INFRASTRUCTURE 98:
NSF CISE/EIA RI and MII PI's Workshop

Snowbird, Utah
July 24 – 26, 1998

A workshop for the principal investigators of NSF/CISE infrastructure awards, sponsored by the CISE Research Infrastructure and CISE Minority Institutions Infrastructure Programs, Division of Experimental and Integrative Activities (EIA) in the Directorate Computer and Information Science and Engineering (CISE) of the National Science Foundation.

Infrastructure for Research and Training on High-Performance Heterogeneous Distributed Database Management

High Performance Database Research Center
School of Computer Science, Florida International University
Miami, FL 33199; rishen@fiu.edu; http://hpdrc.cs.fiu.edu

PI: Naphtali Rishe
Co-PI's: David Barton, Chung-Min Chen, Wei Sun, Frank Urban

Introduction

Florida International University (FIU) is one of the largest majority-minority doctoral-granting universities in the United States. Nearly 70% of our students are minorities. The University has the largest contingent of Hispanic students of any doctoral-granting university in the country, and graduates the most Hispanic engineering students in the Nation. The High Performance Database Research Center (HPDRC) was founded in 1994, and is associated with the School of Computer Science at Florida International University. HPDRC conducts research on database management systems and various applications, leading to the development of new types of database systems and refinement of existing database systems.

The general goals of the project are to provide an infrastructure that will enable FIU's HPDRC to perform heterogeneous database research and to better recruit and retain minority students through their M.S. and Ph.D. degrees. Students participate in in-depth research and training in heterogeneous database integration.

FIU is an urban university whose surrounding community base is substantially comprised of under-represented minorities: 86% of students in the Miami-Dade County Public School are minority (51% Hispanic, 34% Black non-Hispanic, 1% others). One goal of this project is to establish a regional outreach program to attract talented local minority students to FIU. Without the support of this project, those students would otherwise not be able to take advantage of the career and educational opportunities, or would have to attend an out-of-state university (a non-favorable choice of many of the local minority students).

The infrastructure being assembled will provide the students with a networked computing environment on which the research work will be conducted. The ultimate research goal is to develop a heterogeneous database management system, using semantic modeling to integrate and reconcile information from multiple, disparate data sources. Of particular interest are the methodologies to integrate geo-spatial and Web data sources. Geo-spatial data are vital to environmental research and studies (e.g. the Global Warming effect), but are often collected and stored in independently operated organizations. Web data generate new issues in data integration because, unlike traditional databases or data repositories, Web data are usually made available through form-filling interfaces, without divulging the data model behind the scenes. Specific research issues include: heterogeneous data model integration using semantic modeling, specification of Web data sources, geospatial data integration, reconciliation, and fusion (e.g. overlapping raster and vector data), rapid integration methodologies, query processing and optimization, and exploration of mobile agent technology.

HPDRC maintains a WWW page describing its projects and staff at http://hpdrc.cs.fiu.edu.

First Year Accomplishments

Goals, Objectives, and Targeted Activities

The goals of our MII (Minority Institute Infrastructure) grant are to provide an infrastructure that will enable FIU's HPDRC to better recruit minority faculty members, better recruit and retain graduate students through the Ph.D., and to perform more in-depth research and training in database management. Since the grant's inception in the Fall of 1997, we have begun to lay a foundation that will enable us to achieve these goals. The activities we have been engaged in are described in the following sections.

Recruiting Minority Faculty: we are negotiating with two female minority faculty candidates.

Retaining Graduate Students Through the Ph.D.: We are recruiting promising students to take advantage of the funds provided by our MII grant.

Workshop on Next Generation Database Design and Applications: HPDRC held a workshop on Next Generation Database Design and Applications on April 30, 1998 and May 1, 1998. This workshop featured HPDRC-developed technology and facilitated the exchange of ideas with other researchers. Over 50 people attended this workshop. Sessions on medical informatics, advances in database design, GIS and spatial data applications, and semantic/object-oriented database management systems were held. The workshop's keynote speakers were Professor Wesley Chu (UCLA), Mr. Richard Campanella (a Remote Sensing/GIS Specialist with the Institute for Technology Development), Professor Naphtali Rishe (FIU HPDRC), and Mr. Kent Wreder (the Corporate Director of Object Technology for Baptist Health Systems). The HPDRC graduate and undergraduate students supported by our MII grant presented their projects at poster sessions held during this workshop.

AFRL Grant Awarded: On March 30, 1998, HPDRC was awarded $196,000 by the Air Force Research Lab to for a project entitled "Database Query Distribution over Intelligent Networks." The work supported by AFRL will leverage off of the work being funded by NSF, which is described above. We feel that our success in pursuing this award, which will enhance previous work funded by industry, was due largely to the sharpening of our MII-sponsored research to focus on distributed heterogeneous databases. The research for AFRL will focus on query optimization for distributed databases, and will take the variances in network bandwidth between the component databases into account.

Affinity Groups Established: We have established four Affinity Groups modeled after those at the University of Texas, El Paso. The Affinity Groups are made up of faculty members, postdoctoral associates, and graduate and undergraduate students.

Outreach Program to Schools: We have begun an outreach program that will ultimately consist of both visits to FIU and a traveling "show" that includes a presentation geared to the appropriate audience at schools. The presentation is followed by a hands-on demonstration of interesting database projects to which the students can relate, such as advanced "virtual reality" demonstrations and the like. One aspect of this show is viewing a South Florida Landsat image through which it is possible to "fly" by updating the image in real-time from the semantic database in which the Landsat data is stored. Our first visit was to Coral Reef Elementary School, where the entire 5th grade attended a presentation that was conducted by Prof. Urban and Martha Gutierrez, an MII-supported graduate student at FIU HPDRC. The students and all fifth grade teachers attended two separate multimedia presentations on the topics of matter and remote sensing. The presentations were made via a portable computer and projector. The format was informal and included questions and answers. The students videotaped the presentations and showed a lot of interest. A regular series of similar appearances are planned at the K-12 levels, and we feel that these visits will have an important impact on recruiting future scientists to our fields of interest.

TerraFly: We have adapted HPDRC's Semantic Object-Oriented Database (Sem-ODB) to Windows-95 and have produced a first draft version of a CD-ROM edutainment application using Sem-ODB. TerraFly is an interactive vehicle for flying over remotely sensed data; we are presently using data covering South Florida. The TerraFly system implements a Semantic database for the storage and retrieval of all the data used by the system and allows the users to fly over and manipulate the retrieved data. Currently the database contains semantic/textual data, spatial/remote sensed data, and digital data including Landsat and aerial photography data. A friendly graphical user interface that contains several text boxes and drop down menus is provided for ease of use. This application will be distributed to local schools and released as an edutainment product. The exposure generated by this offering should enable us to educate students and the general public about the exciting opportunities available to computer scientists. It is hoped that this will assist us in recruiting a new generation of computer scientists.

Course Development: The courses proposed in our response to the site visit have been submitted to the School of Computer Science's curriculum committee.

Components and Materials Required and Indications of Success

Infrastructure Additions: We have added the following infrastructure using the MII funds:

- 7 Dell PC Workstations – 6 Pentium MMX, 1 Pentium II
- 1 Sun Ultra Enterprise Server
- 1 8mm Tape Storage Device
- 46 GB in additional disk storage

This equipment is being used every day by the student and faculty researchers. The PC workstations have been used to perform database research by the student researchers. The server and storage devices have allowed spatial databases of significant size to be created and used for this research.

REU Supplement: HPDRC requested and was awarded an REU supplement to our MII grant. Many undergraduates, all members of under-represented groups, have been supported, in part, by this REU supplement. Two of these students received their B.S. degrees in the Fall of 1997.

Graduate Students Supported: Six graduate students have been supported, in part, by our MII grant. All but one of these students are members of under-represented groups. One post doc, a Hispanic Female, has been supported, in part, by our MII grant.

Publications: Nineteen papers acknowledging the support of our MII grant have either been published or accepted for publication in journals or conference proceedings.

Evaluation

Degree of Success

Toward the goal of better recruiting and retaining under-represented minority students in our graduate programs, we have successfully increased the enrollment (see previous section for numbers) as a result of the support from our MII grant. Several Affinity Groups and an Outreach Program have been established to enhance our teaching environment and graduate recruitment. We have worked progressively towards the initial design and analysis of the heterogeneous database system. The effort led to the publication of several technical papers. Appropriate facilities (see previous section) have been acquired to support students and the research. These activities and achievements evidence a great success in fulfilling the grant's first-year goals.

Outcome

We have achieved three major research outcomes during the first year.

The first outcome is a comparative study between two common approaches to distributed heterogeneous databases from a global query perspective. The issue is important to the design of the global query processor of a distributed heterogeneous database system. Our results show that a loosely-coupled approach which uses "glueware" to mediate between different DBMS and resolving global queries solely at the application level (namely, SQL), is very cost effective. In most cases, it also yields better performance than a tightly-coupled approach, which centralizes all inter-site join operations at a DBMS that is specifically tailored for this purpose. The major shortcoming of a loosely-coupled system is the lack of overlapping network communication and join computation for an inter-site join operation. This causes long response time as well as completion time for global queries that involve large tables. Based on our research, we have also developed a novel join algorithm, called fragmented join, that is devised to shorten long global queries. The algorithm achieves pipelining at the application level by dividing a large join into several smaller fragmented joins. It greatly reduces the elapsed time to get the result of a global query and uses much less temporary storage space than a sequential, non-fragmented algorithm.

The second outcome is toward the development of a JAVA-based open programming interface for the heterogeneous database system. The object-oriented programming language Java is an ideal companion to an object-oriented database system. We have defined an approach to provide a seamless application programmer interface. It is based on a modular architecture with components for database engines, a communications protocol, and a JAVA API facilitator. The open architecture is flexible, scalable and distributed in nature. We have begun to add Java support to Sem-ODB, the semantic objected-oriented database developed at HPDRC that will be used to power the heterogeneous database system. The Java support will be used to provide an interface to heterogeneous databases, and will be used to make the spatial data stored in Sem-ODB available via the WWW.

Finally, we have adapted SQL, the standard relational database language, to semantic databases. The original purpose of this adaptation was to be compatible with, and be able to communicate with, relational tools. Interestingly, it turned out that the size of a typical SQL program for a semantic database is many times smaller than for an equivalent relational database. The addition of SQL to Sem-ODB will help to facilitate its use as the glue to hold together a distributed heterogeneous database system. In addition, we have developed a CGI-based tool to create web pages that enable a user to pose arbitrary SQL queries to a Sem-ODB database containing both conventional and spatial data. This tool takes as its input a specification of the database schema and creates the pages with little user intervention. A specialized version has been created to generate standard forms and reports.

Impact

The exploding growth and use of the Internet and World Wide Web have enabled users to access huge volumes of data with unprecedented convenience and speed. However, the data sources often diverge in their data model (how the data are organized) and their retrieval interface (how the data can be queried). The deployment of a heterogeneous database will greatly benefit the users in translating isolated, multi-sourced data into integrative information. Focusing on reconciliation of text as well as geospatial data, our project will have a great impact on better facilitating earth scientists in collecting and integrating environmental data (images, maps, and texts) for analysis.

Immediate Impact

Students: The following undergraduates have been supported, in part, by the REU supplement to our MII grant: Enrique Almendral, Abraham Anzardo, Jorge Besada, Annette DeHoyos*, Julie Fernandez, Freddy Haayen, Alexander Hernandez, Jose Iglesias, Luis Llanes, Jose Obando, Sebastian Ojanguren, Michael Olivero, Wilbis Padron, Eduardo Perez, Christian Pesantes, Guido Pozo, Yvonne Ricard, and Jenny Rodriguez*. All of these students are from under-represented groups; those marked with an * received their B.S. degrees in the Fall of 1997. The following graduate students have been supported, in part, by our MII grant: Elma Alvarez*, Debra Davis*, Guillermo Fernandez*, Scott Graham, Martha Gutierrez*, Birago Jones*. Those marked with an * are from under-represented groups. Dr. Maria Martinez, a Hispanic Female, has been supported, in part, as a post doctoral associate by our MII grant.

Publications: 19 publications this year acknowledge the support of our MII award, including:

G. Cao, M. Singhal, Y. Deng, N. Rishe, and W. Sun. "A Delay- Optimal Quorum-Based Mutural Exclusion Scheme with Fault-Tolerance Capability." 1998 IEEE International Conference on Distributed Computing Systems (ICDCS '98), Amsterdam, Netherlands, May 1998, p. 444-451.

M. Chekmasov, N. Rishe, D. Barton, K. Medhyanapu, M. Chekmasova, R. Rodriguez. "Design and Implementation of the ENP Environmental Database using ORACLE." American Society of Photogrammetry and Remote Sensing/ Resource Technology Institute Annual Conference, Tampa, FL March 30-April 4, 1998. pp. 630-638.

C. Chen and N. Rishe, "Development of an Open and Scalable Web-based Information Publishing System." Proceedings of the 36th Annual ACM Southeast Conference, 1998, pp. 163-165.

C. Chen and N. Rishe. "Fragmented join: a pipelined multidatabase join method." International Conference on Parallel and Distributed Processing Techniques and Applications, July 1998.

C. Chen, W. Sun, N. Rishe. "Performance Comparison of Three Approaches of Multidatabase Systems: a Global Query Perspective." IEEE International Performance, Computing and Communication Conference, February 14-16, 1998, Tempe, AZ.

N. Rishe, D. Barton, M. Chekmasov, K. Medhyanapu, S. Graham. M. Chekmasova. "Everglades Data Integration using a Semantic Database System." International Conference Geospatial Information in Agriculture and Forestry, Lake Buena Vista, FL June 1-3, 1998.

N. Rishe, D. Barton, N. Prabakaran, M. Gutierrez, M. Martinez, R. Athauda, A. Gonzalez, S. Graham. "Landsat Viewer: A Tool to Create Color Composite Images of Landsat Thematic Mapper Data." International Conference Geospatial Information in Agriculture and Forestry, June 1998.

N. Rishe, D. Barton, N. Prabhakaran, M. Gutierrez, E. Alvarez, R. Athauda, J. Tola-Rodriguez, A. Gonzalez. "Landsat Data Visualization via the Internet." International Symposium on Spectral Sensing Research (ISSSR), San Diego, CA. December 13-19, 1997.

E. Alvarez, N. Rishe, D. Barton. "Semantic Geographic Information System." ISMM Applied Informatics Journal, in press.

C. Chen, C. Orji, and N. Rishe. "Batch Processing." to appear in Encyclopedia of Electrical and Electronic Engineering, ed. J. Webster, John Wiley & Sons, Feb. 1999.

Y. Ling, W. Sun, N. Rishe, X. Xiang. "A Hybrid Estimator For Selectivity Estimation." To appear in IEEE Transactions on Knowledge and Data Engineering, in press.

K. Liu, W. Meng, C. Yu, G. Trajcevski, and N. Rishe. "Retrieving Most Similar documents from Local Sources." To appear in IEEE Transactions on Knowledge and Data Engineering (Special Issue).

W. Meng, K. Liu, C. Yu, W. Wu, and N. Rishe. "A Statistical Method for Estimating the Usefulness of Text Databases." To appear in IEEE Transactions on Knowledge and Data Engineering (Special Issue on Web Technologies, 1998).

R. Ege, Y. Battikhi, P. Pardo, J. Uppal. "A Modular Java API for Object-Oriented Databases." IEEE Compsac '98. To appear.

W. Meng, K. Liu, C. Yu, X. Wang, Y. Chang, N. Rishe. "Determining Text Databases to Search in the Internet." To appear at VLDB 98.