

## **ELECTRONIC SEISMOLOGIST**

May/June 2003

Thomas J. Owens  
Department of Geological Sciences  
University of South Carolina  
Columbia, SC 29208  
Ph: 803-777-4530  
Fx: 803-777-0906  
Em: [owens@sc.edu](mailto:owens@sc.edu)

### **SEISMOLOGISTS ARE DOING IT!**

Since broaching the idea that we should embrace Information Technology in the geosciences in general and seismology in particular (ES, July/August 2002), the ES has learned something very important about IT ... everybody loves talking about it! And, lately, they've all been talking to the ES. In fact, if procrastination has its own "axis of evil", it would have to be talking about IT, sports, and the weather. But, fortunately for those of us who only talk about IT, there are talented, dedicated folks out there who are actually DOING IT. This month, the ES is pleased to host a column from one of the two big NSF ITR (Information Technology Research) program projects that have been funded recently in the geosciences. The SCEC (Southern California Earthquake Center) ITR project has been underway for about a year, focusing its attention on the use of information technology to build integrated approaches to seismic hazards analysis. Before turning over the podium to the SCEC ITR team, the ES would also like to draw your attention to another NSF ITR project that you should be following. The project is GEON (The Geosciences Network; [www.geon.org](http://www.geon.org)). GEON is a broad-based coalition of Earth scientists and computer scientists who have outlined a vision for the integration of multidisciplinary data in the Earth sciences through a modern information technology framework. Better still, they outlined their vision well enough to get funded to implement this vision. As you can see, it's been a good year for Geosciences IT. For his part, the ES promises to bring to this column not only his usual babbling about IT issues, but legitimate status reports by those who are actually "in the IT trenches"!

### **The SCEC Community Modeling Environment—An Information Infrastructure for System-Level Earthquake Science**

Thomas H. Jordan, Philip J. Maechling, and the SCEC/CME Collaboration  
Southern California Earthquake Center  
University of Southern California  
Los Angeles, CA 90089-0742  
Telephone: +1-213-740-5843  
Email: [tjordan@usc.edu](mailto:tjordan@usc.edu), [maechlin@usc.edu](mailto:maechlin@usc.edu)

In the July/August issue of *SRL*, Tom Owens, writing his first column as the Electronic Seismologist (ES), asked “The Question”: Can we as a community develop the long-term relationships with the computer science community, the financial resources, and the culture necessary to build and maintain a state-of-the-art information technology (IT) structure that will really make a difference in the way we do our science? He discussed the tough practical challenges posed by The Question and highlighted some of the IT efforts now underway by IRIS, EarthScope, and other geoscience organizations. In particular, he mentioned a new IT project by the Southern California Earthquake Center (SCEC) that involves an interesting mix of geoscientists and computer scientists. As a follow-up, he invited us to contribute an article to ES on how this project was formulated, how the collaborations have been set up, and how it’s going. Tall order, but we’ll try!

We begin with a little background. In October 2001, SCEC was awarded a 5-year, \$10-million grant by the NSF Information Technology Research (ITR) Program to develop a Community Modeling Environment (CME)—what IT aficionados would call a “collaboratory”—for system-level earthquake science. In addition to a number of participating SCEC institutions, the project involves IRIS (<http://www.iris.edu>), the USGS’s Pasadena office (<http://pasadena.wr.usgs.gov>), and two major IT research organizations, the Information Sciences Institute (ISI) of the University of Southern California (<http://www.isi.edu>) and the San Diego Supercomputer Center (SDSC) of the University of California, San Diego (<http://www.sdsc.edu>).

The overarching goal of CME project is to improve the information infrastructure for seismic hazard analysis (SHA), which has historically been the major focus of the SCEC’s research program. SHA seeks to describe the maximum level of shaking that can be expected at a specified site on the Earth’s surface due to earthquakes anticipated over a moderately long time span. The design requirements for the CME can be summarized as follows (see the project web site <http://www.scec.org/cme> for a more formal statement):

- (1) Permit users to rapidly prototype new SHA algorithms and products such as probabilistic seismic hazard analysis maps.
- (2) Enable execution of physics-based simulations and data inversions using current computer models of fault-system dynamics, rupture dynamics, wave propagation, and non-linear site response.
- (3) Manage the large datasets produced by these simulations as well as the associated datasets used in their calculations.
- (4) Provide access to SHA products to non-seismologists including engineers, emergency managers, and the general public.

As anyone who has been involved in IT can attest, such statements are very easy to formulate but terribly difficult to achieve. Rule 1 for Progress is *get specific!* We knew we had to boil the general SHA problem down considerably, so we framed the SCEC scientific objectives in terms of the four “computational pathways” diagrammed in Figure 1. The first three represent increasing levels of sophistication in the use of physics-based simulations to forward-model earthquake behaviors, while the fourth concerns a collection of important geoscience inverse problems. Given the limitations of time and money, we decided to focus our 5-year research plan on the first two pathways, for which the scientific methodology is best developed and the results can be most directly applied to SHA.

Pathway 1 portrays the current methodology of SHA, which combines *earthquake forecast models* with *attenuation relationships* to provide estimates (usually probabilistic) of ground-shaking *intensity measures*. The latter might include the peak ground acceleration (PGA), peak ground velocity (PGV), or the response spectral densities at specified frequencies. The earthquake forecast comprises a set of earthquake scenarios, each described by a magnitude, a location, and the probability that the scenario will occur by some future date. The attenuation relationship relates a possible earthquake scenario to the intensity of shaking, usually accounting for the local geologic conditions at each site (e.g., sediment sites tend to shake more than rock sites). The analysis determines the intensity that will be exceeded at some specified probability over a fixed period of time (e.g., PGA with a 10% probability of exceedance during the 50-year span). The results are often presented as hazard maps. Engineers use these maps to design buildings, emergency preparedness officials use them for planning purposes, and insurance companies use them to estimate potential losses.

Pathway 2 begins with an earthquake forecast model, but it employs the scenario events as sources for a physics-based calculation of ground motions. The waves from these sources are propagated using an *anelastic wave model* (AWM), and they excite ground motions at a specified location through a *site response model* (SRM), perhaps nonlinear, that accounts for the near-surface conditions, such as soil rigidity and layering. The results are vector-valued ground displacements as a function of time, from which essentially any intensity measure can be computed. However, in a region like Southern California where the geological structures are highly three-dimensional, the wavefield calculations must be done for each scenario earthquake on very dense grids to get the high frequencies of engineering interest ( $\geq 0.3$  Hz). The computational demands for these simulations can be enormous. One of the principal objectives of our project is to accelerate the use of ground-motion modeling in SHA.

As you can see, the geoscientists on the project are faced with a whole range of research issues. Each component in each pathway must be modeled and verified. The computations require parametric information about the faults and seismic velocity structure of the region, and these features need to be folded with other data (e.g., topography, crustal motions, stresses) into a consistent model of Southern California—the “unified structural representation” of Figure 1. Simulations must be run with a variety of input parameters to test the effect of various assumptions. We hope the CME will automate the construction of these computational pathways and the running of these simulations to a degree not previously available.

For the computer scientists, IT researchers, and software developers, the task of constructing a CME raises many challenges that currently impede progress toward a workable cyberinfrastructure for system-level problems like SHA. The IT issues are too numerous to catalog here, but we can describe a few:

- *Computational requirements.* The models become more complex as a user moves from Pathway 1 toward Pathway 4, requiring computational facilities not readily available to most research groups.
- *Distributed component development.* The CME will involve many components developed by different groups, so that the system must “know” how to incorporate a component into a computational pathway. This type of knowledge must be stored in a knowledge base that can be manipulated by “inference engines” in guiding users towards proper pathway assembly.
- *Data management.* Pathway 2 simulations can easily produce datasets in the gigabyte to terabyte range. Storing these large simulation datasets and identifying how they were

produced and what they represent (their “legacy” and “pedigree”) are essential for their scientific use. Users must be provided with easy and rapid access to the data.

To address these IT challenges, the CME collaboration is adopting appropriate existing technologies as well as performing fundamental IT research. To help us with the CME computational requirements, we have adopted a grid-based architecture that allows us to distribute our computations across multiple computer systems including high-performance systems at collaborating organizations. We are employing knowledge representation and reasoning (KR&R) methods to organize knowledge about our system and to reason across that knowledge. The data management issues—location-transparent storage, access to datasets, the association of metadata with datasets, and the ability to handle very large volumes of data—are being addressed with digital library technologies.

So much for plans and the vision thing... what have we actually accomplished during the first year? An essential first step was to form project teams structured around the Pathway 1 and 2 objectives. Here we obeyed Rule 2 for Progress: *require crosstalk!* A key feature of our organizational strategy was to involve both geoscientists and computer scientists on each development team to facilitate communication and cross-training between these disciplines. The teams have engendered a high degree of cooperation and enthusiasm among the investigators—everyone has a lot to learn, which makes it fun. SCEC also hired one of us (Maechling) as a full-time Information Architect, who will act as project manager and coordinate CME integration.

In the first year, a Pathway-1 team led by Ned Field erected a new object-oriented architecture for seismic hazard analysis, dubbed “OpenSHA” (<http://www.opensha.org>). This Java-based code implements a number of SHA conceptual objects, such as earthquake forecast models (EFM), intensity measure relationships (IMR), and intensity measure types (IMT). The team has thus far incorporated seven different IMRs that are applicable to Southern California and has developed an analysis tool that lets users explore the implications of the IMRs via a web-enable graphical user interface. The API between the IMRs and the analysis tool is very general and flexible, so that any new models can be plugged into the framework without having to change existing code. Along with the codes that calculate seismic hazard analysis curves, we have developed web-based analysis tools that allow the user to explore the implications of combining various EFMs with a number of possible IMRs. OpenSHA will thus provide the platform for integrating the research done by the SCEC working group on Regional Earthquake Likelihood Models (RELM) (<http://www.relm.org>). An overview of the OpenSHA architecture will be presented in a paper by N. Field, T. H. Jordan, and C. A. Cornell, soon to be published in *SRL*.

The OpenSHA framework has provided a very interesting and challenging initial application of our KR&R technology. A Pathway-1 team comprising SCEC scientists and AI researchers from ISI has developed an initial knowledge base for SHA objects such as EFMs and IMRs using a powerful KR&R inference engine named PowerLoom (<http://www.isi.edu/isd/LOOM/PowerLoom>). A web-based tool called DOCKER (Distributed Operations of Code with Knowledge-based descriptions for Earthquake Research) was developed to allow users to define and perform Pathway-1 calculations by accessing the SHA knowledge base. As the user sets up a computational pathway by specifying the hazard-curve parameters, DOCKER checks the user’s selections for consistency by querying the SHA knowledge base and warns the user of inconsistencies. Moving the SHA system into the

calculation of hazard maps will significantly increase the execution time, so this type of consistency checking should prove useful.

During this first year, the Pathway-2 wave-modeling team, comprising researchers from UC Santa Barbara, San Diego State University, Carnegie Mellon University, and UCSD, demonstrated the ability to generate realistic 4D wavefields representing the response of Southern California during an earthquake. The calculations incorporate the latest version of the SCEC 3D Community Velocity Model (CVM-V3.0; see <http://www.scecdc.scec.org/3Dvelocity/3Dvelocity.html>). The codes are finite-difference and finite-element models that propagate seismic waves through the Southern California region using time steps on the order of 1 s. The team has been systematically validating the models by intercomparing synthetic seismograms and comparing synthetics with observed data.

These Pathway-2 calculations furnished the target of our first-year efforts to set up software for computational grids. We have integrated the AWM software with the Globus grid toolkit (<http://www.globus.org>). The Globus toolkit is being adopted by a number of collaborating organizations including the University of Southern California (USC) High Performance Computing Center (HPCC) (<http://www.usc.edu/hpcc>), the San Diego Supercomputer Center (SDSC) (<http://www.sdsc.edu>), and the Pittsburg Supercomputer Center (PSC) (<http://www.psc.edu>) as a part of NSF's National Middleware Initiative (NMI) (<http://www.nsf-middleware.org>). Globus provides many services for distributed computing, include scheduling of jobs, naming and directory services, and data transfer—all performed with network-secure standards. Because we want to distribute computing over a heterogeneous collection of machines, there is no single operating system that can provide the requisite services. By incorporating the Globus grid software into the CME system, we can deal with these issues in a uniform manner.

As part of this effort, we have set up a “SCEC grid testbed,” which now includes two shared memory Sun servers along with a 1000+ CPU Linux cluster. By creating appropriate grid configuration scripts, we can queue our codes to the Sun servers for smaller jobs, or to the Linux cluster for larger jobs. The grid software helps us dispatch and manage our jobs, but it doesn't (yet) eliminate administrative requirements such as applying for allocations on the supercomputers. Our initial experience with integrating the Pathway 2 AWM software into this testbed has demonstrated that the grid software makes distributing and dispatching these codes on a multi-node system significantly easier.

Pathway 2 is also where we have focused our metadata development and our application of digital library technology. Since the datasets produced by our AWMs can be quite large, we don't want to duplicate or regenerate them needlessly. A Pathway-2 team led by IRIS has produced a draft wavefield metadata description which is being reviewed by the modeling team. Our design strategy is to define a minimal set of required metadata for each type of dataset, possibly supplemented with code-specific extensions, which the user will have to fill in before the system will generate a dataset.

Our digital library technology integration began with establishment of a web-accessible CME data repository using the Storage Resource Broker (SRB) technology developed by Reagan Moore and his colleagues at SDSC (<http://www.npaci.edu/DICE/SRB>). CME documents and datasets, along with their metadata, can be published, stored, and accessed via this system.

The first year of the SCEC/CME Project saw significant developments in the area of data visualization. Because the CME will generate enormous quantities of simulation data, more sophisticated techniques for evaluating the data will be needed. A Pathway-2 team led by ISI

developed a visualization capability for the 4D datasets produced by some of the AWM codes, which can display a time varying image of the wavefield as it propagates through the 3D volume. This visualization gives the modelers an alternative to viewing 2D seismograms derived from the wavefield datasets.

Another 3D visualization capability was created last year by the SCEC/CME Summer Intern Program. This group comprised 12 undergraduate and graduate students studying computer science and engineering, Earth science, mathematics, economics, and communications. The challenge put to the students was to use data visualization technology to answer a simple question, “Do most Southern California earthquakes occur on major (mapped) faults?”

To investigate this issue, the students developed a 3D geo-referenced display of active faults and hypocenters for southern California (<http://www.scec.org/geowall>). The fault information was based on a preliminary release of the SCEC Community Fault Model (CFM-A; see <http://structure.harvard.edu/cfma>), and the hypocenters included historic earthquake catalogs as well as recent events acquired in near-real-time from the California Integrated Seismic Network (CISN) server at Caltech (<http://www.trinet.org>). The software was developed in Java 3D and runs on a portable 3D visualization system originally developed by Paul Morin and his colleagues of the Geowall Consortium (<http://geowall.geo.lsa.umich.edu>). The Geowall system is a PC-based system that uses polarized projection for 3D rendering. When the southern California faults and recent hypocenters were plotted in 3D, the students had visual evidence that most of the Southern California earthquakes are *not* occurring on mapped faults.

Given all this activity and progress, we have some ambitious plans for CME Year 2. Our activities will be governed by Rule 3 for Progress: *integrate early and often!* SCEC is assembling a complete CME test bed, based on a distributed set of computers, where CME software components can be collected and integrated. A major task is to develop standardized data storage formats and standard APIs to access these datasets. The largest datasets, such as the 4D wavefield simulations, have first priority because they are expensive to create, store, and access.

We are integrating our Pathway 1 and Pathway 2 calculations to verify that the simulation results become more accurate as we develop more sophisticated, physics-based models. We have identified comparable predictive datasets that can be produced by Pathway 1 and Pathway 2, and we plan to perform comparable simulations. Both Pathway 1 and Pathway 2 codes will be enabled to read in information from the SCEC CVM and CFM through standardized interfaces. We anticipate that the web services model, in which data are automatically accessible to computers via web service calls, will be used extensively in our CME system.

We are porting our AWM codes to run on high performance computing clusters. To make use of clustered computers, we are modifying the codes to employ special programming languages and libraries such as the Message Passing Interface (MPI). Some of our models will be modified to use other parallel programming techniques such as parallel I/O.

We are continuing work on the data visualization. The modelers are working with the visualization experts to develop displays that help viewers extract information from the data. Side-by-side, synchronized, wavefield viewers (for comparing data sets produced by different models) and 3D earthquake sequence animations are currently under development.

And so on... Writing this report reminds us that, no matter what's been accomplished, there's so much yet to do! The CME system has the potential to change SHA, and earthquake science more generally, but our tentative steps during the first year of the SCEC/CME Project have made us well aware that establishing a workable and useful system will be very difficult. Please visit

us at (<http://www.scec.org/cme>) for more information about the project and updates about our current activities. We are interested in your comments and feedback.

Indeed, answering The Question will be a challenge for the entire community.



## Figure Caption

**Figure 1.** Computational pathways that will be facilitated by the information infrastructure developed in the SCEC/CME Project. (1) Current methodology for probabilistic seismic hazard analysis. (2) Ground-motion prediction using an anelastic wave model (AWM) and a site-response model (SRM). (3) Earthquake forecasting using a fault-system model (FSM) and a rupture-dynamics model (RDM). (4) Inversion of ground-motion data for parameters in the unified structural representation (USR), which includes 3D information on active faults, tectonic stresses, and seismic wave speeds.