1. Motivation and Project Overview

The importance of simulating earthquakes is intuitively obvious. For instance, the recent January 16, 1995 Kobe, Japan earthquake was only a magnitude 6.9 event and yet produced an estimated $200 billion loss. Despite an active earthquake prediction program in Japan, this event was a complete surprise. Drastic scenarios similar to those reported this year in Turkey and Taiwan are possible and indeed eventually likely in Los Angeles, San Francisco, Seattle, and other urban centers around the Pacific plate boundary. Over the last three years, we have built a team of Computer and Earthquake scientists from academia and government to initiate a program GEM "General Earthquake Models" [17], aimed at applying the latest computational technology in this area. This thrust contributes to the nationally identified importance of developing new approaches to Geoscience involving advanced instrumentation (EarthScope) and computing. Earthscope is an NSF/EAR/MRE initiative to develop extensive new networks of sensors in the western United States to monitor all aspects of earthquake phenomenology. As yet, there is no corresponding, comprehensive modeling and simulation initiative, a void, which the GEM collaboration proposes to fill. The GEM group includes representatives of several universities (11 are involved in this proposal), multiple government agencies and laboratories (DoE, NASA, NSF, USGS) and is coordinated with the major NSF Southern California Earthquake Center (SCEC) in this area whose outreach services we will use. There is substantial international interest in these problems and GEM works closely with an effort ACES (APEC Co-operation for earthquake Simulation [3]) among several Asia-Pacific nations including Australia, Taiwan, Japan, China, and the USA. This includes Japan's ambitious *Earth Simulator* project involving a 30 teraflop parallel computer and a correspondingly major software and science research effort [23]. Current funding for GEM is at the level of $100K/year from the NSF/SCEC and NSF/EAR, which has allowed important seed projects on which we build.

Earthquake science spans many scales in space and time and needs simulation techniques from partial differential equation, particle dynamics and statistical physics using concepts such as "correlation length" and "critical state" in understanding regional seismicity. GEM can impact all of these from real-time analysis of scientific data from an earthquake; the systematic longer term integration of data from multiple sensors into simulations and the fundamental study of earthquakes as an emergent phenomena in a complex system. This field is an attractive target for computer science due to the intrinsic richness of the applications and the societal importance and also because the use of computers is not yet too extensive and so modern approaches and infrastructure can be used without major distraction from existing legacy approaches. The field is naturally distributed with sensors, scientists and earthquakes scattered around the globe. Thus there is an immediate application of emerging concepts such as computational grids to link large-scale simulations, data and people in a distributed fashion. The computer science research focuses on the issues on building an integrated information infrastructure that can support collaborative distributed scientific research over a range of time scales and computational needs. The work in tools and distributed systems will be driven by three application area timeframes characterized by time scales of hours (post earthquake analysis), 6-12 months (data assimilation and development of new earthquake forecasting approaches) and ten years (fundamental theory). The work will contribute to earth science research in these three timeframes and to computational science where we have defined five thrust areas; distributed collaborative (shared) scientific objects, HPCC simulations including new uses of fast multipole techniques, multi-sensor metadata, data and simulation visualization, and interactive scientific datamining for earthquake pattern analysis. We give a more detailed discussion of the three application timeframes and the five computational science thrust areas after a brief discussion of some earth science issues. The preproposal ends with outreach and management sections.

2. Understanding Earthquakes

There are a variety of valid approaches to trying to understand earthquakes through modeling and data interpretation and GEM intends to be involved in a wide range of them, since it is not clear that any one will provide the best approach for any or all purposes. Perhaps all workers feel that at some level earthquakes might be regarded as either a stochastic nonlinear system, or an example of deterministic chaos, but there is a wide range of opinion concerning whether it is better to focus on the chaotic aspect, the stochastic behavior, or the deterministic properties. One view might be that it is impossible to ever know all of the relevant variables affecting their size, timing, and character of an earthquake, and that this means that we might as well give up trying to understand the physics at any detailed level. Another view, to which we tend to subscribe, is that earthquakes fall broadly into one of several universality classes, whose behavior is governed by one of a small number of fixed points. If this is the case, it will be possible to obtain fundamental understanding of the broad behavior of the system even if the details remain obscure. One can then focus on looking at patterns in earthquake occurrence in both real earthquakes and in earthquake simulations as an optimal way to gain understanding.

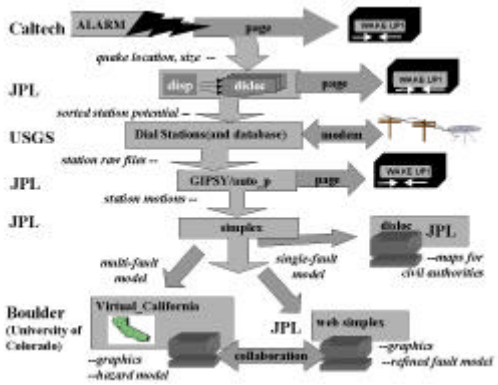
3 GEM Application Timeframes

3.1 Project Methodology

We have divided the earthquake science activity into three teams corresponding to modus operandi (timeframes) with (superficially) very different requirements for the supporting computational infrastructure. Each timeframe captures an aspect of earthquake science that contributes to a core understanding of the field. We will support each timeframe with the same integrated information system and test and evaluate this with special attention to the successful enabling of new and more effective models of collaborative scientific research. We will of course feed lessons back into both computer science and earthquake science and compare with related activities on an ongoing basis. The 3 timeframes are described in more detail in following sections.

3.2 Timeframe 1(minutes/hours): Real Time Science and Data Analysis after an Earthquake

Here we collect together the types of activity exemplified in the figure, which shows how decisions are made in real time as to what data should be gathered and fed into simulations that can aid both the forecasting of possible aftershocks and suggest which further data will be useful. The flow of actions moves down the figure starting with initial notification of an earthquake and followed by decisions as to which data to gather. At the bottom, we have iterative integration of simulations with the earthquake data. We have started to build a Problem Solving Environment enabling this type of interaction with synchronous interaction between the world wide distributed scientists, simulations and data visualizations. However, this effort is in its infancy and the support of this ITR proposal could bring it into a useful reality. Note we are restricting our attention to interactions between scientists but many of the tools and concepts can be applied to support the work of crisis management teams.



3.3 Timeframe 2 (month/years): HPCC Simulations and Data Integration

Achieving the goals of GEM and for instance meeting the challenge of Sec. 3.4 for theoretical understanding, will require the ability to simulate models on time and size scales presently unattainable, and to compare data sets from different models with each other and with observations of real faults. To accomplish these goals we will develop and refine both efficient algorithms such as fast multipole methods and explore acceleration techniques. For initial efforts we propose to develop a standard set of real data for calibration of models or comparison with results of simulations. By using the same real data set for all models and simulations we facilitate comparison of model effectiveness and establish real (rather than theoretical) performance standards for simulation results.

Note that earthquake science is benefiting from the rapid increase in quality and type of data and this is driving the urgent need for new GEM information infrastructure. GPS, InSAR and broadband seismic (TERRASCOPE) data, together with archived (in particular by the SCEC Data Center) and newly developed paleoseismic data can be used in conjunction with the simulation capabilities to establish the relevant model parameters. These parameters include, for example, the current geometry of faults; slip rates at any known point; recurrence intervals and historic variations in slip during earthquakes—leading to estimates of frictional parameters; deformation data leading to estimates of elastic plate thickness and sub-crustal stress; and so forth. Typically, different investigators are expert in the different types of data and a much improved collaborative environment is needed to integrate the different data together. The data is of course critical for earthquake understanding, especially now with the existence of only rudimentary models.

Several of the computer science activities directly support this scenario. We are designing systematic (XML based) metadata for the diverse data types; integrating parallel fast multipole methods into several simulations and designing visualization methods that will support effective viewing of the different data sets.

3.4 Timeframe 3 (years/decade): Fundamental Theory and Complex Systems

The primary goal of the theoretical research is to integrate current earthquake modeling approaches into a more general, comprehensive model, and to provide a theoretical framework through which the data generated by simulating this model can be understood. We expect that this model will cover time scales from seconds, the time associated with a rupture, to centuries, the scale of strain accumulation and release. In addition, this model will contain most of the aspects of real faults; characteristics of wave propagation, frictional behavior and fault interaction will be included. We hope to generate an understanding of earthquakes as a collective effect and relate the structure of quakes and their precursors to related complex systems. In order to develop such a model we need to understand the essential features of fault systems and how omission of selected features affects the physics obtained

from the model. This requires an investigation of a wide range of models to ascertain which aspects of the physics are robust and which aspects rely on model detail. Current models range from cellular automaton versions of single faults, to slider block and elastodynamic models with various friction forces, to stochastic models of fault systems. Each model provides insight into different, but overlapping aspects of fault dynamics.

Implementing this investigation will require the ability to simulate these models on time and size scales presently unattainable, and to compare data sets from different models with each other and with observations of real faults. This requires largely asynchronous collaboration with a sophisticated PSE that can support the rich range of model and data. The PSE will be used for both validation and assimilation. An area of growing importance is scientific datamining from both physical data and simulations to decide what are patterns that could signal a quake and what simplified coarse grain dynamics could describe these patterns [41, 35, 36]. Support of this will be an important thrust of our proposed GEM Information infrastructure.

4 Computer and Computational Science

4.1 Basic Distributed Object Information Technology Architecture

Problem Solving Environments (PSE) have been pursued for many years with the work at Purdue [21] pioneering many important concepts. The increasing power of computers and the increasing capability of distributed object and web technologies are making this approach increasingly attractive for users and system builders. One uses the "Object Web" (CORBA, COM, Java, XML etc.) and a browser based user interface to provide a single integrated view or portal [9] to the resources and tools needed by the scientist. In this proposal we focus on the issues needed to design a single information infrastructure to support multiple timeframes i.e. the development of multiple PSE's built from the same resources. We start with the successful Gateway and WebFlow systems [1,14,15] developed at Syracuse and applied to several DoD and NSF projects. These have a classic three-tier architecture with client, brokers/servers and services in the three layers. High performance is obtained even while using Java and CORBA in the middle tier, by careful separation of control and data. The middle tier provides a flexible control layer implemented with proxies and traditional high performance mechanisms such as Globus [18] and MPI are used for data transfer in the backend. This WebFlow distributed object technology has a powerful dataflow and coarse grain object computing model with all interfaces defined in XML and compatible with community activities [9,19]. In this sense it is more powerful than the earlier NILE system [26] while it is less ambitious than the Common Component Architecture [8], Legion [24] and POOMA [28], which provide a fine grain, object model. Gateway is fully consistent with commodity standards (CORBA and XML) and therefore suitable for the ambitious project proposed here which aims to provide an information infrastructure for a complete application area in three radically different timeframes.

We will implement early prototypes of the information infrastructure on Gateway but expect that one can better use one of the emerging set of "Object Web operating systems" such as E-Speak [12] or Ninja [27] as the basic framework. We will investigate these new possibilities over the next few months. The following sections 4.2 to 4.6 describe activities that provide resources and tools (services) that will be integrated into the Gateway systems and presented as the different portals to the scientists working in the three timeframes of section 3. One research result of this project will be an evaluation of different object web architectures and operating infrastructure in terms of their ability to support scientific PSE's.

4.2 Real-Time HPCC Simulations

Here we use the best known HPCC techniques with optimizations for both the real time needs of the first timeframe and the conventional large-scale analysis of the timeframe of Sec. 3.4. This domain is data-intensive and we expect to make extensive use of the resources and methodology developed by NPACI in this area [25]. Minster on our proposal is Earth Systems Science lead for NPACI. We are also developing collaborations with the Maui Supercomputer Center [28].

An early result of the seed funding for the GEM collaboration was the realization that fast multipole methods could be applied to many of the Green's function simulations [31-34,37]. As the complexity decreases from N^2 to N logN, this dramatically increases the simulation resolution with the maximum number N of elements moving from some tens of thousands to the many millions possible with multipoles on a fast parallel machine. We chose to work from a portable code developed by Salmon and Warren over the last few years [42]. Adapting this approach to our problem has the potential of both allowing realistic problems in fault mechanics to be investigated for the first time as well as presenting the opportunity for innovative modifications to be made to the fast multipole approach. For example, the Green's functions in our fault mechanics problem fall off as $1/r^3$, a faster falloff than in the situations in which the multipole approach has typically been used. Another important difference is that the multipoles are fixed in position but of variable intensity in earthquakes but have the opposite characteristics in

astrophysics. We will develop the appropriate new ansatz's, especially in situations of complex fault geometry, to determine both the order and geometry of the multipoles that are best used.

4.3 Multi-sensor Metadata

A team from JPL has already begun to design prototype XML based metadata [5,25,43] for some of the sensor-based data and we have used this in our early PSE mentioned in Sec. 3.2. As part of this project, we will extend this design to sensor, field and simulation data in a way that we can use it to integrate different data sources into our collaboration, visualization and assimilation tools. Our PSE environment [1] fully supports XML for job and data definition and its dataflow paradigm will allow files with compatible metadata to be exchanged between application components. This will require a hierarchical metadata design and the construction of services based on available XML tools to process data in the different parts of the computational environment. This activity will be linked to related work in our arena (e.g. the Grid Forum [19], NCSA and NPACI) and so be important in designing science wide approaches to metadata.

4.4 Data and Simulation Visualization

One of the aspects of the GEM community and research that is interesting from a science systems perspective (the meta-topic of "how to do science effectively") is the enormous heterogeneity of the data used by earthquake modelers. A major challenge of visualization in this proposal is to create methods of integrating the very different kinds of data and supporting their visualization with a common toolkit. Collaborative visualization will be needed and here we will base our approach on experiments Java systems developed in TangoInteractive [40] and the interesting general analyses of UNC [30] and Wood [44]

A variety of approaches will be investigated for scientific visualization of the data produced by simulations and observations. Because such data in the realm of earthquake processes involves both space and time, we plan to explore using volume rendering with time as a third axis for situations where the spatial coordinates can be adequately represented in the other two dimensions. Adoption of suitable thresholds for transparency of portions of the data set should allow the most interesting portions to be viewed and better understood. We plan to use the new NSF MRI-funded "cave" immersive virtual reality environment (The TAN Cube) at Brown University and similar three dimensional representations to view such models. This environment offers the opportunity to monitor the progress of a computation in real time by simultaneous visualization and computation. We will evaluate whether it is feasible or desirable to steer the computation and/or the visualization of it interactively during simultaneous computation and visualization in order to focus inspection and/or computing resources on the areas that are most interesting.

As discussed in sec. 3.3, an important component of our research will involve comparing simulations and observations. We will focus on ways to evaluate how well the simulations match the data, and we will investigate whether visualization can play an important role in this. For example visual comparisons of simulations and of data for synthetic cases with known degrees of goodness of fit can be rated for the quality of the agreement by teams of observers – including collaborating over a distance. These subjective evaluations can be compared with a variety of statistical measures of agreement to determine the accuracy of the visual perception. Teams of observers with different degrees of experience in the subject matter will be employed to discover how training affects the quality of the evaluation. One possible advantage of the use of visualization to compare simulations with observations is that if it proves to be as reliable as statistical tests, it may be more suitable for comparison in situations where devising appropriate statistical tests is more difficult.

4.5 Interactive Scientific Datamining

Recently a combination of simulation and observational data has been used to identify patterns that could be helpful in forecasting earthquakes [41, 35, 36]. This approach uses techniques first developed in the climate field and computationally involves matrix (eigenvector) analysis combined with visualization of geographic data related in a particular eigenpattern. This initial success highlights the role of scientific datamining as the appropriate way to generalize the classic earthquake related phenomenology to the proposed information infrastructure with many orders of magnitude more data from diverse sources and the corresponding need for a systematic approach. The datamining needs to exploit the hierarchical XML metadata structure proposed in Sec 4.3 and link to the visualization of the last section. Collaborative discussion of possible forecasting approaches (datamining methods and results) seems important. Thus we intend to build a collaboration-aware Java analysis system which can support access to data from simulations and observations and the type of computationally modest calculations found helpful so far. The computer science research will evaluate other data mining approaches and integrate them into the interactive analysis environment as either client side or backend computational resource. Research issues include architecture and integration of scientific datamining in a collaborative object based environment. We can expect interesting datamining algorithms to be needed in this relatively new field. For instance the initial work [41] found

signal to noise was greatly enhanced by assuming that the system is a pure phase dynamical system, ignoring changes in state vector normalization.

4.6 Collaboration over ranges of Distance and Time

We have found a mix of success and failure with initial collaborative systems such as Microsoft NetMeeting, NCSA's Habanero [20] and Syracuse's TangoInteractive [40]. Higher speed networking and quality of service will address some of the difficulties such as variable quality in digital audio video conferencing; here we track the ANL/NCSA Access Grid project. We have been quite successful in educational applications but have yet to develop collaborative computing applications which are both robust and of compelling value. We will use the existing collaboration systems in early experiments but we intend to build much of the collaborative infrastructure from scratch replacing custom protocols and services by those available from infrastructure like Ninja [27]. The timeframes in our proposal illustrates a critical challenge for collaboration systems – namely supporting asynchronous interactions (timeframe 3), real-time synchronous (timeframe 1) and mixtures thereof (timeframe 2). Our web-based PSE approach implies that collaboration is a service that shares web-based distributed objects. However we also need to support several collaborative modes; shared display and both collaboration-aware and collaboration-unaware shared event models. Previous systems have focussed on one of these mechanisms and have not been able to support the needed range of collaboration. Initially we will support these different modes with separate subsystems but will replace this by an integrated system CPW (Collaborative Portal on the Web) based on a generalized shared queued event service. This terminology indicates that our approach is to build first a portal with collaboration as one its services; this will be implemented using XML systematically to define the details of the collaboration and the portal infrastructure (e.g. Ninja's event service) as the building blocks of the collaborative system. We believe this will integrate collaboration directly into the scientific analysis and make it more useful than before.

5 Outreach

We are fortunate to be able to leverage the very successful SCEC [38] outreach program led by Jill Andrews whose mission is to promote earthquake loss reduction and to actively engage the public at large in activities that focus on earthquake-related education, research-based technology development and transfer, and systemic reform. Enhancing current SCEC-funded Web-based education modules now under construction by Andrews and Donnellan [39] with GEM material will complement this general goal. Because the education standards of today strongly encourage an inquiry-based, accessible approach to learning science, this SCEC work has met with enthusiastic acceptance among reviewers from the California Science Implementation Network. The first module on Investigating Earthquakes through Regional Seismicity, along with a second module on Global Positioning Systems (GPS) technology, created at an upper division high school / lower division college level, are being adapted to middle school curricula. A partnership with the GEM principal investigators will certainly enhance the material presented in the existing modules. As a first activity for this proposal we will create a mathematically-oriented Web-based module, using GEM as the illustrative example. This will acquaint high school instructors and students with the concept of an integrated approach to solving computational challenges, and to lead them through an exercise to produce their own earthquake forecast (probability) models.

6 Management and Budget

The management plan is based on our substantial experience with three large NSF center activities – SCEC (Science and Technology center in earthquake science), CRPC (Science and Technology Center for Research in Parallel Computing), and the NCSA Alliance (PACI Partnership in Advanced Computational Infrastructure). There is an overall GEM management structure similar to these activities, which will support this proposal as the major computer science activity with other projects mainly aimed directly at Earth Science. The proposal itself is divided into nine teams, whose leaders form a technical committee, covering the three application areas of Sec. 3, the five computer science thrusts of Sec. 4 and outreach described in Sec. 5. The principal investigators of the proposal will form a steering group that will review important project decisions and interface between the GEM executive board and the technical committee. GEM already meets approximately four times per year, often together with synergistic activities such as SCEC or AGU meetings. We will link the proposal technical discussions and workshops to this meeting series. We propose a total budget of approximately $840K per year for a period of three years. This budget is split roughly equally between computer/computational science work at Boston, Brown, Florida State and USC and Earth Science activities. There is $65K per year in the outreach work led by SCEC/USC. SCEC will play an important management role on the earth science side as USC already has mechanisms to subcontract without overhead and we will use this for the smaller Earth Science sites.