99-1R

INFRASTRUCTURE 99:

NSF CISE EIA RI and MII PIs' Workshop

New Mexico State University

Las Cruces, New Mexico

August 7-9, 1999

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 Computer and Information Science and Engineering (CISE) directorate of the National Science Foundation.

Compiled by: Gopal Gupta (Lab for Logic and Databases, NMSU)

LIST OF RI and MII RECIPIENTS

(Year of award shown in parenthesis)

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-investigators: Wei Sun, Maxim Chekmasov, N. Prabhakaran, Dmitiy Beryoza, Marina Chekmasova

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.

Second 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 been striving to achieve these goals. The activities we have been engaged in are described in the following sections.

Recruiting Minority Faculty: HPDRC is still seeking minority faculty members to hire. We have extended an offer to a minority faculty member. Unfortunately, he got a more attractive offer elsewhere and did not accept our offer. We have trained a female Hispanic Post-doc, Maria Martinez, who took a position as a visiting assistant professor of Electrical and Computer Engineering at FIU. She is an HPDRC-affiliated faculty member and is serving as a role model to our students. Minority SCS faculty members Dawn Holmes and Joslyn Smith are taking part in the NSF-sponsored HPDRC activities with students.

Retaining Graduate Students Through the Ph.D.: Leonardo Loureiro, a Hispanic student at HPDRC, recently received his PhD. Two other minority PhD's, Rosany Rodriguez and Carlos Ibarra, are near completion of their dissertations. Khaled Naboulsi, an MII-supported graduate student, received his PhD this Spring. We are recruiting promising students to take advantage of the funds provided by our MII grant. Professors Jai Navlakha (FIU School of Computer Science) and Luis Martinez-Perez (FIU College of Education) are researching methods to enhance the retention of graduate students through the Ph.D. We have compiled a list of all current and former computer science students at FIU, as well as a list of contacts at other minority institutions. We will use this information to step up our recruiting effort. We have involved high school students in our research through Miami-Dade County Public School's Advanced Academic Internship Program. Their involvement provides us a pipeline of researchers from high school through the Ph.D.

Affinity Groups: Four Affinity Groups modeled after those at the University of Texas, El Paso are in place at HPDRC. The Affinity Groups are made up of faculty members, postdoctoral associates, and graduate and undergraduate students. The following groups continue to pursue research tracks at HPDRC: *Semantic Database Engine Group* – devoted to designing and developing semantic database technology; *Applications Group* – devoted to investigating spatial data technology and applications and GIS; *Heterogeneous Database Group* – devoted to deepening research in distributed heterogeneous databases; and *Semantic-Relational Systems Group* – devoted to making the semantic database technology available to all database users.

Grants Awarded: NSF has granted supplementary funds to HPDRC to expand our research goals under our NSF CREST grant. The additional funding allows us to expand the data visualization focus of our CREST sub-component. The heterogeneous database research described above is being leveraged to provide data to be visualized under our CREST support. N. Rishe is also PI of an NSF HPC grant that was recently awarded to FIU. This grant will provide FIU with an OC-3 connection to the vBNS and Abilene. HPDRC is leading the development of Next Generation Internet applications at FIU. We are presently collaborating with FIU's Experimental Nuclear Physics Group and with FIU's International Hurricane Center on data intensive applications that would not have been feasible without the High Performance Connection. These applications leverage off of the distributed heterogeneous database research supported by this MII award.

Outreach Program to Schools: We have continued to develop 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. In one of our visits, Debra Lee Davis-Chu, an HPDRC graduate student, visited Hammocks Middle School. Three presentations, each geared to individual class levels, were given on the topics of remote sensing, databases and TerraFly. The presentations were made via the computers available in the classroom. The format was semi-formal and included questions and answers. The students were very enthusiastic and

showed a lot of interest. We have compiled a list of contacts for the schools in our area and a regular series of similar appearances are planned at the K-12 levels; we feel that these visits will have an important impact on recruiting future scientists to our fields of interest. (88% of the students of Miami-Dade County Public Schools are members of under-represented minorities, including 53% Hispanic, 33% Black non-Hispanic, and 2% Other.)

HPDRC is hosting seven students from the Miami-Dade Public Schools under the school system's Advanced Academic Internship Program during the 1998-1999 school year. These students have worked alongside researchers at HPDRC and have contributed to the research and development goals of the Center. The students include Shaun Vendryes (South Miami High), Yicheng Huang (Miami Palmetto High), Roy Duque de Estrada (Hialeah-Miami Lakes High), Jan Yang (Southwest Miami High), Jonathan Kaldor (Miami Killian High), Kirk Padilla (Miami Central High), and Pedro Carabeo-Nieva (G. Holmes Braddock High). HPDRC will be hosting additional students during the 1999-2000 school year.

In the summer of 1999, HPDRC has conducted two database seminars at FIU for about 20 county schoolteachers and one seminar, "A Day of Databases," for about 70 high school students from M-DCPS magnet programs.

Course Development: The courses proposed in our response to the site visit have been approved as experimental by the School of Computer Science's curriculum committee. One of the experimental courses was taught during the Spring 1999 semester. Seven students took COP 6993, Global Optimization. The course was summarized as follows. 'Global Optimization algorithms are of importance from both theoretical and practical points of view. While an area of Computer Science, Global Optimization is closely related to mathematical disciplines, such as the Theory of Probability, Operations Research, Cybernetics, and Mathematical Statistics. Optimization algorithms are used in various engineering projects in industry, simulation of processes, and for testing purposes. This allows us to expand students' mathematical background as well as practical programming experience.'

Components and Materials Required and Indications of Success

Infrastructure Additions: We have added the following infrastructure using the MII funds: 1 Sun Ultra 10 PC Workstations, 5 Pentium II PC Workstations, additional disk storage for database research, and additional miscellaneous hardware. This equipment, and the equipment acquired last year, is being used every day by the student and faculty researchers. Additional acquisitions are planned before the end of the current award period.

REU Supplement: HPDRC requested and was granted an REU supplement to our MII grant. The ten students supported, in part, by the REU supplement are detailed in the Immediate Impact section.

Students Supported directly by MII: Twenty-seven graduate and undergraduate students have been directly supported, in part, by our MII grant. The majority of these have been members of under-represented groups. Students supported by out MII grant are detailed in the Immediate Impact section.

Publications: During the past year, we have published or have has accepted 32 items under the support of this grant. Additionally, a patent covering parts of our Semantic Database technology has been allowed: Application Number 08/905,679 Filing Date 08/04/97, Notice of Allowability Date 03/01/99.

Evaluation

Degree of Success

Toward the goal of better recruiting and retaining under-represented minority students in our graduate programs, we have successfully increased supported student enrollment as a result of the support from our MII grant. Several Affinity Groups and an Outreach Program are in place to enhance our teaching environment and graduate recruitment. We have worked progressively towards the 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 goals and a continuing improvement over the first year's results.

Outcome

Scientific data sometimes needs to be distributed across the network. Each site might maintain an individual database with high autonomy. Nevertheless, data from several component databases may need to be combined in order to fulfill the requirements of an application. It is unavoidable that heterogeneity, which is partially caused by different data models or different query interfaces, will exist among component databases. Even if component databases implement the same data model, there is usually a need for us to deal with schematic or semantic conflicts. Such conflicts include structural conflicts, naming conflicts, abstraction conflicts, domain conflicts, etc. In order to meet all these requirements, we are building a heterogeneous database system (HDB) that extends our semantic binary database. Users can access a number of databases with the extended system as if there were only one semantic database because we use the Semantic Object-oriented Data Model to construct a global data schema. The query interfaces of our HDB are exactly the same as those provided by the Semantic Object-Oriented Database (Sem-ODB) that is under development at HPDRC. The query processor and optimizer in our HDB are responsible for decomposing the global queries into one or more sub-queries. The sub-queries are then submitted to the corresponding component databases via ODBC. When the sub-queries are completely processed, their answers are re-assembled and reconciled at the destination site. Our HDB incorporates logical level optimization. The semantic database engine takes care of the physical level optimization by means of previously developed techniques, such as indexing and clustering on spatial-temporal data. In addition to the conflicts at the database level, there are other kinds of heterogeneity at the system level, for example network heterogeneity. Because of this, we have developed a CORBA-based middleware between the HDB layer and the underlying databases. User interfaces and an API facilitate queries of the global schema at the HDB layer. The HDB interprets and accesses the appropriate component databases via the CORBA middleware. CORBA is a well-founded industrial standard with ease of providing basic facilities for a distributed environment.

The native Application Programming Interface (API) for Sem-ODB is expressed in C++; it is the most efficient, flexible, and direct way to access the database. In addition to the C++ API, we are implementing a Java API for Sem-ODB. This will make it easier for application programmers writing in Java to work with the Sem-ODB engine and will also enable some other valuable features. We will make the Java API as close as possible to the C++ API, which will ensure efficiency and allow both APIs to be used interchangeably. The C++ and Java APIs should be able to call each other's functions interchangeably so that a stored procedure could call C++ code, which may, in turn, execute another stored procedure. The Java API will be implemented through calls to the native API. Our approach is not specific to our semantic database system; it can serve as a recipe for adding 'stored procedures' to any database system.

Structured Query Language (SQL) is the standard language used to write queries for relational databases. We are developing Semantic SQL (Sem-SQL) for the Sem-ODB. The detailed syntax of Sem-SQL is the same as the syntax of Open Database Connectivity (ODBC) 2.0 standard SQL, but it is interpreted differently. We are pursuing Sem-SQL in order to further the following goals.

- SQL is a uniform interface provided by almost every database system; it is, perhaps, the most popular database language and it is known by millions of users. The availability of the Sem-SQL interface will significantly enhance the popularity and accessibility of Sem-ODB.
- Sem-SQL also affords us the possibility of supporting ODBC, which is a standard database-access interface. As a result, Sem-ODB can be easily integrated into the heterogeneous multi-organizational information system that we are developing.

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 our MII grant: Enrique Almendral*, Abraham Anzardo*, Michael Armentano*, Jorge Besada*, Joel Delgado, Julie Fernandez, Dario Gonzalez, Freddy Haayen, Alexander Hernandez*, Jose Iglesias, Sheldon Jones, Ying Liu, Wilbis Padron*, Guido Pozo, Patrick Quinlivan, Dario Rivera, and Julio Ruano. All but one of these students are from under-represented groups; those marked with an * received their B.S. degrees during the past year. The following graduate students have been supported, in part, by our MII grant: Elma Alvarez, Debra Davis-Chu, Guillermo Fernandez*, Scott Graham, Martha Gutierrez, Guangyi Li, Khaled Naboulsi#, Philippe Pardo*, Joseph Pontillo, and Rosany Rodriguez. All but three of these students are from under-represented groups; those marked with an * received their Master's degrees during the past year and those marked with a # received their PhD during the past year. The following undergraduates have been supported, in part, by the REU supplement to out MII grant: Abraham Anzardo*, Luis Espinal, Luis Llanes*, Daniel Mendez, Michael Olivero, Jose Obando, Sebastian Ojanguren*, Wilbis Padron*, Oscar Parrales, and Rob Valenti. All of these students are from under-represented groups; those marked with an * received their Bachelor's degrees during the past year.

Publications: 32 publications this year acknowledge the support of our MII award, including:

C.M. Chen, R. Qiu, and N. Rishe. "A Pipelined Query Processor for Distributed Multidatabases." Proceedings of the IEEE Southeast Conference, Apr. 24-26, 1998 Orlando FL. pp. 132-133.

D.L. Davis-Chu, E. Alvarez, and N. Rishe. "The Creation of a System for 3D Satellite and Terrain Imagery." The 13th Internation Conference in Applied Geologic Remote Sensing, Vancouver, British Columbia, Canada, March 1-3, 1999, pp. II-329 - II-336.

W. Meng, K. Liu, C. Yu, W. Wu, and N. Rishe. "Estimating the Usefulness of Search Engines." Proceedings of the 15th International Conference on Data Engineering, March, 1999.

W. Meng, C. Yu, W. Wang, and N. Rishe. "Performance Analysis of Three Text-join Algorithms." IEEE Transactions on Knowledge and Data Engineering, Vol 10, No 3, May/June 1998, pp. 477-492.

N. Prabhakaran, S. Sridhar, and N. Rishe. "A Two Phase Digital Ortho Photo Mosaicking System." International Conference on Imaging Science, Systems, and Technology (CISST '99), June 28 - July 1, 1999, Las Vegas, Nevada, pp. 151-154.

N. Prabhakaran, V. Maddineni, and N. Rishe. "Spatial Overlay of Vector Data on Raster Data in a Semantic Object-Oriented Database Environment." International Conference on Imaging Science, Systems, and Technology (CISST '99), June 28 - July 1, 1999, Las Vegas, Nevada, pp. 100-104.

N. Rishe, K. Naboulsi, O. Wolfson, and B. Ehlmann. "An Efficient Web-based Semantic SQL Query Generator." The 19th IEEE International Conference on Distributed Computing Systems Workshop. Austin, TX, May 31 - June 5, 1999, pp. 23-30.

N. Rishe, O. Wolfson, and K. Naboulsi. "Report Generators." Encyclopedia of Electrical and Electronics Engineering, John Wiley & Sons, 1999. V. 18, pp. 500-513.

O. Wolfson, L. Jiang, A. Sistla, S. Chamberlain, N. Rishe, M. Deng. "Databases for Tracking Mobile Units in Real Time." 7th International conference on Database Theory (ICDT99), Jerusalem, Israel, January 10-12, 1999, pp.169-186.

C. Yu, K. Liu, W. Wu, W. Meng, and N. Rishe. "Finding the Most Similar Documents across Multiple Text Databases." Proc. of the IEEE Conference on Advances in Digital Libraries (ADL'99), Baltimore, Maryland, May 1999, pp. 150-162.