

SCEC Earthworks Science Gateway: Interactive Configuration and Automated Execution of Earthquake Simulations on the TeraGrid

Philip Maechling¹, Joanna Muench², Hunter Francoeur¹, David Okaya¹, Yifeng Cui³, Ewa Deelman⁴, Gaurang Mehta⁴, Reagan Moore³, Thomas Jordan¹

- (1) Southern California Earthquake Center, Los Angeles, CA, 90089
{maechlin, francoeu, okaya, tjordan}@usc.edu
- (2) Incorporated Research Institutions for Seismology, Seattle, WA 98105
{joanna}@iris.washington.edu
- (3) San Diego Supercomputer Center, La Jolla, CA 92093
{yfcui, moore}@sdsc.edu
- (4) USC Information Sciences Institute, Marina Del Rey, CA 90292
{deelman, gmehta}@isi.edu

Abstract

The SCEC Earthworks Science Gateway is a portal-based scientific workflow system designed to help members of the SCEC geoscience community perform computationally-intensive, geophysical research using TeraGrid resources. The SCEC Earthworks Science Gateway allows users to configure and execute earthquake wave propagation simulations using well validated geophysical models and high performance simulation software. The Earthworks system generates a set of data products including surface seismograms and ground motion maps. Users access the SCEC Earthworks system through a web-based portal built using the GridSphere Portlets engine. Users can configure, submit, and monitor wave propagation simulations. They can also submit verification and validation simulations to ensure new versions of the Geoscientific codes work properly. Users can also access the resulting simulation data products. All steps in the wave propagation simulations including mesh generation, wave propagation, and post processing are run using a scientific workflow system based on the Pegasus workflow management system, Condor DAGMan, and the Globus toolkit. Long term data storage is provided by the Storage Resource Broker (SRB).

1. Introduction

Geoscientists working within the Southern California Earthquake Center (SCEC) community have developed sophisticated, well-validated, computationally-intensive, earthquake wave propagation simulation codes. These simulation capabilities are of significant value to the broad SCEC community. However, the knowledge required to configure and run these simulations on high-performance computing

systems presents a significant barrier to entry for some members of the community that are not frequent users of high-performance computing programs and computer systems. In addition, users must often run a whole series of codes in a particular sequence in order to produce the data product of interest.

To expand the SCEC community access to these simulation capabilities, researchers and software developers on the SCEC Community Modeling Environment (SCEC/CME) are developing the SCEC Earthworks Science Gateway. The SCEC Earthworks portal allows users to configure, submit and monitor model simulations, as well as access the resulting simulation data products. Through the use of grid-based workflow technology, the Earthworks system automates the sequential processing steps needed to produce common data products. The SCEC Earthworks portlet-based User Interface allows users to browse simulation data products, save configurations, as well as share simulations with other users.

The SCEC Earthworks Science Gateway represents an important step in enabling wider SCEC community access to high-performance computing resources. We believe that it lowers the barrier to entry into high-performance computing for scientists and that it will facilitate the research goals of the SCEC community by allowing members of the community to share metadata-driven seismic simulations and their resulting data products.

2. Scientific Basis

Physics-based earthquake simulations can potentially provide enormous practical benefits for assessing and mitigating earthquake risks through seismic hazard analysis. For example, recent advances in 3D geological velocity models and anelastic wave propagation

simulation codes make it possible to perform highly accurate earthquake simulations that produce suites of synthetic seismograms for scenario earthquakes. If the scenario earthquakes are historic events, for which observed seismograms are available, the synthetics can be compared to the observed seismograms to validate the geological models and processing codes. If the scenario events are anticipated future events, the seismograms that are produced can be used by scientists and building engineers to better understand the strong ground motions that may be observed in the future.

These highly accurate earthquake simulations are performed using high-performance simulation software called Anelastic Wave Models (AWM). AWM's compute the propagation, interference, and attenuation of seismic waves that travel from a fault rupture to a target site. The results are vector-valued ground displacements as a function of time, from which essentially any intensity measure can be computed.

The SCEC Earthworks Science Gateway is a system designed to compute and distribute ground-motion simulations for use in risk assessment and earthquake-engineering analysis. Such catalogs are needed, for example, as input to research done at NSF's earthquake engineering research centers and its Network for Earthquake Engineering Simulation (NEES).

The SCEC Earthworks system uses the 3D SCEC Community Velocity Model V3 [1] and the AWM-Olsen software [2]. This combination of velocity model and AWM software is used extensively on the SCEC's TeraGrid simulations including the large-scale SCEC TeraShake simulations [3]. While the SCEC Earthworks system is not designed to run simulations on the TeraShake scale (more than 1 billion mesh points), it will enable SCEC scientists to run the well-validated and highly scaleable AWM-Olsen code on the TeraGrid through an easy to use portal-based interface.

Given the broad interest in these AWM computations, providing a facility for scientists to both run their own simulations and view the results of other simulations will enhance the research aims of the SCEC community.

3. Architectural Overview

The SCEC Earthworks Science Gateway architecture is driven by two, sometimes competing, requirements. First, the system requires a simple, accessible, and easy-to-use interface. Second, the system must

support large-scale data management and computational capabilities in a secure and reliable manner in a heterogeneous computing environment that includes SCEC, USC, and the TeraGrid computing facilities.

To address the first requirement, we have implemented a portal-based user interface using the GridSphere portlet framework. Through custom design portlets, users can configure, save, submit, monitor, and access the results of complex earthquake simulations.

To address the second requirement and support the end-to-end analysis that involves many inter-dependent computational steps, we model the AWM simulations as scientific workflows. We then utilize the grid-based SCEC/CME workflow system that has been developed over several years on the SCEC/CME Project to configure, submit, and execute the SCEC Earthworks workflows [4]. The workflows then execute on the TeraGrid and on other grid-based computing resources that are shared on the SCEC grid. The workflow system used by SCEC is composed of the Pegasus workflow mapper and manager [5] and of the Condor DAGMan workflow executor [7].

The SCEC Earthwork system combines the GridSphere portal tools with the Pegasus/DAGMan software stack to provide a scientific gateway to the TeraGrid as shown in Figure 1.

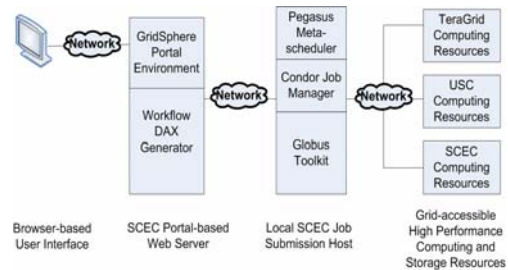


Figure 1: The SCEC Earthworks Science Gateway provides a portal based interface for configuring and running scientific workflows on the TeraGrid.

4. Portal User Interface

The SCEC Earthworks portal-based User Interface is designed to make it easy to configure and specify an earthquake simulation. An example screenshot from this SCEC Earthworks Portal-based interface is shown in Figure 2. The user interface is designed simplify the specification of simulation parameters as

well as to make the workflow steps and the technology supporting them invisible to users.

Through the SCEC Earthworks portal, users configure, submit and monitor wave propagation simulations. They can also search for data products of their own or shared simulations. Portlets running within the Earthworks portal allow users to save workflow definitions, easing the task of repeating simulations or comparing results of minor modifications to a simulation. This feature also eases the important process of validating simulation codes against standards.

The AWM-Olsen finite difference model underlying SCEC Earthworks is part of an active research initiative. Therefore the inputs required to run a simulation may change. To accommodate these expected changes, the definitions of model inputs are contained within XML descriptor files, enabling researchers to improve the underlying simulations without re-writing the portal interface.

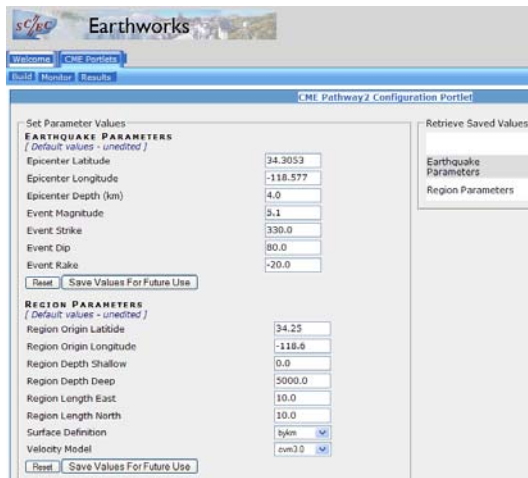


Figure 2: The SCEC Earthworks Science Gateway portlet-based User Interface allows SCEC scientists to configure, and save, the parameters that define an earthquake simulation. The UI also allows the user to submit the simulation for execution on the TeraGrid and to track its progress and view the resulting data products.

The GridSphere portlet engine allows users to be given different levels of access. In the SCEC Earthworks Gateway, only privileged users may run simulations. A user can elect to share a simulation, or keep the results private. A single TeraGrid allocation, awarded to the SCEC Earthworks Gateway project, supplies the required service units to run simulations. This

decreases the administrative overhead of managing multiple allocations. Total simulation size, based on the product of the number of grid-points times the number of timesteps, is limited to conserve resources.

Non-privileged users can search and download simulation data products from all shared simulations. The search capabilities include search by product type (e.g. seismogram, PGV map) and meta-driven search for simulation data products. Users can build complex queries over multiple parameters to discover data products for simulations relating to their individual area of research, such as event location or a specific source description. Seismogram data products can be compared to observed seismograms retrieved through the Incorporated Research Institutions in Seismology (IRIS) Data Handling Interface (DHI), which provides access to observed data including earthquake catalogs and seismograms.

5. On-Demand Verification Capabilities

When earthquake wave propagation simulation codes, or data analysis codes, are integrated into an automated processing system such as SCEC Earthworks, it is important to ensure that the automatic results are equivalent to the results produced when the codes are run manually by scientists. Also, when modifications are made to any of the scientific codes, the codes must be re-verified before the results are scientifically useful. In order to address the need for code and system verification, we have integrated what we term “on-demand verification” capabilities into the system.

The SCEC Earthworks on-demand verification capabilities are implemented through the use of verification problems. Earthworks allows the user to select and specify simulation problems that have either known analytic solutions or well-accepted reference solutions. Earthworks then will execute one or more verification problems and the results produced by the Earthworks workflow are automatically compared against known-good results. Differences between the new and the known-good results indicate a problem with the system while results that match help to revalidate the system.

6. Workflow Construction

All steps in the wave propagation simulations are run using a grid-based workflow system incorporating the Pegasus workflow management system, the Condor DAGMan

workflow engine, and the Globus toolkit. Using Java and an API provided by collaborators from ISI, SCEC Earthworks provides an interface for creating abstract scientific workflows that represent all of the steps needed to run a wave propagation simulation.

Parameters collected through the Earthworks UI are transformed into workflow instructions called an abstract workflow, or DAX (Directed Acyclic Graph in XML format). A DAX specifies the sequence of computations, and the data dependencies in the workflow. However, a DAX is resource independent in that it does not specify which specific computers will be used, or the physical location of the files used in the workflow. These specifics will be provided by the Pegasus at a later stage in the workflow processing.

The Earthworks backend submits the DAX to the Pegasus system that determines where appropriate computational resources are available on the SCEC grid [6]. The computational resource providers on the SCEC grid include local SCEC machines, ISI computers, USC High Performance and Communications systems, and TeraGrid systems. Pegasus also selects appropriate physical files to use in the workflow by querying a Replica Location Service (RLS) [8].

Once Pegasus translates the abstract workflow description (the DAX) into a concrete workflow (a DAG), the DAG is submitted to DAGMan, Condor-G, and then workflow jobs are sent to the site-specific schedulers where they are handed off to Globus.

A SCEC Earthworks user can monitor the progress of submitted workflows using data obtained from the scheduler on the original submit host. This data is then translated and displayed to the user through the Earthworks web portal interface. Users also have the option of canceling a workflow.

7. Metadata Management

Metadata management is a key to understanding the provenance of each simulation run by the system, so the SCEC Earthworks system maintains comprehensive metadata about each data product produced by the system. The user provides the initial metadata for the simulation as they define the simulation parameters. Then, additional metadata is added by each processing step in the workflow. The metadata attributes describing each data product resulting from a workflow are registered as attribute-name, attribute-value pairs within a

Metadata Catalog Service (MCS) [9]. MCS stores metadata in a database and provides an API that enables data product search and discovery through metadata searches.

8. Data Products

The SCEC Earthworks produces a collection of data products for each wave propagation simulation. When a SCEC Earthworks workflow successfully completes, the resulting data products are placed both on disk at SCEC and in SCEC's Storage Resource Broker [10] (SRB)-based digital library at SDSC. Metadata describing each data product is registered with the both MCS and the SRB. The data products include three component synthetic seismograms and peak ground motion maps. Access to these products is provided through the portal interface and the SRB.

9. Conclusions

The SCEC Earthworks Gateway project represents an important step towards making the computing power of the TeraGrid available to scientists. The underlying architecture of the science gateway removes the burden of understanding the complexity of grid computing from the user. The integration of on-demand verification into the system helps to build confidence in the system results. We believe that the SCEC Earthworks system provides capabilities needed by many SCEC researchers and that it will broaden the SCEC community using the high performance computing capabilities of the TeraGrid.

10. Acknowledgements

This work was supported by the SCEC Community Modeling Environment Project which is funded by the National Science Foundation (NSF) under contract EAR-0122464 (The SCEC Community Modeling Environment (SCEC/CME): An Information Infrastructure for System-Level Earthquake Research). This research was supported in part by the Southern California Earthquake Center (SCEC). SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. The SCEC contribution number for this paper is 989.

11. References

[1] Olsen, K.B., Day, S.M., and C.R. Bradley (2003). Estimation of Q for long-period (>2 s) waves in the Los Angeles Basin, *Bull. Seis. Soc. Am.* 93, 627-638. Olsen, 1994;

- [2] Magistrale, H., S. Day, R. Clayton and R.W. Graves (2000) The SCEC southern California reference three-dimensional seismic velocity model version 2 *Bulletin of the Seismological Society of America* 90 no. 6B
- [3] Olsen, K. B., S. M. Day, J. B. Minster, Y. Cui, A. Chourasia, M. Faerman, R. Moore, P. Maechling, and T. Jordan (2006), Strong shaking in Los Angeles expected from southern San Andreas earthquake, *Geophys. Res. Lett.*, 33, L07305, doi:10.1029/2005GL025472.
- [4] Maechling, P., H. Chalupsky, M. Dougherty, E. Deelman, Y. Gil, S. Gullapalli, V. Gupta, C. Kesselman, J. Kim, G. Mehta, B. Mendenhall, T. Russ, G. Singh, M. Spraragen, G. Staples, K. Vahi (2005) Simplifying Construction of Complex Workflows for Non-Expert Users of the Southern California Earthquake Center Community Modeling Environment, *ACM SIGMOD Special issue on Scientific Workflows*, Record Vol. 34 No. 3, 24-30
- [5] Deelman, E. G. Singh, M-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob, D. S. Katz. "Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems" *Scientific Programming Journal*, Vol 13(3), 2005, Pages 219-237
- [6] Deelman, E., J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, K. Blackburn, A. Lazzarini, A. Arbre, R. Cavanaugh, and S. Koranda, "Mapping Abstract Complex Workflows onto Grid Environments," *Journal of Grid Computing*, vol. 1, pp. 25-39, 2003
- [7] Frey, J. T. Tannenbaum, I. Foster, M. Livny, and S. Tuecke, "Condor-G: A Computation Management Agent for Multi-Institutional Grids.," *Cluster Computing*, vol. 5, pp. 237-246, 2002.
- [8] Chervenak, A. E. Deelman, et al., "Giggle: A Framework for Constructing Scalable Replica Location Services," *Proceedings of Supercomputing 2002 (SC2002)*, Baltimore, MD. 2002.
- [9] Deelman, E. G. Singh, M. P. Atkinson, A. Chervenak, N. P. Chue Hong, C. Kesselman, S. Patil, L. Pearlman, M.-H. Su "Grid-Based Metadata Services," 16th International Conference on Scientific and Statistical Database Management (SSDBM04), 21-23 June 2004 Santorini Island Greece.
- [10] Baru, C. et al., "The SDSC Storage Resource Broker," *Proceedings of Proc. CASCON'98 Conference*, 1998.