Complexity Computational Environment: Data Assimilation SERVOGrid

Andrea Donnellan
Jay Parker
Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena, CA 91109

Geoffrey Fox
Marlon Pierce
Indiana University
501 N. Morton, Suite 224
Bloomington, IN 47404

John Rundle
University of California
Davis, CA 95616

Dennis McLeod
University of Southern California
Los Angeles, CA 90089

Abstract-*We are using Web (Grid) service technology to demonstrate the assimilation of multiple distributed data sources (a typical data grid problem) into a major parallel high-performance computing earthquake forecasting code. Such a linkage of Geoinformatics with Geocomplexity demonstrates the value of the Solid Earth Research Virtual Observatory (SERVO) Grid concept, and advance Grid technology by building the first real-time large-scale data assimilation grid. Here we develop the next steps for both the SERVO concept and the identified need for a Solid Earth problem-solving environment. We use a challenging motivating problem of importance to NASA namely integrating NASA space geodetic observations with numerical simulations of a changing earth.*

I. Introduction

We are developing a Complexity Computational Environment (CCE) to enable the study of earthquakes. We are developing the environment to manage and integrate data and simulations and also provide data understanding and mining tools that integrate XML metadata and large-scale federated database repositories.

Earthquakes are one of the most important contributors to time- and space-dependent changes in the earth's surface observed by NASA space geodetic satellites and systems. Observation of phenomena associated with these sudden and extreme events, together with analysis via modeling and numerical simulations, are critical if we are to find answers to two fundamental questions as posed by NASA's Solid Earth Science Working Group: 1) *What are the motions of the Earth and the Earth's interior, and what information can be inferred about the Earth's internal processes?* 2) *How is the Earth's surface being transformed, and how can such information be used to predict future changes?*

This project will result in the necessary applied research and infrastructure development to carry out efficient performance of complex models on high-end computers using distributed heterogeneous data. The system will enable an ease of data discovery, access, and usage from the scientific user point of view, as well as provide capabilities to carry out efficient data mining. In this project, we focus on the development and use of data assimilation techniques to support the evolution of numerical simulations of earthquake fault systems, together with NASA space geodetic and other datasets. Our eventual goal is to develop the capability to forecast the earthquakes in fault systems such as those in California.

II. Integration of Data and Models

The last five years has seen unprecedented growth in the amount and quality of space geodetic data collected to characterize geodynamical crustal deformation in earthquake prone areas such as California and Japan. The Southern California Integrated Geodetic Network (SCIGN), the growing EarthScope Plate Boundary Observatory (PBO) network, and data from Interferometric Synthetic Aperature Radar satellites are examples. The generality and importance of Grid applications exhibiting

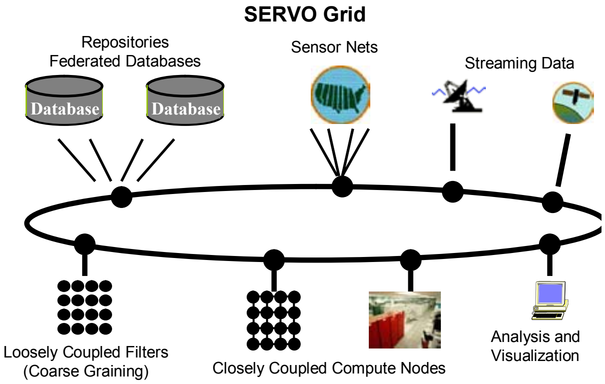

Fig. 1: SERVO Grid with typical Grid Nodes

this "data deluge" has been stressed by Hey and Trefethen (http://www.grid2002.org) [1].

Many of the techniques applied here grow out of the modern science of *dynamic data-driven complex nonlinear systems*. The natural systems we encounter in life are complex in their attributes and behavior, nonlinear in fundamental ways, and exhibit properties over a wide diversity of spatial and temporal scales. The most destructive and largest of the events produced by these systems are typically called *extreme events*, and are the most in need of forecasting and mitigation. The systems that produce these extreme events are *dynamical systems*, because their configurations evolve as forces change in time from one definable state of the system in its *state space* to another. Since these events emerge as a result of the rules governing the temporal evolution of the system, they constitute *emergent* phenomena produced by the dynamics. Moreover, extreme events such as large earthquakes are examples of *coherent space-time structures*, because they cover a definite spatial volume over a limited time span, and are characterized by physical properties that are similar or coherent over space and time.

We project major advances in the understanding of complex systems from the expected increase in data. The work here will result in the merging of parallel complex system simulations with federated database and datagrid technologies to manage heterogeneous distributed data streams and repositories (Figure 1). By the year 2006, we plan to have developed software packages that can ingest broad classes of data into dynamical models that have predictive capability based on new pattern analysis algorithms. The resultant Complexity Computational Environment (CCE) advances both the needed SERVO Grid and Solid Earth Science problem-solving Environment.

Integration of multi-disciplinary models is a critical goal for both physical and computer science in all approaches to complexity, which one typically models as a heterogeneous hierarchical structure. As one moves up the hierarchy, new abstractions are introduced and a process that we term coarse graining is defined for deriving the parameters at the higher scale from those at the lower. Multi-scale models are derived by various methods that mix theory, experiment and phenomenology and are illustrated by multigrid, fast multipole and pattern dynamics methods successfully applied in many fields including Earth Science. We believe that it is essential to explicitly recognize and support coarse-graining in the CCE. It is not only a critical scientific step but also allows one to classify parts of the CCE that really require high-end computing resources from those (like the averaging of fine grain data and simulations) that can be performed on more cost effective loosely coupled Grid facilities.

Such a multiscale integration project for earth science requires the linkage of: data grids and high performance computing (Figure 1). Data grids must manage data sets that are either too large to be stored in a single location or else are geographically distributed by their nature (such as data generated by distributed sensors). The computational requirements of data grids are often loosely coupled and so are embarrassingly parallel. Large-scale simulations require closely coupled systems whether clusters or traditional massively parallel computers. We will support both styles of computing and design the environment to make good use of the different resources. The modeler will be allowed to specify the linkage of descriptions across scales as well the criterion to be used to decide at which level to represent the system. The CCE will support a multitude of distributed data sources, ranging over federated database, sensor, satellite data and simulation data, all of which may be stored in various locations with various technologies in various formats. Web Service software component models will be used systematically and conform to the emerging Open Grid Services Architecture.

III.CCE Architecture and Infrastructure

Our architecture is built on modern Grid and Web Service technology whose broad academic and commercial support should lead to sustainable solutions that can track the inevitable technology change. The architecture of the CCE consists of distributed, federated data systems, data filtering and coarse graining applications, and high performance applications that require coupling. All pieces (the data, the computing resources, and so on) are specified with URIs and described by XML metadata. The associated infrastructure for federation and information management is described elsewhere.

A. Web Services

We use Web services to describe the interfaces and communication protocols needed to build our CCE. Web services, generally defined, are the constituent parts of an XML-based distributed service system. Standard XML schemas are used to define implementation independent representations of the service's invocation interface

(WSDL): the messages (SOAP) exchanges between two applications. Interfaces to services may be discovered through XML-based repositories. Numerous other services may supplement these basic capabilities, including message level security and dynamic invocation frameworks that simplify client deployment. Implementations of clients and services can in principle be implemented in any programming language (such as Java, C++, or Python), with interoperability obtained through XML's neutrality.

One of the basic attributes of Web services is their loose integration. One does not have to use SOAP, for example, as the remote method invocation procedure. There are obviously times when this is desirable. For example, a number of protocols are available for file transfer, focusing on some aspect such as reliability or performance. These services may be described in WSDL, with WSDL ports binding to appropriate protocol implementations, or perhaps several such implementations. In such cases, negotiation must take place between client and service. We are currently investigating the use of SOAP as an intermediate negotiation protocol in such quality of service scenarios.

Our approach to Web services divides them into two major categories: core and application. Core services include general tasks such as file transfer and job submission. Application services consist of metadata and core services needed to create instances of scientific application codes. Application services may be bound to particular host computers and core services needed to accomplish a particular task.

Two very important investigations are currently underway under the auspices of the Global Grid Forum. The first is the merging of computing grid technologies and Web services (i.e. grid Web services). The current focus here is on describing transitory (dynamic, or stateful) services. The second is the survey of requirements and tools that will be needed to orchestrate multiple independent (grid) Web services into aggregate services. These will have direct implications on our application and core services, and we are following these developments. We leverage such technologies where appropriate.

B. XML-Based Metadata Services

In general, we view the CCE as a distributed object environment. All constituent parts (data, computing resources, services, applications, etc.) are named with universal resource identifiers and described with XML metadata. The challenges faced in assembling such a system include a) resolution of URIs into real locations and service points; b) simple creation and posting of XML metadata nuggets in various schema formats; c) browsing and searching XML metadata units.

XML descriptions (schemas) can be developed to describe everything: computing service interfaces, sensor data, application input decks, CCE user profiles, and so on. Because all metadata are described by some appropriate schema, which in turn derive from the XML schema specification, it is possible to build tools that dynamically create custom interfaces for creating and manipulating individual XML metadata pieces. We have taken initial steps in this direction with the development of a "Schema Wizard" tool.

After metadata instances are created, they must be stored persistently in distributed, federated databases. On top of the federated storage and retrieval systems, we must build organizational systems for the data. This requires the development of URI systems for hierarchically organizing metadata pieces, together with software for resolving these URIs and creating internal representations of the retrieved data. It is also possible to define multiple URIs for a single resource, with URI links pointing to the "real" URI name. This allows metadata instance to be grouped into numerous hierarchical naming schemes.

Finally, on top of the URI resolving system we plan to build access, discovery, and information conversion tools. URI names naturally match to XPath search queries. We are exploring browsing implementations that make use of the hierarchical nature of URIs coupled with cataloguing languages such as RSS make this possible.

C. Federated Database Systems and Associated Tools

Our goal is to provide interfaces through which users transparently access a heterogeneous collection of independently operated and geographically dispersed databases, as if they formed a large virtual database [2,3]. There are five main challenges associated with developing a meta-query facility for earthquake science databases: (1) Define a basic collection of concepts and inter-relationships to describe and classify information units exported by participating information providers (a *"geophysics meta-ontology"*), in order to provide for a linkage mechanism among the collection of databases. (2) Develop a "meta-query mediator" engine to allow users to formulate complex meta-queries. (3) Develop methods to translate meta-queries into simpler derived queries addressed to the component databases. (4) Develop methods to collect and integrate the results of derived queries, to present the user with a coherent reply that addresses the initial meta-query. (5) Develop generic software engineering methodologies to allow for easy and dynamic extension, modification, and enhancement of the system.

We use the developing Grid Forum standard data repository interfaces to build data understanding and data mining tools that integrate the XML and federated database subsystems. Data understanding tools enable the discovery of information based upon descriptions, and the conversion of heterogeneous structures and formats into CCE compatible form. The data mining applied here focuses on insights into patterns across levels of data abstraction, and perhaps even to mining or discovering new pattern sequences and corresponding issues and concepts.

IV. Data Assimilation and Mining Infrastructure

Solid earth science models must define an evolving, high-dimensional nonlinear dynamical system and the problems are large enough that they must be executed on high performance computers. Data from multiple sources must be ingested into the models to provide constraints and methods must be developed to increase throughput and efficiency in order for the constrained models to be run on high-end computers. We have worked on a variety of nonlinear threshold dynamic fault system models, including the slider block model, the Traveling Density Wave model, and realistic fault system-level models. Coarse-grained field data that will be obtained by NASA observations include GPS and InSAR. Local seismicity rate is also significant, and available from public archives and real-time sources.

A. Data Assimilation

Data assimilation is the process by which observational data is incorporated into models to set these parameters, and to "steer" or "tune" them in real time as new data becomes available. The result of the data assimilation process is a model that is maximally consistent with the observed data and is useful in ensemble forecasting. Data assimilation methods must be used in conjunction with the dynamical models as a means of developing an ensemble forecast capability.

In order to automate the modeling process we are to developing approaches to data assimilation based initially on the most general, and currently the most used, method of data assimilation, the use of *T*angent linear and *A*djoint *M*odel *C*ompilers (TAMC). Such models are based on the idea that the system state follows an evolutionary path through state space as time progresses, and that observations can be used to periodically adjust model parameters, so that the model path is as close as possible to the path represented by the observed system.

Our research in the areas of ensemble forecasting and data assimilation is in three areas: 1) Static assimilation with event search for optimal history using cost function; 2) assessment of applicability of existing TAMC methods; and 3) use of the Genetic Algorithm (GA) approach. All three of these tasks require that a cost function be defined quantifying the fit of the simulation data to the historic data through time. These data types include earthquake occurrence time, location, and moment release, as well as surface deformation data obtained from InSAR, GPS, and other methods. The data will ultimately be assimilated into viscoelastic finite element and fault interaction models and into a Potts model in which data must be assimilated to determine the universal parameters that defined the model. Once these parameters are determined, the Potts model represents the general form of a predictive equation of system evolution.

B. Datamining: Pattern Analysis as a General Course Graining

In many solid earth systems, one begins by partitioning (tesselating) the region of interest with a regular lattice of boxes or tiles. Physical simulations demonstrate that the activity in the various boxes is *correlated* over a length scale ξ, which represents the appropriate length scale for the fluctuations in activity. The *analysis* (as opposed to the simple *identification*) of space-time patterns of earthquakes and other earth science phenomena is a relatively new field. Several techniques of pattern analysis have been recently developed and are being applied: *Eigenpattern Method ("Karhunen-Loeve" or "Principal Components")* uses catalogs of seismic activity to define matrix operators on a coarse-grained lattice of sites within a spatial region. *Hidden Markov Models (HMM)* provide a framework in which given the observations and a few selected model parameters it is possible to objectively calculate a model for the system that generated the data, as well as to interpret the observed measurements in terms of that model. *Wavelet-Based Methods* emphasizes the use of wavelets to 1) search for scale dependent patterns in coarse-grained space-time data, and 2) analyze the scale dependence of the governing dynamical equations, with a view towards developing a reduced set of equations that are more computationally tractable yet retain the basic important physics. Automated *Ensemble Forecasting* will be carried out using the data assimilation and pattern recognition.

While pattern dynamics offers system characterization modeling on an empirical level, and fast multipoles and multigrid methods automatically calculate useful levels of coarse graining, we do not neglect physics-based techniques that enable abstractions at various scales through direct modeling. The advantages of such methods are numerous: clues to direct physical relationships, insights into potential coarse-graining and coupling, and meaningful interpolation and visualization. For fault systems, boundary elements, finite elements and hybrid coupled systems are the most promising. Implementation of these technologies is required to enable realistic simulations using space-based and other data.

Acknowlegment

We thank the graduate students, Anne Chen, Sang-Soo Song, and Shan Gao from USC for their work on the database systems. We thank Robert Shcherbakov from UC

Davis for his work assessing the data assimilation methods. Portions of this work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA.

REFERENCES

[1] *Grid Computing: Making the Global Infrastructure a Reality* edited by Fran Berman, Geoffrey Fox and Tony Hey, John Wiley & Sons, Chichester, England, ISBN 0-470-85319-0, March 2003.

[2] *Federated Database-Systems For Managing Distributed, Heterogeneous, And Autonomous Databases,* Sheth AP, Larson JA, Computing Surveys, 22 (3): 183-236, Sep 1990.

[3] *Federating Neuroscience Databases*, Computing the Brain: A Guide to Neuroinformatics (editors Arbib, M., and Grethe, J.), Academic Press, 2001.

[4] *Interoperation, Mediation, and Ontologies* Wiederhold, G., Proceedings of the International Symposium on Fifth Generation Computer Systems, Tokyo, Japan, pages 33-84, December 1994.

[5] *Integrating and Accessing Heterogeneous Information Sources in TSIMMIS*, Garcia-Molina, H., et. al., Proceedings of AAAL.