Taming Torrents of Data

THE VITERBI SCHOOL'S INFORMATION SCIENCES INSTITUTE LEADS SEARCH FOR SCALEABLE KNOWLEDGE DISCOVERY THROUGH GRID WORKFLOWS

A growing number of scientific fields suffer from a stifling embarrassment of riches. Data pile up faster than researchers can analyze them. At the Viterbi School's Information Sciences Institute, computer scientists are addressing that problem by building the prototype of a system that will automate scientific workflows.

Yolanda Gil leads the newly funded $13.8 million Windward Project, aimed at "Scaleable Knowledge Discovery through Grid Workflows."

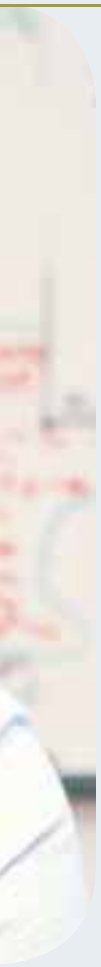
Gil says that in fields like climatology, high-energy physics and seismic modeling, "our ability to gather data is surpassing our ability to analyze it. Our data warehouses are becoming data graveyards."

In a sense, Windward will bring to analysis of scientific problems an approach that is similar to that of industrial engineering, where engineers create optimal workflows, so that raw material and machinery combine in the most efficient fashion to create products. But in today's world of scientific research, the product is not a physical item like an automobile or computer; rather, it is more often a model or an understanding. Efficient workflows to create it are equally critical, and because the raw material is information instead of matter, it is much easier to automate.

Gil and ISI collaborator Ewa Deelman co-chaired a National Science Foundation workshop on the subject in May 2006.

"Significant scientific advances today are achieved through complex distributed scientific computations," their overview for this workshop noted. "These computations, often represented as workflows of executable jobs and their associated dataflow, may be composed of thousands of steps that integrate diverse models and data sources."

The workshop held out the possibility that computer science would be able to channel this waterfall of data into orchestrated workflows, leading to recommendations for "basic work in computer science to create a science of workflows." The workshop suggested that scientists proactively



ISI's Yolanda Gil leads the $13.8 million Windward Project.

build workflow architecture into their research plans.

"Workflow representations that capture scientific analysis at all levels should become the norm when complex distributed scientific computations are carried out," concluded the overview.

Windward is an effort by Gil, who is principal investigator and project leader of the ISI Interactive Knowledge Capture research group, Deelman, and two fellow ISI project leaders, Paul Cohen and Carl Kesselman. They believe they can accomplish this ambitious task by integrating two longtime ISI specialties, artificial intelligence (AI) and grid computing.

AI tries to give computers power to respond accurately and appropriately to changing and novel circumstances, bringing multiple concerns to bear on the problem of making the right choice from a number of alternatives.

Cohen will build on his work at the ISI Center for Research in Unexpected Events, which has focused on AI systems for complex data analysis. He has been working specifically in the area of AI analysis of scientific data for years, publishing papers on "Intelligent Assistance for Computational Scientists: Integrated Modeling, Experimentation, and Analysis" 10 years ago, with work on planning systems going even farther back.

Cohen has also studied the history of science in certain fields to try to see patterns in the process of discovery. It is work that underlies the researchers' approach.

In order for AI systems to automate processes and provide assistance to scientists in defining workflows of complex computations, they need to have the world carefully structured and described.

Gil has long been active in developing the semantic web, which creates a digital universe that AI can explore and understand, and which will be a building block of the Windward system.

Previous AI systems have been much smaller than the regional, national and even intercontinental data structures needed to do workflow science.

This is where grid computing and Deelman and Kesselman come in. Since 1996, Kesselman has been perfecting the Globus software that allows multiple users in multiple locations secure, easy and transparent access, not just to raw data, but also to resources (computers) to process the data.

Linking to grid computing software, Deelman and her collaborators have developed a workflow management system, called Pegasus, that maps large numbers of computations to distributed resources while optimizing the overall performance of the application.

Deelman will continue to evolve Pegasus, which has already been successfully used in applications in the fields of astronomy, earthquake science, gravitational-wave physics and others.

> "...our ability to gather data is surpassing our ability to analyze it. Our data warehouses are becoming data graveyards."

The AI and grid computing groups at ISI have been collaborating in the area of scientific workflows for several years now, with notable results in earthquake science, in joint work with the Southern California Earthquake Center.

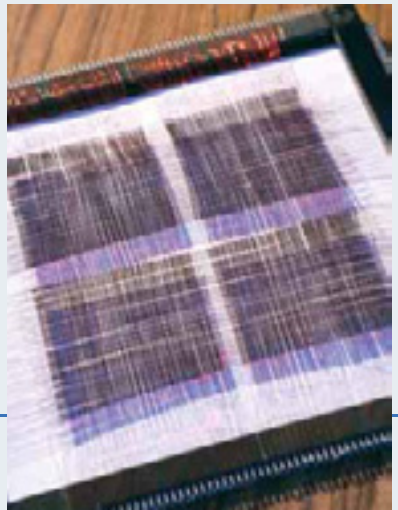
In the Windward project, they will develop new workflow techniques to represent complex algorithms and their subtle differences, so that they can be automatically selected and configured to satisfy the stated application requirements.

They will also investigate mechanisms to support autonomous and robust execution of concurrent workflows over continuously changing data.

In addition, they will develop learning techniques to improve the performance of the workflow system by exploiting an episodic memory of prior workflow executions.

Gil, with David DeRoure and Jim Hendler, co-edited a January 2004 special issue of IEEE Intelligent Systems journal on "E- Science," putting forth many of the ideas Windward will develop.

Funding for the project comes from the Air Force Research Laboratory. //



Computer science is helping researchers manage massive data flows.