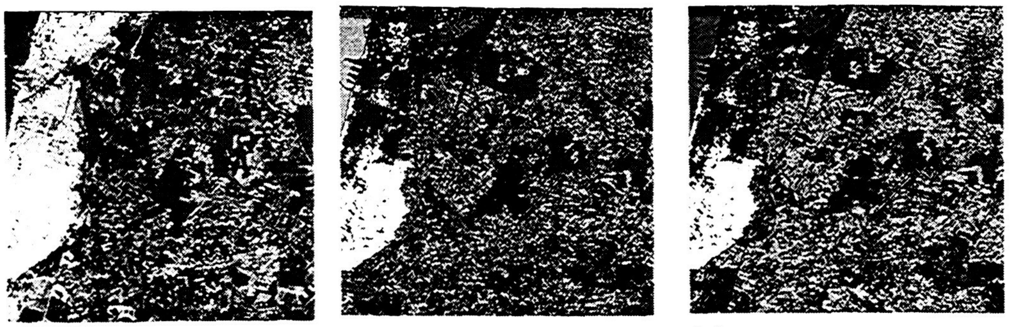
At CESDIS, Kostas Kalpakis and colleagues (including Burt Edelson, Pat Gary, Nabil Adam, Susan Hoban, Tarek El-Ghazawi, Neil Helm, Lee Foster) are working on completing the tasks outlined in the first phase of the GLIN project with the Law Library of the Library of Congress. In parallel a project plan for the second phase is being developed in cooperation with the Law Library. Meanwhile, Dr. Kalpakis is conducting a preliminary investigation on system architectures for the second phase of GLIN. He gave a presentation on this subject at the 3rd Annual GLIN Director's meeting in September. ■

For further information, visit GLIN's web page: http://lcweb2.loc.gov/glin/glin.html)

T-RACKS

Work on a new project within the Beowulf program, called T-Racks, has begun at CESDIS. T-Racks will be a secondary storage system that will be developed and evaluated within the operational environment of the ESS CAN machine. The system will have a Terabyte of spinning storage and provide a gigabyte/s aggregate external bandwidth. The system is distinguished from other secondary and mass storage systems because it will be assembled entirely from COTS hardware and will run enhancements to the Linux operating system integrating publicly available parallel file systems. The work is sponsored jointly by DARPA and NASA and will result in a trivially replicatable system. ■

Figure 1: Classification of Landsat-TM Images

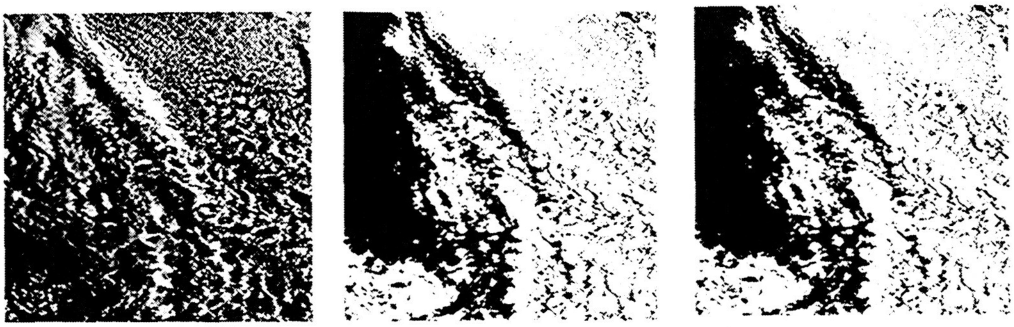


(a) Landsat-TM Original Image, Ashdod, Israel

(b) Probabilistic Neural Network (PNN) Classification

(c) Nearest Neighbor Classification (NN)

Figure 2: Classification of AVHRR Images



(a) AVHRR Original Image

(b) Probabilistic Neural Network (PNN) Classification

(c) Nearest Neighbor (NN) Classification

Metadata Extraction

(continued from page 1)

"characterization modules" for extracting metadata (from remotely sensed imagery). These modules rely mainly on supervised (image) classification techniques, such as neural networks, nearest neighbor searching, decision tree classification, etc. To further enhance the IIFS's performance with regard to metadata extraction, we have also pursued unsupervised clustering modules (based on principles of robust statistical estimation).

Extracting metadata from remotely sensed images in a fast and relatively accurate manner can be adequately achieved through a combination of probabilistic neural networks (PNN) and backpropagation-trained neural networks. Both techniques require a user-supplied training set, which can be obtained, for example, via photo-interpretation. Given its fast training time, the PNN serves to establish initially an "optimal" training set that is representative of the classes present in a given (set of) scene(s), after which a trained backpropagation network is invoked in its feed forward mode to classify these scenes.

The PNN's learning process is based on nonparametric density estimation techniques. Specifically, the network estimates the probability density function of a newly introduced test pattern by computing for each class (that is present in the training set) a sum of Gaussians centered at each individual training pattern that belongs to the class, and evaluated at the test pattern. The pixel (or test pattern) is assigned to that class for which the above computation is the highest.

The PNN becomes computationally intensive if other than training/testing of a (relatively) small number of pixels it should be used to classify a very large number of patterns (e.g., millions of pixels). On the other hand, the PNN lends itself naturally to a single instruction multiple data (SIMD) parallelization. Thus, to reduce significantly its run-time, we have implemented a parallel version of the PNN on the massively parallel machine, the 16K processing element MasPar, MP-2. As an example, it takes the parallel version approximately 30sec (net data processing unit (DPU) time) to classify a 7-band 512 X 512 Landsat image against a training set of ~1500 patterns, as opposed to hours it takes a sequential version to run on a SUN workstation.

We have also parallelized nearest neighbor (NN) classification scheme(s) and run it on the MasPar. The run-time and classification results appear to be comparable to those of the PNN. The module can be readily incorporated into the IIFS, thereby further extending the system's overall capability.

Having generated training sets from a number of Landsat-TM images we ran the PNN and the NN-searching module to classify these and other Landsat-TM scenes. The characterization vectors obtained serve to populate the database of the IIFS and help meet relevant prospective queries regarding the data's contents. Figure 1 above provides a flavor of the classification maps obtained due to the previously described supervised schemes. It depicts (a) band 3 of a 512 X 512 Landsat-TM (sub)image of Ashdod, Israel, (b) its classification due to the PNN, and (c) its classification due to NN-searching. Note the good segmentation to Urban, Agriculture, Water, and Barren (e.g., beach sand) classes. Likewise, Figure 2 demonstrates that the above procedures can be applied to AVHRR images, classifying them to cloud height categories. It depicts (a) band 2 of a 512 X 512 AVHRR subimage, (b) its PNN-based classification, and (c) its classification due to NN-searching.

Finally, it should be noted that in cases where merely an approximation of the data's contents is required, classifying only a small sample of the image would suffice. Specifically, we have shown theoretically and empirically, that for a typical 2,984 X 4,320 Landsat-TM scene quadrant, it suffices to classify merely a sample of ~17,000 pixels to obtain a characterization vector whose individual components are within 1% of the corresponding class frequencies in the entire

(continued on page 5)

Metadata Extraction

(continued from page 4)

image. This feature offers extended flexibility, as far as planning/scheduling of the concurrent tasks the IIFS is expected to handle.

To improve the performance (i.e., accuracy) of the above discussed classification modules (currently, it is estimated that the PNN provides ~70% accuracy), we have been investigating the possibility of combining spectral intensities with spatial information due to wavelet processing.

The underlying assumptions that we have been attempting to verify are that mixels (i.e., "border-line"/mixed pixels, in a spatial/spectral sense) contribute significantly to the overall misclassification of an image, and that functions of wavelet parameters will indicate how to single out these questionable pixels. Once detected, the classification of such pixels can be deferred to a post-processing stage at which time other sophisticated schemes (e.g., relaxation-based, linear mixture modeling, etc.) could be invoked to yield an improved overall accuracy.

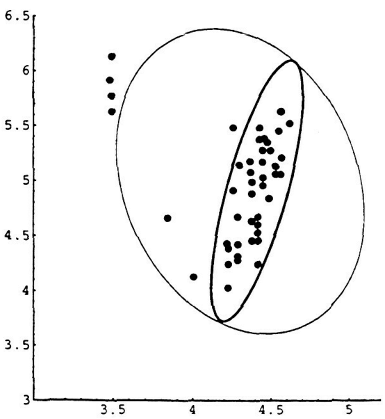
In addition to the above described supervised classification methods, we have been pursuing automated, unsupervised clustering of multispectral images in feature space. Although the output of such schemes should be further processed (to enable querying of the data by content), unsupervised clustering is of valid interest as it provides a "first cut" segmentation (to homogeneous regions) of a multispectral image without relying on a representative training set as a starting point.

Numerous unsupervised clustering algorithms have been proposed and applied to various problem domains. Unfortunately, many of the strategies for partitional clustering may result in methods that are iterative and that depend heavily on assumptions made with regard to underlying probability density functions, a priori numbers of clusters, etc. Also, these methods could be severely affected by outliers, i.e., contaminated data due to noise, encoding errors, etc. Pursuing a mode seeking approach, however, could alleviate most of the above phenomena. In particular, processing (hyper)histograms in a discrete, multidimensional space seems especially suitable for mode seeking-based clustering of multispectral images. Such an approach, however, requires careful consideration, as far as the memory and run-time constraints of storing and searching multidimensional histograms are concerned.

Alternatively, we adopt principles based on robust (statistical) estimation (RE) to construct a mode seeking clustering scheme in a continuous domain. (Inherent to robust estimators is their low sensitivity to outlying/noisy data which leads to much improved performance compared to classical estimators.) Specifically, we have employed a variant of the minimum volume ellipsoid (MVE) estimator which achieves the above objective by finding hyperellipsoids that "best" enclose subsets of the data in d-dimensional space. As demonstrated in Figure 3 (courtesy Rousseeuw & Leroy '87), an MVE estimator clearly achieves better results than does a maximum likelihood estimator (MLE), for example, in the presence of outliers.

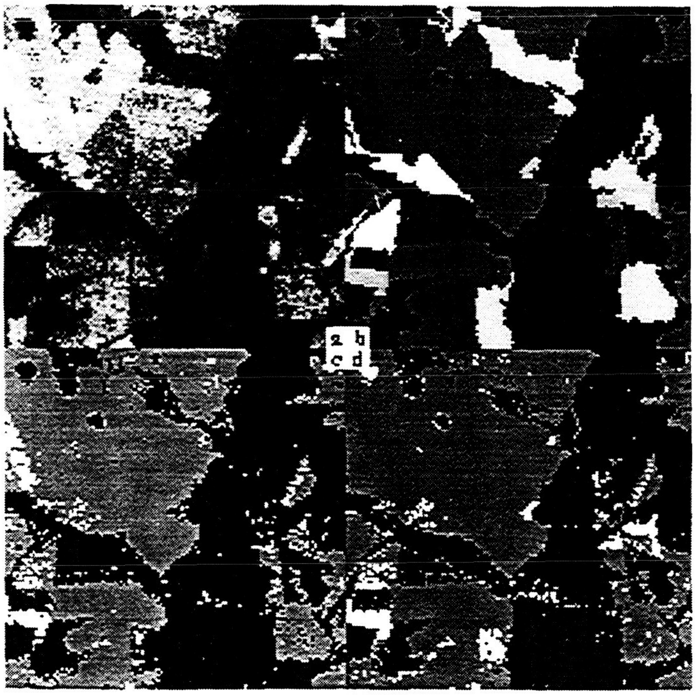
We have implemented an MVE-based algorithm and tested its performance on synthetic and real data. Preliminary empirical results for remotely sensed images (having 3 and 4 bands) show good promise with close to 70% accuracy rate versus corresponding ground truth. Also, we have demonstrated that combining the current clustering approach in feature space with a spatial region growing approach could enhance the overall performance of the RE-based scheme. Figure 4 below illustrates this idea.

Figure 3: MVE Estimator vs. MLE



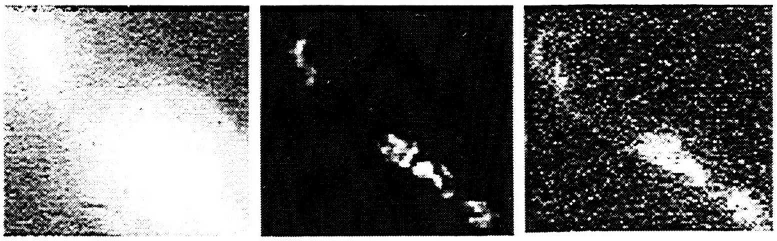
Finally, in an attempt to greatly reduce the computational burden for d >> 1, which is absolutely vital for processing images acquired by instruments such as Landsat-TM (d = 7), MODIS (Moderate-Resolution Imaging Spectrometer, d = 36), and even AVIRIS (Airborne Visible Infrared Imaging Spectrometer, d = 224), we are implementing a parallel version of the clustering scheme on the MasPar. ■

Figure 4: RE-based Clustering of Remotely Sensed Data



(a) band 5 of a 128 x 128 Landsat-TM (sub) image of Ridgely, MD, (b) its ground truth, (c) a result due to RE-based clustering, and (d) a result due to RE and region growing combined.

Figure 5: Hubble Space Telescope Faint Object Camera (HST/FOC) Ultraviolet Image of the R-Aquarii Symbiotic Star System



(a.) **(b.)** **(c.)**

(a.)Dec. 1992 Pre-Hubble Fix FOC Image (l=253 nm.)
(b.)Dec. 1992 Maximum Entropy Restored HST/FOC Image (l=253 nm.)
(c.)Sep. 1994 Post-Hubble Fix FOC Image (l=231 nm.)

Image Processing and Restoration
(continued from page 1)
low/high pass filtering, edge detection, pattern recognition, erosion and dilation etc. The algorithms are generally independent of the sensor system which accumulated the data and in some cases independent of the noise and the class of object under study.

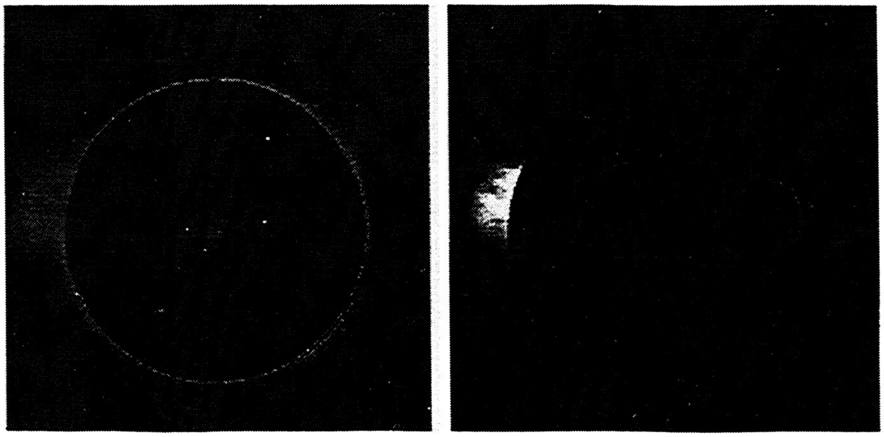
Model-based image processing seeks to combine, in an optimal fashion, all prior knowledge related to the problem under study. Prior knowledge refers to any and all known information about the optical response, detector response, imaging process, noise statistics and class of object being viewed. Some examples would be maximum entropy deconvolution of the system's optical response in the presence of noise, optimal detection of point sources in extremely noisy backgrounds with a statistical matched filter based on the optical response function and prior noise statistics.

A number of model-based image processing methods have been researched, developed and successfully applied to space flight data. The methods include maximum entropy image restoration, image deconvolution, spectral restoration and hyperspectral restoration methods. All these algorithms attempt to utilize all the prior knowledge related to the system and the sensor detection process to extract the maximum scientific content from the observed dataset. The image restoration algorithm (MEM) uses a maximum entropy algorithm with both convolution constraint and a flux normalization constraint. If the region of data which is corrupted is less than the size of the optical PSF, then MEM is very robust with respect to reconstructing the corrupted region. The image deconvolution algorithm (PMEM) is also a maximum entropy algorithm, but with an explicit maximum likelihood constraint and an implicit convolutional constraint along with flux conservation. PMEM attempts to find that object which has maximal entropy subject to the constraint that the residual noise distribution must match the a priori noise distribution. PMEM is very robust with respect to reconstruction of low signal to noise background structure. Also developed are phase retrieval algorithms to recover unknown optical parameters from observed data. The recovered optical parameters are used to model the optical system to generate calculated noise-free point spread functions to use in the image restoration algorithms. All the algorithms are currently coded in MPL on a MasPar MP-2 and are based on parallel FFT techniques.

Figure 5 is an example of a restoration of an ultraviolet Hubble Space Telescope (HST) Faint Object Camera image of the symbiotic star system R-Aquarii. The restoration was performed using a calculated optical point spread function. The leftmost image is the blurred noisy image taken in December 1992; the middle image is the result of applying a maximum entropy algorithm with maximum likelihood constraints. The rightmost image is the same object taken with the corrected telescope in 1994.

Figure 6 shows an example of applying maximum entropy techniques to a set of eight images of the solar corona. Four images are observed at the spectral line of Fe XIV (l=530.3 nm.) (line emission) and the other 4 images are observed slightly shifted off the spectral line (continuum emission). The line emission images contain both the stray light (scattered and diffracted) and the coronal features of interest and are dominated by the stray light. The continuum emission images contain only the stray light. The 4 line emission images are averaged and differenced from the average of the 4 continuum emission images. The difference image is the leftmost image of Figure 6. The rightmost image shows the result of applying maximum entropy to the set of 8 images to separate the line emission from the continuum emission. Shown is the separated line emission restored image.■

Figure 6: Preliminary Solar and Heliospheric Observatory (SOHO) Large Angle Spectrometric Coronagraph (LASCO) Image of the solar corona (data courtesy of Naval Research Laboratory - LASCO/SOHO)



(a.) Difference Image **(b.) Maximum Entropy Restored Imag**

mage Registration

continued from page 1)

)igital image registration, important in nany applications of image processing, uch as medical imagery, robotics, visual nspection, and remotely sensed data pro:essing is defined as the process which determines the most accurate match between wo or more images acquired at the same or t different times by different or identical ensors. Registration provides the "relaive" orientation of two images (or one imge and other sources, e.g., a map), with espect to each other, from which the absoute orientation into an absolute reference ystem can be derived. Currently, image egistration of remote sensing images is nost often accomplished by a human opertor "manually" selecting reference points 1 two images, and these points become the nput to compute the deformation model etween the two datasets. Our work on auomatic registration focuses on the speed of ch an algorithm and on its ability to hanle multi-sensor data. These two requireents brought us to the utilization of multisolution wavelet transforms to perform ch a task. Similarly to a Fourier transrm, wavelet transforms provide a timeequency representation of a signal, which n be inverted for later reconstruction. owever, the wavelet representation alws a better spatial localization as well as better division of the time-frequency ane than a Fourier transform, or than a indowed Fourier Transform. Choosing avelets to perform image registration is stified for the following reasons:

Iulti-resolution wavelets, largely used r compression and browing, are used as a st step to bring the multiple types of data the same resolution without losing sigficant information and without blurring e higher resolution data.

urther multi-resolution wavelet decomsition highlights strong image features at e lower resolution, thus eliminating weak gher resolution features.

he multi-resolution iterative search foses progressively towards the final transmation with a search interval decreasing d an accuracy increasing at each iteran. This algorithm achieves higher accuies with higher speeds than a full search full resolution.

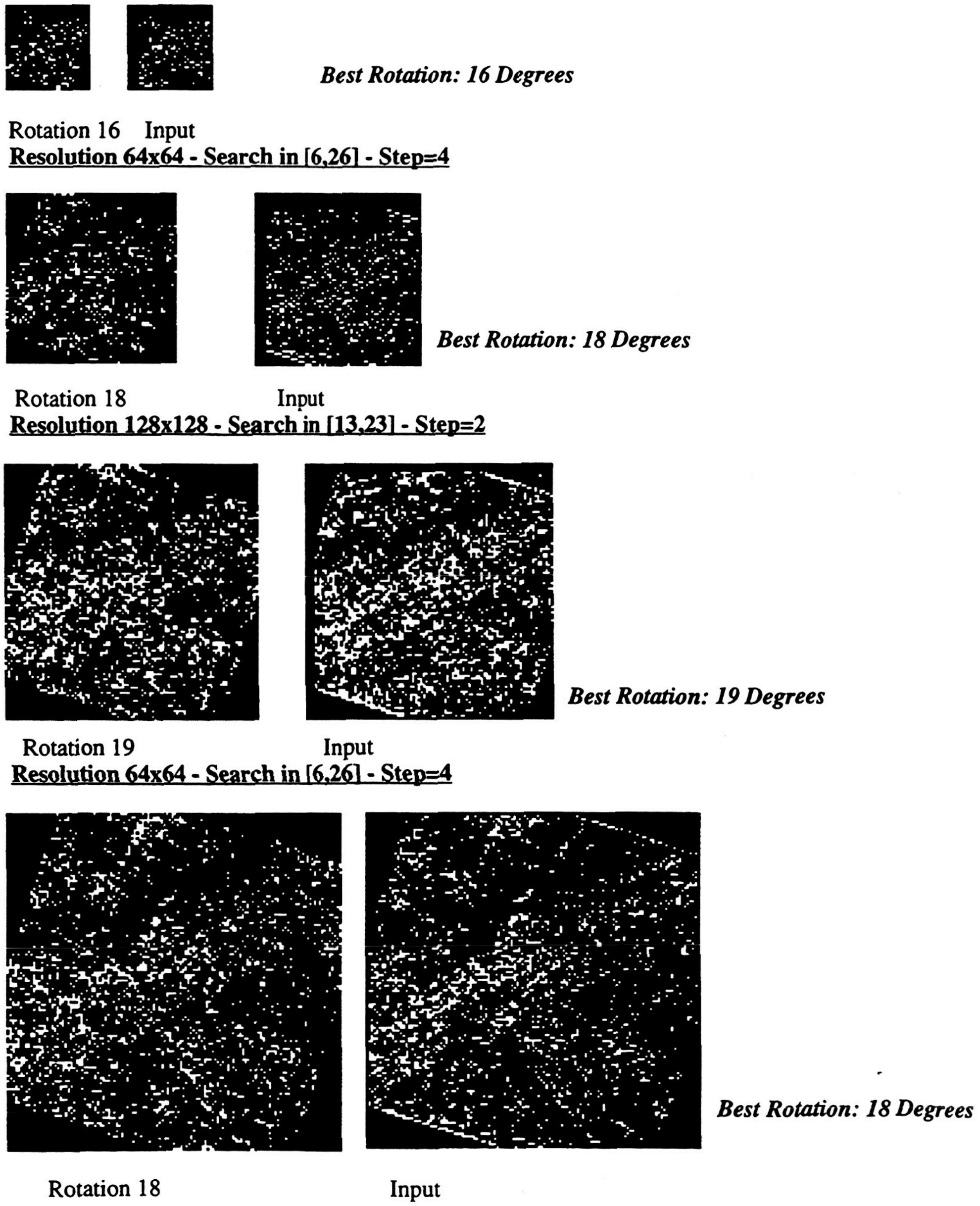
•Multi-resolution wavelet transforms can be implemented very easily on a massively parallel computer.

Our wavelet-based image registration algorithm utilizes the main features extracted by a Multi-Resolution Analysis (MRA) wavelet decomposition to perform an iterative registration of remote sensing images. These features are chosen as the maximum values of the high-frequency sub-bands of the wavelet decomposition. The algorithm searches for a composition of rotations and translations (will be extended to affine transformations). At first the set of all possible transformations is searched for in a small size image, and progressively the first approximations of the transformation are refined using larger and larger images. At each level, the search focuses on an interval around the "best" transformation found at the previous level, and the accuracy of the search increases when going from low resolution to high resolution. Results are illustrated in Figure 7.

The final step of this algorithm, still under development, will involve the use of this affine transformation to locate and match automatically a few accurate reference points in the high resolution data and refine locally the previous transformation.

Other research avenues are also being studied, which focus on the use of other features and other search strategies.■

Figure 7: Wavelet-Based Registration (Search for Rotations)



Resolution 32x32 - Search in [0,90] - Step=8 degrees

Best Rotation: 16 Degrees

Rotation 16 Input

Resolution 64x64 - Search in [6,26] - Step=4

Best Rotation: 18 Degrees

Rotation 18 Input

Resolution 128x128 - Search in [13,23] - Step=2

Best Rotation: 19 Degrees

Rotation 19 Input

Resolution 64x64 - Search in [6,26] - Step=4

Best Rotation: 18 Degrees

Rotation 18 Input

RECENT PUBLICATIONS OF THE IMAGE PROCESSING GROUP

Chan, A. K., Chui, C., Le Moigne, J., Lee, H. J., Liu, J. C., & El-Ghazawi, T. A. (1995). *The Performance Impact of Data Placement for Wavelet Decomposition of Two-Dimensional Image Data on SIMD Machines.* Proceedings of Frontiers'95 Fifth Symposium on the Frontiers of Massively Parallel Computation, McLean, VA, Feb. 6-9 (Also published as CESDIS Tech. Rep. No. 94-125).

El-Ghazawi, T., & Le Moigne, J. (1996). *Wavelet Decomposition on High-Performance Computing Systems.* 1996 International Conference on Parallel Processing (ICPP96).

El-Ghazawi, T., & Le Moigne, J. (1994, August). Multi-Resolution Wavelet Decomposition on the MasPar Massively Parallel System. *International Journal of Computers and their Applications, 1 (1),* 24-30 (Also published as CESDIS Tech. Rep. No. 94-122).

Hollis, J. M., Lyon, R. G., Dorband, J.E., Feibelman, W. A. (1997, January 20). Motion of the Ultraviolet R-AQUARII Jet, January 20, 1997, *Astrophysical Journal.*

Hollis, J. M., Lyon, R. G., Dorband, J. E., & Feibelman, W. A. (1995). Movie Showing Motion in the R Aqr Jet From October 1991 to October 1993 Using HST FOC Data. *B.A.A.S., 27,* 815.

Le Moigne, J., Campbell, W. J., & Cromp, R. F. 1996, June). *An Automated Parallel Image Registration Technique of Multiple Source Remote Sensing Data.* (Tech. Rep. No. 96-182). (Also submitted to IEEE Transactions on Geoscience and Remote Sensing).

Le Moigne, J., & Tilton, J. C. (1995, May). Refining Image Segmentation by Integration of Edge and Region Data. *IEEE Transactions on Geoscience and Remote Sensing, 33 (3),* 605-615.

Le Moigne, J. (1995). *Towards a Parallel Registration of Multiple Resolution Remote Sensing Data.* Proceedings IGARSS'95, Firenze, Italy, July 10-14.

Lyon, R. G., Dorband, J. E., & Hollis, J. M. (1995, July). Characterization of a Maximum Entropy Image Reconstruction Algorithm. *Proceedings of SPIE, 2564.*

Lyon, R. G., & Hollis, J. M., Dorband, J. E. A Maximum Entropy Method with A-Priori Maximum Likelihood Constraints, April 1, 1997, *Astrophysical Journal.*

Lyon, R.G., Dorband, J.E., Hollis, J.M. Hubble Space Telescope Faint Object Camera Calculated Point Spread Functions, March 10, 1997, *Journal of Applied Optics.*

Mount, D. M., Netanyahu, N. S., Silverman, R., & Wu, A. (1995, August 10-14). *Chromatic Approximate Nearest Neighbor Searching: A Query Sensitive Approach.* Proceedings of the Seventh Canadian Conference on Computational Geometry, Quebec City, Quebec, Canada.

Mount, D. M., & Netanyahu, N. S. (1994, July). Computationally Efficient Algorithms for High-Dimensional Robust Estimators. *Computer Vision Graphics and Image Processing - Graphical Models and Image Processing, 56 (4),* 289-303.

Netanyahu, N. S., Tilton, J. C., & Gualtieri, J. A. (1995, July 10-14). *Unsupervised, Robust Estimation-based Clustering of Remotely Sensed Images.* Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Florence, Italy.

Netanyahu, N. S., & Cromp, R. F. *Random Sampling of Remotely Sensed Images for Efficient Metadata Extraction.* (Internal ISTB Manuscript, April 1995).

Netanyahu, N. S., & Weiss, I. (1994, October 9-13). Analytic Outlier Removal in Line Fitting. *Proceedings of the Twelfth IAPR International Conference on Pattern Recognition, Volume II, Jerusalem, Israel* (pp. 406-408).

Short, N. M., Cromp, R. F., Campbell, W. J., Tilton, J. C., Le Moigne, J., Fekete, G., Netanyahu, N. S., Ligon, W., & Wichman, K. (1995, December). Mission to Planet Earth: AI Views the World. *IEEE Expert,* (Also published in AAAI Workshop on AI Technologies for Environmental Applications, Seattle, Washington, 1994, July 31-August 4, 1-15).

CESDIS
NASA Goddard Space Flight Center
Code 930.5
Greenbelt, MD 20771

APPENDIX B

CESDIS Seminars

Seminar Announcement

Wednesday, August 7th
Building 28, Room W230F - 1:30 p.m.
Hosted by Nabil Adam

How Reliable Can We Make Software?

Alfred V. Aho
Department of Computer Science
Columbia University

In the 1940's John von Neumann showed how we can make more reliable hardware from unreliable components by using redundancy and Claude Shannon showed how we can achieve more reliable communication over a noisy channel using error detecting and correcting codes. Today, redundancy and error detecting and correcting codes are routinely used to improve the reliability of hardware systems. This presentation will examine the question of how reliable can we make software systems and why hardware reliability techniques have not been successful in improving the reliability of software. This talk will also discuss software engineering techniques that are being used by leading software developers to improve the robustness of the software development process.

Alfred V. Aho became professor and chair of the Computer Science Department at Columbia University in 1995. From 1991 to 1995 he was General Manager of the Information Sciences and Technologies Research Laboratory at Bellcore in Morristown, New Jersey. The work of this laboratory was directed at advancing the national information networking infrastructure. From 1987 to 1991 he was Director of the Computer Science Research Center at AT&T Bell Laboratories, Murray Hill, New Jersey. Inventions of this center include the UNIX operating system and the C and C++ programming languages.

Dr. Aho received a B.A.Sc. in Engineering Physics from the University of Toronto and a Ph.D. in Electrical Engineering (Computer Science) from Princeton University. Upon graduating from Princeton, Dr. Aho joined Bell Laboratories in 1967 as a Member of Technical Staff in the Computer Techniques Research Department, and in 1980, was appointed Head of the Computing Principles Research Department. He has also been an adjunct professor of Computer Science at Stanford University and at the Stevens Institute of Technology.

Dr. Aho's personal research is centered on multimedia information systems, database systems and query languages, programming languages and their compilers, algorithms and the theory of computing. He has published more than sixty technical papers in these areas and ten textbooks that are widely used worldwide in computer science research and education. He is a coinventor of the AWK programming language and other UNIX system tools.

For further information regarding directions, access to NASA Goddard Space Flight Center, or meeting with Dr. Aho, please contact Georgia Flanagan at 301-286-2080

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Wednesday, June 11, 1997
Building 28, Room E210, 10:00 a.m.
Hosted by Bill Campbell & Jon Robinson (Code 935/Hughes STX)

Eagles on the GIS: Satellite Tracking Harpy Eagles and Mapping Their Rain Forests in Venezuela and Panama

Dr. Eduardo Alvarez-Cordero
Harpy Eagle Program, The Peregrine Fund

Focused on the world's most powerful eagle, the seven-foot wingspan Harpy, this program integrates biological and social approaches to protect the eagles and the biological diversity of the rain forests they inhabit. Besides the generalized threat of habitat loss, we have confirmed that many local populations of the Harpy Eagle are jeopardized by shooting mortality. Our international team works with the local people in partnership with various agencies in Venezuela and Panama to boost and demonstrate in-country capacity for conservation. The approach combines the latest computer-based technologies to collect and manage information, and distribute environmental knowledge.

Five years ago, with the help of NASA and NBS (National Biological Service) we pioneered the use of satellite telemetry in Latin America by tracking Harpy Eagles with ARGOS transmitters. We utilize another satellite-based technology, the Global Positioning System or GPS, to survey the regional habitat, creating digital maps of the layout of eagle's nests and the network of roads accessing the rain forest. By exchanging this information with government institutions we have pieced together the first computerized Geographic Information System (GIS) for the Harpy Eagle. As we regularly monitor the eagles and their nest sites, this digital database is constantly being updated, and the results shared with local forest managers. At a recent meeting to review the design of Mesoamerican Biological Corridors, the government of Panama presented a proposal based on our joint research on the Harpy Eagle.

Eduardo Alvarez-Cordero, the team's leader in his 6th year with The Peregrine Fund, recently completed his Ph.D. at the University of Florida. Based on the emerging concept of the virtual office, he coordinates the field projects from his home in Gainesville FL. Rafael Alvarez, another Venezuelan trained as the program's Field Manager, is based in Guri, and has become a world expert on these eagles.

The program's success is based on creative collaboration with select Research & Conservation Associates. Peter E. Kung (based in Logan, UT) is a biologist with extensive wildlife biology experience who contributes GPS surveying expertise. He has mapped over 2,500 km of roads and access trails cut into the most remote rain forests in southeastern Venezuela. Gustavo Martinez of Puerto Ordaz, Venezuela, provides information systems advice to squeeze top-level performance from the mixed bag of hardware and software available to the group. Luis Enrique (Kike) Arnal, a photographer an expert climber from Caracas, Venezuela initially delivered the on-rope training to climb the towering trees where the eagle's nest. For the past four years Kike has volunteered to create an impressive pictorial record of our work. The Harpy Eagle Team plans to expand their approach to several other Latin American countries. Their results are now the basis for information campaigns, ranging from in-country slide presentations at logging camps and indigenous schools to a series of formal lectures, posters, and publications in hard-copy and in the internet.

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Tuesday, July 2, 1996
Building 28, Room E210
11:00 a.m.
Hosted by Yair Amir

Measures for Performance Scalability in Networks of Workstations and Servers (NOWS)

Amnon Barak
Computer Science Institute
The Hebrew University
http://www.cs.huji.ac.il/mosix

This talk is about performance scalability by efficient resource utilization in a scalable Network of Workstations and Servers (NOWS). The main tool for performance scalability is a preemptive process migration, that could transparently move any process from one platform to another. This mechanism could be controlled by various algorithms, in order to take advantage of available network wide resources. For example, load-balancing and load-sharing algorithms could distribute the load evenly among the platforms to improve the overall performance. A memory sharing algorithm could allow a platform which has exhausted its main memory to use available free memory in another platform, instead of paging to a disk. Communication optimization algorithms could migrate communicating processes to a common site, to benefit from a shared memory communication. Yet other algorithms could be developed for network RAM and shared memory, by migrating a process to its data instead of bringing the data to the process.

Some of the above algorithms have already been implemented in MOSIX, a multicomputer enhancement of UNIX for NOWS. MOSIX is featuring preemptive process migration, dynamic load balancing and memory sharing. It is run on clusters of Intel x86, Pentium and Pentium-Pro based workstations, file and CPU servers. The machines in a cluster are connected by Ethernet, Fast Ethernet, and Myrinet. As a result, MOSIX shows a dramatic performance improvement in the execution of multiple processes over other networks of workstations. The presentation will describe the unique algorithms of MOSIX and its performance.

For further information regarding directions,
access to NASA Goddard Space Flight Center,
or meeting with Dr. Barak,
please contact Georgia Flanagan at 301-286-2080

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Monday, March 17, 1997
Building 28, Room E210
1:30 p.m.
Hosted by Yelena Yesha

The CCUBE Constraint Object-Oriented Database System: An Overview (Joint work with Victor Segal)

Alex Brodsky
Dept. of Information and Software Systems Engineering
George Mason University

Constraints provide a flexible and uniform way to conceptually represent diverse data capturing spatio-temporal behavior, complex modeling requirements, partial and incomplete information, etc., and have been used in a wide variety of application domains. Constraint databases have recently emerged to deeply integrate data captured by constraints in databases.

In this talk I will describe the development of the first constraint object-oriented database system, CCUBE, and the challenges in the area of constraint databases. The CCUBE system is designed to be used for both implementation and optimization of high-level constraint object-oriented query languages such as LyriC or constraint extensions of OQL, and for directly building software systems requiring extensible use of constraint database features.

The focal point of our work is achieving the right balance between expressiveness, complexity and representation usefulness, without which the practical use of the system would not be possible. To that end, CCUBE constraint calculus guarantees polynomial time data complexity, and, furthermore, is tightly integrated with monoid comprehensions to allow deep global optimization.

Dr. Brodsky is Assistant Professor of Information and Software Systems Engineering at George Mason University (GMU). His research interests include Constraint Databases and Programming, Database and Knowledge-base systems, and Spatio-temporal Information Systems. Alex came to GMU from IBM T.J. Watson research center. Prior to that, he worked for the Israeli Aircraft Industries and Israeli Defense Forces.

Alex received the B.Sc. in Mathematics and Computer Science from the Hebrew University in 1982, and the M.Sc. and Ph.D. in Computer Science from the Hebrew University in 1983 and 1991 respectively. He has received National Science Foundation Research Initiation Award, a grant from the Office of Naval Research, and Eshkol and Leibnitz fellowships.

Dr. Brodsky is an invited member of the ACM Strategic Directions in Computing Research Group in Constraint Programming, in which he represents the area of Constraint Databases, and in the group in Electronic Commerce and Digital Libraries. Dr. Brodsky's work on constraint databases led to the development of the first Constraint Object-Oriented Database System, CCUBE, implemented at GMU.

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Wednesday, December 18, 1996
Building 28, Room E210
2:00 p.m.
Hosted by Donald Becker

AC and the T3D

William Carlson *and* Jesse Draper
Supercomputing Research Center

We discuss our efforts to obtain the highest possible performance from the T3D during using our AC research compiler. As a node compiler, AC (which is based on the GNU C Compiler infrastructure) often produces code that is 20% to 300% faster than code produced by SCC. In addition, the flexibility of the compiler allows researchers to learn how to achieve the best node-level performance. AC also provides a mechanism for accessing memory on distant nodes of the T3D system using a simple pointer and array syntax extension. Experiments show promising potential for this mechanism providing an efficient whole-machine programming model. Currently, its performance surpasses optimized library-call communication.

William Carlson graduated from Worcester Polytechnic Institute in 1981 with a BS degree in Electrical Engineering. He then attended Purdue University, receiving the MSEE degree in 1983 and a Ph.D. in Electrical Engineering in 1988. From 1988 to 1990, Dr. Carlson was an Assistant Professor at the University of Wisconsin-Madison, where his work centered on performance evaluation of advanced computer architectures. Since 1990 he has been with the IDA Center for Computing Sciences, where his work focuses in the areas of operating systems, languages, and compilers for parallel and distributed computers.

Jesse Draper graduated from Rice University in 1972 with a B.A. in chemistry and from the University of Virginia in 1980 with a Ph.D. in English. From 1978-1987 he worked at the Institute for Computer Science and Technology of the National Bureau of Standards. Since 1987 he as been at the IDA Center for Computing Sciences, where his work has focused on languages and compilers for parallel computing.

For further information regarding directions,
access to NASA Goddard Space Flight Center,
or meeting with Drs. Carlson and Draper,
please contact Georgia Flanagan at 301-286-2080

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Thursday, August 15th
Building 28, Room E210 - 2:00 p.m.
Hosted by Yelena Yesha

A Global Ocean Model for Climate Change Applications on Massively Parallel Processors

Robert Chervin
National Center for Atmospheric Research (NCAR)*
Climate and Global Dynamics Division

The Parallel Ocean Program (POP) was developed at Los Alamos National Laboratory (LANL) by Smith, Malone and Dukowicz based on the earlier Parallel Ocean Climate Model (POCM) of Semtner and Chervin for application in distributed memory, massively parallel computing environments. The model uses second-order finite differencing in space and leapfrog finite differencing in time. An implicit technique is used for the free-surface calculation. The model uses a generalized curvilinear coordinate system which permits displacing the north pole to a location such that the convergence of meridians does not place an excessive restriction on the allowable time step.

A multi-institutional (LANL, JPL, NCAR and Cray Research, Inc.) effort was organized to implement and optimize POP for the Cray T3D. The aim was to create a single FORTRAN90 version of the model capable of executing well on either the CM-5 or the T3D. The major differences are contained in a small set of stencil routines which directly use the specific communication primitives for each machine. Other considerations included input/output and the use of the cache memory on the T3D.

One particular version of POP for climate change applications was derived from a previous 2/3 degree (on average) version and includes increased latitudinal resolution near the equator to resolve the strong tropical current systems. Ocean spinup and testing has begun on the 512 processing element T3D at the Pittsburgh Supercomputing Center. Initial results are promising.

Dr. Chervin is a research scientist in the Climate Change Research Section within the Climate and Global Dynamics Division at the National Center for Atmospheric Research, Boulder, Colorado. His research interests include the simulation of climate and climate change with physically based, three-dimensional models of the climate system, the application of objective statistical testing procedures to the analysis of such simulations, the causes of interannual variability, air-sea interaction and climate predictability. Recently, he has also developed a strong interest in supercomputing techniques to advance these research endeavors.

Dr. Chervin is a pioneer in the restructuring and conversion of atmospheric and ocean general circulation models for parallel execution on modern supercomputers. He has lectured extensively on this topic at a variety of research centers and universities, spanning eleven countries on four continents. He has received the Gordon Bell Award for Parallelism (honorable mention) and a Gigaflop Performance Award from Cray Research, Inc. and, most recently, the inaugural Cray Research Computerworld/Smithsonian Information Technology Leadership Award for Breakthrough Computational Science.

For further information regarding directions, access to NASA Goddard Space Flight Center, or meeting with Dr. Chervin, please contact Georgia Flanagan at 301-286-2080

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

**The National Center for Atmospheric Research is sponsored by the National Science Foundation.*

Seminar Announcement

Tuesday, August 6th
Building 28, Room W230F - 2:00 p.m.
Hosted by Milton Halem

Research and Development Opportunities in Information Services (IS)*

Mel Ciment
National Science Foundation

This talk will sketch information technology R&D industry trends; identify main sector drivers of competitiveness; identify key contributions of R&D to sector; and make R&D policy recommendations. A key finding is the recognition that there is a disconnect between the information services (IS) field from the information technologies R&D world, a disconnect which is cultural, educational and technological. Most industrial IRM/IS managers come from, and work, in cultures that do not value and invest in R&D. In the federal government, many agencies procure huge information systems associated with various non-technical functions agencies without the benefit of sufficient interaction with each other, or with the R&D community. This gap between the industrial IS segment and the R&D performers in the information technologies (IT) sector, is of some concern to vendors, users and R&D performers and government supporters of R&D in information technologies. The newly formed Applications Council of the National Science & Technology Council Committee on Computing, Information, and Communications is working with various federal bodies to address these issues.

* (The opinions expressed in this talk are solely the author's and do not necessarily represent the position or opinion of the National Science Foundation)

Dr. Melvyn Ciment, is Deputy Assistant Director, of the directorate for Computer and Information Science and Engineering (CISE), National Science Foundation (NSF). From October 1995-96, Dr. Ciment was on special assignment as a Visiting Scientist with the Department of Computer Science at the University of Maryland and served as a consultant to the Council on Competitiveness (COC) and contributed to a COC study on R&D in the Information Technologies sector. For nearly a year, he served as Acting Assistant Director until July 1, 1994. The CISE Directorate is responsible for the NSF High Performance Computing and Communications (HPCC) Program which incorporates the NSFNET and the four NSF supercomputer centers and research and infrastructure programs in: computer and computational research; information robotics and intelligent systems; advanced scientific computing; microelectronics information processing systems; networking and communications research, and infrastructure; and cross-disciplinary activities. These programs are a major portion of the NSF involvement in the Administration's initiative to advance the National Information Infrastructure. In May 1996, Dr. Ciment was appointed to Chair the newly formed Applications Council of the White House National Science & Technology Council's Committee on Computing, Information and Communications.

For further information regarding directions, access to NASA Goddard Space Flight Center, or meeting with Dr. Ciment, please contact Georgia Flanagan at 301-286-2080

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Wednesday, September 25, 1996
Building 28, Room E210 - 3:00 p.m
Hosted by Tarek El-Ghazawi

Origin and Evolution of Desert Landforms: The Case of the Western Desert of Egypt

Farouk El-Baz
Center for Remote Sensing
Boston University

The Western Desert of Egypt is the driest place on Earth, where it rains only once every 20 to 50 years. However, climate conditions were vastly different during the Pleistocene, where wet episodes alternated with dry periods. The basic landforms of the desert were created by fluvial action. Minor modifications were made by eolian action during dry climates. This is being revealed by detailed study of satellite images, particularly SIR-A and SIR-C images; the radar is capable of penetrating through dry sand to reveal buried courses of dry rivers and streams. This information is critical to the development of this desert, as well as for comparative studies of features on Mars.

Dr. Farouk El-Baz is Director of the Center for Remote Sensing at Boston University. He received a B.Sc. degree in chemistry and geology from Ain Shams University in Cairo, Egypt (1958), an M.S. degree in geology from Missouri School of Mines and Metallurgy (1961), and a Ph.D. degree in geology from the University of Missouri-Rolla after performing research at M.I.T. (1964).

He taught geology at Assiut University (1958-1960) and the University of Heidelberg, Germany (1964-1966). From 1967 to 1972 he participated in the Apollo program as Supervisor of Lunar Science Planning, where he served as secretary of the Lunar Landing Site Committee and chairman of astronaut training in visual observations and photography.

Starting in 1973 he established the Center for Earth and Planetary Studies of the Air and Space Museum of the Smithsonian Institution, Washington DC, which he directed during the next 9 years. In 1982, he became Vice President of Science and Technology at Itek Optical System, and in 1986 he joined Boston University. He served as Science Advisor (1978-1981) to the late Anwar Sadat, President of Egypt. His research interests center on the applications of remote sensing to archaeology, geography and geology.

*For further information regarding directions,
access to NASA Goddard Space Flight Center,
or meeting with Dr. El-Baz,
please contact Georgia Flanagan at 301-286-2080*

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Tuesday, January 28th, 1997
Building 28, Room E210, 10:30 a.m.
Hosted by Dr. Milt Halem

A Revolutionary General Extension for Nonlinear Kalman Filtering

Simon Julier
IDAK Industries

In this talk I will describe a radical alternative to the extended Kalman filter (EKF) that is probably more accurate, far easier to implement, and potentially much faster. The method allows arbitrarily complex nonlinear transformations to be applied directly to mean and covariance estimates without the need for linearization, i.e., without the derivation of Jacobians. As a consequence, the method is applicable to non-differentiable (including discontinuous) functions that simply cannot be used with the EKF.

This alternative approach, called Unscented Filtering, is having a dramatic impact on control and estimation applications around the world. Surprisingly, its use has spread less quickly in the US and is presently restricted to a handful of universities and a couple of aerospace companies. Throughout Europe and Australia, however, the new filtering algorithm is being used in a wide spectrum of applications including satellite attitude determination, fault detection, self-localization for autonomous vehicles, star tracking, and many others.

In this talk I will argue that the new filtering algorithm renders linearization/EKF approaches completely obsolete. I strongly encourage anyone who uses EKF-based methods to attend.

Simon Julier recently completed his doctoral studies at the University of Oxford, UK, where he studied the application of the new filtering algorithm to ultra-high fidelity vehicle models. In addition to his theoretic contributions to the field of nonlinear estimation, he has also developed a variety of very useful practical tools for evaluating and tuning nonlinear filters. He is now employed by IDAK Industries as head of their Advanced Systems branch.

For further information regarding directions,
access to NASA Goddard Space Flight Center,
or meeting with Dr. Julier,
please contact Georgia Flanagan at 301-286-2080.

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.htmlCESDIS

Seminar Announcement

Friday, August 9th
Building 28, Room E210 - 10:30 a.m.
Hosted by Jacqueline Le Moigne

A Poly-Algorithmic Approach for Achieving Scalable Performance on MPPs

Ashfaq A. Khokhar
Department of Electrical Engineering *and*
Department of Computer and Information Sciences
University of Delaware

Parallel computers hold enormous promise for achieving high performance at a reasonable cost for many application areas. However, exploiting the potential of current parallel machines requires a detailed understanding of parallel algorithms and programming, as well as an intimate knowledge of the underlying architecture. A strategic question in designing algorithms and software for parallel systems is how to achieve high performance without requiring users to have extensive parallel processing expertise or, conversely, how to make parallel computers easy to use without sacrificing performance.

In this talk we will present our research efforts geared towards achieving the above goal for the application area of computer vision and image processing. The goal of our research is to develop the means of providing better usability of high performance computing systems and to develop tools that operate seamlessly over a wide range of computing and communications platforms. A poly-algorithmic approach will be presented for achieving wide range of scalability on such platforms. We will substantiate the approach by presenting implementation results for a variety of application algorithms including basic communication primitives, spatial filtering of images, list ranking (contour ranking) of image edge maps, and 2-D FFT and Wavelet Transforms.

Ashfaq A. Khokhar received his B.S. in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 1985 and his M.S. in computer engineering from Syracuse University, in 1989. He received his Ph.D. in computer engineering from University of Southern California, in 1993. After his Ph.D., he spent two years as a Visiting Assistant Professor in the Department of Computer Sciences and School of Electrical and Computer Engineering at Purdue University. He joined University of Delaware in 1995, where he is Assistant Professor in the Department of Electrical Engineering and Department of Computer and Information Sciences. His research interests include parallel algorithms and architectures, parallel computation models, and high performance computing for computational geometry, image understanding, and multimedia applications.

For further information regarding directions, access to
NASA Goddard Space Flight Center, or meeting with Dr. Khokhar,
please contact Georgia Flanagan at 301-286-2080

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Tuesday, February 11th, 1997
Building 28, Room E210, 10:00 a.m.
Hosted by Dr. Nabil Adam

Image Restoration from Atmospheric Blur and Its Application to Satellite Imagery

Norman S. Kopeika
Department of Electrical and Computer Engineering
Ben-Gurion University of the Negev

Atmospheric blur derives from atmospheric turbulence and from small angle forward scattering of light by aerosols. Each is characterized by it's own modulation transfer function (MTF). The first phenomenon causes random wavefront tilt typically on the order of tens and hundreds or microradians. Blur from the second phenomenon derives from random scatter at angles typically on the order of tens, hundreds, and thousands of microradians. Scatter at larger angles is manifested as attenuation and gives rise to atmospheric path radiance which varies with wavelength and weather conditions. This latter phenomenon causes difficulties in comparisons of satellite imagery of surface detail and characteristics at different wavelengths and at different times. An atmospheric Wiener filter has been developed which corrects simultaneously for all three atmospheric degradations - turbulence blur, aerosol blur, and atmospheric path radiance. Average turbulence MTF and aerosol MTF used to describe average atmospheric MTF. The jitter or variance of turbulence MTF describes the variance of atmospheric MTF and is considered as a source of noise in addition to the usual instrumentation noise. Implementation of the atmospheric Wiener filter with a pc results in enhancement of imagery at high spatial frequencies, but selectively, so that enhancement at high spatial frequencies characterized by higher turbulence "noise" and instrumentation noise is less than at other high spatial frequencies. The atmospheric MTF after restoration is much broader and higher at high spatial frequencies than that without the atmospheric Wiener filter restoration. The MTF broadening is manifested in resolution of much small detail. Restoration is essentially complete. The MTF increase at high spatial frequencies is manifested as improved contrast resulting from atmospheric path radiance correction. Examples of implementation to NOAA AVHRR satellite imagery are presented, where turbulence MTF is evaluated from weather data.

For further information regarding directions,
access to NASA Goddard Space Flight Center,
or meeting with Dr. Kopeika,
please contact Georgia Flanagan at 301-286-2080.

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Tuesday, June 24, 1997
Building 28, Room E210, 10:00 a.m.
Hosted by Jacqueline Le Moigne

On Information Extraction Principles for Hyperspectral Data

David Landgrebe
Purdue University

Means for optimally analyzing hyperspectral data has been the topic of a study since 1986 . The point of departure for this study has been that of signal theory and the signal processing principles that have grown primarily from the communication sciences area over the last half century. The basic approach has been to seek a more fundamental understanding of high dimensional signal spaces in the context of the remote sensing problem, and then to use that knowledge to extend the methods of conventional multispectral analysis to the hyperspectral domain in an optimal or near optimal fashion. The purpose of this presentation is to outline what has been learned so far in this effort.

The introduction of hyperspectral sensors which produce much more complex data than those previously should provide much enhanced abilities to extract useful information from the data stream they produce. However, it is also the case that this more complex data requires more complex and sophisticated data analysis procedures if their full potential is to be achieved. Much of what has been learned about the necessary procedures is not particularly intuitive, and indeed, in many cases is counter-intuitive. In this presentation, we shall attempt not only to illuminate some of these counter-intuitive aspects, but to make them understandable and therefore acceptable.

Dr. Landgrebe holds the BSEE, MSEE, and PhD degrees from Purdue University. He is presently Professor of Electrical and Computer Engineering at Purdue. His area of specialty in research is communication science and signal processing, especially as applied to Earth observational remote sensing. His contributions over the last 25 years in that field have related to the design from a signal processing point of view of multispectral imaging sensors, suitable spectral and spectral/spatial analysis algorithms, methods for designing and training classifier algorithms, and overall systems analysis. He was one of the originators of the multispectral approach to Earth observational remote sensing in the 1960's, was instrumental in the inclusion of the MSS on board Landsat 1, 2, and 3, and hosted and chaired the NASA meeting at which the bands and other key parameters were selected for the Thematic Mapper. He has been a member of a number of NASA and NRC advisory committees for this area since the 1960's.

He was President of the IEEE Geoscience and Remote Sensing Society for 1986 and 1987 and a member of its Administrative Committee from 1979 to 1990. He received that Society's Outstanding Service Award in 1988. He is a co-author of the text, Remote Sensing: The Quantitative Approach, and a contributor to the book, Remote Sensing of Environment, and the ASP Manual of Remote Sensing (1st edition). He has been a member of the editorial board of the journal, Remote Sensing of Environment, since its inception.

Dr. Landgrebe is a Fellow of the Institute of Electrical and Electronic Engineers, a Fellow of the American Society of Photogrammetry and Remote Sensing, and a member of the American Society for Engineering Education, as well as Eta Kappa Nu, Tau Beta Pi, and Sigma Xi honor societies. He received the NASA Exceptional Scientific Achievement Medal in 1973 for his work in the field of machine analysis methods for remotely sensed Earth observational data. In 1976, on behalf of the Purdue's Laboratory for Applications of Remote Sensing which he directed, he accepted the William T. Pecora Award, presented by NASA and the U.S. Department of Interior. He was the 1990 individual recipient of the William T. Pecora Award for contributions to the field of remote sensing. He was the 1992 recipient of the IEEE Geoscience and Remote Sensing Society's Distinguished Achievement Award.

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Thursday, July 11, 1996
Building 28, Room W230F
1:30 p.m.
Hosted by Yelena Yesha

High Throughput Computing on Clusters of Workstations

Miron Livny
Department of Computer Science
University of Wisconsin at Madison
miron@cs.wisc.edu

For many experimental scientists, scientific progress and quality of research are strongly linked to processing capacity. While some of them have to rely on exotic computing hardware, the computing needs of most scientists can be satisfied by commodity CPUs and memory. For more than a decade we have been developing, implementing, and deploying, software tools that can efficiently and effectively harness the capacity of hundreds of workstations. The workstations can be scattered throughout the globe and may be owned by different individuals, groups, or institutions. Using our software, a scientist may simultaneously and transparently exploit the capacity of workstations he/she is not even aware exist.

Our tools are based on a novel layered approached to Resource Allocation and Management in a Meta-Computing environment. Depending on the characteristics of the application, the user, and/or the environment, different layers are employed to provide a comprehensive set of Resource Management services. In the talk we will outline the overall architecture of the software tools we developed, and discuss the interaction between the different layers. We will outline the principals that have been guiding our work and the lessons we have learned from deploying our tools in real-life production environments. The role of resource owners, system administrators, and resource consumers, in such environments will be discussed.

Miron Livny received a B.S. degree in Physics and Mathematics in 1975 from the Hebrew University and M.Sc. and Ph.D. degrees in Computer Science from the Weizmann Institute of Science in 1978 and 1984, respectively. Since 1983 he has been on the Computer Sciences Department faculty at the University of Wisconsin-Madison, where he is currently a Professor of Computer Sciences.

Dr. Livny's research focuses on data management, visualization, and exploration systems, and meta computing. His recent work includes the Condor meta-computing environment, the DEVise data visualization and exploration system, the ZOO experiment management environment, quality controlled lossy image compression, processing data on tapes, and data clustering.

For further information regarding directions,
access to NASA Goddard Space Flight Center,
or meeting with Dr. Livny,
please contact Georgia Flanagan at 301-286-2080

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Friday, June 6, 1997
Bldg. 28, Room W230F, 1:30 p.m.
Hosted by Donald Becker

A Scalable Library For Pseudorandom Number Generation: Theory and Practice

Michael Mascagni
University of Southern Mississippi
(http://www.ncsa.uiuc.edu/Apps/CMP/RNG/mascagni)

Providing high quality pseudorandom numbers for parallel computers supplies many deep and fascinating mathematical problems as well as unique software engineering challenges. One of the more practical issues in parallel pseudorandom number generation is finding methods that provide portability and reproducibility across architectures. We summarize some recent developments in the design and analysis of pseudorandom number generators for parallel computers that are portable and reproducible. These results are the basis for a DARPA sponsored project for the development of a scalable library for pseudorand om number generation that is based at the University of Illinois, Urbana-Champaign. It is hoped that this scalable library will be the seed for a more comprehensive problem solving environment for Monte Carlo computations on scalable platforms.

Dr. Michael Mascagni is the Coordinator of the Ph.D. program in Scientific Computing, and Associate Professor of Mathematics at the University of Southern Mississippi (http://usm.edu). In addition, he is running the Programming Environment and Training academic program at the Naval Oceanographic Office's Major Shared Resource Center center at the Stennis Space Center (http://www.navo.hpc.mil). Dr. Mascagni recieved his Ph.D. in Mathematics from NYU's Courant Institute of Mathematical Sciences. He then moved to DC as a National Research Council Post-Doctoral Fellow at the National Institutes of Health in Bethesda. Following, he accepted a staff position at the Institute for Defense Analyses's new Supercomputing Research Center (now called the IDA Center for Computing Sciences) in Bowie, Maryland. Recently, he moved to Hattiesburg, Mississippi. His research interests are Monte Carlo methods, numerical analysis, parallel computing, random number generation, and scalable libraries.

*For further information regarding directions,
access to NASA Goddard Space Flight Center,
or meeting with Dr. Mascagni,
please contact Georgia Flanagan at 301-286-2080*

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Thursday, August 1st
Building 28, Room W230F - 10:30 a.m.
Hosted by Yelena Yesha

Correspondence and Translation for Heterogeneous Data

Tova Milo
Computer Science Department
Tel-Aviv University

A primary motivation for new database technology is to provide support for the broad spectrum of multimedia data available notably through the network. These data are stored under different formats: SQL or ODMG (in databases), SGML or LaTex (documents), DX formats (scientific data), Step (CAD/CAM data), etc. In this work, we provide a formal foundation to facilitate the integration of such heterogeneous data and the maintenance of heterogeneous replicated data.

Our solution is based on using a simple *middleware* data model that serves as a basis for the integration task, and *declarative rules* for specifying the integration. A main contribution is in identifying cases where one declarative set of rules can be used for the full specification of an integration task (derivation of correspondences, transformation of data from one world to the other, incremental integration of a new bulk of data from one world or the other). The talk will discuss both the theoretical foundation and the implementation of a system based on it.

Joint work with Serge Abiteboul and Sophie Cluet.

Tova Milo is currently a Prof. in the Computer Science Department at Tel-Aviv University. She holds a PhD from the Hebrew University in Jerusalem. Her main interests include databases and electronic documents, object oriented databases, heterogeneous databases, database theory, and database languages and models.

For further information regarding directions, access to NASA Goddard Space Flight Center, or meeting with Dr. Milo, please contact Georgia Flanagan at 301-286-2080

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Tuesday, August 13th
Building 28, Room W230F - 1:30 p.m.
Hosted by Jacqueline Le Moigne

Wreath Product Group-Based Correlation Applications to Problems in Signal Processing

Gagan Mirchandani
Department of Computer Science & Electrical Engineering
University of Vermont, Burlington

This talk builds on work done by others on Wreath Product (WP) Groups and their potential application to problems in signal processing. A brief introduction to Wreath Product Groups, and their Spectral Representation will be given. For the cyclically based WP-group, the associated multiresolution spectrum, it's invariance properties and a fast transform will be described. WP-based convolution will be introduced as also it's extension to correlation.

Many simulation results with WP-based correlation and standard correlation applied to image data will be shown. Possible use in image coding and a 2-D WP-based transform will be described.

Gagan Mirchandani is a Professor in the Department of Computer Science & Electrical Engineering, The University of Vermont, Burlington, VT. His field of interest is Digital Signal Processing with current applications in Image Coding and Registration.

Professor Mirchandani obtained his B.Sc, M.Sc and Ph.D from Worcester Poly, Syracuse and Cornell respectively, all in Electrical Engineering.

For further information regarding directions, access to NASA Goddard Space Flight Center, or meeting with Dr. Mirchandani, please contact Georgia Flanagan at 301-286-2080

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Wednesday, January 15, 1997
Building 28, Room E210, 1:30 p.m.
Hosted by Jacqueline Le Moigne

On Discrete and Discrete-Time Wavelets with Optimum Time-Frequency Resolution

Joel M. Morris
Computer Science and Electrical Engineering Department
University of Maryland Baltimore County
http://engr.umbc.edu/~itl/EE/Faculty/Morris/Morris.html

The Wavelet Transform (WT), due to its constant relative bandwidth (Q) property, provides good time-resolution for high-frequency signal components and good frequency resolution for low-frequency signal components. Moreover, it is possible to construct an orthonormal wavelet basis with good time-frequency resolution. Until recently, however, little attention has been paid to the design of orthonormal wavelets with optimum time-frequency resolution. The first treatment was limited to Daubechies' family of orthonormal wavelets, i.e., minimum phase and least asymmetric wavelets.

We have recently proposed new classes of compactly-supported orthonormal wavelets with optimum time-frequency resolution. The approach is to search over all orthonormal wavelet bases of L2(R)and l2 generated by a real FIR filter of length N (i.e., among all possible orthonormal wavelets generated by multiresolution analysis). The optimization problem was formulated as a constrained optimization problem whose global solution was obtained numerically via adaptive simulated annealing. We will present results for discrete and discrete-time wavelets that were optimized for the three performance measures of minimum time-duration, minimuin bandwidth, and minimum time-bandwidth product.

*For further information regarding directions,
access to NASA Goddard Space Flight Center,
or meeting with Dr. Morris
please contact Georgia Flanagan at 301-286-2080*

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Monday, June 16, 1997
Building 28, Room E210, 1:30 p.m.
Hosted by Rick Lyon (UMBC/CESDIS)

Signal Processing in Optical Time Domain Reflectometer

Timothy Murphy
GN Nettest

This presentation discusses the motivation and methods for testing communication optical fibers. The optical time domain reflectometer provides a distance dependent picture of the transmissive and reflective properties of the fiber by launching a pulse of light down the fiber and observing the returned signal. Besides showing the location of reflections, the signal also displays non-reflective losses as drops in the backscatter level. The signal processing software accompanying the test hardware interprets the signal for the user by providing a table of reflective and lossy "event" locations and magnitudes. Certain aspects of the signal become easier to detect when its wavelet decomposition is examined.

Timothy Murphy is currently a Systems Engineer for GN Nettest, Utica NY. Mr. Murphy has approximately 14 years experience in the instrumentation and aerospace fields and has worked on a diversity of projects including FIR and IIR filters, electro-optic characterization and testing of large format CCD arrays, rapid implementation of DSP technology for Fast Fourier Transforms (FFTs), CCD scanner for DNA fingerprinting, analysis of electro-optic performance of absorption and Raman spectroscopy detectors, and radiation shielding for space flight star tracker CCDs. Mr. Murphy currently holds an MS and a BS in Electrical Engineering from Columbia University.

For further information regarding directions,
access to NASA GSFC, or meeting with
Tim Murphy, please contact Georgia Flanagan
at 301-286-2080 or georgia@cesdis.usra.edu

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Tuesday, July 16, 1996
Building 28, Room W230F
10:30 a.m.
Hosted by Yelena Yesha

People, Interfaces and Systems

Anthony Norcio
Computer Science Department
University of Maryland, Baltimore County
http://umbc7.umbc.edu/~norcio

This talk will discuss a series of studies that have examined the critical issues concerning people in complex systems and intelligent interfaces. With respect to complex systems, an overview and experimental results of two projects will be presented; the first is the SCR project and the second is an experimental project which used a CASE tool with a dynamic simulator. With respect to intelligent interfaces, the use of neural networks and fuzzy logic has been examined as underling methodologies for dynamic user modeling and user classification in adaptive interfaces. An overview of these studies and their results will be presented. Finally, future directions and applications of this work in the areas of multimedia environments, intelligent tutoring systems, and voice systems will be discussed.

Dr. Norcio's research interests are in the general areas of intelligent human-computer interfaces and software design. His current work focuses on designing and constructing cognitive/performance users models that can form the rules that underlie adaptive interfaces to complex software environments. Dr. Norcio has also studied alternative design methodologies for specifying complex software systems. This work examined design, code, test, and change data for constructing formal specifications of information hiding modules. He has also examined the cognitive processes that are involved in designing, comprehending, and maintaining software systems. The purpose of these studies was to identify the underlying processes that transcend the syntax of any specific programming language.

Dr. Norcio regularly serves on planning and program committees for national and international conferences. He has also served as the Scientific Advisor to the Computer Science Division of the Office of Naval Research and currently serves as a Computer Scientist at the Naval Research Laboratory.

Dr. Norcio also directs the USER Lab (User System Environment Research Laboratory) and has an extensive list of publications.

*For further information regarding directions,
access to NASA Goddard Space Flight Center,
or meeting with Dr. Norcio,
please contact Georgia Flanagan at 301-286-2080*

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Tuesday, April 29, 1997
Building 28, Room W230F
10:30 a.m.
Hosted by Milt Halem

A GRAPHIC VOYAGE through COLOR & FORM

Unraveling the Creative Process by Giving Technology a Soul

Arthur Secunda

Where I live and work makes me think about nature's elements and what corresponds to their equivalents in terms of tactility, viscosity, density and reality. The qualities of the materials we normally use in art to express wind, sun, rain, rocks, earth, flame often lack the *sincerity* that these timeless elements merit. So how do we load them with the power to communicated the truth of what we attempt to express?

In the age when we are bombarded with too many banal images, even the rendering of flesh, hair and eyeballs requires an inventive or provocative presentation in order to engage the viewer with authority. Originality is the heart and soul of the matter.

To be memorable, the inner life of an image (its forms and colors) have to be explored, created, and recreated abstractly to a point where the materials used speak as one with the structure and content. A perfect marriage occurs when the colors, the "pigment", and the subject are so interwoven that they are inseparable and project a clear statement of the artists intent.

I seek to incorporate new, incongruous shapes, forms and materials in my own art in order to articulate the truth of my formal language and style. Though I believe in *taste* and a *meaning*, even *narrative*, these descriptions belong to the critic rather than the artist. The *spirit* of a landscape or the *inner expression* in a gesture is akin to discovering the "beat" or the soul in art.

Using new technologies and the cornucopia of materials and media available is not only challenging, it takes on a life of its own and teaches on several levels simultaneously. I like to combine materials. I need to mix the media: I utilize paper with rocks, epoxy with photos, sand with acrylic; I draw, computerize, digitize, paint, tear and glue. The creative experience comes about by blending my heart to my hand, my senses, my mind and my materials. It is a challenge that I joyfully accept as being all in a days work (or a lifes) work.

Dadaist Tristan Tzara said that philosophy is a question of whether life is seen from the standpoint of *God* or from an *Idea*. He said that everything is unreal until we designate it with our sense of reality. This is what I try to do in my art - integrate a personal, visible embodiment to what I see and feel around me.

For further information regarding directions,
access to NASA Goddard Space Flight Center,
or meeting with Mr. Secunda,
please contact Georgia Flanagan at 301-286-2080

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Monday, July 29, 1996
Building 28, Room E210 - 11:00 a.m.
Hosted by Jacqueline Le Moigne

Retrieval by Content in Symbolic-Image Databases

Aya Soffer
Computer Science and Electrical Engineering Department
University of Maryland, Baltimore County
http://www.cs.umbc.edu/~soffer/

Methods for integrating symbolic images into the framework of a database management system will be described in this talk. Symbolic images are images where the set of objects that may appear in them is known a priori, and where these objects are represented by graphical symbols. We propose two approaches for storing and indexing symbolic images in order to support retrieval by content of these images. The classification approach preprocesses all images and attaches a semantic classification and an associated certainty factor to each object found in the image. The abstraction approach describes each object in the image by using a vector consisting of the values of some of its features (e.g., shape, genus, etc.). For each approach, we describe methods to input, store, index, and query symbolic images using that approach.

A method for specifying pictorial queries to a symbolic image database will be described in detail. A pictorial query specification consists of a query image and a similarity level that must hold between the query image and database images. The similarity level specifies the *contextual similarity* (how well does the content of one image match that of another) as well as the *spatial similarity* (the relative locations of the matching symbols in the two images). Algorithms for retrieving all database images that conform to a given pictorial query specification will be presented.

A system that we have developed which uses these methods to perform retrieval by content of map images will be demonstrated. An example query to this is system is "find all map images containing camping sites within 3 miles of fishing sites".

Aya Soffer received the B.S degree in computer science from the Hebrew University of Jerusalem in 1986, and the M.S and Ph.D degrees in computer science from the University of Maryland at College Park in 1992 and 1995, respectively. She is currently a research assistant professor in the computer science and electrical engineering department at the University of Maryland Baltimore County. She also has an appointment as a research scientist at the Center of Excellence in Space Data and Information Sciences (CESDIS), at Goddard Space Flight Center and at the Center for Automation Research (CfAR) at the University of Maryland College Park. Her research interests include pictorial information systems, document analysis and recognition, digital libraries, spatial databases, geographic information systems, and non-traditional database systems.

For further information regarding directions, access to
NASA Goddard Space Flight Center, or meeting with Dr. Soffer,
please contact Georgia Flanagan at 301-286-2080

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

Seminar Announcement

Monday, January 27th, 1997
Building 28, Room E210, 1:30p.m.
Hosted by Donald Becker

TreadMarks: Shared Memory Computing on Networks of Workstations

Willy Zwaenepoel
Department of Computer Science
Rice University

By building on existing infrastructure, networks of workstations (NOWs) provide a low-cost, low-risk entry into the parallel computing arena. Furthermore, using a shared memory programming model, existing sequential codes can be parallelized with much less programmer effort than when using message passing. The combination of the two, shared memory programming on a network of workstations, thus provides excellent leverage both for hardware and software investments. It also provides some measure of portability between SMPs and NOWs.

We have developed a runtime package, called TreadMarks, that provides a shared memory image to processes executing on different workstations. The package is relatively portable, and runs on most common Unix platforms (DEC, HP, IBM, Intel, SGI, and SUN). No kernel modifications or special privileges are required to run TreadMarks programs. C, C++, and Fortran are supported, using standard compilers and linkers.

While the appeal of shared memory programming has been well known for some time, early software implementations have suffered from poor performance due to excessive communication. We have developed a number of techniques to address the communication problem. In this talk I will discuss the two principal techniques: lazy release consistency and multiple-writer protocols, and contrast them to the sequential consistency and single-writer protocols used in conventional systems.

I will demonstrate the programmability and efficiency of TreadMarks by discussing the parallelization of a couple of applications, one from operations research (mixed integer programming) and one from computational biology (genetic linkage analysis). The latter was recently used in the discovery of a linkage for Parkinson's disease by researchers at NIH.

Willy Zwaenepoel received the B.S. from the University of Gent, Belgium, in 1979, and the M.S. and Ph. D. from Stanford University, in 1980 and 1984. He is a currently a Professor of Computer Science and of Electrical and Computer Engineering at Rice University, where he has been on the faculty since 1984. His interests are in all aspects of parallel and distributed computing.

*For further information regarding directions,
access to NASA Goddard Space Flight Center,
or meeting with Dr. Zwaenepoel,
please contact Georgia Flanagan at 301-286-2080.*

http://cesdis.gsfc.nasa.gov/admin/cesdis.seminars/seminar.html

APPENDIX C

CESDIS Technical Reports

Anurag Acharya, University of Maryland College Park

TR-96-177	**Tuning the Performance of I/O -Intensive Parallel Applications**	**Anurag Acharya, Joel Saltz, Alan Sussman**	**March 1996**

Getting good I/O performance from parallel programs is a critical problem for many application domains. In this paper, we report our experience tuning the I/O performance of four application programs from the areas of satellite-data processing and linear algebra. After tuning, three of the four applications achieve application-level I/O rates of over 100 MB/s on 16 processors. The total volume of I/O required by the programs ranged from about 75 MB to over 200 GB. We report the lessons learned in achieving high I/O performance from these applications, including the need for code restructuring, local disks on every node and knowledge of future I/O requests. We also report our experience in achieving high performance on peer-to-peer configurations. Finally, we comment on the necessity of complex I/O interfaces like collective I/O and strided requests to achieve high performance.

Nabil Adam, Rutgers University

TR-97-190	**Electronic Commerce and Digital Libraries: Towards a Digital Agora**	**Nabil Adam, Yelena Yesha**	**January 1997**

Electronic commerce (EC) and digital libraries (DL) are two increasingly important areas of computer and information sciences with different user requirements but similar infrastructure requirements. In exploring strategic directions, we examine both requirements of the global information infrastructure that are necessary prerequisite for EC and DL [2], and specific requirements of EC and DL within the global infrastructure.

Both EC and DL are concerned with systems that support the creation of information sources and with the movement of information across global networks. EC supports effective and efficient business interactions and transactions that take place on behalf of consumers, sellers, intermediaries, and producers, while DL supports effective and efficient interaction among knowledge seekers. A digital library may require the transactional aspects of EC to manage the purchasing and distribution of its content while a digital library can be used as a resource in electronic commerce to manage products, services, providers and consumers. EC and Dl share a common infrastructure in the networking, security, searching and advertising, negotiating and matchmaking, contracting and ordering, billing, payment, production, distribution, accounting, and customer service mechanisms that support such distributed information systems [31].

In a generic EC/DL model, providers (information providers, merchants, retailers, wholesalers) make multimedia objects available to consumers (customers, information seekers, users) in exchange for payment. An EC/DL system itself is characterized as a collection of distributed autonomous sites (servers) that work together to give the consumer the appearance of a single cohesive collection. Each site may store a large number of multimedia objects (documents, images, video, audio, software, structured data). This content may be stored in a variety of formats and on a variety of media such as disk, tape or CD-ROM and typically originates from a variety of providers who may wish to control its use (retrieval or modification) or to add value. Consumers are assumed to have a wide variety of domain expertise and computer proficiency which must be taken into account by designers of EC/DL systems.

Section 2 examines EC and DL research requirements in six key subareas, which section 3 provides case studies that describe three electronic commerce research projects (USC-ISI, CommerceNet, First Virtual) and six digital libraries projects sponsored by an NSF/ARPA/NASA initiatives.

TR-97-194	**Globalizing Business, Education, Culture Through the Internet**	**Nabil Adam, Baruch Awerbuch, Jacob Slonim, Peter Wegner, Yelena Yesha**	**February 1997**

Globalization occurs at both the national and international levels. Infrastructure is initially developed and regulated at the national level, since most utilization of the telecommunication infrastructure is within rather than among nations. Many of the technical and social questions arising at the national level are relevant to international globalization, while some issues such as interoperability among heterogeneous multilingual components occur primarily at the international level.

The technology of globalization is being driven by commercial incentives for improving the efficiency of business enterprises as well as societal concerns with improving the quality of life. We examine electronic commerce to illustrate business enterprises and education to illustrate the impact of globalization on the quality of life.

Underlying globalization is a set of technologies for human-computer interaction, finding and filtering information, security, negotiating and matchmaking, integration and interoperability, and networking. We discuss a few of these technologies.

TR-97-199	**Information Extraction based Multiple-Category Document Classification for the Global Legal Information Network**	**Nabil Adam, Richard D. Holowczak**	**March 1997**

This paper describes a prototype application of an information extraction (IE) based document classification system in the international law domain. IE is used to determine if a set of concepts for a class are present in a document. The syntactic and semantic constraints that must be satisfied to make this determination are derived automatically from a training corpus. A collection of IE systems are arranged in a classification hierarchy and novel documents are guided down the hierarchy based on a subset of the Global Legal Information Network domain.

TR-97-200	**The Global Legal Information Network ("GLIN")**	**Nabil Adam, Burt Edelson, Tarek El-Ghazawi, Milt Halem, Kostas Kalpakis, Nick Kosura, Rubens Medina, Yelena Yesha**	**December 1996**

The current globalization of the marketplace generates a greater need for cultures to learn more about one another so that decisions regarding international transactions or associations are based on trustworthy information. Additionally, many nations feel a sense of commonality not only with their immediate neighbors but also with distant trading or cultural partners. These expanding bonds help fuel the growth of common markets and greater cultural ties. Information, particularly legal information, is an essential element of these international ties because critical issues surrounding such relationships are resolved using this information. Legal researchers no longer can rely solely on the laws of a single nation to solve a legal problem; they must be able to access the law of several nations.

Fortunately, information technology has made possible faster, more accurate searches of larger and more current volumes of information. The result has been broader researching capabilities in the area of multinational comparative legal studies. Additionally, legal researchers appear to be expanding their language capabilities, as reflected in other nations. This technology may find application to worldwide databases within our lifetimes due to the great progress that has been made in machine translation.

TR-97-201	**Modeling and Analysis of Workflows Using Petri Nets**	**Nabil Adam, Vijayalakshmi Aturi, Wei-Kuang Huang**	**April 1997**

A workflow system, in its general form, is basically a heterogeneous and distributed information system where the tasks are performed using autonomous systems. Resources, such as databases, labor, etc. are typically required to process these tasks. Prerequisite to the execution of a task is a set of constraints that reflect the applicable business rules and user requirements.

In this paper we present a Petri Net (PN) based framework that (1) facilitates specification of workflow applications, (2) serves as a powerful tool for modeling the system under study at a conceptual level, (3) allows for a smooth transition from the conceptual level to a testbed implementation and (4) enables the analysis, simulation and validation of the system under study before proceeding to implementation. Specifically, we consider three categories of task dependencies: control flow, value, and external (temporal).

We identify several structural properties of PN and demonstrate their use for conducting the following type of analyses: (1) identify inconsistent dependency specifications among tasks; (2) test for workflow safety, i.e. test whether the workflow terminates in an acceptable state; (3) for a given starting time, test whether it is feasible to execute a workflow with the specified temporal constraints.

Don Becker, CESDIS

TR-96-168	**Evaluation of Segmented Ethernet Interconnect Topologies for the Beowulf Parallel Workstation**	**Chance Reschke, Thomas Sterling, Donald J. Becker, Daniel Ridge, Phillip R. Merkey, Odysseas Pentakalos, Michael R. Berry**	**January 1996**

The Beowulf Parallel Workstation exploits the Pile-of-PC (PopC) approach integrating mass market PC-based subsystems to achieve superior performance and order of magnitude disk capacity and bandwidth gain over conventional scientific workstations at comparable cost. Disk access intensive applications are susceptible to performance degradation due to interconnect bandwidth limitations. This paper investigates the potential of a parallel segmented Ethernet network topology limited to two ports per node. A simple row-column segmented Ethernet network topology is shown to deliver significant and sometimes dramatic sustained throughput advantage over the more common multidrop connection scheme for both basic network packet traffic and end-to end file copies.

Alok Choudhary, Syracuse University

TR-96-185	**Accessing Sections of Out-of-Core Arrays Using an Extended Two-Phase Method**	**Rejeev Thakur, Alok Choudhary**	**January 1995**

In out-of-core computations, data needs to be moved back and forth between main memory and disks during program execution. In this paper, we propose a technique called the Extended Two-Phase Method, for accessing sections of out-of-core arrays efficiently. This is an extension and generalization of the Two-

Phase Method for reading in-core arrays from files, which was previously proposed in [7, 3]. The Extended Two-Phase Method uses collective I/O requests into fewer larger requests, eliminating multiple disk accesses for the same data and reducing contention for disks. We describe the algorithms for reading as well as writing array sections. Performance results on the Intel Touchstone Delta for many different access patterns are presented and analyzed. It is observed that the Extended Two-Phase Method gives consistently good performance over a wide range of access patterns.

Burt Edelson, George Washington University

TR-97-200	**The Global Legal Information Network ("GLIN")**	**Nabil Adam, Burt Edelson, Tarek El-Ghazawi, Milt Halem, Kostas Kalpakis, Nick Kosura, Rubens Medina, Yelena Yesha**	**December 1996**

The current globalization of the marketplace generates a greater need for cultures to learn more about one another so that decisions regarding international transactions or associations are based on trustworthy information. Additionally, many nations feel a sense of commonality not only with their immediate neighbors but also with distant trading or cultural partners. These expanding bonds help fuel the growth of common markets and greater cultural ties. Information, particularly legal information, is an essential element of these international ties because critical issues surrounding such relationships are resolved using this information. Legal researchers no longer can rely solely on the laws of a single nation to solve a legal problem; they must be able to access the law of several nations.

Fortunately, information technology has made possible faster, more accurate searches of larger and more current volumes of information. The result has been broader researching capabilities in the area of multinational comparative legal studies. Additionally, legal researchers appear to be expanding their language capabilities, as reflected in other nations. This technology may find application to worldwide databases within our lifetimes due to the great progress that has been made in machine translation.

Tarek El-Ghazawi, George Washington University

TR-96-170	**Parallel Input/Output Issues in Sparse Matrix Computations**	**Sorin G. Nastea, Tarek El-Ghazawi, Ophir Frieder**	**January 1996**

Sparse matrix computations have many important industrial applications and are characterized by large volumes of data. Due to the lagging input/output (I/O) technology, compared to processor technology, the negative impact of input/output could be challenging to the overall performance of such class of applications. In this work, we empirically investigate the performance of typical parallel file system options for performing parallel I/O operations in sparse matrix applications. We introduce a dynamic scheduling method to further hide I/O latency. We also investigate the impact of parallel I/O on the overall performance of sparse-matrix vector multiplications.

Our experimental results using the Intel Paragon and standard matrix data will show that using our technique, tangible performance gains can be attained, beyond what asynchronous system calls alone may offer. For some data sets, it is possible to significantly ease the I/O bottleneck through latency hiding and amortization to a limit that can preserve the scalability characteristics of the computations. The results will also empirically uncover the pros and cons associated with the different parallel file system calls found on parallel systems, such as the Intel Paragon.

TR-96-176	**MIPI: Multi-level Instrumentation of Parallel Input/Output**	**Michael R. Berry, Tarek A. El-Ghazawi**	**April 1996**

As the disparity between processing speed and I/O capabilities of systems continues to grow, the need for useful tools to analyze I/O performance and characteristics becomes vital. Multi-level instrumentation and analysis techniques allow the I/O performance of applications running on parallel systems to be effectively measured. The MIPI (Multi-level Instrumentation of Parallel Input/Output) tool presented here combines multi-level instrumentation of I/O with post-processing, analysis, and graphical data display for use in systems performance analysis. MIPI allows instrumentation and inter-level analysis at the application, file system, and device driver level. This tool provides users and designers of parallel (and serial) systems with the capability to obtain and view integrated information from multiple system views of I/O an activity. This information provides a better understanding of the I/O workload behaviors that dominate performance. Such understanding can translate into design principles and considerations that lead to high performance/cost implementations.

TR-97-200	**The Global Legal Information Network ("GLIN")**	**Nabil Adam, Burt Edelson, Tarek El-Ghazawi, Milt Halem, Kostas Kalpakis, Nick Kosura, Rubens Medina, Yelena Yesha**	**December 1996**

The current globalization of the marketplace generates a greater need for cultures to learn more about one another so that decisions regarding international transactions or associations are based on trustworthy information. Additionally, many nations feel a sense of commonality not only with their immediate neighbors but also with distant trading or cultural partners. These expanding bonds help fuel the growth of common markets and greater cultural ties. Information, particularly legal information, is an essential element of these international ties because critical issues surrounding such relationships are resolved using this information. Legal researchers no longer can rely solely on the laws of a single nation to solve a legal problem; they must be able to access the law of several nations.

Fortunately, information technology has made possible faster, more accurate searches of larger and more current volumes of information. The result has been broader researching capabilities in the area of multinational comparative legal studies. Additionally, legal researchers appear to be expanding their language capabilities, as reflected in other nations. This technology may find application to worldwide databases within our lifetimes due to the great progress that has been made in machine translation.

Konstantinos Kalpakis, University of Maryland Baltimore County

TR-97-200	**The Global Legal Information Network ("GLIN")**	**Nabil Adam, Burt Edelson, Tarek El-Ghazawi, Milt Halem, Kostas Kalpakis, Nick Kosura, Rubens Medina, Yelena Yesha**	**December 1996**

The current globalization of the marketplace generates a greater need for cultures to learn more about one another so that decisions regarding international transactions or associations are based on trustworthy information. Additionally, many nations feel a sense of commonality not only with their immediate neigh-

bors but also with distant trading or cultural partners. These expanding bonds help fuel the growth of common markets and greater cultural ties. Information, particularly legal information, is an essential element of these international ties because critical issues surrounding such relationships are resolved using this information. Legal researchers no longer can rely solely on the laws of a single nation to solve a legal problem; they must be able to access the law of several nations.

Fortunately, information technology has made possible faster, more accurate searches of larger and more current volumes of information. The result has been broader researching capabilities in the area of multinational comparative legal studies. Additionally, legal researchers appear to be expanding their language capabilities, as reflected in other nations. This technology may find application to worldwide databases within our lifetimes due to the great progress that has been made in machine translation.

Jacqueline Le Moigne, CESDIS

TR-96-171	**The Use of Wavelets for Remote Sensing Image Registration and Fusion**	**Jacqueline Le Moigne, Robert F. Cromp**	**February 1996**

With the new trend of smaller missions in which sensors will be carried on separate platforms, the amount of remote sensing data to combined will increase tremendously, and will require fast and automatic image registration and fusion. Image registration techniques will help develop "ready to use" global datasets from multi-instrument/multi-platform/multi-temporal observations, while image fusion will provide new image products summarizing some basic understanding of the original data. These methods will find applications in numerous domains such as Earth Science data analysis, map updating, and space exploration.

Our work on image registration and fusion focuses on the speed of such methods and on their ability to handle multi-sensor data. These two requirements brought us to the utilization of multi-resolution wavelet transforms to perform such tasks. Our registration algorithm utilizes a wavelet based multi-resolution search to determine the best transformation between two or more images to be registered. As of now, the algorithm searches for rotation, translation or a composition of both. This algorithm has been tested successfully on uni-sensor images - Landsat -Thematic Mapper (TM), Advanced Very High Resolution Radiometer (AVHRR), and Geostationary Operational Environmental Satellite (GOES) data, as well as multisensor data such as Modis Airborne Simulator (MAS) with Landsat-TM data. The second step in the combination of the data deals with the fusion of the data. This fusion can be considered at two levels; either the fusion occurs after registration of the original data and before any further analysis, or each individual dataset is analyzed independently and then a composite image is created. Both approaches may be considered utilizing a wavelet-based approach. Some preliminary results on image fusion are presented.

TR-96-182	**An Automated Parallel Image Registration Technique of Multiple Source Remote Sensing Data**	**Jacqueline Le Moigne, William J. Campbell, Robert F. Cromp**	**June 1996**

With the increasing importance of multiple platform/multiple remote sensing missions, the integration of digital data from disparate sources has become critical to the success of these endeavors. New remote sensing systems will generate enormous amounts of data representing multiple-time or simultaneous observations of the same features by spread over multiple platforms, resulting in a tremendous amount of data that must be different sensors. Also, with the new trend of smaller missions, these sensors will be combined. In most available systems, the standard approach to image registration is to manually choose, in both input and reference images, some well defined Control Points or Tie-Points, and then to compute the parameters of a deformation model. The main difficulty lies in locating and matching the control points which, in the current manual approach, is labor intensive. In our work, we show how maxima of wavelet coefficients can form the basic features for an automatic registration of multiple resolution data. After developing a parallel implementation of wavelet decomposition on a Single Instruction Multiple Data (SIMD) massively parallel computer, the MasPar MP-2, our wavelet-based registration algorithm is tested successfully with data from the NOAA Advanced Very High Resolution Radiometer (AVHRR) , the Landsat/ Thematic Mapper (TM) as well as from the Geostationary Operational Environmental Satellite (GOES). We feel this development is a significant step towards accurate automated image registration.

Richard G. Lyon, University of Maryland Baltimore County

TR-97-196	**Hubble Space Telescope Faint Object Camera Calculated Point-Spread Functions**	**Rick Lyon, Jan M. Hollis, John E. Dorband**	**March 1997**

A set of observed noisy Hubble Space Technology Faint Camera point-spread functions used to recover the combined Hubble and Faint Object Camera wave-front error. The low-spatial-frequency wave-front error is parameterized in terms of a set of 32 annular Zernike polynomials. The midlevel and higher spatial frequencies are parameterized in terms of set of 891 polar-Fourier polynomials. The parameterized wave-front error is used to generate accurate calculated point-spread functions, both pre- and post-COSTAR (corrective optics space telescope axial replacement), suitable for image restoration at arbitrary wavelengths. We describe the phase-retrieval-based recovery process and the phase parameterization. Resultant calculated precorrection and postcorrection point-spread functions are shown along with an estimate of both pre- and post-COSTAR spherical aberration.

TR-97-197	**Motion of the Ultraviolet R Aquarii Jet**	**Rick Lyon, Jan M. Hollis, John E. Dorb, W.A. Feibelman**	**January 1997**

We present evidence for subarcsecond changes in the ultraviolet (~2550 Å) morphology of the inner 5 arcseconds of the R Aqr jet over a 2 yr. period. These data were taken with the Hubble Space Telescope (HST) Faint Object Camera (FOC) when the primary mirror flow was still affecting observations. Images of the R Aqr stellar jet were successfully restored to the original design resolution by completely characterizing the telescope-camera point spread function (PSF) with the aid of phase-retrieval techniques. Thus, a noise-free PSF was employed in the final restorations which utilized the maximum entropy method (MEM). We also present recent imagery obtained with the HST/FOC system after the COSTAR correction mission that provides confirmation of the validity of our restoration methodology. The restored results clearly show that the jet is flowing along the northeast (NE)-southwest (SW) axis with a prominent helical-like structure evident on the stronger NE side of the jet. Transverse velocities increase with increasing distance from the central source, providing a velocity range of 36-235 km s -1. From an analysis of proper motions of the two major ultraviolet jet components, we detect an ~40.2 yr. event separation of this apparent enhanced material ejection occurring probably at periastron which is consistent with the suspected ~44 yr. binary period; this same analysis shows that the jet is undergoing magnetic effects. The restoration computations and the algorithms employed demonstrate that mining of flawed HST data can be scientifically worthwhile.

TR-97-198	**A Maximum Entropy Method with a Priori Maximum Likelihood**	**Rick Lyon, Jan M. Hollis, John E. Dorband**	**April 1997**

Implementations of the maximum entropy method for data reconstruction have almost universally used the approach of maximizing the statistic S - l X2 where S is the Shannon entropy of the reconstructed distribution and X2 is the usual statistical measure associated with agreement between certain properties of the reconstructed distribution and the data. We develop here an alternative approach which maximizes the entropy subject to the set of constraints the X2 be at a minimum with respect to the reconstructed distribution. This in turn modifies the fitting statistics to be S - I• X2 where I is now a vector. This new method provided a unique solution to both the well-posed and ill-posed problem, provides a natural convergence criterion which has previously been lacking in other implementations of maximum entropy, and provides the most conservative (least informative) data reconstruction result consistent with both maximum entropy and maximum likelihood methods, thereby mitigating against over-interpretation of reconstruction results. A spectroscopic example is shown as a demonstration.

Daniel Menascé, George Mason University

TR-97-202	**Pythia: A Performance Analyzer of Hierarchical Mass Storage Systems**	**Odysseas Pentakalos, Daniel Menascé, Yelena Yesha**	**July 1997**

Hierarchical mass storage systems are becoming more complex each day and there are many possible ways of configuring them. The options range from the type an number of devices to be used to their connectivity. An extensible object-oriented performance analyzer, called Pythia, was designed and implemented to allow users to easily investigate the most cost-effective configurations for a given workload. One of the most important reasons to build such a tool is to provide a simple way through which queuing analytic models can be used for performance prediction and system sizing of mass storage systems. The tool incorporated a modeling wizard component that is capable of automatically building a queuing network model from a mass storage system representation defined through a graphic editor. Thus, the user of the tool does not need to know queuing network modeling techniques to use it.

TR-96-174	**Analytical Performance Modeling of Hierarchical Mass Storage Systems**	**Odysseas I. Pentakalos, Daniel A. Menascé, Milt Halem, Yelena Yesha**	**March 1996**

Mass storage systems are finding greater use in scientific computing research environments for retrieving and archiving the large volumes of data generated and manipulated by scientific computations. This paper presents a queuing network model that can be used to carry out capacity planning studies of hierarchical mass storage systems. Measurements taken on a Unitree mass storage system and a detailed workload characterization provided the workload intensity and resource demand parameters for the various types of read and write requests. The performance model developed here is based on approximations to multi-class Mean Value Analysis of queuing networks. The approximations were validated through the use of discrete event simulation and the complete model was validated and calibrated through measurements. The resulting model was used to analyze three different scenarios: effect of workload intensity increase, use of file compression at the server and client, and use of file abstractions.

TR-96-175	**Analytical Modeling of Distributed Hierarchical Mass Storage Systems with Network-Attached Storage Devices**	**Odysseas I. Pentakalos, Daniel A. Menascé, Yelena Yesha**	**March 1996**

Network attached storage devices improve I/O performance by separating control and data paths and eliminating host intervention during data transfer. Devices are attached to both a high speed network for data transfer and to a slower network for control messages. Hierarchical mass storage systems use disks to cache the most recently used files and a combination of robotic and manually mounted tapes to store the bulk of the files in the file system. This paper shows how queuing network models can be used to assess the performance of distributed hierarchical mass storage systems that use network attached storage devices. Simulation was used to validate the model. The analytical model was used to analyze many different scenarios including the variation of the number of network attached disks, network attached tapes, and file servers. The model was also used to compare a network attached device based mass storage system to an equivalent host attached device based mass storage system under the same workload.

TR-96-181	**An Object Oriented Performance Analyzer of Hierarchical Mass Storage Systems**	**Odysseas I. Pentakalos, Daniel A. Menascé, Yelena Yesha**	**June 1996**

Hierarchical mass storage systems (HMSS) provide high storage capacity at a cost per megabyte comparable to magnetic tapes with access times per megabyte comparable to magnetic disks. There are many ways of configuring an HMSS including using RAID disks and network attached devices. This paper discusses the design and implementation of a tool that uses queuing network (QN) models to conduct performance prediction studies of HMSSs. The QN model uses MVA-based approximations to deal with RAID devices and network attached tapes. The tool is object oriented and easily extensible to accommodate new technologies as they appear in the market. An important component of the tool is a modeling wizard that automatically builds a QN models form a graphical description of the mass storage system architecture. Results provided by the tool include file transfer times and throughputs per workload as well as an indication of the bottleneck device for each workload.

TR-97-188	**Pythia and Pythia/ WK: Tools for the Performance Analysis of Mass Storage Systems**	**Odysseas I. Pentakalos, Daniel A. Menascé, Yelena Yesha**	**January 1997**

The constant growth on the demands imposed on hierarchical mass storage systems creates a need for frequent reconfiguration and upgrading to ensure that the response times and other performance metrics are within the desired service levels. This paper describes the design and operation of two tools, Pythia and Pythia/WK, that assist system managers and integrators in making cost-effective procurement decisions. Pythia automatically builds and solves an analytic model of a mass storage system based on a graphical description of the architecture of the system and on a description of the workload imposed the system. The use of a modeling wizard to perform this conversion unique among analytic performance tools. Pythia/WK uses clustering algorithms to characterize the workload from the log files of the mass storage system. The resulting workload characterization is used as input to Pythia.

TR-97-192	**Automated Clustering-Based Workload Characterization**	**Odysseas I. Pentakalos, Daniel A. Menascé, Yelena Yesha**	**August 1996**

The demands placed on the mass storage systems at various federal agencies and national laboratories are continuously increasing in intensity. This forces system managers to constantly monitor the system, evaluate the demand placed on it, and tune it appropriately using either heuristics based on experience or analytic models. Performance models require an accurate workload characterization. This can be a laborious and time consuming process. In previous studies [1,2], the authors used k- means clustering algorithms to characterize the workload imposed on a mass storage system. The result of the analysis was used as input to a performance prediction tool developed by the authors to carry out capacity planning studies of hierarchical mass storage systems [3]. It became evident from our experience that a tool is necessary to automate the workload characterization process.

Phillip Merkey, CESDIS

TR-96-168	**Evaluation of Segmented Ethernet Interconnect Topologies for the Beowulf Parallel Workstation**	**Chance Reschke, Thomas Sterling, Donald J. Becker, Daniel Ridge, Phillip R. Merkey, Odysseas Pentakalos, Michael R. Berry**	**January 1996**

The Beowulf Parallel Workstation exploits the Pile-of-PC (PopC) approach integrating mass market PC-based subsystems to achieve superior performance and order of magnitude disk capacity and bandwidth gain over conventional scientific workstations at comparable cost. Disk access intensive applications are susceptible to performance degradation due to interconnect bandwidth limitations. This paper investigates the potential of a parallel segmented Ethernet network topology limited to two ports per node. A simple row-column segmented Ethernet network topology is shown to deliver significant and sometimes dramatic sustained throughput advantage over the more common multidrop connection scheme for both basic network packet traffic and end-to end file copies.

TR-96-169	**The Performance of Earth and Space Science Applications on the Convex Exemplar Scalable Shared Memory Multiprocessor**	**Thomas Sterling, Phillip Merkey, Daniel Savarese**	**January 1996**

The Earth and space sciences (ESS) community embodies a set of rich and diverse computational challenges from static, regular, and embarrassingly parallel to dynamic, unstructured, and tightly coupled, requiring in many cases teraflops scale performance for some of their larger problems. The shared memory cache coherent architecture offers the prospect of superior programmability and efficiency compared to distributed memory systems. The Convex Exemplar is among the most recent of the commercial scalable cache coherent architectures. Extensive studies have been performed to characterize the operational properties of the Exemplar and determine how well suited it is to ESS applications. This paper presents the findings of this investigation and demonstrates the strengths and limitations of this class of architecture It is shown that global cache coherence can be employed effectively to simplify programming and data migration but that the basic problem of locality

TR-97-193	**An Empirical Evaluation of the Convex SPP-1000 Hierarchical Shared Memory System**	**Thomas Sterling, Phillip Merkey, Daniel Savarese, Kevin Olson**	**August 1996**

Cache coherency in a scalable parallel computer architecture requires mechanisms beyond the conventional common bus based snooping approaches which are limited to about 16 processors. The new Convex SPP-1000 achieves cache coherency across 128 processors through a two-level shared memory. NUMA structure employing directory based and SCI protocol mechanisms. While hardware support for managing a common global name space minimizes overhead costs and simplifies programming, latency considerations for remote accesses may still dominate and can under unfavorable conditions constrain scalability. This paper provides the first published evaluation of the SP-1000 hierarchical cache coherency mechanisms from the perspective of measured latency and its impact on basic global flow control mechanisms. scaling of a parallel science code, and sensitivity of cache miss rates to system scale. It is shown that global remote access latency is only a factor of seven greater than that of local cache miss penalty and the scaling of a challenging scientific application is not severely degraded by the hierarchical structure for achieving consistency across the system processor caches.

Matthew O'Keefe, University of Minnesota

TR-96-178	**Interactive Smooth-Motion Animation of High Resolution Ocean Circulation Calculations**	**Aaron Sawdey, Derek Lee, Thomas Ruwart, Paul Woodward, Matthew O'Keefe, Rainer Bleck**	**April 1996**

In this paper we describe how our recent high resolution North Atlantic ocean circulation calculations were visualized and animated. The calculations, performed on the Cray T3D at the Pittsburgh Supercomputer Center, required display technology beyond today's limited resolution screens. We describe how we post-process the calculation datasets, the hardware required, and how interactive smooth-motion animation is achieved at full model resolution.

TR-96-179	**A comparison of data-parallel and message-passing versions of the Miami Isopycnic Coordinate Ocean Model (MICOM)**	**Rainer Bleck, Sumner Dean, Matthew O'Keefe, Aaron Sawdey**	**June 1995**

A two-pronged effort to convert a recently developed ocean circulation model written in Fortran-77 for execution on massively parallel computers is described. A data-parallel version was developed for the CM-5 manufactured by Thinking Machines, Inc., while a message-passing version was developed for both the Cray T3D and the Silicon Graphics ONYX workstation. Since the time differentiation scheme in the ocean model is fully explicit and does not require solution of elliptic partial differential equations, adequate machine utilization has been achieved without major changes to the original algorithms. We developed a partitioning strategy for the message passing version that significantly reduces memory requirements and increases model speed. On a per-node basis (a T3D node is one Alpha processor, a CM-5 node is one Sparc chip and four vector units), the T3D and CM-5 are found to execute our "large" model version consisting of 511 x 511 horizontal mesh points at roughly the same speed.

TR-96-180	**Instrumenting a Unix Kernel for Event Tracing**	**Steven R. Soltis, Matthew T. O'Keefe, Thomas M. Ruwart**	**October 1995**

Characterizing the performance of the storage subsystem on computers requires measurements of the operating system as well as its hardware. While instruments such as bus analyzers can successfully measure hardware devices, precise timing measurements of the operating system's routines can be difficult. Measuring at a system call level may not give enough information to characterize each component of execution time in the kernel, so finer-grain measurements are required. This paper describes a method to measure performance of kernel routines by instrumenting the operating system to perform event tracing. Several of the Silicon Graphics operating system IRIX 5.3 device drivers have been instrumented to call a trace routine which captures the time, a location identifier, and data passed to the routine. This information is stored in buffers within kernel space. The data is extracted from kernel memory by reading from a special purpose software device and analyzed to give performance statistics. This paper describes the trace routine and the software device driver in detail. We will show the perturbation caused by the tracing facility affects the system performance only slightly.

Odysseas Pentakalos, University of Maryland Baltimore County

TR-97-202	Pythia: A Performance Analyzer of Hierarchical Mass Storage Systems	Odysseas Pentakalos, Daniel Menascé, Yelena Yesha	July 1997

Hierarchical mass storage systems are becoming more complex each day and there are many possible ways of configuring them. The options range from the type an number of devices to be used to their connectivity. An extensible object-oriented performance analyzer, called Pythia, was designed and implemented to allow users to easily investigate the most cost-effective configurations for a given workload. One of the most important reasons to build such a tool is to provide a simple way through which queuing analytic models can be used for performance prediction and system sizing of mass storage systems. The tool incorporated a modeling wizard component that is capable of automatically building a queuing network model from a mass storage system representation defined through a graphic editor. Thus, the user of the tool does not need to know queuing network modeling techniques to use it.

TR-96-168	Evaluation of Segmented Ethernet Interconnect Topologies for the Beowulf Parallel Workstation	Chance Reschke, Thomas Sterling, Donald J. Becker, Daniel Ridge, Phillip R. Merkey, Odysseas Pentakalos, Michael R. Berry	January 1996

The Beowulf Parallel Workstation exploits the Pile-of-PC (PopC) approach integrating mass market PC-based subsystems to achieve superior performance and order of magnitude disk capacity and bandwidth gain over conventional scientific workstations at comparable cost. Disk access intensive applications are susceptible to performance degradation due to interconnect bandwidth limitations. This paper investigates the potential of a parallel segmented Ethernet network topology limited to two ports per node. A simple row-column segmented Ethernet network topology is shown to deliver significant and sometimes dramatic sustained throughput advantage over the more common multidrop connection scheme for both basic network packet traffic and end-to end file copies.

TR-96-174	Analytical Performance Modeling of Hierarchical Mass Storage Systems	Odysseas I. Pentakalos, Daniel A. Menascé, Milt Halem, Yelena Yesha	March 1996

Mass storage systems are finding greater use in scientific computing research environments for retrieving and archiving the large volumes of data generated and manipulated by scientific computations. This paper presents a queuing network model that can be used to carry out capacity planning studies of hierarchical mass storage systems. Measurements taken on a Unitree mass storage system and a detailed workload characterization provided the workload intensity and resource demand parameters for the various types of read and write requests. The performance model developed here is based on approximations to multi-class Mean Value Analysis of queuing networks. The approximations were validated through the use of discrete event simulation and the complete model was validated and calibrated through measurements. The resulting model was used to analyze three different scenarios: effect of workload intensity increase, use of file compression at the server and client, and use of file abstractions.

TR-96-175	**Analytical Modeling of Distributed Hierarchical Mass Storage Systems with Network-Attached Storage Devices**	**Odysseas I. Pentakalos, Daniel A. Menascé, Yelena Yesha**	**March 1996**

Network attached storage devices improve I/O performance by separating control and data paths and eliminating host intervention during data transfer. Devices are attached to both a high speed network for data transfer and to a slower network for control messages. Hierarchical mass storage systems use disks to cache the most recently used files and a combination of robotic and manually mounted tapes to store the bulk of the files in the file system. This paper shows how queuing network models can be used to assess the performance of distributed hierarchical mass storage systems that use network attached storage devices. Simulation was used to validate the model. The analytical model was used to analyze many different scenarios including the variation of the number of network attached disks, network attached tapes, and file servers. The model was also used to compare a network attached device based mass storage system to an equivalent host attached device based mass storage system under the same workload.

TR-96-181	**An Object Oriented Performance Analyzer of Hierarchical Mass Storage Systems**	**Odysseas I. Pentakalos, Daniel A. Menascé, Yelena Yesha**	**June 1996**

Hierarchical mass storage systems (HMSS) provide high storage capacity at a cost per megabyte comparable to magnetic tapes with access times per megabyte comparable to magnetic disks. There are many ways of configuring an HMSS including using RAID disks and network attached devices. This paper discusses the design and implementation of a tool that uses queuing network (QN) models to conduct performance prediction studies of HMSSs. The QN model uses MVA-based approximations to deal with RAID devices and network attached tapes. The tool is object oriented and easily extensible to accommodate new technologies as they appear in the market. An important component of the tool is a modeling wizard that automatically builds a QN models form a graphical description of the mass storage system architecture. Results provided by the tool include file transfer times and throughputs per workload as well as an indication of the bottleneck device for each workload.

TR-97-188	**Pythia and Pythia/ WK: Tools for the Performance Analysis of Mass Storage Systems**	**Odysseas I. Pentakalos, Daniel A. Menascé, Yelena Yesha**	**January 1997**

The constant growth on the demands imposed on hierarchical mass storage systems creates a need for frequent reconfiguration and upgrading to ensure that the response times and other performance metrics are within the desired service levels. This paper describes the design and operation of two tools, Pythia and Pythia/WK, that assist system managers and integrators in making cost-effective procurement decisions. Pythia automatically builds and solves an analytic model of a mass storage system based on a graphical description of the architecture of the system and on a description of the workload imposed the system. The use of a modeling wizard to perform this conversion unique among analytic performance tools. Pythia/WK uses clustering algorithms to characterize the workload from the log files of the mass storage system. The resulting workload characterization is used as input to Pythia.

TR-97-189	**Analytical Modeling of Robotic Tape Libraries Using Stochastic Automata**	**Odysseas I. Pentakalos, Tugrul Dayar, A.B. Stephens**	**January 1997**

This paper presents results of preliminary work concerning the analytical modeling of a robotic tape library using stochastic automata. This model is a first approximation which we use to test the efficiency and accuracy of stochastic automata networks as a modeling tool. The effects of interactive get and put requests, migration, and purging at online and nearline layers of a tertiary storage device are investigated. The discussion in this paper sheds further light to the merits and limitations of stochastic automata networks as a modeling methodology for evaluation the performance of complex real life applications.

TR-97-192	**Automated Clustering-Based Workload Characterization**	**Odysseas I. Pentakalos, Daniel A. Menascé, Yelena Yesha**	**August 1996**

The demands placed on the mass storage systems at various federal agencies and national laboratories are continuously increasing in intensity. This forces system managers to constantly monitor the system, evaluate the demand placed on it, and tune it appropriately using either heuristics based on experience or analytic models. Performance models require an accurate workload characterization. This can be a laborious and time consuming process. In previous studies [1,2], the authors used k- means clustering algorithms to characterize the workload imposed on a mass storage system. The result of the analysis was used as input to a performance prediction tool developed by the authors to carry out capacity planning studies of hierarchical mass storage systems [3]. It became evident from our experience that a tool is necessary to automate the workload characterization process.

Terrence Pratt, CESDIS

TR-96-183	**Using High Performance Fortran for Earth and Space Science Applications**	**Terrence W. Pratt**	**May 1996**

High Performance Fortran (HPF) provides an industry standard set of extensions to Fortran 90 for parallel computing. This report analyzes the potential of HPF as a language for writing large-scale scientific applications codes for Earth and space science applications. The report includes the history and current status of the language and its implementations. A critical analysis is provided of the HPF design and its implications for programmers of ESS applications. Experience in rewriting sample applications from the HPCC/ESS CAN Benchmark suite is described, with an eye toward the kinds of difficulties that scientists might encounter in rewriting larger codes in HPF. Recommendations for follow-on activities are given.

Udaya Ranawake, University of Maryland Baltimore County

TR-96-184	**Performance Evaluation of Piecewise Parabolic Method on Convex Exemplar SPP1000**	**Udaya A. Ranawake**	**June 1996**

We consider the parallel implementation of an Euler equation solver using the piecewise parabolc method (PPM) on a HP-Convex Exemplar SPP1000. The parallelization is accomplished by dividing the computational grid into rectangular regions surrounded by a border of ghost points four zones wide. We argue that this approach results in low communication overhead and effective utilization of the cache on a wide variety of parallel computer systems. The performance of several implementations based on the shared memory programming model is compared against a PVM based implementation and a straightforward parallelization using the Convex Fortran compiler. The different versions based on the shared memory paradigm utilize different memory class addressing schemes in order to determine the best memory layout for the shared data structures. A calculation on a 480 by 1920 grid using the shared memory version of the program delivers 41 Mflops per node on all processors of a 16 processor Exemplar. We also discuss the programming effort involved in implementing this code.

Daniel Reed, University of Illinois

TR-96-172	I/O, Performance Analysis, and Performance Data Immersion	Daniel A. Reed, Tara Madhyastha, Ruth A. Aydt, Christopher L. Elford, Will H. Scullin, Evgenia Smirni	February 1996

A large and important class of national challenge applications are irregular, with complex, data dependent execution behavior, and dynamic, with time varying resource demands. We believe the solution to the performance optimization conundrum is integration of dynamic performance instrumentation and on-the-fly performance data reduction with configurable, malleable resource management algorithms, and a real-time adaptive control mechanism that automatically chooses and configures resource management algorithms based on application request patterns and observed system performance. Within the context of parallel input/output optimization, we describe the components of such a closed-loop control system based on the Pablo performance analysis environment, a portable parallel file system (PPFS), and virtual environments for study of dynamic performance data and interactive control of file system policies.

Chance Reschke, (formerly of) CESDIS

TR-96-168	Evaluation of Segmented Ethernet Interconnect Topologies for the Beowulf Parallel Workstation	Chance Reschke, Thomas Sterling, Donald J. Becker, Daniel Ridge, Phillip R. Merkey, Odysseas Pentakalos, Michael R. Berry	January 1996

The Beowulf Parallel Workstation exploits the Pile-of-PC (PopC) approach integrating mass market PC-based subsystems to achieve superior performance and order of magnitude disk capacity and bandwidth gain over conventional scientific workstations at comparable cost. Disk access intensive applications are susceptible to performance degradation due to interconnect bandwidth limitations. This paper investigates the potential of a parallel segmented Ethernet network topology limited to two ports per node. A simple row-column segmented Ethernet network topology is shown to deliver significant and sometimes dramatic sustained throughput advantage over the more common multidrop connection scheme for both basic network packet traffic and end-to end file copies.

Joel Saltz, University of Maryland College Park

TR-96-177	Tuning the Performance of I/O -Intensive Parallel Applications	Anurag Acharya, Joel Saltz, Alan Sussman,	March 1996

Getting good I/O performance from parallel programs is a critical problem for many application domains. In this paper, we report our experience tuning the I/O performance of four application programs from the areas of satellite-data processing and linear algebra. After tuning, three of the four applications achieve application-level I/O rates of over 100 MB/s on 16 processors. The total volume of I/O required by the programs ranged from about 75 MB to over 200 GB. We report the lessons learned in achieving high I/O performance from these applications, including the need for code restructuring, local disks on every node and knowledge of future I/O requests. We also report our experience in achieving high performance on peer-to-peer configurations. Finally, we comment on the necessity of complex I/O interfaces like collective I/O and strided requests to achieve high performance.

Linda Shapiro, University of Washington

TR-96-186	**A Visual Database System for Image Analysis on Parallel Computers and its Project**	**Linda Shapiro, Steven Tanimoto, James Ahrens**	**October 1996**

The goal of this work was to create a design and prototype implementation of a database environment that is particular suited for handling the image, vision and scientific data associated with the NASA's EOS Amazon project. We are focusing on a data model and query facilities that are designed to execute efficiently on parallel computers. A key feature of the environment is an interface which allows a scientists to specify high-level directives about how query execution should occur. Using the interface does not require an understanding of the intricate details of parallel scheduling.

Aya Soffer, University of Maryland Baltimore County

TR-96-187	**Image Categorization Using Texture Features**	**Aya Soffer**	**December 1996**

A method for categorization images using texture features is presented. The goal of categorization is to find all images from the same category as a given query image. Some example categories are floor plans, music notes, satellite images, houses, and fingerprints. The hypothesis that two images that have the same texture features are likely to belong to the same category, is examined. A new texture feature termed N x M -gram is presented in detail. It is based on N -grams, a technique that is commonly used for determining similarity of text documents. Intuitively an N x M -gram is a small pattern in an image. The notion of N x M -gram is defined and the process of computing an image profile in terms of its N x M -gram is explained. Three similarity measures for N x M -grams are compared to each other as to results of categorization using other well known texture features such as cooccurrence matrices, local standard deviation, Laws texture features, and grey level distribution features. The results show that for our test images N X M -gram based methods were more successful in finding images from the same category as a given query image than other texture features.

Thomas Sterling, (formerly of) CESDIS

TR-96-168	**Evaluation of Segmented Ethernet Interconnect Topologies for the Beowulf Parallel Workstation**	**Chance Reschke, Thomas Sterling, Donald J. Becker, Daniel Ridge, Phillip R. Merkey, Odysseas Pentakalos, Michael R. Berry**	**January 1996**

The Beowulf Parallel Workstation exploits the Pile-of-PC (PopC) approach integrating mass market PC-based subsystems to achieve superior performance and order of magnitude disk capacity and bandwidth gain over conventional scientific workstations at comparable cost. Disk access intensive applications are susceptible to performance degradation due to interconnect bandwidth limitations. This paper investigates the potential of a parallel segmented Ethernet network topology limited to two ports per node. A simple row-column segmented Ethernet network topology is shown to deliver significant and sometimes dramatic sustained throughput advantage over the more common multidrop connection scheme for both basic network packet traffic and end-to end file copies.

TR-96-169	**The Performance of Earth and Space Science Applications on the Convex Exemplar Scalable Shared Memory Multiprocessor**	**Thomas Sterling, Phillip Merkey, Daniel Savarese**	**January 1996**

The Earth and space sciences (ESS) community embodies a set of rich and diverse computational challenges from static, regular, and embarrassingly parallel to dynamic, unstructured, and tightly coupled, requiring in many cases teraflops scale performance for some of their larger problems. The shared memory cache coherent architecture offers the prospect of superior programmability and efficiency compared to distributed memory systems. The Convex Exemplar is among the most recent of the commercial scalable cache coherent architectures. Extensive studies have been performed to characterize the operational properties of the Exemplar and determine how well suited it is to ESS applications. This paper presents the findings of this investigation and demonstrates the strengths and limitations of this class of architecture It is shown that global cache coherence can be employed effectively to simplify programming and data migration but that the basic problem of locality sensitivity still demands direct programmer involvement to achieve effective cache behavior.

TR-97-193	**An Empirical Evaluation of the Convex SPP-1000 Hierarchical Shared Memory System**	**Thomas Sterling, Phillip Merkey, Daniel Savarese, Kevin Olson**	**August 1996**

Cache coherency in a scalable parallel computer architecture requires mechanisms beyond the conventional common bus based snooping approaches which are limited to about 16 processors. The new Convex SPP-1000 achieves cache coherency across 128 processors through a two-level shared memory. NUMA structure employing directory based and SCI protocol mechanisms. While hardware support for managing a common global name space minimizes overhead costs and simplifies programming, latency considerations for remote accesses may still dominate and can under unfavorable conditions constrain scalability. This paper provides the first published evaluation of the SP-1000 hierarchical cache coherency mechanisms from the perspective of measured latency and its impact on basic global flow control mechanisms. scaling of a parallel science code, and sensitivity of cache miss rates to system scale. It is shown that global remote access latency is only a factor of seven greater than that of local cache miss penalty and the scaling of a challenging scientific application is not severely degraded by the hierarchical structure for achieving consistency across the system processor caches.

Alan Sussman, University of Maryland College Park

TR-96-177	**Tuning the Performance of I/O -Intensive Parallel Applications**	**Anurag Acharya, Joel Saltz, Alan Sussman**	**March 1996**

Getting good I/O performance from parallel programs is a critical problem for many application domains. In this paper, we report our experience tuning the I/O performance of four application programs from the areas of satellite-data processing and linear algebra. After tuning, three of the four applications achieve application-level I/O rates of over 100 MB/s on 16 processors. The total volume of I/O required by the programs ranged from about 75 MB to over 200 GB. We report the lessons learned in achieving high I/O performance from these applications, including the need for code restructuring, local disks on every node and knowledge of future I/O requests. We also report our experience in achieving high performance on peer-to-peer configurations. Finally, we comment on the necessity of complex I/O interfaces like collective I/O and strided requests to achieve high performance.

Yelena Yesha, CESDIS and University of Maryland Baltimore County

TR-97-202	**Pythia: A Performance Analyzer of Hierarchical Mass Storage Systems**	**Odysseas Pentakalos, Daniel Menascé, Yelena Yesha**	**July 1997**

Hierarchical mass storage systems are becoming more complex each day and there are many possible ways of configuring them. The options range from the type an number of devices to be used to their connectivity. An extensible object-oriented performance analyzer, called Pythia, was designed and implemented to allow users to easily investigate the most cost-effective configurations for a given workload. One of the most important reasons to build such a tool is to provide a simple way through which queuing analytic models can be used for performance prediction and system sizing of mass storage systems. The tool incorporated a modeling wizard component that is capable of automatically building a queuing network model from a mass storage system representation defined through a graphic editor. Thus, the user of the tool does not need to know queuing network modeling techniques to use it.

TR-96-173	**Optimal Allocation of Replicated Data in Tree Networks**	**A. B. Stephens, David M. Lazoff, Yelena Yesha**	**February 1996**

The problem of placing replicated data in a tree network in order to maximize data availability in the presence of communication link failures is analyzed. Previous results are extended by considering tree networks in which link failure probabilities are small but nonuniform, and in which access requests do not have a uniform distribution throughout the system. We prove that optimal locations of non-replicated data must occur at a weighted median of the tree. We prove that optimal placements of replicated data for read requests must allocate copies to the leaves of a weighted k-tree core. and that optimal placements of replicated data for write request must form a pure copy subtree of minimal weighted status. We then discuss how these results hold as well under an alternative network reliability model based on site failures.

TR-96-174	**Analytical Performance Modeling of Hierarchical Mass Storage Systems**	**Odysseas I. Pentakalos, Daniel A. Menascé, Milt Halem, Yelena Yesha**	**March 1996**

Mass storage systems are finding greater use in scientific computing research environments for retrieving and archiving the large volumes of data generated and manipulated by scientific computations. This paper presents a queuing network model that can be used to carry out capacity planning studies of hierarchical mass storage systems. Measurements taken on a Unitree mass storage system and a detailed workload characterization provided the workload intensity and resource demand parameters for the various types of read and write requests. The performance model developed here is based on approximations to multiclass Mean Value Analysis of queuing networks. The approximations were validated through the use of discrete event simulation and the complete model was validated and calibrated through measurements. The resulting model was used to analyze three different scenarios: effect of workload intensity increase, use of file compression at the server and client, and use of file abstractions.

TR-96-175	**Analytical Modeling of Distributed Hierarchical Mass Storage Systems with Network-Attached Storage Devices**	**Odysseas I. Pentakalos, Daniel A. Menascé, Yelena Yesha**	**March 1996**

Network attached storage devices improve I/O performance by separating control and data paths and eliminating host intervention during data transfer. Devices are attached to both a high speed network for data transfer and to a slower network for control messages. Hierarchical mass storage systems use disks to cache the most recently used files and a combination of robotic and manually mounted tapes to store the bulk of the files in the file system. This paper shows how queuing network models can be used to assess the performance of distributed hierarchical mass storage systems that use network attached storage devices. Simulation was used to validate the model. The analytical model was used to analyze many different scenarios including the variation of the number of network attached disks, network attached tapes, and file servers. The model was also used to compare a network attached device based mass storage system to an equivalent host attached device based mass storage system under the same workload.

TR-96-181	**An Object Oriented Performance Analyzer of Hierarchical Mass Storage Systems**	**Odysseas I. Pentakalos, Daniel A. Menascé, Yelena Yesha**	**June 1996**

Hierarchical mass storage systems (HMSS) provide high storage capacity at a cost per megabyte comparable to magnetic tapes with access times per megabyte comparable to magnetic disks. There are many ways of configuring an HMSS including using RAID disks and network attached devices. This paper discusses the design and implementation of a tool that uses queuing network (QN) models to conduct performance prediction studies of HMSSs. The QN model uses MVA-based approximations to deal with RAID devices and network attached tapes. The tool is object oriented and easily extensible to accommodate new technologies as they appear in the market. An important component of the tool is a modeling wizard that automatically builds a QN models form a graphical description of the mass storage system architecture. Results provided by the tool include file transfer times and throughputs per workload as well as an indication of the bottleneck device for each workload.

TR-97-188	**Pythia and Pythia/ WK: Tools for the Performance Analysis of Mass Storage Systems**	**Odysseas I. Pentakalos, Daniel A. Menascé, Yelena Yesha**	**January 1997**

The constant growth on the demands imposed on hierarchical mass storage systems creates a need for frequent reconfiguration and upgrading to ensure that the response times and other performance metrics are within the desired service levels. This paper describes the design and operation of two tools, Pythia and Pythia/WK, that assist system managers and integrators in making cost-effective procurement decisions. Pythia automatically builds and solves an analytic model of a mass storage system based on a graphical description of the architecture of the system and on a description of the workload imposed the system. The use of a modeling wizard to perform this conversion unique among analytic performance tools. Pythia/WK uses clustering algorithms to characterize the workload from the log files of the mass storage system. The resulting workload characterization is used as input to Pythia.

TR-97-190	**Electronic Commerce and Digital Libraries: Towards a Digital Agora**	**Nabil Adam, Yelena Yesha**	**January 1997**

Electronic commerce (EC) and digital libraries (DL) are two increasingly important areas of computer and information sciences with different user requirements but similar infrastructure requirements. In exploring strategic directions, we examine both requirements of the global information infrastructure that are necessary prerequisite for EC and DL [2], and specific requirements of EC and DL within the global infrastructure.

Both EC and DL are concerned with systems that support the creation of information sources and with the movement of information across global networks. EC supports effective and efficient business interactions and transactions that take place on behalf of consumers, sellers, intermediaries, and producers, while DL supports effective and efficient interaction among knowledge seekers. A digital library may require the transactional aspects of EC to manage the purchasing and distribution of its content while a digital library can be used as a resource in electronic commerce to manage products, services, providers and consumers. EC and Dl share a common infrastructure in the networking, security, searching and advertising, negotiating and matchmaking, contracting and ordering, billing, payment, production, distribution, accounting, and customer service mechanisms that support such distributed information systems [31].

In a generic EC/DL model, providers (information providers, merchants, retailers, wholesalers) make multimedia objects available to consumers (customers, information seekers, users) in exchange for payment. An EC/DL system itself is characterized as a collection of distributed autonomous sites (servers) that work together to give the consumer the appearance of a single cohesive collection. Each site may store a large number of multimedia objects (documents, images, video, audio, software, structured data). This content may be stored in a variety of formats and on a variety of media such as disk, tape or CD-ROM and typically originates from a variety of providers who may wish to control its use (retrieval or modification) or to add value. Consumers are assumed to have a wide variety of domain expertise and computer proficiency which must be taken into account by designers of EC/DL systems.

Section 2 examines EC and DL research requirements in six key subareas, which section 3 provides case studies that describe three electronic commerce research projects (USC-ISI, CommerceNet, First Virtual) and six digital libraries projects sponsored by an NSF/ARPA/NASA initiatives.

TR-97-192	**Automated Clustering-Based Workload Characterization**	**Odysseas I. Pentakalos, Daniel A. Menascé, Yelena Yesha**	**August 1996**

The demands placed on the mass storage systems at various federal agencies and national laboratories are continuously increasing in intensity. This forces system managers to constantly monitor the system, evaluate the demand placed on it, and tune it appropriately using either heuristics based on experience or analytic models. Performance models require an accurate workload characterization. This can be a laborious and time consuming process. In previous studies [1,2], the authors used k- means clustering algorithms to characterize the workload imposed on a mass storage system. The result of the analysis was used as input to a performance prediction tool developed by the authors to carry out capacity planning studies of hierarchical mass storage systems [3]. It became evident from our experience that a tool is necessary to automate the workload characterization process.

TR-97-194	**Globalizing Business, Education, Culture Through the Internet**	**Nabil Adam, Baruch Awerbuch, Jacob Slonim, Peter Wegner, Yelena Yesha**	**February 1997**

Globalization occurs at both the national and international levels. Infrastructure is initially developed and regulated at the national level, since most utilization of the telecommunication infrastructure is within rather than among nations. Many of the technical and social questions arising at the national level are relevant to international globalization, while some issues such as interoperability among heterogeneous multilingual components occur primarily at the international level.

The technology of globalization is being driven by commercial incentives for improving the efficiency of business enterprises as well as societal concerns with improving the quality of life. We examine electronic commerce to illustrate business enterprises and education to illustrate the impact of globalization on the quality of life.

Underlying globalization is a set of technologies for human-computer interaction, finding and filtering information, security, negotiating and matchmaking, integration and interoperability, and networking. We discuss a few of these technologies.

TR-97-200

The Global Legal Information Network ("GLIN")

**Nabil Adam,
Burt Edelson,
Tarek El-Ghazawi,
Milt Halem,
Kostas Kalpakis,
Nick Kosura,
Rubens Medina,
Yelena Yesha**

December 1996

The current globalization of the marketplace generates a greater need for cultures to learn more about one another so that decisions regarding international transactions or associations are based on trustworthy information. Additionally, many nations feel a sense of commonality not only with their immediate neighbors but also with distant trading or cultural partners. These expanding bonds help fuel the growth of common markets and greater cultural ties. Information, particularly legal information, is an essential element of these international ties because critical issues surrounding such relationships are resolved using this information. Legal researchers no longer can rely solely on the laws of a single nation to solve a legal problem; they must be able to access the law of several nations.

Fortunately, information technology has made possible faster, more accurate searches of larger and more current volumes of information. The result has been broader researching capabilities in the area of multinational comparative legal studies. Additionally, legal researchers appear to be expanding their language capabilities, as reflected in other nations. This technology may find application to worldwide databases within our lifetimes due to the great progress that has been made in machine translation.

CENTER OF EXCELLENCE IN SPACE DATA AND INFORMATION SCIENCES
CODE 930.5
NASA GODDARD SPACE FLIGHT CENTER
GREENBELT, MD 20771

TECHNICAL REPORT SERIES ORDER FORM

301-286-4403 Internet: cas@cesdis1.gsfc.nasa.gov

Number of Copies Requested	Report Number	Title
	TR-90-01	Analyzing a CSMA/CD Protocol through a Systems of Communicating Machines Specification (Raymond E. Miller)
	TR-90-02	Altruistic Locking (Kenneth Salem)
	TR-90-03	Modeling the Logical Structure of Flexible Manufacturing Systems with Petri-Nets (P. David Stotts)
	TR-90-04	On the Bit-Complexity of Discrete Solutions of PDEs: Compact Multigrid (John Reif)
	TR-90-05	Rules and Principles of Scientific Data Visualization (Hikmet Senay)
	TR-90-06	Changes in Connectivity in Active Contour Models (Ramin Samadani)
	TR-90-07	Designing C++ Libraries (James M. Coggins)
	TR-90-08	Stabilization and Pseudo-Stabilization (Raymond E. Miller)
	TR-90-09	Coordinating Multi-Transaction Activities (Kenneth Salem)
	TR-90-10	Bounding Procedure Execution Times in a Synchronous (P. David Stotts)
	TR-90-11	VISTA: Visualization Tool Assistant for Viewing Scientific Data (Hikmet Senay)

____________	TR-90-12	Model-Driven Image Analysis to Augment Databases (Ramin Samadani)
____________	TR-90-13	Interfacing Image Processing and Computer Graphics Systems Using an Artificial Visual System (James M. Coggins)
____________	TR-90-14	Protocol Verification: The First Ten Years, The Next Ten Years; Some Personal Observations (Raymond E. Miller)
____________	TR-90-15	Coverability Graphs for a Class of Synchronously Executed Unbounded Petri Net (P. David Stotts)
____________	TR-90-16	Compositional Analysis and Synthesis of Scientific Data Visualization Techniques (Hikmet Senay)
____________	TR-90-17	Evaluation of an Elastic Curve Technique for Automatically Finding the Auroral Oval from Satellite Images (Ramin Samadani)
____________	TR-90-18	Anticipated Methodologies in Computer Vision (James M. Coggins)
____________	TR-90-19	Synthesizing a Protocol Converter from Executable Protocol Traces (Raymond E. Miller)
____________	TR-90-20	YTRACC: An Interactive Debugger for YACC Grammars (David P. Stotts)
____________	TR-90-21	Finding Curvilinear Features in Speckled Images (Ramin Samadani)
____________	TR-90-22	Multiscale Geometric Image Descriptions for Interactive Object Definition (James M. Coggins)
____________	TR-90-23	Testing Protocol Implementations Based on a Formal Specification (Raymond E. Miller)
____________	TR-90-24	A Mills-Style Iteration Theorem for Nondeterministic Concurrent Program (P. David Stotts)
____________	TR-90-25	A Computer Vision System for Automatically Finding the Auroral Oval from Satellite Images (Ramin Samadani)
____________	TR-90-26	Multiscale Vector Fields for Image Pattern Recognition (James M. Coggins)
____________	TR-90-28	Generalizing Hypertext (P. David Stotts)
____________	TR-90-29	Evaluation of an Elastic Curve Technique for Automatically Finding the Auroral Oval from Satellite Images (Ramin Samadani)

______________	TR-90-30	Interactive Object Definition in Medical Images Using Multiscale, Geometric Image Descriptions (James M. Coggins)
______________	TR-90-31	Two New Approaches to Conformance Testing of Communication Protocols (Raymond E. Miller)
______________	TR-90-32	Increasing the Power of Hypertext Search with Relational Queries (P. David Stotts)
______________	TR-90-33	A Multiscale Description of Image Structure for Segmentation of Biomedical Images (James M. Coggins)
______________	TR-90-34	Temporal Hyperprogramming (P. David Stotts)
______________	TR-90-35	Biomedical Image Segmentation Using Multiscale Orientation Fields (James M. Coggins)
______________	TR-90-36	Programmable Browsing Semantics in Trellis (P. David Stotts)
______________	TR-90-37	Image Structure Analysis Supporting Interactive Object Definition (James M. Coggins)
______________	TR-90-38	Separating Hypertext Content from Structure in Trellis (P. David Stotts)
______________	TR-90-39	Hierarchy, Composition, Scripting Languages, and Translators for Structured Hypertext (P. David Stotts)
______________	TR-90-40	Browsing Parallel Process Networks (P. David Stotts)
______________	TR-90-41	aTrellis: A System for Writing and Browsing Petri-Net-Based Hypertext (P. David Stotts)
______________	TR-91-42	Generating Test Sequences with Guaranteed Fault Coverage for Conformance Testing of Communication Protocols (Raymond E. Miller)
______________	TR-91-43	Specification and Analysis of a Data Transfer Protocol Using Systems of Communicating Machines (Raymond E. Miller)
______________	TR-91-44	An Exact Algorithm for Kinodynamic Planning in the Plane (John Reif)
______________	TR-91-45	Place/Transition Nets with Debit Arcs (P. David Stotts, Parke Godfrey)
______________	TR-91-46	Adaptive Prefetching for Disk Buffers (Kenneth Salem)
______________	TR-91-48	Adaptive Control of Parameters for Active Contour Models (Ramin Samadani)Visual System (James M. Coggins)
______________	TR-91-50	Structured Dynamic Behavior in Hypertext (David Stotts)

	TR-91-51	BLITZEN: A Highly Integrated Massively Parallel Machine (John Reif)
	TR-91-52	Efficient Parallel Algorithms for Optical Computing with the DFT Primitive (John Reif)
	TR-91-53	This Technical Report has been superceded by TR-92-87 A Minimization-Pruning Algorithm for Finding Elliptical Boundaries in Images with Non-Constant Background and with Missing Data (Ramin Samadani)
	TR-91-54	A Functional Meta-Structure for Hypertext Models and Systems (P. David Stotts)
	TR-91-55	An Optimal Parallel Algorithm for Graph Planarity (John Reif)
	TR-91-56	A Randomized EREW Parallel Algorithm for Finding Connected Components in a Graph (Hillel Gazit and John Reif)
	TR-91-57	Study of Six Linear Least Square Fits (Eric Feigelson)
	TR-91-58	Fast Computations of Vector Quantization Algorithms (John Reif)
	TR-91-59	Probabilistic Diagnosis of Hot Spots (Kenneth Salem)
	TR-91-60	Multi-Media Interaction with Virtual Worlds (Hikmet Senay)
	TR-91-61	Image Compression Methods with Distortion Controlled Capabilities (John Reif)
	TR-91-62	Management of Partially-Safe Buffers (Kenneth Salem)
	TR-91-63	Non-Deterministic Queue Operations (Kenneth Salem)
	TR-91-64	Dynamic Adaptation of Hypertext Structure (P. David Stotts)
	TR-91-65	Scientific Data Visualization Software: Trends and Directions (James Foley)
	TR-91-66	Planar Separators and the Euclidean Norm (Hillel Gazit)
	TR-91-67	A Deterministic Parallel Algorithm for Planar Graphs Isomorphism (Hillel Gazit)
	TR-91-68	A Deterministic Parallel Algorithm for Finding a Separator in Planar Graphs (Hillel Gazit)
	TR-91-69	An Algorithm for Finding a Separator in Planar Graphs (Hillel Gazit)

____________	TR-91-70	Optimal EREW Parallel Algorithms for Connectivity Ear Decomposition and st- Numbering of Planar Graphs (Hillel Gazit)
____________	TR-91-71	An Optimal Randomized Parallel Algorithm for Finding Connecting Components in a Graph (Hillel Gazit)
____________	TR-91-72	Modified Version of Generating Minimal Length Test Sequences for Conformance Testing of Communication Protocols (Raymond Miller)
____________	TR-91-73	Adaptive Image Segmentation Applied to Extracting the Auroral Oval from Satellite Images (Ramin Samadani)
____________	TR-91-74	Parallel Programming on the Silicon Graphics Workstation Using the Multiprocessing Library (Cynthia Starr)
____________	TR-91-75	Placing Replicated Data to Reduce Seek Delays (Kenneth Salem)
____________	TR-92-76	CESDIS Annual Report; Year 3
____________	TR-92-77	Generating Test Sequences with Guaranteed Fault Coverage for Conformance Testing of Communication Protocols (Raymond E. Miller)
____________	TR-92-78	On the Generation of Minimal Length Conformance Tests for Communication Protocols (Raymond E. Miller)
____________	TR-92-79	A Knowledge Based System for Scientific Data Visualization (Hikmet Senay)
____________	TR-92-80	On Generating Test Sequences for Combined Control and Data Flow for Conformance Testing of Communication Protocols (Raymond E. Miller)
____________	TR-92-81	MR-CDF: Managing Multi-Resolution Scientific Data (Kenneth Salem)
____________	TR-92-82	Adaptive Block Rearrangement (Kenneth Salem)
____________	TR-92-83	A Markov Field/Accumulator Sampler Approach to the Atmospheric Temperature Inversion Problem (Noah Friedland)
____________	TR-92-84	Adaptive Snakes: Control of Damping and Material Parameters (Ramin Samadani)
____________	TR-92-85	Faults, Errors and Convergence in Conformance Testing of Communication Protocols (Raymond E. Miller, Sanjoy Paul))
____________	TR-92-86	Research Issues for Communication Protocols (Raymond E. Miller)

____________	TR-92-87	This Technical Report Supercedes TR-91-53 A Minimization-Pruning Algorithm for Finding Elliptical Boundaries in Images with Non-Constant Background and with Missing Data (Ramin Samadani)
____________	TR-92-88	Summary Report of the CESDIS Workshop on Scientific Database Management (Kenneth Salem)
____________	TR-92-89	Structural Analysis of a Protocol Specification and Generation of a Maximal Fault Coverage Conformance Test Sequence (Raymond E. Miller)
____________	TR-92-90	Kernel-Control Parallel Versus Data Parallel: A Technical Comparison (Terrence Pratt)
____________	TR-92-91	Efficient Synchronization with Minimal Hardware Support (James H. Anderson)
____________	TR-92-92	A Fine-Grained Solution to the Mutual Exclusion Problem (James H. Anderson)
____________	TR-92-93	On the Granularity of Conditional Operations (James H. Anderson, Mohamed G. Gouda)
____________	TR-92-94	CESDIS Annual Report; Year 4
____________	TR-93-95	Image Analysis by Integration of Disparate Information (Jacqueline Le Moigne)
____________	TR-93-96	A Virtual Machine for High Performance Image Processing (Douglas Smith)
____________	TR-93-97	Generating Maximal Fault Coverage Conformance Test Sequences of Reduced Length for Communication Protocols (Raymond E. Miller)
____________	TR-93-98	Bounding the Performance of FDDI (Raymond E. Miller)
____________	TR-93-99	Report on the Workshop on Data and Image Compression Needs and Uses in Scientific Community (Stephen R. Tate)
____________	TR-93-100	Fine Grain Dataflow Computation without Tokens for Balanced Execution (Thomas Sterling)
____________	TR-93-101	Implementing Extended Transaction Models Using Transaction Groups (Kenneth Salem)
____________	TR-93-102	Adaptive Block Rearrangement Under UNIX (Kenneth Salem)
____________	TR-93-103	The Realities of High Performance Computing and Dataflow's Role in It: Lessons from the NASA HPCC Program (Thomas Sterling)

______________	TR-93-104	Summary Report of the CESDIS Seminar Series on Earth Remote Sensing (Jacqueline Le Moigne)
______________	TR-93-105	Space-Efficient Hot Spot Estimation (Kenneth Salem)
______________	TR-93-106	DQDB Performance and Fairness as Related to Transmission Capacity (Raymond E. Miller)
______________	TR-93-107	Deadlock Detection for Cyclic Protocols Using Generalized Fair Reachability Analysis (Raymond E. Miller)
______________	TR-94-108	Summary Report of the CESDIS Seminar Series on Future Earth Remote Sensing Missions (Jacqueline Le Moigne)
______________	TR-94-109	Generalized Fair Reachability Analysis for Cyclic Protocols: Part 1 (Raymond E. Miller)
______________	TR-94-110	CESDIS Annual Report; Year 5
______________	TR-94-111	This Technical Report has been superceded by TR-94-129. I/O Performance of the MasPar MP-1 Testbed (Tarek A. El-Ghazawi)
______________	TR-94-112	Parallel Registration of Multi-Sensor Remotely Sensed Imagery Using Wavelet Coefficients (Jacqueline Le Moigne)
______________	TR-94-113	Paradise - A Parallel Geographic Information System (David De Witt)
______________	TR-94-114	Computer Assisted Analysis of Auroral Images Obtained from High Altitude Polar Satellites (Ramin Samadani)
______________	TR-94-115	2Q: A Low Overhead High Performance Buffer Management Replacement Algorithm (Theodore Johnson)
______________	TR-94-116	Sensitivity Analysis of Frequency Counting (Theodore Johnson)
______________	TR-94-117	Client-Server Paradise (David De Witt)
______________	TR-94-118	Performance Characteristics of a 100 MegaByte/second Disk Array (Matthew T. O'Keefe)
______________	TR-94-119	Compiler and Runtime Support for Out-of-Core HPF Programs (Alok Choudhary)
______________	TR-94-120	Use of Subband Decomposition for Management of Scientific Image Databases (Kathleen G. Perez-Lopez)
______________	TR-94-121	Client-Server Paradise (David DeWitt)
______________	TR-94-122	Multi-Resolution Wavelet Decomposition on the MasPar Massively Parallel System (Jacqueline LeMoigne and Tarek A. El-Ghazawi)

______________	TR-94-123	An Initial Evaluation of the Convex SPP-1000 for Earth and Space Science Applications (Thomas Sterling, Phillip Merkey)
______________	TR-94-124	Runtime Support for Parallel I/O in PASSION (Alok Choudhary)
______________	TR-94-125	The Performance Impact of Data Placement for Wavelet Decomposition of Two Dimensional Image Data on SIMD Machines (Jacqueline LeMoigne and Tarek El-Ghazawi)
______________	TR-94-126	Planet Photo-Topography Using Shading and Stereo (Charles XiaoJian Yan)
______________	TR-94-127	Highly Scalable Data Balanced Distributed B-trees (Theodore Johnson)
______________	TR-94-128	Index Replication in a Distributed B-tree (Theodore Johnson)
______________	TR-94-129	Characteristics of the MasPar Parallel I/O System (Tarek El-Ghazawi)
______________	TR-94-130	PASSION: Parallel and Scalable Software for Input-Output (Alok Choudhary)
______________	TR-94-131	Development of a Data Reduction Expert Assistant (Glenn Miller)
______________	TR-94-132	Multivariate Statistical Analysis Software Technologies for Astrophysical Research Involving Large Data Sets (S. G. Djorgovski)
______________	TR-94-133	The Grid Analysis and Display System (GrADS) (James Kinter)
______________	TR-94-134	An Interactive Environment for the Analysis of Large Earth Observation and Model Data Sets (Kenneth Bowman and Robert Wilhelmson)
______________	TR-94-135	A Distributed Analysis and Visualization System for Model and Observational Data (Robert Wilhelmson)
______________	TR-94-136	VIEWCACHE: An Incremental Database Access Method for Autonomous Interoperable Databases (Nicholas Roussopoulos)
______________	TR-94-137	Topography from Shading and Stereo (Berthold Horn)
______________	TR-94-138	Experimenter's Laboratory for Visualized Interactive Science (Elaine Hansen)
______________	TR-94-139	A Land-Surface Testbed for EOSDIS (William Emery)

____________	TR-94-140	High Performance Compression of Science Data (James Storer)
____________	TR-94-141	SAVS: A Space and Atmospheric Visualization Science System (Edward P. Szuszcwicz)
____________	TR-94-142	Interactive Interface for NCAR Graphics (Bill Buzbee)
____________	TR-94-143	McIDAS-eXplorer: A Tool for Analysis of Planetary Data (Sanjay Limaye)
____________	TR-94-144	Software-based Fault Tolerance (Jonathan Bright)
____________	TR-95-145	AstroNet: A Tool Set for Simultaneous, Multi-Site Observations of Astronomical Objects (Supriya Chakrabarti)
____________	TR-95-146	Refining Image Segmentation by Integration of Edge and Region Data (Jacqueline Le Moigne and James Tilton)
____________	TR-95-147	An Approximate Performance Model of a Unitree Mass Storage System (Odysseas I. Pentakalos, Daniel A. Menasce, Milt Halem and Yelena Yesha)
____________	TR-95-148	Unsupervised, Robust Estimation-based Clustering of Remotely Sensed Images (Nathan S. Netanyahu, James C. Tilton and J. Anthony Gualtieri)
____________	TR-95-149	Knowledge Discovery from Structural Data (Diane J. Cook, Lawrence B. Holder and Surnjani Djoko)
____________	TR-95-150	Online Data Compression in a Mass Storage File System (Odysseas I. Pentakalos and Yelena Yesha)
____________	TR-95-151	A User's Guide to Pablo® I/O Instrumentation (Ruth A. Aydt)
____________	TR-95-152	Input/Output Characteristics of Scalable Parallel Applications (Phyllis E. Crandall, Ruth A. Aydt, Andrew A. Chien, Daniel A. Reed)
____________	TR-95-153	Towards a Parallel Registration of Multiple Resolution Remote Sensing Data (Jacqueline Le Moigne)
____________	TR-95-154	PPFS: A High Performance Portable Parallel File System (James V. Huber Jr., Christopher L. Elford, Daniel A. Reed, Andrew A. Chien, David S. Blumenthal)
____________	TR-95-155	An Approximate Performance Model of a Unitree Mass Storage System (Odysseas I. Pentakalos, Daniel A. Menasce, Milt Halem, Yelena Yesha)
____________	TR-95-156	Communication Strategies for Out-of-Core Programs on Distributed Memory Machines (Rajesh Bordawekar and Alok Choudhary)

____________	TR-95-157	Optimal Allocation for Partially Replicated Database Systems on Ring Networks (A. B. Stephens, Yelena Yesha and Keith Humenik)
____________	TR-95-158	Minimizing Message Complexity of Partially Replicated Data on Hypercubes (Keith Humenik, Peter Matthews, A. B. Stephens and Yelena Yesha)
____________	TR-95-159	Two Approaches for High Concurrency in Multicast-Based Object Replication (Theodore Johnson and Lionel Maugis)
____________	TR-95-160	Designing Distributed Search Structures with Lazy Updates (Theodore Johnson and Padmashree Krishna)
____________	TR-95-161	The Proceedings of The Petaflops Frontier Workshop-February 6, 1995 (Thomas Sterling and Michael J. Mac Donald)
____________	TR-95-162	Findings of the Second Pasadena Workshop on System Software and Tools for High Performance Computing Environments (Thomas Sterling, Paul Messina and Jim Pool)
____________	TR-95-163	An Experimental Study of Input/Output Characteristics of NASA Earth and Space Sciences Applications (Michael R. Berry and Tarek El-Ghazawi)
____________	TR-95-164	An Analytic Model of Hierarchical Mass Storage Systems with Network-Attached Storage Devices (Daniel A. Menasce, Odysseas I. Pentakalos and Yelena Yesha)
____________	TR-95-165	Analytical Performance Modeling of Hierarchical Mass Storage Systems (Odysseas I. Pentakalos, Daniel Menasce, Milt Halem and Yelena Yesha)
____________	TR-95-166	CESDIS Annual Report; Year 6
____________	TR-95-167	CESDIS Annual Report; Year 7
____________	TR-96-168	Evaluation of Segmented Ethernet Interconnect Topologies for the Beowulf Parallel Workstation (Chance Reschke, Thomas Sterling, Donald J. Becker, Daniel Ridge, Phillip Merkey, Odysseas Pentakalos and Michael R. Berry)
____________	TR-96-169	The Performance of Earth and Space Science Applications on the Convex Exemplar Scalable Shared Memory Multiprocessor (Thomas Sterling, Phillip Merkey and Daniel Savarese)
____________	TR-96-170	Parallel Input/Output Issues in Sparse Matrix Computations (Sorin G. Nastea, Tarek El-Ghazawi and Ophir Frieder)
____________	TR-96-171	The Use of Wavelets for Remote Sensing Image Registration and Fusion (Jacqueline Le Moigne and Robert F. Cromp)

____________	TR-96-172	I/O, Performance Analysis, and Performance Data Immersion (Daniel A. Reed, Tara Madhyastha, Ruth A. Aydt, Christopher L. Elford, Will H. Scullin, Evgenia Smirni)
____________	TR-96-173	Optimal Allocation of Replicated Data in Tree Networks (A. B. Stephens, David M. Lazoff and Yelena Yesha)
____________	TR-96-174	Analytical Performance Modeling of Hierarchical Mass Storage Systems (Odysseas I. Pentakalos, Daniel A. Menascé, Milt Halem, Yelena Yesha)
____________	TR-96-175	Analytical Modeling of Distributed Hierarchical Mass Storage Systems with Network-Attached Storage Devices (Odysseas I. Pentakalos, Daniel A. Menascé, and Yelena Yesha)
____________	TR-96-176	MIPI: Multi-level Instrumentation of Parallel Input/Output (Michael R. Berry and Tarek A. El-Ghazawi)
____________	TR-96-177	Tuning the Performance of I/O-Intensive Parallel Applications (Anurag Acharya, Mustafa Uysal, Robert Bennett, Assaf Mendelson, Michael Beynon, Jeff Hollingsworth, Joel Saltz, Alan Sussman)
____________	TR-96-178	Interactive Smooth-Motion Animation of High Resolution Ocean Circulation Calculations (Aaron Sawdey, Derek Lee, Thomas Ruwart, Paul Woodward, Matthew O'Keefe, Rainer Bleck)
____________	TR-96-179	A comparison of data-parallel and message-passing versions of the Miami Isopycnic Coordinate Ocean Model (MICOM) (Rainer Bleck, Sumner Dean, Matthew O'Keefe, Aaron Sawdey)
____________	TR-96-180	Instrumenting a Unix Kernel for Event Tracing (Steven R. Soltis, Matthew T. O'Keefe, Thomas M. Ruwart)
____________	TR-96-181	An Object Oriented Performance Analyzer of Hierachical Mass Storage Systems (Odysseas I. Pentakalos, Daniel A. Menascé, Yelena Yesha)
____________	TR-96-182	An Automated Parallel Image Registration Technique of Multiple Source Remote Sensing Data (Jacqueline LeMoigne, William J. Campbell, Robert F. Cromp)
____________	TR-96-183	Using High Performance Fortran for Earth and Space Science Applications (Terrence W. Pratt)
____________	TR-96-184	Performance Evaluation of Piecewise Parabolic Method on Convex Exemplar SPP1000 (Udaya A. Ranawake)
____________	TR-96-185	Accessing Sections of Out-of-Core Arrays Using an Extended Two-Phase Method (Alok Choudhary, Rajeev Thakur)

____________ TR-96-186 A Visual Database System for Image Analysis on Parallel Computers and its Application to the EOS Amazon Project (Linda Shapiro, Steven Tanimoto, James Ahrens

____________ TR-96-187 Image Categorization Using Texture Features (Aya Soffer)

____________ TR-97-188 Pythia and Pythia/WK: Tools for the Performance Analysis of Mass Storage Systems (Odysseas Pentakalos, Daniel Menascé, Yelena Yesha)

____________ TR-97-189 Analytical Modeling of Robotic Tape Libraries Using Stochastic Automata (Tugrul Dayar, Odysseas Pentakalos, A. B. Stephens)

____________ TR-97-190 Electronic Commerce and Digital Libraries: Towards a Digital Agora (Nabil Adam, Yelena Yesha)

____________ TR-97-192 Automated Clustering-Based Workload Characterization (Odysseas Pentakalos, Daniel Menascé, Yelena Yesha)

____________ TR-97-193 An Empirical Evaluation of the Convex SPP-1000 Hierarchical Shared Memory System (Thomas Sterling, Daniel Savarese, Phillip Merkey, Kevin Olson)\

____________ TR-97-194 Globalizing Business, Education, Culture Through the Internet (Nabil Adam, Baruch Awerbuch, Jacob Slonim, Peter Wegner, Yelena Yesha)

____________ TR-97-195 CESDIS Annual Report; Year 8

____________ TR-97-196 Hubble Space Telescope Faint Object Camera Calculated Point-Spread Functions (Rick Lyon, Jan M. Hollis, John Dorband)

____________ TR-97-197 Motion of the Ultraviolet R Aquarii Jet (Rick Lyon, Jan M. Hollis, John Dorband, W.A. Feibelman)

____________ TR-97-198 A Maximum Entropy Method with a Priori Maximum Likelihood Constraints (Rick Lyon, Jan M. Hollis, John Dorband)

____________ TR-97-199 Information Extraction Based Multiple-Category Document Classification for the Global Legal Information Network (Nabil Adam, Richard Holowczak)

____________ TR-97-200 The Global Legal Information Network (GLIN) (Nabil Adam, Burt Edelson, Tarek El-Ghazawi, Milt Halem, Kostas Kalpakis, Nick Kozura, Rubens Medina, Yelena Yesha)

____________ TR-97-201 Modeling and Analysis of Workflows Using Petri Nets (Nabil Adam, Vijayalakshmi Atluri, Wei-Kuang Huang)

____________ TR-97-202 Pythia: A Performance Analyzer of Hierarchical Mass Storage Systems (Odysseas Pentakalos, Daniel Menascé, Yelena Yesha)

NAME____

ADDRESS______________________________________

____

____

PHONE_______________________________________

E-MAIL_______________________________________

AREAS OF RESEARCH INTEREST____________________

APPENDIX D

CESDIS Personnel and Associates

CESDIS ADMINISTRATIVE OFFICE

301-286-4403 fax: 301-286-1777

Individual extensions will roll over to another number or phonemail, if the party called does not answer. Please allow sufficient rings for this to happen.

U. S. Mail Address

CESDIS
Code 930.5
NASA Goddard Space Flight Center
Greenbelt, MD 20771

Federal Express/UPS Address

CESDIS
Building 28, Room W223
NASA Goddard Space Flight Center
Greenbelt, MD 20771

DIRECTOR

Yelena Yesha	CESDIS: 301-286-4108	yesha@cesdis.edu
	UMBC: 410-455-3542	yeyesha@cs.umbc.edu

SENIOR AND STAFF SCIENTISTS

Donald Becker	301-286-0882	becker@cesdis.edu
Jacqueline Le Moigne	301-286-8723	lemoigne@nibbles.gsfc.nasa.gov
Les Meredith	301-286-8830	les@usra.edu
Phillip Merkey	301-286-3805	merk@cesdis.edu
Terry Pratt 301-286-0880	pratt@cesdis.edu	

TECHNICAL PERSONNEL

Richard Burk	301-286-0881	rick@cesdis.edu
Oktay Dogramaci	301-286-7992	oxd@cesdis.edu
Dan Ridge	301-286-3062	newt@cesdis.edu

ADMINISTRATIVE PERSONNEL

Nancy Campbell	301-286-4099	campbell@cesdis.usra.edu
Georgia Flanagan	301-286-2080	georgia@cesdis.usra.edu
Joyce Hines	301-286-0913	joyce@cesdis.usra.edu
Jillian Lusaka	301-286-8755	jillian@cesdis.usra.edu
Michele Meyett	301-286-4403	shelly@cesdis.usra.edu
Annemarie Murphy	301-286-8951	murphy@cesdis.usra.edu

UNIVERSITY/INDUSTRY PROJECT PERSONNEL

Associated Technical Consultants

P. O. Box 20
Germantown, MD 20875-0020

Murray Felsher	301-428-0557	felsher@tmn.com

Brown University

Department of Computer Science
Providence, RI 02912

Peter Wegner	401-863-7632	pw@cs.brown.edu
Stanley Zdonik		sbz@cs.brown.edu

Clemson University

Department of Electrical and Computer Engineering
Clemson, SC 29634-0915

Walt Ligon	803-656-1224	walt@eng.clemson.edu

Fran Stetina and Associates

Bowie, MD 20715

Fran Stetina	301-286-0769	stetina@gsti.com

George Mason University

Department of Computer Science
Fairfax, VA 22030-4444

Daniel Menascé	703-993-1537	menasce@cs.gmu.edu

George Mason University

Institute for Computational Science and Informatics
Fairfax, VA 22030

Rainald Lohner	703-993-4075	rlohner@science.gmu.edu

George Washington University

Department of Electrical Engineering and Computer Science
Washington, DC 20052

Tarek El-Ghazawi	CESDIS: 301-286-8178 GWU: 202-994-5507	tarek@seas.gwu.edu

George Washington University

Institute for Applied Space Research
Washington, DC 10052

Burt Edelson	202-994-5509	edelson@seas.gwu.edu
Neil Helm	202-994-1431	helm@seas.gwu.edu

Georgia Institute of Technology

Broadband and Wireless Networking Laboratory
Atlanta, GA 30332

Ian Akyildiz	404-894-5141	ian@ee.gatech.edu

Johns Hopkins University

Department of Computer Science
Baltimore, MD 21218

Yair Amir		yairamir@cs.jhu.edu

Ohio State University

Department of Computer and Information Science
Columbus, OH 43210

Mukesh Singhal	614-292-5839	singhal@cis.ohio-state.edu

Oregon Graduate Institute

Department of Computer Science and Engineering
Data-Intensive Systems Center
Portland, OR 97291

David Maier	503-690-1154	maier@cse.ogi.edu

Rutgers University

Center for Information Management, Integration, and Connectivity
180 University Avenue
Newark, NJ 07102

Nabil Adam	973-353-5239	adam@adam.rutgers.edu

Stanford University

Department of Computer Science
Stanford, CA 94305-9045

Serge Abiteboul	415-725-4802	abitebou@db.stanford.edu

Syracuse University

Northeast Parallel Architectures Center
111 College Place
Syracuse, NY 13244-4100

Geoffrey Fox	315-443-1723	gcf@nova.npac.syr.edu

University of Florida

Department of Computer and Information Science
Gainesville, FL 32611-2024

Theodore Johnson	904-392-1492	ted@cis.ufl.edu

University of Illinois

Department of Computer Science
Urbana, IL 61801

Daniel Reed	217-333-3807	reed@oboe.cs.uiuc.edu

University of Illinois

National Center for Supercomputing Applications
152 C.A.B., 605 E. Springfield Avenue
Champaign, IL 61820

Dinshaw Balsara	217-244-1481	u10956@ncsa.uiuc.edu

University of Illinois at Chicago

Department of Electrical Engineering and Computer Science
851 S. Morgan Street
Chicago, IL 60607-7053

Ouri Wolfson	312-996-6770	wolfson@eecs.uic.edu

University of Maryland Baltimore County

Department of Computer Science and Electrical Engineering
5401 Wilkens Avenue
Baltimore, MD 21228-5398

Susan Hoban	301-286-7980	shoban@pop900.gsfc.nasa.gov
Kostas Kalpakis	410-455-3143	kalpakis@cs.umbc.edu
Richard Lyon	301-286-4302	lyon@jansky.gsfc.nasa.gov
Udaya Ranawake	301-286-3046	udaya@neumann.gsfc.nas.gov
Aya Soffer	301-286-2439	aya@cesdis.edu
Russell Turner	401-455-3965	turner@cs.umbc.edu
Sushel Unninayar	301-286-2757	sushel@cesdis.edu

University of Maryland College Park

Center for Automation Research
Computer Vision Laboratory
College Park, MD 20742-3275

Nathan Netanyahu	301-286-4652	nathan@nibbles.gsfc.nasa.gov

University of Maryland College Park

Department of Computer Science
College Park, MD 20742

Anurag Acharya		acha@cs.umd.edu
Joel Saltz	301-405-2684	saltz@cs.umd.edu
Alan Sussman	301-405-3360	als@cs.umd.edu

University of Minnesota

Department of Electrical Engineering
4-174 EE/CSci Building
Minneapolis, MN 55455

Matthew O'Keefe	612-625-6306	okeefe@ee.umn.edu

University of Pennsylvania

Department of Computer and Information Science
Moore School/D2
Philadelphia, PA 19104

O. Peter Buneman	215-898-7703	peter@cis.upenn.edu

University of Rochester

Department of Physics and Astronomy
Bausch and Lomb Building
Rochester, NY 14627-0171

Adam Frank	716-275-1717	afrank@alethea.pas.rochester.edu

University of Texas at Arlington

Department of Computer Science Engineering
Box 19015
Arlington, TX 76019

Diane Cook	817-273-3606	cook@cse.uta.edu

University of Texas at Austin

Department of Aerospace Engineering and Engineering Mechanics
Austin, TX 78712-1085

Hans Mark	512-471-5077	betty_richardson@asemailgate.ae.utexas.edu

University of Toronto

Computer Science Department
Room 374C, 10 King's College Road
Toronto, Ontario M5S 3G4, Canada

Jacob Slonim	416-946-3335	jslonim@cs.toronto.edu

University of Virginia

Department of Computer Science
Thornton Hall
Charlottesville, VA 22903

James French	804-982-2213	french@cs.virginia.edu

University of Washington

Department of Astronomy
FM-20
Seattle, WA 98195

George Lake	206-543-7106	lake@astro.washington.edu
Derek Richardson	206-543-0206	dcr@astro.washington.edu

University of Washington

Department of Computer Science and Engineering
FR-35
Seattle, WA 98195

Linda Shapiro	206-543-2196	shapiro@cs.washington.edu

University of Wisconsin

Department of Computer Science
1210 W. Dayton Street
Madison, WI 53706

David DeWitt	608-263-5489	dewitt@cs.wisc.edu
Miron Livny	608-262-0856	miron@cs.wisc.edu

CESDIS ASSOCIATES

Brandeis University

Computer Science Department
Waltham, MA 02254-9110

James Storer	617-736-2714	storer@cs.brandeis.edu

University of California

Computer Engineering Department
Room 225, Applied Sciences
Santa Cruz, CA 95064

Glen Langdon	408-459-2212	langdon@cse.ucsc.edu

University of Maryland College Park

Department of Computer Science
College Park, MD 20742

Joel Saltz	301-405-2684	saltz@cs.umd.edu

University of North Carolina

Department of Computer Science
Sitterson Hall
Chapel Hill, NC 27599-3175

David Stotts	919-962-1833	stotts@cs.unc.edu

University of Waterloo

Department of Computer Science
Waterloo, Ontario N2L 3G1, Canada

Kenneth Salem	519-888-4567, ext 3485	kmsalem@zonker.uwaterloo.ca

ALPHABETICAL DIRECTORY OF CESDIS PERSONNEL

Abiteboul, Serge	415-725-4802	abitebou@db.stanford.edu
Acharya, Anurag		acha@cs.umd.edu
Adam, Nabil	973-353-5239	adam@adam.rutgers.edu
Akyildiz, Ian	404-894-5141	ian@ee.gatech.edu
Amir, Yair		yairamir@cs.jhu.edu
Balsara, Dinshaw	217-244-1481	u10596@ncsa.uiuc.edu
Becker, Don	301-286-0882	becker@cesdis.edu
Buneman, O. Peter	215-898-7703	peter@cis.upenn.edu
Burk, Richard	301-286-0881	rick@cesdis.edu

Campbell, Nancy	301-286-4099	campbell@cesdis.usra.edu
Cook, Diane	817-273-3606	cook@cse.uta.edu
DeWitt, David	608-263-5489	dewitt@cs.wisc.edu
Dogramaci, Oktay	301-286-7992	oxd@cesdis.edu
Edelson, Burt	202-994-5509	edelson@seas.gwu.edu
El-Ghazawi, Tarek	CESDIS: 301-286-8178 GWU: 202-994-5507	tarek@seas.gwu.edu
Felsher, Murray	301-428-0557	felsher@tmn.com
Flanagan, Georgia	301-286-2080	georgia@cesdis.usra.edu
Fox, Geoffrey	315-443-1723	gcf@nova.npac.syr.edu
Frank, Adam	716-275-1717	afrank@alethea.pas.rochester.edu
French, James	804-982-2213	french@cs.virginia.edu
Helm, Neil	202-994-1431	helm@seas.gwu.edu
Hines, Joyce	301-286-0913	joyce@cesdis.usra.edu
Hoban, Susan	301-286-7980	shoban@pop900.gsfc.nasa.gov
Johnson, Theodore	904-392-1492	ted@cis.ufl.edu
Kalpakis, Kostas	410-455-3143	kalpakis@cs.umbc.edu
Lake, George	206-543-7106	lake@astro.washington.edu
Langdon, Glen	408-459-2212	langdon@cse.ucsc.edu
Le Moigne, Jacqueline	301-286-8723	lemoigne@nibbles.gsfc.nasa.gov
Ligon, Walt	803-656-1224	walt@eng.clemson.edu
Livny, Miron	608-262-0856	miron@cs.wisc.edu
Lohner, Rainald	703-993-4075	rlohner@science.gmu.edu
Lusaka, Jillian	301-286-8755	jillian@cesdis.usra.edu
Lyon, Richard	301-286-4302	lyon@jansky.gsfc.nasa.gov
Maier, David	503-690-1154	maier@cse.ogi.edu
Mark, Hans	512-471-5077	betty_richardson@asemailgate.ae.utexas.edu
Menascé, Daniel	703-993-1537	menasce@cs.gmu.edu
Meredith, Les	301-286-8830	les@usra.edu
Merkey, Phillip	301-286-3805	merk@cesdis.edu
Meyette, Michele	301-286-4403	shelly@cesdis.usra.edu
Murphy, Annemarie	301-286-8951	murphy@cesdis.usra.edu
Netanyahu, Nathan	301-286-4652	nathan@nibbles.gsfc.nasa.gov
O'Keefe, Matthew	612-625-6306	okeefe@ee.umn.edu
Pratt, Terry	301-286-0880	pratt@cesdis.edu
Ranawake, Udaya	301-286-3046	udaya@neumann.gsfc.nasa.gov
Reed, Daniel	217-333-3807	reed@oboe.cs.uiuc.edu
Richardson, Derek	206-543-0206	dcr@astro.washington.edu
Ridge, Dan	301-286-3062	newt@cesdis.edu
Salem, Kenneth	519-888-4567, ext 3485	kmsalem@zonker.uwaterloo.ca
Saltz, Joel	301-405-2684	saltz@cs.umd.edu
Shapiro, Linda	206-543-2196	shapiro@cs.washington.edu

Singhal, Mukesh	614-292-5839	singhal@cis.ohio-state.edu
Slonim, Jacob	416-946-3335	jslonim@cs.toronto.edu
Soffer, Aya	301-286-2439	aya@cesdis.edu
Stetina, Fran	301-286-0769	stetina@gsti.com
Storer, James	617-736-2714	storer@cs.brandeis.edu
Stotts, David	919-962-1833	stotts@cs.unc.edu
Sussman, Alan	301-405-3360	als@cs.umd.edu
Unninayar, Sushel	301-286-2757	sushel@cesdis.edu
Wegner, Peter	401-863-7632	pw@cs.brown.edu
Wolfson, Ouri	312-996-6770	wolfson@eecs.uic.edu
Yesha, Yelena	CESDIS: 301-286-4108	yesha@cesdis.edu
	UMBC: 410-455-3542	yeyesha@cs.umbc.edu
Zdonik, Stanley		sbz@cs.brown.edu