EarthCube and Earthquake Science

Andrea Donnellan, Jay Parker
NASA Jet Propulsion Laboratory

Geoffrey Fox, Marlon Pierce
Pervasive Technology Institute
Indiana University

Dennis McLeod
Department of Computer Science
University of South California

John Rundle
Department of Physics
University of California, Davis

Lisa Grant Ludwig
Program in Public Health
University of California, Irvine

Charles Meertens
UNAVCO, Boulder, Colorado

Introduction

The field of Earthquake Science is a rich potential partner in the EarthCube effort to provide cyberinfrastructure for the geo-sciences. As the inadequate preparation and response to recent major earthquakes in Haiti, Chile, and Japan have shown, the field is ripe for transformation: formerly isolated groups must work more effectively with each other. Data providers need to better understand how their data are consumed and fused with other data sources by downstream geophysicists. Geophysicists must understand how to relate their work to emergency planners and responders. Experts focused on the processes of particular areas of the globe must find ways to translate their knowledge to other regions and other research teams. All must be focused on identifying and tackling grand challenges that span areas of expertise. Collaboration alone is not enough: the field needs a common framework designed to foster the desired connections.

This white paper is based on the authors' decade of experience collaborating on the QuakeSim project. QuakeSim is a multi-source, synergistic, data-intensive computing infrastructure for modeling earthquake fault models individually and as part of complex interacting systems. Numerous and growing online data sources from NASA, USGS, NSF, and other resources provide an exceptional opportunity to integrate varied data sources to support comprehensive efforts in data mining, analysis, simulation, and forecasting. The primary focus of QuakeSim is fault modeling and its application to earthquake forecasting, but the developed technology can support a wide array of science and engineering applications. QuakeSim development has resulted in a number of successes but has also identified a number of key challenges related to data, computational infrastructure, and model, analysis, and visualization infrastructure.

The key challenges that we see in the community are the need for more open processes, particularly in the creation of data products, and the greater integration and accountability of different groups on each other. The solutions are partially technical and partially sociological. EarthCube should provide forward-looking technologies that, when integrated, will enable new discoveries, but it is equally important that EarthCube provides community governance, or at least explores community governance models, that increase the reliance and accountability of earthquake scientists and cyberinfrastructure specialists on each other.

Data Products

Models are requiring an increasing number of types of data to guide them. The data are of many different forms and sizes. Fault data, for example, yield information about fault geometry, slip rates, and earthquake recurrence. At the other end of the spectrum interferometric radar data tend to be in large binary files on the order of 1 GB. Key data types used by QuakeSim models are fault data, GPS time series and velocities, seismicity data, seismic moment tensor information, and interferometric synthetic aperture radar (InSAR) images. Data products, rather than the raw data are used in each of these instances.

The largest volume data is expected to be Synthetic Aperture Radar inferograms InSAR recording changes in regions over time. Currently InSAR data comes from uninhabited aerial vehicles (UAVSAR from JPL) or satellites (WInSAR from UNAVCO). The situation could be revolutionized by the approval of the DESDynI-R Mission (Deformation Ecosystem and Dynamics of Ice– Radar) recommended in the Earth Science Decadal Survey. DESDynI would produce around a terabyte of data per day. This data is analyzed (as by QuakeSim for recent earthquakes) to find rate of changes, which are then used in simulations that can lead to better understanding of fault structures and their slip rates.

Data products often change with time as new processing techniques or new interpretations become available. One key challenge is keeping up with improved solutions as they become available. Ideally, there is a feedback loop between models that may identify issues with the data and reprocessing of data.

Another issue is that data products, even for the same data type are not standardized, and are often not available for machine interfaces. This requires manual input, or often, at best scraping of web pages for information. While this is not the right approach, it is often a necessary approach. Standardized service interfaces are needed for interfacing data with modeling and visualization tools.

Data product needs for earthquake science:

- All data products should be coupled with self-describing network service interfaces. A great deal of useful data and metadata about earthquakes, for example, is bound in human-readable web pages instead of machine readable formats (e.g., ontologies).

- Services should be documented, published, and discoverable.
- Services for analyzing lower level data products should also be designed with the same approach: they are generating products that may be consumed downstream.
- Data formats should be standardized through community use cases.

Data Providers

Data providers face several challenges as well. Earthquake science is driven by observational data and the growing array of sensors. The next generation of data management must address several problems.

- Better understanding of how data products are consumed.
- More computing power, and more flexible computing infrastructure, for handling increased data volumes and for producing custom data products.
- Data provenance and documentation: how are the data created? Is the processing of the original raw data fully explained and up to date?
- Validation: Making sure independently that the data are correct.
- Identification of newer reprocessed data products.

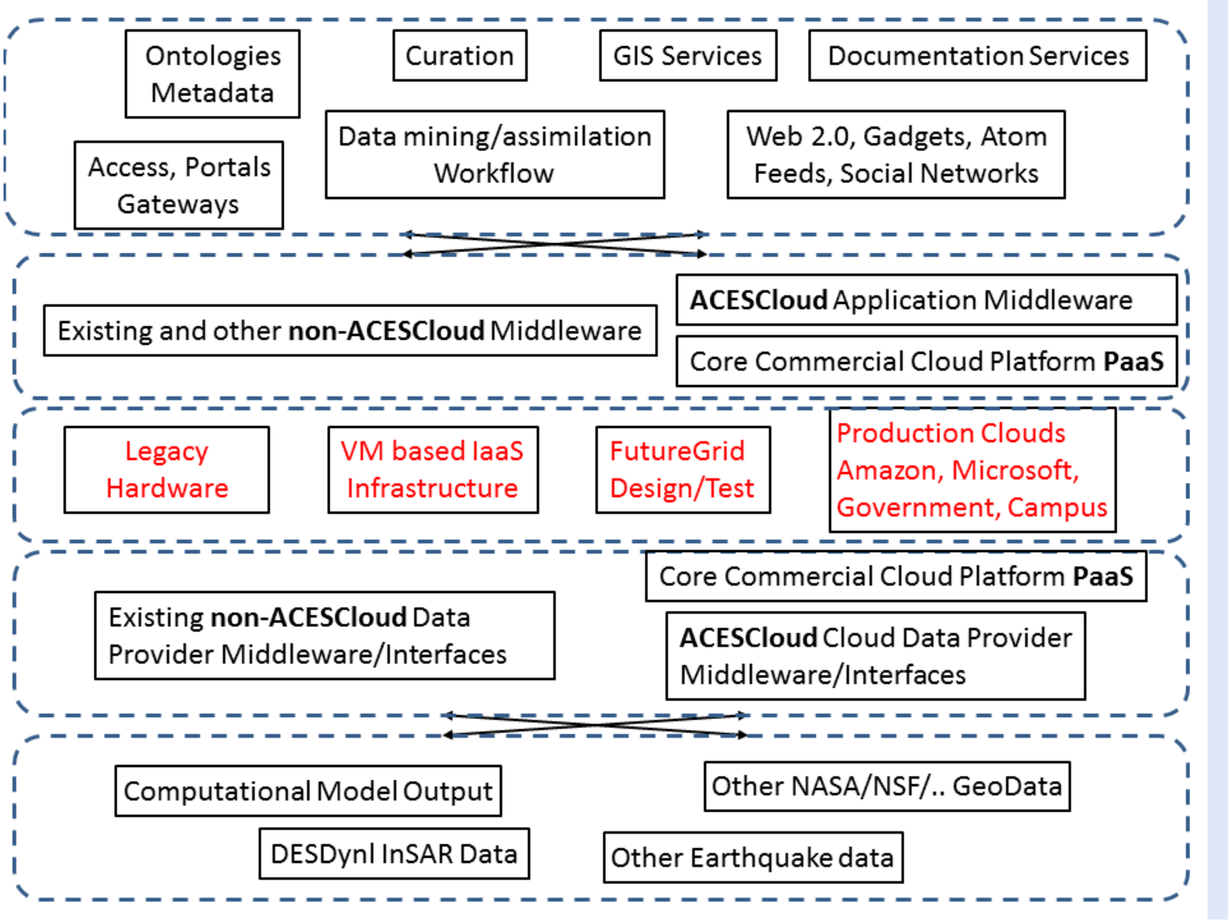
NASA has a developed a system whereby data products are classified by level, starting from raw data through higher level data labeled Level-0, Level-1, etc. as the data products become of greater utility, but at the same time depart farther and farther from the raw data as the data are processed. Data products needs to be validated and there needs to be an expectation that data may be reprocessed as new processing techniques become available for science analysis identifies an issue with the data product.

Modelers can also provide simulated data, which at times should be archived and made available to the broader community. An example of this is the Primary Reference Earth Model (PREM). A methodology needs to be developed to identify the level of confidence of the simulated data, and a decision process is necessary for deciding between intermediate models and final validated models. The models and simulation products must be carefully documented and described. These, as with more standard data products should be available through web services.

Computational Infrastructure and Cloud Computing

Increasing data volumes, complexity of data processing algorithms, and more comprehensive models continually drive a need for more and more compute power. Additionally, as large data sets are accessed it is necessary to either keep the data close to the models or have extremely broad band connections between the data sources and computers carrying out the models. There are many research issues here as there is no real consensus on good architectures for data intensive applications.

Large jobs are currently run on supercomputers, which reside on high performance computing HPC facilities. We expect large parallel simulations to continue to run on such systems which are of course moving inexorably to exascale performance. However most data analysis applications are well suited to cloud environments and can exploit the elasticity and costperformance advantages of clouds. The real-time needs of major events can require especially elastic computing on demand. We have also found that MapReduce and its iterative extensions (http://www.iterativemapreduce.org) are well suited for data analysis and offer a platform interoperable between clouds and traditional HPC platforms. We have designed a cloud platform for earth science with a layered architecture given below and aimed at supporting the international collaboration ACES or APEC Cooperation for Earthquake Simulation.


Ontologies Metadata
Curation
GIS Services
Documentation Services
Access, Portals Gateways
Data mining/assimilation Workflow
Web 2.0, Gadgets, Atom Feeds, Social Networks
ACESCloud Application Middleware
Existing and other non-ACESCloud Middleware
Core Commercial Cloud Platform PaaS
Legacy Hardware
VM based IaaS Infrastructure
FutureGrid Design/Test
Production Clouds Amazon, Microsoft, Government, Campus
Core Commercial Cloud Platform PaaS
Existing non-ACESCloud Data Provider Middleware/Interfaces
ACESCloud Cloud Data Provider Middleware/Interfaces
Computational Model Output
Other NASA/NSF/.. GeoData
DESDynI InSAR Data
Other Earthquake data

The goal is to capture both data and data processing pipelines using sustainable hardware and support both new environments and legacy systems using virtual machines. Data is accessible from resources via Cloud-style interfaces using Amazon S3, MS Azure REST interfaces as the core as these APIs are the best chance for sustainable access. We also need to follow developments such as the Swift OpenStack Object store which is very promising but it is unclear whether this, Lustre style wide area file systems or HDFS Hadoop style data parallel file systems will end up as most effective. Higher level Geographical Information Systems GIS, search, metadata, ontology services are built on these base storage services. Virtual clusters will then implement data processing pipelines/workflows/dataflows typically implemented by iterative MapReduce.

Visualization tools are increasingly necessary for understanding data and models. Users are frequently hampered by tools being licensed products or not run in the same environment in which the data or models are stored or processed. This results in the need to move large volumes of information and often requires an additional reprocessing or reformatting step before visualization can take place. Open source tools are not yet mature enough. They are sometimes not maintained or are incomplete. An investment in open source visualization tools will result in much greater scientific efficiency. Both data management and simulation tools would benefit from a redesign of the underlying computing infrastructure.

Portals, Web Services, Workflows, and Visualization

Service-oriented architectures supported by cloud-based infrastructure as a service are well established and should be adopted by EarthCube. Data discovery and delivery services can be combined with analysis and simulation services in community workflows, and rich user interfaces (Web or otherwise) can be built on top of all these. A principal challenge to this approach is sustainability. The entropy of large systems will inevitably cause failures in an open system.

- Network services must reside on stable hosting services.
- Workflows must be preserved and recoverable.
- Results obtained by Web-based computing must be reproducible.

The growing size of data sets in earthquake science and elsewhere also presents another key architectural challenge: data must be kept close to the computation (or vice-versa). This is obviously true for computation, but visualization (including interactive visualization) also faces this problem.

OpenQuake Infomall

We have re-examined aspects of our portal QuakeSim http://www.quakesim.org/ and its services in light of both recent technology developments and experience from its use during recent serious earthquakes. Highlights of particularly relevant technology developments include rise in importance of lightweight clients (smartphones, tablets), rich Web 2.0 collaborative sites and Cloud backends supporting inter alia Software as a Service. Further open source (community) approaches to software development have blossomed. Recent disasters have emphasized the requirement for essentially real-time response. This was presumably always a requirement but only recently has the internet interfaced cloud resources made it clearly possible. Another important feature of Earthquake response is that the science use of portal should leverage and help related disaster response resources which continue to expand in number and functionality. This suggests that commercial and open source help motivated by societal reasons should be readily available. Looking at portal support for crises, there are many features that are common both to different modalities of disasters and indeed to military command and control systems. Such basic features include collaborative tools including sharing of real time data from sensors and web-cams with information supported on web-based GIS (Geographic Information Systems).

Further the major Internet companies especially Google and Microsoft have collected a remarkable amount of information relevant for crises. This varies from collections of scholarly papers (say on earthquake prediction) to the lists of all geo-located entities in the world. There are nice examples of Web 2.0 resources dedicated to the recent tragic earthquake in Japan.

1. http://japan.person-finder.appspot.com/?lang=en and
2. http://www.google.com/crisisresponse/japanquake2011.html.

We have re-examined QuakeSim in light of the above ideas and identified the features of the next generation portal – we call this the OpenQuake Infomall to emphasize the role of community contributions. The OpenQuake Infomall includes an electronic exchange of data and tools of relevance to Earthquake response and science. It is open so it motivates people to contribute new tools in an interoperable fashion. OpenQuake will support and establish the needed interface standards to promote this. The OpenQuake Infomall will collaborate with the Internet giants so that their data and base tools in GIS/collaboration areas are exploited. It will examine for gaps in the commodity offerings and put its efforts into filling these. Areas of clear importance include specialized data not compiled commercially (such as fault data and some sensors); simulation and data mining tools and further customizations of base tools to the needs of OpenQuake. One example of latter is the scholarly literature searches customized to particular earthquake regions and features. OpenQuake will offer a convenient web-based workflow engine allowing quick analyses on demand in the cloud with a powerful visualization front end. Other tools include data mining/analysis and portals to simulations such as those predicting aftershocks. Both data archives and links to real time data should be present.