Managing Large Scale Data for Earthquake Simulations

Marcio Faerman[1], Reagan Moore[2], Bernard Minister[3], and Philip Maechling[4]

[1]San Diego Supercomputer Center
9500 Gilman Drive, La Jolla, CA, USA
mfaerman@gmail.com
[2]San Diego Supercomputer Center
9500 Gilman Drive, La Jolla, CA, USA
moore@sdsc.edu
[3]Scripps Institution of Oceanography
9500 Gilman Drive, La Jolla, CA, USA
jbminster@ucsd.edu
[4]University of Southern California
Los Angeles, CA, USA
maechlin@usc.edu

Abstract. The Southern California Earthquake Center digital library publishes scientific data generated by seismic wave propagation simulations. The output from a single simulation may be as large as 47 Terabytes of data and 400,000 files. The total size of the digital library is over 130 Terabytes with nearly three million files. We examine how this scale of simulation output can be registered into a digital library built on top of the Storage Resource Broker data grid. We also examine the multiple types of interactive services that have been developed for accessing and displaying the multi-terabyte collection.

1 Introduction

The Southern California Earthquake Center (SCEC) [1], in collaboration with the San Diego Supercomputer Center (SDSC), the Scripps Institution of Oceanography, the Information Sciences Institute, the Incorporated Research Institutions for Seismology, and the U.S. Geological Survey, is developing the Southern California Earthquake Center Community Modeling Environment (SCEC/CME) [2] under a five-year grant from the National Science Foundation's Information Technology Research (ITR) Program jointly funded by the Geosciences and Computer and Information Science & Engineering Directorates. Recent advances in earthquake science, combined with the increased availability of Terascale computing resources, have made it practical to create fully three-dimensional (3D) simulations of fault-system dynamics. These physics-based simulations can potentially provide enormous practical benefits for assessing and mitigating earthquake risks through seismic hazard analysis.

The SCEC/CME system is an integrated geophysical simulation modeling framework that automates the process of selecting, configuring, and executing models of earthquake systems. By facilitating the investigation, modification, and adoption of these physics-based models, the SCEC/CME can improve the scientist's system-level understanding of earthquake phenomena and can substantially improve the utilization of seismic hazard analysis. The collaboration generates a wide variety of data products derived from diverse earthquake simulations. The datasets are archived in the SCEC Community Digital Library, which is supported by the San Diego Supercomputer Center (SDSC) Storage Resource Broker (SRB) [3], for access by the earthquake community. The digital library provides multiple access mechanisms needed by geophysicists and earthquake engineers.

In this paper we describe the large scale data management issues facing the SCEC/CME collaboration and the activities related to the development of the SCEC Community Digital Library.

2 SRB Support for SCEC Digital Library

The SRB data grid provides the distributed data management needed for a shared collection that resides on multiple storage systems. The underlying capabilities can be characterized as data virtualization, trust virtualization, latency management, collection management, and federation management [4]. Data virtualization is the management of data collection properties independently of the storage repositories. It is achieved by providing two levels of indirection between the application code and the underlying storage system. The SRB supports registration of digital entities, which may be files, URLs, SQL command strings, directories, objects in object ring buffers, and tables in databases. Digital entities (including files) are logically named, enabling the creation of a collection hierarchy that spans storage repositories. The properties of the digital entities (owner, size, creation date, location, checksum, replicas, aggregation in containers) are managed as state information by the SRB independently of the storage repository, and stored in a central metadata catalog. Storage resources are logically named, enabling collective operations such as load leveling across storage repositories. Finally, the SRB provides a uniform set of operations for interacting with the storage systems, and support for multiple access protocols including digital library systems such as DSpace [5] and Fedora [6]. The two levels of indirection make it possible for all access mechanisms to be used across any of the supported storage systems (Unix and Windows file systems, archival storage systems, object ring buffers, and databases).

The SRB architecture is shown in Figure 1. The system is highly modular, enabling the addition of new storage system drivers without requiring modification to the clients. Similarly, new application clients can be added without having to change the storage system drivers. The top level indicates the set of access clients that are supported. The bottom level identifies the types of storage systems that can be accessed. The databases that can be used to manage the state information are listed.

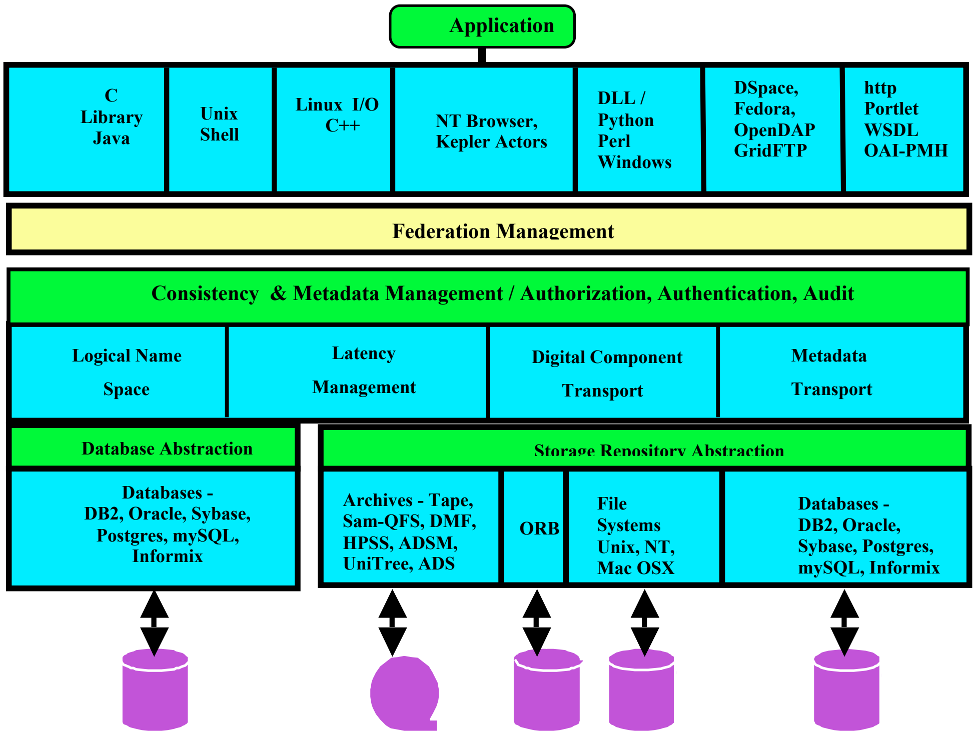

Application
C
Library
Java
Unix
Shell
Linux I/O
C++
NT Browser,
Kepler Actors
DLL /
Python
Perl
Windows
DSpace,
Fedora,
OpenDAP
GridFTP
http
Portlet
WSDL
OAI-PMH
Federation Management
Consistency & Metadata Management / Authorization, Authentication, Audit
Logical Name Space
Latency Management
Digital Component Transport
Metadata Transport
Database Abstraction
Storage Repository Abstraction
Databases - DB2, Oracle, Sybase, Postgres, mySQL, Informix
Archives - Tape, Sam-QFS, DMF, HPSS, ADSM, UniTree, ADS
ORB
File Systems Unix, NT, Mac OSX
Databases - DB2, Oracle, Sybase, Postgres, mySQL, Informix

Figure 1. Block architecture of the Storage Resource Broker

Trust virtualization is the management of authentication, authorization, and audit trails independently of the storage repositories. The SRB accomplishes this by assigning ownership of the files to the SRB data grid. The SRB manages distinguished names for the users of the data grid independently of the storage system. Access controls are then managed as constraints imposed between the logical resource name, the logical file name, and the SRB distinguished user names. These controls remain invariant as the files are moved between storage systems under SRB control. The SRB supports multiple access roles, including a curator who manages the publication of data and metadata in the SCEC Digital Library at /home/sceclib.scec. Access roles include read permission, write permission, creation of audit trails, and data grid administrator. Public access to the SCEC digital library is managed through a "public" account that is given read access permission. Access controls can be set on users, metadata, and storage resources.

Latency management provides support for parallel I/O, bulk file operations, and remote procedures to minimize the number of messages sent over wide area networks. Files are moved between storage systems using TCP/IP and large data blocks. Parallel I/O streams are used for large files (greater than 10 MBs in size). Small files are aggregated before transport, and usually are kept aggregated in containers in the storage system. Remote procedures are used to support metadata extraction, file manipulation, and file filtering operations directly at the remote storage system. Bulk operations include the ability to register directory structures into SRB collections.

Collection management provides the mechanisms needed to manage a metadata catalog that resides in a chosen vendor database. The mechanisms include support for bulk metadata import, import and export of XML files, dynamic SQL generation, extensible schema, user-defined metadata, synchronization of master/slave catalogs, and attribute based queries.

Federation management provides support for synchronizing the logical name spaces between two or more independent data grids. In practice, projects that manage data distributed on an intercontinental scale build separate data grids in each continent. They then federate the data grids to enable data sharing between the independently managed systems. Thus the SCEC digital library can be federated with another digital library elsewhere in the world.

The SRB is implemented as peer-to-peer application-level server software that is installed at each storage system. The SRB server runs under a dedicated user account. A central metadata catalog tracks updates to all state information. Each peer server is able to forward requests to other peer servers within the data grid. All communication between users and the metadata catalog and between servers is authenticated. The results of operations are managed as metadata associated with each file.

The SRB is used in multiple national and international projects to implement persistent archives, digital libraries and data grids [7]. For the SCEC digital library, the SRB data grid manages a shared collection that is distributed across storage systems at SDSC, San Diego State University, Scripps Institution of Oceanography, and University of Southern California. SDSC provides storage space to SCEC for over 130 Terabytes of tape archive, and on average 4 Terabytes of on-line disk space. Cache disk space to support applications on average is about 5 Terabytes. A public read-only account is available to access SCEC collections registered in the Storage Resource Broker (SRB), from a web browser, using the URL: http://www.sdsc.edu/SCEC/.

3 Generating Synthetic Data – Earthquake Simulations

Research conducted by the SCEC/CME collaboration includes TeraShake, a set of large scale earthquake simulations occurring on the southern portion of the San Andreas Fault. The southern portion of the San Andreas Fault, between Cajon Creek and Bombay Beach has not seen a major event since 1690, and has therefore accumulated a slip deficit of 5-6 m. The potential for this portion of the fault to rupture in a single magnitude 7.7 event is a major component of seismic hazard in southern California and northern Mexico. TeraShake is a set of large-scale finite-difference (fourth-order) simulations of such an event based on Olsen's Anelastic Wave Propagation Model (AWM) code [8], and conducted in the context of the Southern California Earthquake Center Community Modeling Environment (CME). The fault geometry is taken from the 2002 USGS National Hazard Maps. The kinematic slip function is transported and scaled from published inversions for the 2002 Denali (M7.9) earthquake. The three-dimensional crustal structure is represented by the SCEC Community Velocity model. The 600km x 300km x 80km simulation domain extends from the Ventura Basin and Tehachapi region to the north and to Mexicali and Tijuana to the south. It includes all major population centers in southern California, and is modeled at 200m resolution using a rectangular, 1.8 giganode, 3000 x 1500 x 400 mesh. The simulated duration is 250 seconds, with a temporal resolution of 0.01seconds, maximum frequency of 0.5Hz, for a total of 22,728 time steps.

3.1 Simulation Support

The simulation was run at SDSC on 240 processors of the IBM Power4, DataStar machine [9,10]. Validation runs conducted at one-sixteenth (4D) resolution have shown that this is the optimal configuration in the trade-off between computational and I/O demands. Each time step produced a 20.1GByte mesh snapshot of the entire ground motion velocity vectors. A 4D wavefield containing 2,000 time steps, amounting to 39 Tbytes of data, was stored at SDSC. Surface data was archived for every time step for synthetic seismogram engineering analysis, totaling 1 Tbyte. The data was registered with the SCEC Digital Library supported by the SDSC Storage Resource Broker (SRB). Data collections were annotated with simulation metadata, which allowed data discovery operations on metadata-based queries. The binary output is described using Hierarchical Data Format headers. Each file was fingerprinted with MD5 checksums to preserve and validate data integrity. Data access, management and data product derivation are provided through a set of SRB APIs, including java, C, web service and data grid workflow interfaces. High resolution visualizations of the wave propagation phenomena are generated as one of the principle analysis methods. The surface velocity data is analyzed and displayed as point seismograms, spectral acceleration, and the resulting surface displacement.

The large-scale TeraShake simulation stretched SDSC resources across the board with unique computational as well as data challenges. The code was enhanced so it would scale up to many 100s of processors and run for the very large mesh size of 1.8 billion points. SDSC computational experts worked closely with the AWM model developer to port the code to the IBM Power4 platform and resolve parallel computing issues related to the large simulation. These issues include MPI and MPI I/O performance improvement, single-processor tuning and optimization, etc. In particular special techniques were introduced that reduced the code's memory requirements, enabling the full TeraShake simulation to run. This testing, code validation, and performance scaling analysis required 30,000 allocation hours on DataStar to prepare for the final production run. The final production case, covering a mesh of 3000*1500*400 cells, was run on 240 processors for 4 days and produced 47 Terabytes of data on the GPFS parallel file system of IBM Power4. The mapping of the mesh onto the processors is shown in Figure 2.

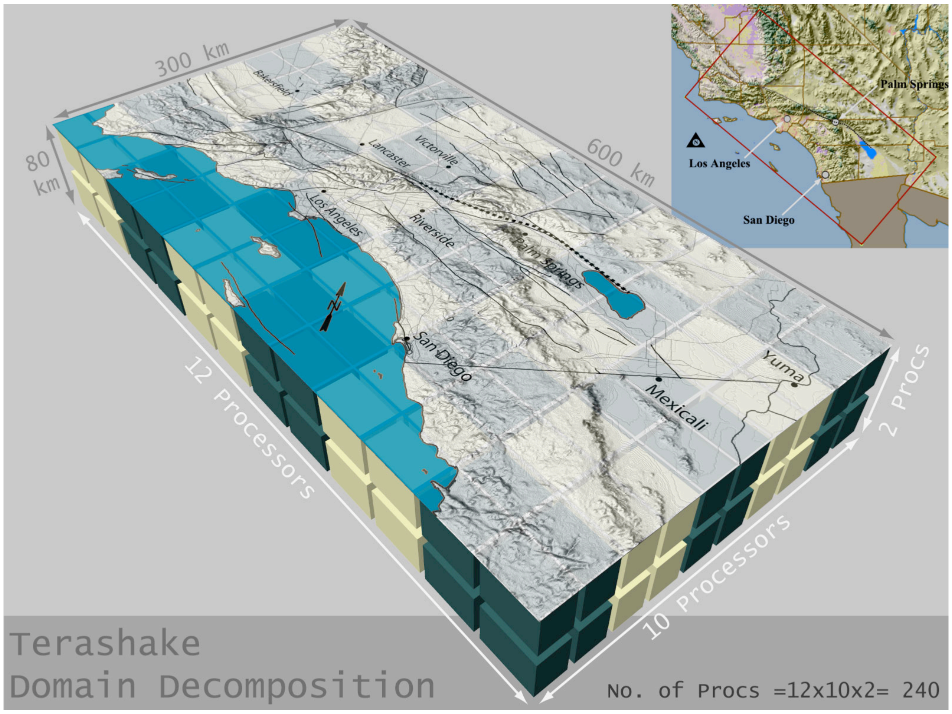


Figure 2 – TeraShake domain decomposition on SDSC IBM Power4 DataStar P655 nodes. The computation was parallelized across 240 processors.

SDSC's computational collaboration effort was supported through the NSF-funded SDSC Strategic Applications Collaborations (SAC) and Strategic Community Collaborations (SCC) programs. TeraShake is a great example of why these programs are so important for the academic computational scientists. These programs allow SDSC computational experts to develop close collaborations with academic researchers who use SDSC's supercomputers and facilitate the scientists' research and enable new science like TeraShake. The particular example of Terashake SAC/SCC work also provides lasting value, with an enhanced code that gives increased scalability, performance, and portability. This optimized code is now available to the entire earthquake community for future large-scale simulations.

3.1.1 TeraShake Production Runs

Four TeraShake production runs were conducted. All the runs were executed on 240 processors, determined to be the most efficient processor setup, balancing parallelism and inter-processor communication overhead. Volume data was generated at each 10th step in the first run. Surface data was generated and archived for every time step in all runs. Checkpoint files were created at each 1000th step in case restarts were required due to reconfigurations or eventual run failures. Table 1 shows the details about the setup of each run. For the TeraShake mesh size of 3000x1500x400, 230 GigaBytes of memory were required.

Table 1 – TeraShake production runs

Run	Rupture Direction	Number of Steps	Surface Output	Volume Output	Checkpoint Files
1	NW-SE	20,000	1TB	40TB	3TB
2	SE-NW	22,728	1TB	-	3TB
3	NW-SE	22,728	1TB	-	3TB
4	SE-NW	22,728	1TB	5TB	3TB

The first run consumed 18,000 CPU hours, executing during a period of about 4 days. The second and third runs consumed 12,000 CPU hours. The additional execution time of the first run refers to the I/O operations required to produce the volume output amounting to 40TB. Based on an analysis of the data from the first run, the amount of volume data generated was reduced to cover specific areas of interest, typically contained within the checkpoint files. This substantially decreased the amount of storage required for the additional runs.

3.1.2 Run-time data management

The AWM parallel I/O routines wrote the output data to the DataStar GPFS parallel disk cache, containing the simulation steps snapshots. As soon as the files started being generated, post-processing, visualization and archiving were also started. The simulation output data were transferred from the GPFS disk cache to the SDSC permanent archives during run-time, to maintain adequate disk cache availability to other DataStar users and for the TeraShake simulation itself. The volume output was archived in about 5 days, finishing one day after the completion of the first run. The data was registered at the SCEC Community Digital Library, supported by the SDSC Storage Resource Broker.

The data management was highly constrained by the massive scale of the simulation. The amount of disk space that was available on DataStar disk was only 36 TBs, while the simulation was expected to generate nearly 43 TBs. As the application ran, the data had to be moved in parallel onto the archival storage systems at SDSC to guarantee enough disk space for the simulation to complete. Two archival storage systems were used to ensure the ability to move 10 TBs per day, for a sustained data transfer rate over 120 MB/sec. Files were checked for completion (expected file size) and then moved off of DataStar through a script. Files were moved into a Sun Sam-QFS disk using a Sanergy client. Once on the Sam-QFS disk, the files then automatically migrated to tape in a silo. Files were also moved into the IBM High Performance Storage System (HPSS) to increase the sustainable archiving rate. The files in the archives were then registered into the SRB data grid, making it possible to access the data from both the SCEC portal and through "C" library calls for generation of derived data products.

Before the output data were written out to the file system, MD5 data digests were generated in parallel at each processor, for each mesh sub-array in core memory. The MD5 data digests allow the data content to have unique fingerprints, which can be used to verify the integrity of the simulation data collections. The parallelization approach substantially decreased the time to generate the data digests for several Terabytes of data.

The Terashake application generated an average of 133,000 files per simulation, or approximately 400,000 files in total. The management of this material was facilitated through the use of the SDSC Storage Resource Broker data grid (SRB). Each simulation was organized as a separate sub-collection in the SRB data grid. The sub-collections were published through the SCEC community digital library, which is able to retrieve both the raw data and derived data products from the SRB data grid.

The files were labeled with metadata attributes that defined the time step in the simulation, the velocity component, the size of the file, the creation date, the grid spacing, and the number of cells. General properties of the simulation such as the source characterization were associated as metadata for the simulation collection. Integrity information was associated with each file (MD5 checksum) as well as existence of replicas. To mitigate risk of data loss, selected files are replicated either onto multiple storage media or onto multiple storage systems.

For long term preservation, use of an Archival Information Package [11] is desired, in which all of the associated provenance metadata is written into the storage system along with the data files. The approach that is being followed is to use the HDF version 5 technology [12] to create a separate HDF header file for each of the output files, store the HDF header files into the archives, and use the SRB data grid to track which HDF header file is associated with each output file. This requires that the metadata used to track the file locations be carefully backed up through dumps of the database that manages the SCEC digital library catalog. Alternatively, we plan to replicate the SCEC digital library catalog onto a separate database at USC, providing two synchronized catalogs that are capable of referencing the output from the simulations.

Derived fields such as velocity magnitude and cumulative peak velocity were generated during run-time, immediately after the velocity field files were created. These derived data products were initially used in visualization renderings.

We started visualizing the TeraShake results as soon as the initial files were produced, using the SDSC Scalable Visualization Toolkit. The run-time visualizations allowed the geophysicists to have a first glance of the simulation results and hence, to evaluate whether the simulations were taking the expected course.

3.2 Simulation Access

The SCEC digital library includes the digital entities (simulation output, observational data, visualizations), metadata about each digital entity, and services that can be used to access and display selected data sets. The services have been integrated through the SCEC portal into seismic-oriented interaction environments. An initial assessment of the user interfaces was done with Professor Steve Day, San Diego State University, as part of a seminar:

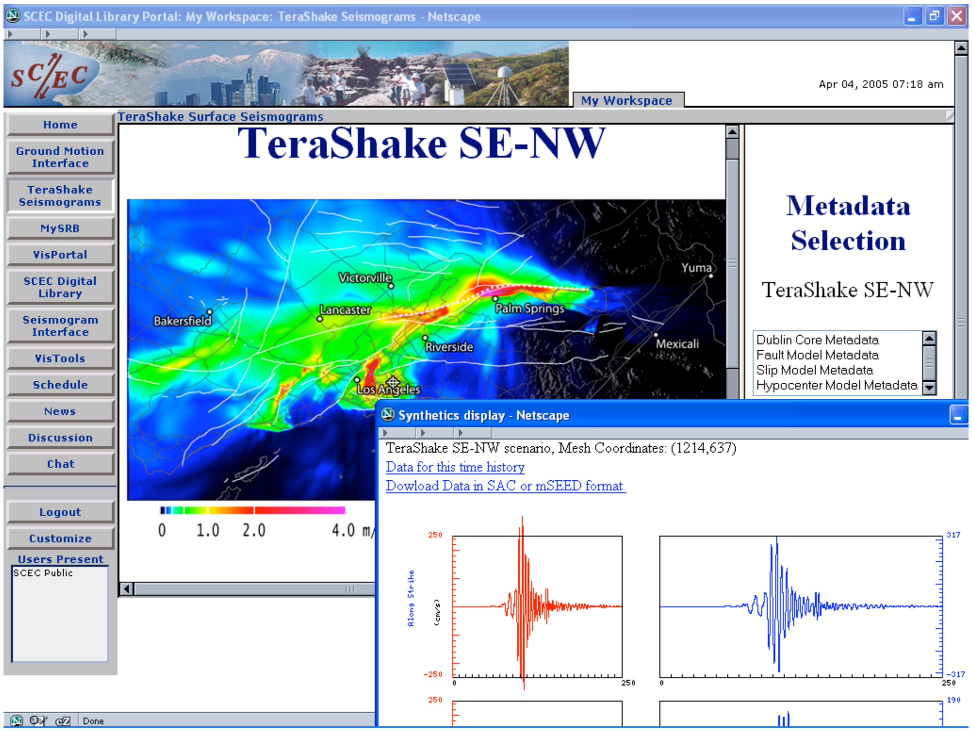
> “I have been navigating the on-line scenarios, especially TeraShake, totally absorbed, for the past half hour--it is fascinating to be able to ask questions about the behavior of these scenarios as you move around the LA basin to different sediment depths, and then answer the questions with waveforms from the ground motion interface almost instantaneously. This interface is already a tremendous tool for conveying the simulation results to the scientists doing them, and now I'm really looking forward to showing the engineers how it might help change, almost revolutionize, their way of thinking about ground motion as well.”

In addition to the services, SRB APIs are used to develop additional interfaces to the SCEC collection. Examples include the DDDS (USC) data discovery interface based on the Jargon java interface to the SRB, and the SOSA (IRIS) synthetic data access interface that is also based on the Jargon java interface.

3.2.1 Earthquake Scenario-Oriented Interfaces

The SCEC/CME collaboration and SDSC have developed digital library interfaces which allow users to interactively access data and metadata of ground motion collections, within a seismic oriented context. The Earthquake Scenario-Oriented Interfaces (available at the URLs: http://sceclib.sdsc.edu/TeraShake, http://sceclib.sdsc.edu/LAWeb) built upon the WebSim seismogram plotting package developed by seismologist Kim Olsen.

The surface seismograms are accessed through the SCEC seismogram service within the SCEC portal. A researcher can then select an earthquake simulation scenario and select a location on the surface, by pointing and clicking over an interactive cumulative peak velocity map. The portal accesses the correct file within the SCEC community digital library, and displays all three velocity components for the chosen seismogram. Users can use this web application, shown in Figure 3, to


SCEC Digital Library Portal: My Workspace: TeraShake Seismograms - Netscape

Apr 04, 2005 07:18 am

TeraShake Surface Seismograms

TeraShake SE-NW

Metadata Selection

TeraShake SE-NW

Dublin Core Metadata
Fault Model Metadata
Slip Model Metadata
Hypocenter Model Metadata

Synthetics display - Netscape

TeraShake SE-NW scenario, Mesh Coordinates: (1214,637)

Data for this time history

Dowload Data in SAC or mSEED format

Figure 3 – User interaction with the TeraShake Surface Seismograms portlet of the SCECLib Portal.

interact with the full surface resolution (3000 x 1500) data of a TeraShake scenario, amounting to 1TB per simulation. The seismograms, also shown in Figure 3, display velocities exceeding 3 m/s on a location near Los Angeles (latitude: 34.0326, longitude: -118.078), for a southeast to northwest earthquake rupture scenario.

The Los Angeles Basin interface allows users to browse in real time earthquake engineering parameter metadata associated with each receiver location, by just moving the mouse over a LA Basin map. The real time metadata attributes include geographical location coordinates, sediment depth and distance to the fault.

Other derived data products were also of interest, including the surface velocity magnitude as a function of time, the peak ground velocity as a function of space and time, displacement vector field and spectra information. These derived data products were registered into the SRB data grid for access by visualization programs and for further data analysis.

Several programs were developed to derive information from multi-terabyte velocity vector datasets produced by wave propagation simulations. The derived products have been used in data analysis, using interactive interfaces and visualizations of earthquake simulations. The data derivations include:

- Creation of seismogram collections for TeraShake simulations for on line interactive browsing and plotting,
- Computation of velocity magnitude,
- Computation of cumulative peak velocity magnitude,
- Computation of displacement vector field,
- Computation of displacement magnitude,
- Computation of data statistics used in color mapping calibrations for visualization renderings.

The derived data products have been stored and registered in the SCEC Community Library for interaction within the project and with external collaborators.

4. Future Data Management Challenges

The challenges in the future are related to management and governance of the collection. The assertion that the data are being reliably maintained requires continual checks of the integrity of the data. The amount of data stored exceeds the size that can be reliably sustained based on current commercial storage system bit-error rates. This means that procedures must be periodically executed to verify data checksums. If a checksum verification fails, then the data need to be updated from a second copy, or regenerated by re-running the simulation from a checkpoint file. Based on current storage and computation costs at the San Diego Supercomputer Center, the storage of 1 Terabyte of data on tape for a year (about $600) costs the same as the use of 500 CPU-hours of computer time. Thus the cost of storing a replica of the simulation output (surface data and checkpoint files) for six years is about the same as the cost of the original simulation. The cost of recomputation from a checkpoint is about 1/12 of the cost of the total run. This means that it is cheaper to recompute if less than 2 errors are found in the stored data per year. For more than 2 errors in the data per year, it is cheaper to store a replica.

As more sophisticated simulations are executed in the future, the SCEC digital library will need to be updated to provide the state-of-the-art results. This is the driving factor behind the creation of digital libraries of simulation output. The research of the future depends upon the ability to compare new approaches with the best approaches from the past, as represented by the digital holdings that the community has assembled.

Acknowledgements

The TeraShake research is funded by the National Science Foundation Information Technology Research program through the award EAR 0122464 (The SCEC Community Modeling Environment (SCEC/CME): An Information Infrastructure for System-Level Earthquake Research). The SDSC Strategic Application Collaboration is funded by NSF through the award SCI 0438741. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation or the U.S. government.

References

1. SCEC Project: http://www.scec.org
2. SCEC/CME: http://www.scec.org/cme
3. The SDSC Storage Resource Broker, http://www.sdsc.edu/srb/
4. Moore, R., Rajasekar, A., Wan, M.: Data Grids, Digital Libraries and Persistent Archives: An Integrated Approach to Publishing, Sharing and Archiving Data. Special Issue of the Proceedings of the IEEE on Grid Computing, Vol. 93, No.3, (2005) 578-588
5. DSpace digital library, http://www.dspace.org/
6. Fedora digital object repository middleware, http://www.fedora.info/
7. Rajasekar, A., Wan, M., Moore, R., Jagatheesan, A., Kremenek, G.: Real Experiences with Data Grids - Case studies in using the SRB. International Symposium on High-Performance Computer Architecture, Kyushu, Japan (2002)
8. Olsen, K.B.: Simulation of three-dimensional wave propagation in the Salt Lake Basin. Ph.D. Thesis, University of Utah (1994)
9. Minster, J.-B., Olsen, K. B., Moore, R. W., Day, S., Maechling, P., Jordan, T. H., Faerman, M., Cui, Y., Ely, G., Hu, Y., Shkoller, B., Marcinkovich, C., Bielak, J., Okaya, D., Archuleta, R., Wilkins-Diehr, N., Cutchin, S., Chourasia, A., Kremenek, G., Jagatheesan, A., Brieger, L., Majumdar, A., Chukkapalli, G., Xin, Q., Moore, R. L., Banister, B., Thorp, D., Kovatch, P., Diegel, L., Sherwin, T., Jordan, C., Thiebaux, M., Lopez, J.: The SCEC TeraShake Earthquake Simulation. Eos Trans. AGU, 85 (2004) 46, Fall Meet. Suppl., Abstract SF31B-05
10. Tooby, P.: TeraShake: SDSC Simulates The 'Big One. HPCWire, Vol. 13, No. 50, (2004) available online at: http://www.tgc.com/hpcwire/hpcwireWWW/04/1217/108981.html
11. OAIS – Reference Model for an Open Archival Information System, http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html
12. HDF5 – Hierarchical Data Format version 5, http://hdf.ncsa.uiuc.edu/HDF5/