ITR/AP: The SCEC Community Modeling Environment—An Information Infrastructure for System-Level Earthquake Research

C.1. Problem Statement: The Need for Information Technology Research in Earthquake Science

In the last several years, notable advances have been made in two distinct areas of earthquake science: (1) the *dynamics of fault rupture*—what happens on a time scale of seconds to minutes when a single fault breaks during a given earthquake—and (2) the *dynamics of fault systems*—what happens within a network of many faults on a time scale of years to centuries to produce the sequencing of earthquakes in a given region. Combined with the increased availability of terascale computing resources, these advances in geophysics make it possible for the first time to create fully three-dimensional (3D) simulations of fault-rupture and fault-system dynamics. Such simulations are crucial to gaining a fundamental understanding of earthquake phenomena. However, even as the challenges of understanding the geophysics are being met, we are faced with a new problem. Constructing a system-level earthquake simulation from models of constituent phenomena and executing that simulation on suitable computing platforms becomes increasingly complex due to (1) the difficulty of selecting and configuring compatible simulation models that are appropriate for the geophysics problem being studied, and (2) mapping those models onto computing resources for execution. This complexity will limit the ability of non-experts to perform accurate earthquake modeling, restricting the community that can benefit from earthquake simulation and ultimately slowing down progress in earthquake science. We propose to address this problem by developing an integrated modeling framework that automates the process of selecting, configuring, and executing models of earthquake systems. We will achieve this ambitious goal via an innovative integration of knowledge representation, knowledge acquisition, Grids, and digital libraries. The proposed research will be conducted by a collaboration between leading researchers in each of these information technology areas with earthquake scientists associated with the Southern California Earthquake Center.

C.1.a. The Opportunity for Improved Seismic Hazard Analysis

Physics-based simulations can potentially provide enormous practical benefits for assessing and mitigating earthquake risks through seismic hazard analysis. Seismic hazard analysis (SHA) seeks to describe the shaking that can be expected at a given point of the Earth's surface due to earthquakes that are likely to occur over a specified time period [1]. From the perspective of deterministic physics, this calculation requires the coupling of models that represent a series of complex physical processes:

- *Unified Structural Representation* (USR): a self-consistent, 3D characterization of active faults and material properties (e.g., seismic velocities) needed to describe regional deformation and seismogenic processes.
- *Fault System Model* (FSM): an evolving representation of the regional stress and deformation fields, capable of predicting when individual fault segments will rupture.
- *Rupture Dynamics Model* (RDM): a dynamical description of the nonlinear stress/displacement interactions across a rupturing fault as a function of space and time.
- *Anelastic Wave Model* (AWM): a computation of the propagation, interference, and attenuation of the seismic waves that travel along complex paths from a fault rupture to a target site.
- *Site Response Model* (SRM): a dynamical description of ground excitation in the near-surface environment at a target site, which, for strong ground motions, often involves significant nonlinearities.

At present, practitioners are forced to make many simplifications and approximations in the application of SHA to earthquake engineering and risk mitigation. Fault system models that include stress interactions are in their early stages of development, and their predictive value has not been demonstrated. Therefore, the FSM is usually replaced by long-term forecasts of the earthquake potential on individual faults derived from historical seismicity catalogs and/or geologic and geodetic slip rates. The RDM is usually replaced by an isotropic source whose strength depends only on earthquake magnitude and thus ignores many important features of real fault ruptures, such as strong azimuthal variations in radiation. The AWM is approximated in terms of an empirical "attenuation relationship" that depends only on source distance, depth, and magnitude, which does not account for wave-interference effects [2]. The SRM is approximated by an empirical function of the rock properties local to the site, rather than a dynamical calculation based on near-surface structure. Recent analyses have attempted to quantify other effects, such as basin depth [3], but a major new study by the Southern California Earthquake Center (SCEC) has concluded that, "any model that attempts to predict ground motion with only a few parameters will have substantial intrinsic variability. Our best hope for reducing such uncertainties is via waveform modeling based on the first principles of physics." [4]

C.1.b. Project Goals and Requirements

SCEC has embarked on an ambitious program to develop physics-based models of earthquake processes and integrate these models into a new scientific framework for seismic hazard analysis and risk management [5]. The success of this program will depend on the construction of a *Community Modeling Environment*, in which the appropriate simulation models will be developed, documented, and maintained on-line for application by SCEC, earthquake researchers elsewhere around the world, and end-users of earthquake information. This environment will function as a virtual collaboratory for the purposes of knowledge quantification and synthesis, hypothesis formulation and testing, data conciliation and assimilation, and prediction. It will greatly facilitate the system-level understanding of earthquake phenomena, and it has the potential to improve substantially the utilization of SHA in reducing earthquake losses.

The goal of the proposed project is to construct an information infrastructure for the SCEC Community Modeling Environment that can satisfy the following four requirements:

R1. Capture and manipulate the knowledge that will permit a variety of users with different levels of sophistication to configure complex computational pathways for (a) rapid prototyping of new SHA algorithms and (b) construction, validation, and dissemination of new SHA products.

R2. Enable execution of physics-based simulations and data inversions that incorporate advances in fault-system dynamics, rupture dynamics, wave propagation, and non-linear site response. These simulations must be capable of resolving the stress-interaction scales, including the inner frictional scales of faulting (< 100 m), as well as the seismic frequencies of engineering interest (> 1 Hz).

R3. Manage large, distributed collections of simulation results, as well as the large sets of geologic, geodetic and seismologic data required to validate the simulations and constrain parameter values.

R4. Provide access to SHA products and methodologies to end-users outside of the SCEC community, including practicing engineers, emergency managers, decision-makers, and the general public. This access must include intelligent handling of queries, workbench environments for user-configured calculations, visualization tools, detailed information about the legacy and pedigree of SCEC products, and mechanisms for educating end-users about SHA methodology.

In the context of the SHA problem and the extended SCEC community, these requirements present an interesting set of challenges for information technology:

- Heterogeneity and multiplicity of the models. Many object types must be manipulated and a number of algorithms must be employed, but the pathway to a computational result is contingent on factors that must be evaluated at each step along the way. Moreover, algorithm inputs and outputs can be very complex and take different meanings for different users, so that this may require not just syntactic translation but also semantic mappings and conversion.
- Distributed development of the models. Models are being developed by different organizations with differing expertise and computational resources, requiring data management and execution environments that span autonomous administration domains.
- Computational requirements of simulation models. A physics-based approach to SHA requires that a large range of potential scenarios be examined, each at much higher spatio-temporal resolution than current simulations. This erects a set of *Grand Challenge* computational problems that require terascale class computational resources to solve [6].
- Requirements for model management. Evolution of the digital representation is driven by physical events, simulation output results, and remote-sensing input, requiring management of multiple versions. Furthermore, the SCEC approach to SHA uses an integrated analysis of seismic, geodetic and geologic data; thus the modeling environment must manage heterogeneous data collections, including a range of existing collections and archives.
- Diversity of user community. Products of SHA are of interest to a wide range of users, from geophysicists developing the analysis, civil and structural engineers designing buildings and structures, to city planners, news media, and disaster response teams. Thus, unsophisticated users may need to manipulate sophisticated models.

To address these challenges, we propose to develop fundamental new approaches to model generation and simulation management by bringing together four distinct computer science disciplines: (a) knowledge representation and reasoning, (b) interactive knowledge acquisition, (c) digital libraries and information management, and (d) Grid resource-sharing environments. The overall architecture of the proposed SCEC Community Modeling Environment in illustrated in **Figure 1** and described in more detail in §C.3. While we focus on the application of this new computer science to the practical problems of earthquake science, the techniques and integrated modeling framework that will be developed in this project will be directly applicable to many other disciplines.

The challenges described above become particularly acute when “all hell breaks loose” during a big earthquake. Modern seismic information systems, such as TriNet in Southern California, can, within a few minutes, locate regional earthquakes and produce preliminary maps of ground shaking to guide emergency response [7]. A long-term goal is to have the capability for automatically configuring an expanded set of modeling resources to assimilate seismic, geodetic, and geologic data as they are acquired in real time, to create new products such as ground-motion and damage predictions, and to distribute the output to multidisciplinary teams scattered across the region. These teams will need the means for jointly visualizing, manipulating, and modifying the products and for communicating the results to non-specialists—engineers, emergency managers, government officials, and the media. All of these operations will have to be done under potentially stressful conditions using distributed, multiply-connected computational systems that are robust to major regional disruptions in power, communications, and transportation. *The proposed project does not directly address the issues of real-time operations and robustness.* However, the SCEC Community Modeling Environment, as envisaged here, should greatly facilitate the development of these capabilities by USGS scientists and others responsible for real-time seismic information systems in Southern California and elsewhere.

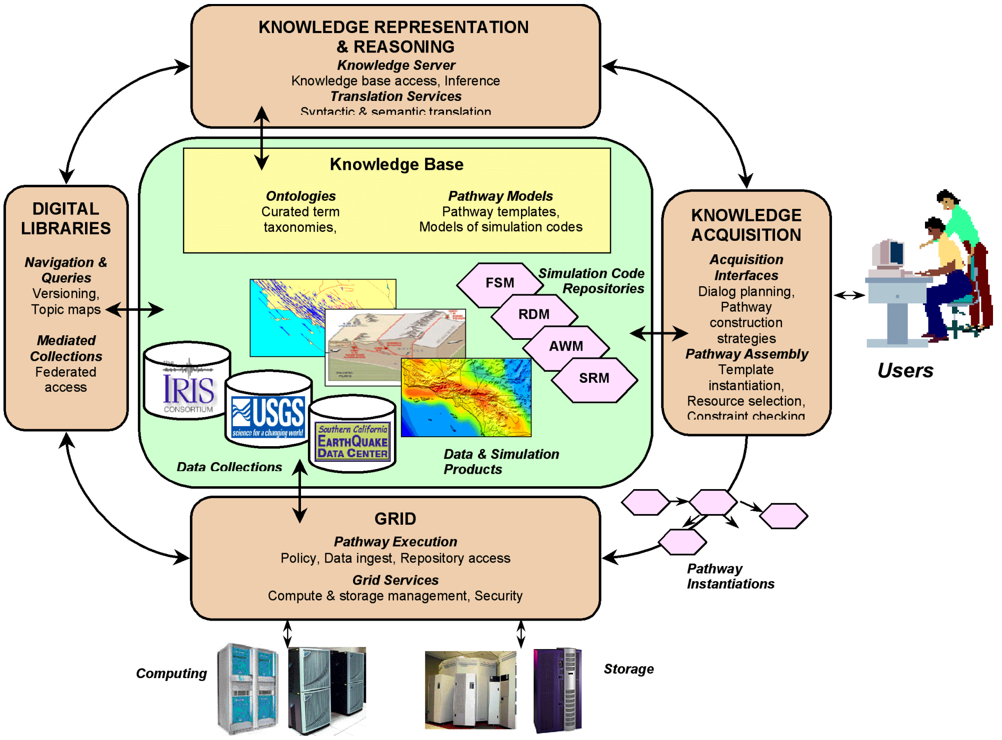

Figure 1. Proposed architecture for SCEC s Community Modeling Environment. Knowledge-rich community models will be developed and curated jointly by geophysicists and knowledge engineers to produce a comprehensive knowledge base of earthquake science. Knowledge representation and reasoning techniques will enable semantic translations and term mappings as well as complex inference. The Grid will provide access to computing and storage resources to execute complex computational pathways. The community models will be grounded in data collections and software repositories through digital libraries technology. This knowledge-rich environment will enable interactive knowledge acquisition tools to guide unsophisticated users in constructing complex computational pathways that result in sophisticated simulations and thus more accurate seismic hazard analysis.

C.1.c. The Need for a Large, Integrated Project

SCEC is a consortium of 40 universities and research organizations funded by the NSF, USGS, and other government agencies to (1) gather new information about earthquakes in Southern California, (2) integrate the available knowledge into a comprehensive and predictive understanding of earthquake phenomena, and (3) communicate this understanding to engineers, emergency managers, government officials, and the general public. Recent estimates by the Federal Emergency Management Agency (FEMA) ascribe nearly half of the national earthquake risk to Southern California, with one-quarter concentrated in Los Angeles county alone [8]. SCEC thus serves a high-risk population of more than 20 million people as its regional center for earthquake information and coordinated earthquake studies. This coordination is essential to the development of the comprehensive data sets, consensus models, and consistent scientific judgements needed for public policy in earthquake risk management and mitigation.

Southern California is a superb natural laboratory for understanding the fundamentals of earthquake processes, endowed with many active faults and diverse tectonic regimes astride the rapidly deforming Pacific-North America plate boundary. Data on earthquakes in this part of the world are outstanding, and the integration of this information into a comprehensive and predictive understanding of earthquake behavior requires the resources of a multidisciplinary consortium capable of system-level research. SCEC coordinates its activities across a scientific community that includes over 150 professional scientists, as well as many postdocs and graduate students. In addition to a national distribution of research universities, the organizations participating in SCEC include the U. S. Geological Survey (USGS) and the California Division of Mines and Geology (CDMG), which have statutory responsibilities for characterizing earthquake hazards on the national and state level, respectively.

We have formed a SCEC/IT Partnership to develop an advanced information infrastructure for system-level earthquake science in Southern California. Our partnership comprises SCEC, USC's Information Sciences Institute (ISI), the San Diego Supercomputer Center (SDSC), the Incorporated Institutions for Research in Seismology (IRIS, a 97-institution consortium), and the U.S. Geological Survey. The funding requested in this proposal will support four project elements:

- Fundamental IT research by ISI and SDSC on how to integrate knowledge representation and reasoning, knowledge acquisition, digital libraries and information management, and Grid resource sharing into a methodology that will support the SCEC Community Modeling Environment.
- Application of this new methodology by SCEC, USGS, and IRIS scientists to SHA for the purpose of reducing earthquake losses in Southern California.
- Transfer of the methodology to other regions and extension to other Earth science problems by IRIS and (at no cost to NSF) the USGS.
- Use of products from the SCEC Community Modeling Environment to educate students at all levels and inform the general public about earthquake hazards.

The scope of this effort clearly requires a large, integrated project involving an extended collaboration among many disciplines and research organizations that spans both earthquake and computer sciences.

C.2. Geoscience Approach and Research

C.2.a. Computational Pathways in Seismic Hazard Analysis

The SCEC/IT partnership will develop the information infrastructure to facilitate the four "computational pathways" diagrammed in **Figure 2**. The first three represent increasing levels of sophistication in the use of physics-based simulations to forward-model earthquake behaviors, while the fourth represents a collection of important seismological inverse problems. The research outlined in this proposal will emphasize the first two, where progress during a five-year project is expected to be rapid and most directly applicable to SHA.

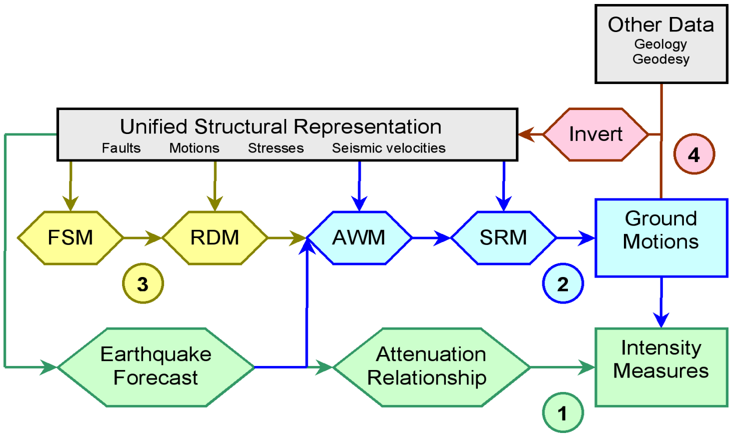

Figure 2. Computational pathways to be facilitated by the information infrastructure developed in the proposed project. (1) Current methodology in probabilistic seismic hazard analysis. (2) Ground-motion prediction using an anelastic wave model (AWM) and a site-response model (SRM). (3) Earthquake forecasting using a fault-system model (FSM) and a rupture-dynamics model (RDM). (4) Inversion of ground-motion data for parameters in the unified structural representation (USR), which includes 3D information on active faults, tectonic stresses, and seismic wave speeds.

Pathway 1 (shown in green in Figure 2) is the current methodology of probabilistic seismic hazard analysis (PSHA), which combines an *earthquake forecast model* with *attenuation relationships* to provide probabilistic estimates of *intensity measures*. The latter might include the peak ground acceleration (PGA), peak ground velocity (PGV), or the response spectral densities at specified frequencies. The earthquake forecast comprises a set of earthquake scenarios, each described by a magnitude, a location, and the probability that the scenario will occur by some future date (e.g., a Poisson distribution). The attenuation relationship is a relatively simple analytical expression that relates each earthquake scenario to the intensity of shaking (e.g., PGA) at each site of interest; it usually accounts for the local geologic conditions at each site (e.g., sediment sites tend to shake more than rock sites). The analysis determines the intensity that will be exceeded at some specified probability over a fixed period of time (e.g., PGA with a10% probability of exceedance during the 50-year life span of a building). The results are often presented as hazard maps [9], and engineers use these maps to design buildings, emergency preparedness officials use them for planning purposes, and insurance companies use them to estimate potential losses.

Pathway 2 (in blue) begins with an earthquake forecast model, but it employs the scenario events as sources for a physics-based calculation of ground motions. The waves from these sources are propagated using an *anelastic wave model* (AWM), and they excite ground motions at a specified location through a *site response model* (SRM) that accounts for the near-surface conditions, such as soil rigidity and layering. The results are vector-valued ground displacements as a function of time, from which essentially any intensity measure can be computed. However, in a region like Southern California where the geological structures are highly three-dimensional, the wavefield calculations must be done for each scenario earthquake on very dense grids to get the high frequencies of engineering interest (> 1 Hz), and the computational demands for these simulations can be enormous. One of the principal objectives of this proposal is to accelerate the use of ground-motion modeling in SHA. In particular, we seek the means to compute and distribute comprehensive catalogs of ground-motion simulations for use in risk assessment and earthquake-engineering analysis. Such catalogs are needed, for example, as input to research done at NSF's earthquake engineering research centers and its Network for Earthquake Engineering Simulation (NEES).

Pathway 3 (in yellow), when linked into Pathway 2, is the "full physics" calculation, in which the tectonic stresses in a *fault system model* (FSM) evolving over long time scales (years to centuries) cause spontaneous failures on fault segments. The details of these ruptures, which develop on very short time scales (seconds), are simulated by a *rupture dynamics model* (RDM), and the resulting fault displacements are used as input to Pathway 2. Fault system models capable of producing synthetic catalogs of earthquakes have been developed under various restrictive assumptions [10], but their ability to reliably predict seismicity sequences over extended intervals has not been fully evaluated. (Given the crudeness of the models, their accuracy is likely to be low.) Fully dynamical, 3D numerical simulations of spontaneous fault rupture are now feasible, and, properly tuned, these simulations have been shown to reproduce observed ground motions for large earthquakes, such as the 1992 Landers earthquake in Southern California [11].

Pathway 4 (in red) comprises a variety of important seismological inverse problems, which include using the ground motions observed in real earthquakes to image the fault rupture process (source inversion) and the 3D variations in seismic wave velocities and attenuation factors (structural inversion). At present, these inversions are usually done using 1D propagation models (e.g., for source imaging and surface-wave inversions) or simplified physics, such as asymptotic ray theory (e.g., for source location and travel-time tomography). Full use of broadband seismographic recordings in these inverse problems is currently limited by the difficulties in managing the forward calculations corresponding to Pathways 2 and 3. Furthermore, with very few exceptions, inverse problems considered to date explain ground motions using *kinematic* modeling, wherein the distribution of displacements or stress drops on a prescribed fault surface is related linearly to seismograms through a wave propagation model. *Dynamic* inversions in which observed seismograms are *assimilated* into a spontaneous fault rupture model with self-organizing geometry present seismological and computational problems that will take many years to solve, although initial efforts show considerable promise [11].

C.2.b. SCEC Research on Earthquake Simulation

A wide variety of data and model components are being developed as part of SCEC's basic research program funded by the NSF and USGS. SCEC scientists are currently engaged in active research on all four computational pathways described in Figure 2. Examples of recent SCEC results using a prototype for Pathway 2 [12] are shown in **Figure 3**. Under the next five years of NSF/USGS funding, SCEC will sponsor disciplinary working groups in seismology, tectonic geodesy, and earthquake geology to conduct data-gathering activities and develop disciplinary infrastructure, including field programs, centralized data processing, and the distribution of data products. Project-oriented focus groups will coordinate interdisciplinary research in four primary areas: (1) unified structural representation, which will combine geologic and seismic information into a coherent picture of subsurface structure, (2) fault-system modeling, including both the kinematical and dynamical behavior of the Southern California fault system, (3) earthquake simulation, including rupture dynamics, wave propagation, and site response, and (4) seismic hazard analysis. A new generation of SHA algorithms will be developed under the SCEC/USGS Working Group on

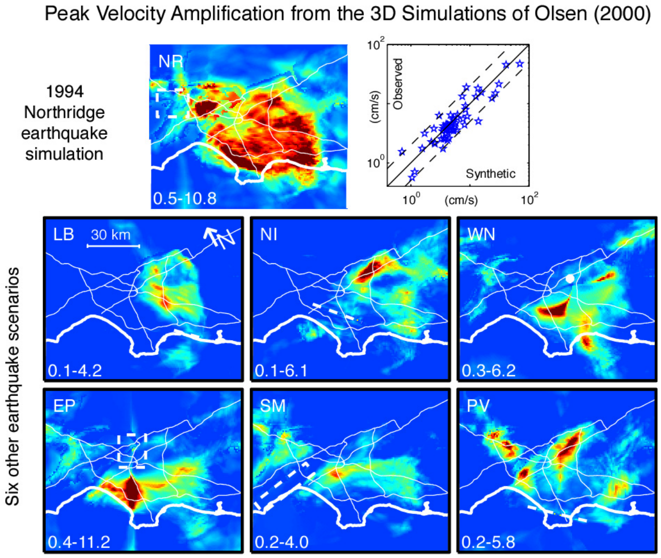

Figure 3. Seismic wavefields to frequencies of 0.5 Hz simulated for earthquakes in the Los Angeles region, showing deviations in peak ground velocity (PGV) from a 1D attenuation relationship in which PGV depends only on distance from the source [12]. Warmer colors express regions of higher PGV, which are primarily due to source directivity and basin effects. Top panels show the PGV anomalies predicted for the 1994 Northridge earthquake (left) and the fit to observations (right). Bottom panels show the simulation results for six future-earthquake scenarios: Long Beach (LB), Newport-Inglewood (NI), Whittier Narrows (WN), Elysian Park (EP), Santa Monica (SM) and Palos Verdes (PV). Faults for these earthquake simulations are indicated as white dashed lines (strike-slip faults) and boxes (thrust faults).

Regional Earthquake Likelihood Models (RELM) [13].

While the geoscience research sponsored under this proposal will leverage heavily on the SCEC efforts, it will not focus on specific data bases or computational algorithms, but rather on three end-to-end aspects of system-level earthquake science. Activities funded that would be funded under this proposal are: (1) the collaborations between geoscientists and computer scientists needed to develop the SCEC Community Modeling Environment, as described in §C.3, (2) testing of the Community Modeling Environment to validate the computer-science methodology, and (3) application of the Community Modeling Environment to produce SHA products of value to end-users. An example of (1) is the extensive work on data structures that will be required to ensure syntactic interoperability. A preliminary object-oriented model of seismological data, FISSURES, has been constructed by the IRIS Data Management Center [14]. SCEC will collaborate with IRIS and its other IT partners to extend the FISSURES model to other data types (geologic, geodetic) and to simulation inputs and outputs, including the 3D structural information about faults and elastic wave velocities contained in the USR.

Scientists from SCEC, IRIS, and the USGS will also create instantiations of the computational pathways described in Figure 2. For some of the simulation components (e.g., finite-difference and finite-element computations of AWM's), the methodologies have been verified by intercomparisons of simulation results [15], while for others the methodology remains the subject of active research (e.g., finite-difference, finite-element, and boundary-integral-element codes for RDM's), and verification will be necessary. The choice of the most appropriate algorithm often depends on the geological situation (e.g., soil type for SRM's or the geometry of faulting for RDM's). The availability of different algorithms with different parameter configurations will provide a multiplicity of computational pathways to arrive at an "answer." Automating the selection of these pathways will be required for rapid earthquake response, as well as for generating the many thousands of scenario simulations needed for SHA.

C.2.c. Examples of Pathway Complexity

The IT challenges of SHA can be illustrated by the standard calculations in Pathway 1. SHA practitioners come up with different earthquake-forecast models or attenuation relationships based on different assumptions and/or data types (e.g., historical seismicity versus geological fault data). In addition, each model is based on uncertain parameter values (e.g., the average recurrence interval of earthquakes on a specific fault). Understanding and dealing with these epistemic uncertainties remains a significant problem. The traditional approach has been via logic trees , where each step in the hazard analysis has branches representing viable alternatives, which are combined to produce a best estimate of the hazard level. Some of the specific issues are:

- *Parameter range constraints:* Models are developed to account for phenomena within certain parameter ranges, but users often forget or may just be unaware of these constraints. For example, models are often built to account for earthquakes of magnitude 7.5 or less. Although they can provide a result for a magnitude of 9, it is not clear that the result is meaningful or that the model developers would stand by them.
- *Parameter approximations and settings:* Models often require parameters that users end up approximating when their values are not readily available. For example, a model may require shear-wave velocity as a parameter,

while the user only knows that the terrain is hard rock. Since the two are related, the user may use some idiosyncratic rules of thumb that are less accurate than other approximations more consistent with the model.

- *Interacting constraints:* A problem that has largely been ignored is covariance among different logic tree choices. For example, at one step a magnitude might be assigned to a scenario earthquake and at a later step a recurrence interval assigned. However, the two may be related in that a larger magnitude might imply a longer recurrence interval. Another example is that a choice in the earthquake-forecast model may determine what are sensible choices of an attenuation relationship, say, one appropriate to a strike-slip earthquake, rather than a reverse-faulting event. Users who unwittingly violate such constraints may produce very inaccurate results.
- *Pathway traceability:* Traceability of the information and sources used in hazard estimation is a major issue. Such documentation is needed to maintain reproducibility and to allow future exploration of alternative model/parameter choices. For example, different historical earthquake catalogs can produce different results, although it may not be clear whether the catalogs themselves or other input data are responsible for the discrepancy.
- *Computational demands:* Although the Pathway-1 calculations are relatively simple, the computer time needed to generate a complete regional hazard map can easily exceed a day on a typical desktop computer. Even these modest demands can discourage useful sensitivity analysis. The other pathways we seek to incorporate into SHA will have far greater computational requirements. Therefore, access to rapid computational services is highly desirable.

Although each step in the analysis may involve only a handful of viable models or parameter settings within a model, navigating the labyrinth of logic tree branches and testing for compatibility is not currently possible with existing tools. The work outlined in this proposal would enable end-users such as earthquake engineers to run the simulations directly themselves and with some guarantee of correctness and accuracy, while maintaining a trace of the sources and parameters used to generate the resulting hazard estimates.

Despite the fact that automating SHA in its most general form is still a formidable task, it has the following characteristics that make us confident that significant progress can be made within the scope of this proposal:

- *Relatively smaller size of components in library*: We anticipate that there will be only a few dozen components in the software library that can be used to run simulations of seismic sources and strong-motion wavefields.
- *Relatively small computational pathways*: Only a handful of models will be selected from the library in order to setup a SHA assessment. We will focus primarily on the computational pathway that deals with the calculation of wavefields generated by realistic source models and propagated through realistic geological models.
- *Relatively well understood model descriptions*: The models can be characterized with a relatively small set of parameters. Efforts by earthquake scientists are currently underway to represent these model parameters as Java classes, currently by earthquake scientists themselves. This will provide an excellent starting point for our proposed ontology development efforts.
- *Availability of models*: Since models and databases used in SHA are available from a variety of sites, they often contain slight variations of the same information. To remedy this situation, there is already an ongoing effort within SCEC—the RELM project—to establish well-maintained community databases to minimize such data heterogeneity, as well as to create a repository of validated code. This will facilitate enormously our efforts to develop the computation infrastructure needed to access such information in real time from the host institution.

The next sections describe our approach in more detail.

C.3. Architecture and Approach

Our goal is to develop an integrated environment in which a broad user community encompassing geoscientists, civil and structural engineers, educators, city planners, and disaster response teams can have access to powerful physics-based simulation techniques for seismic hazard analysis. To achieve this goal, the environment must provide a means for describing, configuring, instantiating, and executing complex computational pathways that result from the composition of various earthquake simulation models. The proposed architecture, illustrated in Figure 1, brings together research from several distinct computer science disciplines, with each area addressing one of the requirements (R1-R4) stated earlier:

1. **Knowledge representation and reasoning** techniques to manage the heterogeneity of the models and capture the complex relationships between the physical processes and the model algorithms, between the algorithms and the simulation codes, and between the simulation codes and the data products. Knowledge-based inference will be used to apply these representations to the problems of pathway construction, constraint checking, execution planning, and information access.
2. **Grid technologies** to enable access to distributed simulation codes and resources for the timely execution of the simulation scenarios defined by users, specifically by integrating high-performance computing resources into the execution environment available to the modeling framework. By providing mechanisms for the discovery,

access, and management of distributed computation and storage resources, Grids address the distributed nature of the developers, resources and user environments.

3. **Digital library** technology to manage the collections of data and simulation code repositories and handle multiple versions of the models. Knowledge-based data management tools will provide an infrastructure for mediating access to existing seismic data catalogs and information repositories, as well as incorporating new collections of data generated by the simulations.
4. **Interactive knowledge acquisition** techniques to enable users with a range of sophistication to configure computational pathways. Knowledge acquisition tools support this activity by selecting appropriate simulation software and input data files from the code and information repositories available. These tools hide implementation details and present users with structured dialogues that guide them to provide information required to set up each simulation while resolving the constraints among the simulation models and their inputs.

The resulting environment will consist of an integrated set of services, knowledge bases, databases and tools which together implement the SCEC community modeling environment. In the following sections, we describe the contributions of each computer science technology in more detail.

C.3.a. Knowledge Representation and Reasoning: Managing Community Models

Earthquake models are very complex due to the complexity of geologic faults, the non-linearity of the underlying physics, uncertainty about initial conditions, computational complexity of the requisite numerical methods, and many other factors. Model implementations have to address additional complexities such as the representation of large and complex data sets, representation and storage of simulation outputs, handling of large sets of input and modeling parameters, execution requirements on high-performance computing infrastructure, etc. For almost all simulation models and codes in use today, these characteristics are completely opaque and implicit, and are at best communicated via textual publications or “word of mouth”.

To achieve the vision of a *virtual collaboratory* where distributed and heterogeneous community model components can be rapidly located, assembled, configured, and run to solve an analysis task at hand, we need software tools to support users in the various stages of this process. These tools need to “know” all the pertinent characteristics of model components to facilitate tasks such as automated configuration, constraint checking, input/output translation, execution planning, etc. To support the representation of complex and heterogeneous model characteristics as well as the necessary reasoning with these characteristics and associated constraints, we propose to use a knowledge-based approach that builds upon (1) the creation and use of shared *ontologies* (2) modeling and inference techniques from the area of knowledge representation and reasoning (KR&R), and (3) ontology translation technology.

The heart of our approach is the creation of curated knowledge base for earthquake physics, an activity that parallels knowledge sharing efforts in other scientific disciplines, such as UMLS and GO [16,17,18]. Building on existing efforts within SCEC to develop consensus community models, geoscientists and knowledge engineers will collaborate to develop a knowledge base (KB) that will contain (1) terminology about the domain of earthquake science represented as ontologies of relevant terms [19,20,21], (2) procedural knowledge models representing pathway templates and idealized descriptions of executable simulation code, and (3) detailed models of the Unified Structural Representation that will capture the surface and subsurface structure of Southern California, including include the fault systems and seismic velocity models that will make possible the physics-based simulations in Pathways 2-4. The structure of the proposed KB is shown in **Figure 4**.

In our approach, each community model component will be annotated with a set of logical descriptions represented in a formal KR language to describe the pertinent characteristics of the component or its *profile*. Such a profile will represent things such as the type of the model, simulation technique used, modeling assumptions, parameters and their ranges, parameter dependencies, input data requirements, characteristics of the output, data formats used, compute power requirements, etc. Note that the representation of these characteristics requires the use of an adequately expressive KR language. For example, we might need to express for some model that “to simulate ground motions with frequencies greater than 0.5 Hz, a computer with more than 3 Gbytes of RAM is required.” We will also need adequate logical reasoning power to allow the system to conclude that a 16-processor computer with 500Mbytes RAM per processor will satisfy this requirement if the model has been appropriately parallelized and can be run on the particular platform.

The knowledge base will also contain *pathway templates* that describe, for example, what combinations of simulation models can be used to obtain intensity measures of a certain type or accuracy. These pathway templates will express the constraints that must be met when individual simulations are combined to configure a specific pathway. The templates are used by interactive knowledge acquisition tools to generate a specification of required data, software and resources for use by the Grid execution environment.

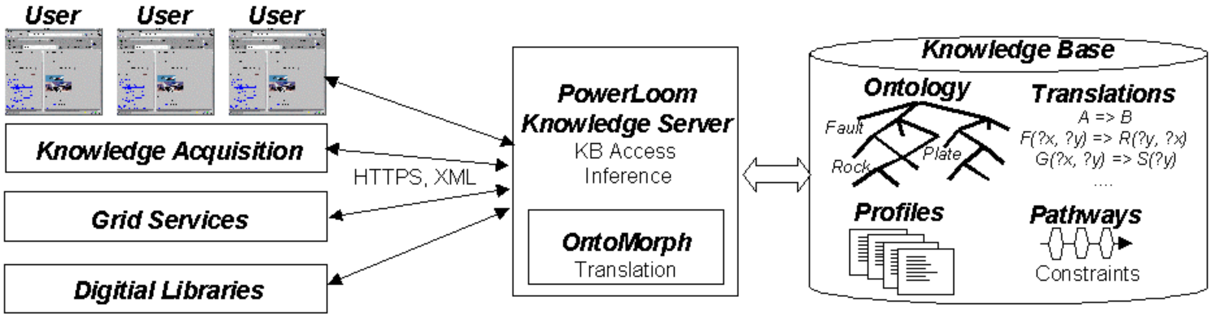

Figure 4. Overview of the SCEC knowledge base and its integration into the modeling environment.

To be able to compare meaningfully the profiles of components developed by different and potentially widely distributed groups, the descriptions of these profiles must draw from the same ontology or representational vocabulary [22]. For example, the previous section mentioned that, in the context of Pathway 1, the fault type used by the earthquake forecasting model must match the fault type controlling the attenuation relationships used to predict intensity measures. This can only be done if both models use the same (or compatible) terms to describe such fault-type restrictions. A major thrust of our work will be the development of such a shared ontology adequate for the area of earthquake science. This ontology needs to cover relevant *physical phenomena and events* such as types of terrain and faults, types of earthquakes and their occurrence, temporal and spatial relationships, etc. It also needs to cover pertinent aspects of the *models* describing or simulating such physical phenomena, as well as aspects of their *implementation* in a particular piece of computer code and their *execution* on some computer platform. The resulting ontology will not only support the various knowledge-based tools, but it will also drive consensus building among the scientific community—a very desirable result.

To create and manipulate KBs, we will build upon the PowerLoom KR&R system developed at ISI to provide the necessary representation and reasoning services. PowerLoom is designed for high expressivity (it uses a representation language based on first-order predicate calculus) but without sacrificing scalability to large knowledge bases. PowerLoom has been successfully applied within the context of various DARPA research programs [23, 24] and, to date, has been distributed to over 100 sites and universities world-wide. As illustrated in Figure 4**,** KBs are integrated into the modeling environment via network-enabled *knowledge servers*, which respond to knowledge and inferencing requests formatted as XML over HTTP connections. As part of this project, we will integrate the knowledge servers into the Grid environment described below, providing for knowledge service discovery via Grid information services and secure knowledge inquiry via the Grid Security Infrastructure.

Access to the PowerLoom knowledge-base and inference services will be central to many of the functions performed by other components. For example, its subsumption reasoner can check whether the input constraints of one component subsume (or are logically entailed by) the output description of another component, which will be used by the Pathway Assembly tool (described in §C.3.d) to check whether two components are compatible. Descriptions of available compute and storage capabilities will be used by Grid services to generate adequate execution plans. Finally, digital libraries can use access to the ontology to organize collections and facilitate search and navigation.

The distributed and heterogeneous nature of the community modeling environment will make it necessary to deal with issues of *translation*. For example, it will often be required to translate data formats, change resolution, perform unit conversions, etc., to match the output of one component to the input requirements of another. Some of these translations can be directly supported by the KR&R system; others will need to rely on special-purpose translation components, e.g., if very large data sets or outputs need to be translated. Such translators can become community components themselves and be leveraged in pathway configuration tasks. More difficult translation issues arise from ontology evolution, independent ontology development, or sometimes simply lack of consensus. While ontologies are intended to provide a stable, universally agreed upon vocabulary, experience shows that they often need to change or evolve to handle new situations or capture newly found consensus or understanding. Moreover, communities sometimes cannot agree and develop different ontologies covering the same subject area. This can cause significant maintenance or translation problems for logical descriptions (such as our model annotations) based on different ontology versions. We will build upon our OntoMorph ontology translation system [25] to handle some of these translation and maintenance problems. For example, whenever the ontology evolves, we can provide a set of translation mappings to automatically translate descriptions based on an older version of the ontology into its new format. This mechanism can also be used to export the ontology into a different representation syntax for use by other systems. For example, we can generate *Topic Maps* [26] to support digital library systems or translate into a different KR language for use with different KR&R technology.

The domain of earthquake science provides a variety of interesting KR&R research challenges, such as having to represent and reason at widely different temporal and spatial scales and resolutions and having to deal with a number of difficult translation problems. There also is the need to reason with partial, approximate, and incomplete information; e.g., to handle only partially satisfiable constraints, or to weigh different soft constraints against each other. This is necessary, since the model components developed by different groups of scientists are heterogeneous and will not necessarily be designed to fit together smoothly, nor might their requirements be completely or correctly specified; however, such assumptions are commonly made by traditional configuration approaches [27,28].

C.3.b. The Grid: High Performance Computing for Pathway Execution

Grids [29] are an emerging technology for creating and maintaining *virtual organizations*—multi-organizational collaborations that share distributed resources to solve problems of common interest [30]. In many ways, SCEC is a prototypical virtual organization. SCEC participants are drawn from many different organizations, and the computers, earthquake catalogs, simulation codes, and other resources used by SCEC to address problems of geophysical modeling are geographically distributed. Actual simulation code may be stored in repositories under the control of the author, while the computers being used to execute these codes may be located within a national facility such as at San Diego Supercomputer Center, or be part of the soon to be deployed Distributed Terascale Facility. Input data may come from one of the existing earthquake catalogs, or may be the result of a previous simulation run. Modeling in this environment requires the ability to pull all of these disparate resources into a single integrated computation. Grids provide these mechanisms.

At its most basic, Grid infrastructure furnishes the fundamental services needed to locate, access, and manage shared resource. These services include:

- A common security infrastructure that provides single sign-on access to all accessible resources [31,32].
- Information services that enable users and applications to discover the existence of resources available to the SCEC community, as well as to determine characteristics of those resources, such as how many processors are available on compute resources, or how much disk space is available on storage resources [33]. As part of this project, we will make configuration and status information provided by the information service available through the knowledge service, so that it can be used by the knowledge acquisition components of the modeling framework to help guide pathway instantiation.
- Uniform resource management protocols that enable allocation, monitoring and control of a variety of different resource types, including computers [34], storage systems [35] and networks [36].

These services are among those that are provided by Globus Grid toolkit [37]. We will use Globus as the infrastructure on which we build the execution environment of our integrated modeling environment. Globus services are being widely used by a number of different Grid oriented projects, in the US, Europe, and Asia. Much like TCP/IP and DNS provide the basic services on which can be built higher-level capabilities, such as the World Wide Web, Globus services are designed to provide the low-level mechanisms on which higher level, domain specific services are constructed. Within the scope of this proposal, we will not be developing new Globus services, but rather creating higher level functionality that builds on the core Grid infrastructure defined by Globus.

Figure 5 illustrates the basic structure of the execution enviroment. Within this component of the modeling framework, instantiated computational pathways, specifying the names and configurations of simulation models to be run, must be mapped to an execution plan, which specifies the sequence of execution steps to be taken, the

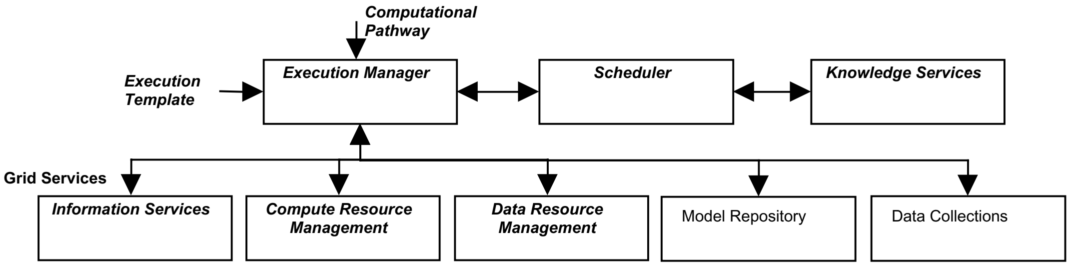

Figure 5. Basic structure of the proposed execution environment. Components to be developed as part of this proposal include the execution manager and a scheduler that can exploit knowledge of the simulation path to facility mapping of execution onto computing resources.

storage systems or collections containing input data, the executables to be used for each simulation model, the computer to be used to execute the model, and the storage system on which to place output data. We will develop a domain specific scheduler to perform this task. To simplify the process, the schededuler will convert the computational pathway into an scripted execution plan, using a set of pre-defined computational structures (such as parameter sweep, processing pipelines, and parallel execution). Information about model components available via the knowledge base will be used to help guide scheduling decisions. Our initial scheduler will be simple, using application-level scheduling heurisics [38,39,40] that exploit the structure of the execution templates. More sophisticated mappings can be obtained by querying the knowledge service to determine characteristics of the various simulation models, along with information about the computational environment. In later years of the project, we propose to apply planning techniques such as those used by the knowledge acquisition tools to instantiate computational pathways to create execution schedules. Near the middle of the project, we will integrate pathway instantiation and exceuction scheduling so that we can better guide pathway instantation based on resource availability and to perform replanning in case of resource failure or unplanned unavalibility.

The second major component of the execution environment is the *execution manager*. The execution manager is responsible for interacting with the scheduler to obtain an execution plan, and then to interact with the underlying Grid services to cause the plan to be followed. The execution manager will activly monitor the progress of the execution plan, dealing with events such as resource failure. Initial versions of the execution manager will employ simple failure recovery strategies, such as restart. In the second half of the project, the execution manager will be augmented to go back to the scheduler and re-plan the execution schedule to work around resource failure.

C.3.c. Knowledge-Based Collection Management: Simulation Code and Data Repositories

The purpose of performing a simulation is to produce data for further analysis. Given the distributed nature of the SCEC environment and the volume and number of data products that will be produced, we need a mechanism for managing and discovering information that lives in existing catalogs, such as IRIS, as well as new data products that are generated within the simulation environment. Digital libraries are currently being used to manage digital objects for a wide range of application domains [41, 42, 43, 44], and we plan to apply this technology to the problem of managing seismic simulation data within the context of our proposed simulation environment. Specifically, we will use the SDSC Storage Resource Broker to provide mechanisms for building distributed collections of simulation output, extensible metadata catalogs for managing collection attributes, and interoperability mechanisms for accessing data stored in archives, file systems, and databases.

A number of interesting research problems arise in the integration of digital library technology with Grids, knowledge bases, and knowledge acquisition, as we propose to do. Specifically, we need to maintain sufficient information about the construction and execution of specific computational pathways so as to facilitate queries based on the means by which the information was produced in order to replicate the production of identical or similar data analyses. This will require coupling information produced by Execution manager and pathway assembly tool (described below) into the managed collection.

Another important issue is how to manage interoperability for data and metadata exchange between the multiple data collections accessed by the community models. The research issues include development of mediation interfaces to the storage resources and information catalogs accessed by the community models, integration of the ability to manage collections of simulation output with the Grid execution environment.

Finally, we observe that a current area of active research in the digital library community is the incorporation of knowledge into a digital library by mapping domain concepts to the digital library metadata attributes [45]. The goal is to support concept-based queries against the collection in order to identify relevant data sets that have specified relationships described within the knowledge base. We will explore use of the ISO 13250 Topic Map standard as a potential syntax for representing the mapping between collection attributes and ontologies. As discussed in §C.3.a, we propose to use OntoMorph to translate between the knowledge representation used within the knowledge base and topic maps. This will provide us with a mechanism by which the collection management and knowledge bases can be integrated.

Another important research topic is the characterization of completeness and closure for the mapping between the knowledge ontologies and the simulation collections. Completeness is the identification of the necessary set of attributes required to represent each of the concepts expressed within the ontology. Without a complete set of attributes, it will not be possible to apply the constraints specified for each pathway. Closure is the identification of the necessary relationships needed to describe the inherent knowledge within the simulation collection. If relationships imposed by a particular choice of input data or simulation algorithms are not captured, then artifacts or anomalies may be introduced into the collection when further processing is done. The development of mechanisms to specify completeness and closure is domain-dependent. Through close interaction between SCEC geoscientists and IT researchers, we will continually evaluate the SCEC Community Modeling Environment for consistency relative to the completeness and closure properties.

C.3.d. Interactive Knowledge Acquisition: A Pathway Assembly Tool

In the previous sections, we described technology that defines the modeling and simulation infrastructure. In this section, we turn our attention to how the end-user interacts with this infrastructure and describe a Pathway Assembly tool that will enable unsophisticated users to compose computational pathways involving complex simulations. Targeted users might include building designers and engineers, emergency preparedness officials, or insurance companies. Since these users are not programmers or knowledge engineers, the tool needs to provide an easy to use acquisition interface that guides them to (1) select from a library of simulation models those that will address their requirements, (2) setup the input parameters required by the simulation code, taking into account constraints specified by the model developers, (3) coordinate interacting constraints across models, as described in §C.2.c. The major components of this Pathway Assembly tool are shown in **Figure 6** and described in the rest of this section.

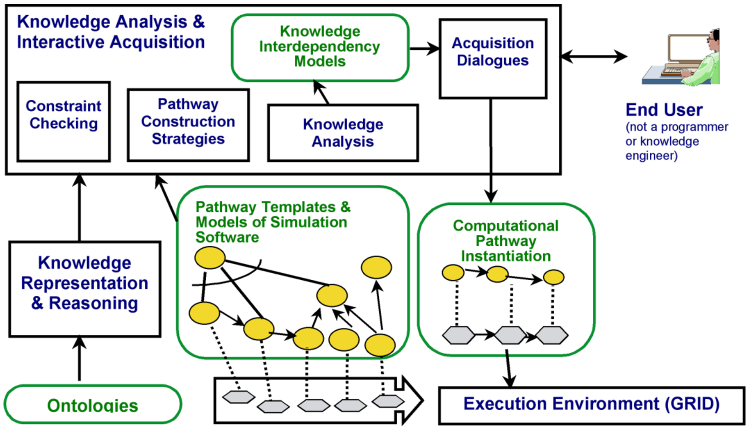

Figure 6. A Pathway Assembly Tool. Users who are not programmers or knowledge engineers will use this tool to setup a computational pathway corresponding to an end-to-end simulation by (1) selecting simulation models from a library, (2) setting up the input parameters required, and (3) coordinating interacting constraints across models in the overall computational pathway.

For the work on this proposal, we will cast the composition of the computational pathway as a planning problem [46,47]. The user will be prompted for requirements in terms of the desired intensity measures and accuracy of the results of the simulation, as well as the computational resources available. These will be turned into planning goals and resources. Pathway templates will be considered plan decomposition fragments that the user weaves together and instantiates. The Pathway Assembly tool will also include a suite of Pathway Construction Strategies that capture typical requirements and dataflow interactions among models within a pathway. Inputs and outputs of the individual components and their associated constraints will be expressed as required preconditions and expected effects. For example, this will enable the tool to detect that one of the inputs required by a ground-motion model is an earthquake forecast model, which can only be generated by an earthquake forecasting simulation. In contrast with other planning tools, our system will be able to exploit knowledge-rich structures that describe tasks and goals [48,49,50,51]. The Pathway Assembly tool will be developed as a plan-authoring environment, where the tool is helping the user by checking that the preconditions, effects, goals, resources are adequately handled [52,53]. This planning framework fits well the paradigm of the Grid execution environment with respect to handling resources, scheduling, and bringing up the need for replanning when the authored plan is not feasible.

We will build on previous research on the EXPECT architecture for knowledge analysis and interactive acquisition [54,55,56,57], which has been used to acquire planning knowledge. The central thesis of EXPECT is that if the tool understands how individual pieces of knowledge relate to each other, then it can understand how new knowledge fits and what additional knowledge is missing and, as a result, guide users in adding it. To this end, EXPECT analyzes a knowledge base and automatically derives *Interdependency Models* between problem-solving knowledge (procedures to achieve goals and tasks), ontologies (object models), and factual knowledge (data). Interdependency Models can be used to plan acquisition dialogues that organize the interaction with users into meaningful topics and sequences of related questions. EXPECT has been used to develop an interactive tool for acquisition of planning constraints, ranging from simple value preferences and bounds to complex procedural constraints. For example, geophysicists can express simple constraints such as “compute the Coulomb stress change for all faults parallel to the San Andreas Fault”, or more complex constraints such as “compute the maximum Coulomb stress change seen on all faults with a strike within 30 degrees of the local strike of the San Andreas fault, and a dip steeper than 60 degrees, knowing that only through-going faults breaking the entire seismogenic zone contribute significantly to the hazard change." EXPECT has been used for other planning domains including air-campaign planning, special operations, and logistics. EXPECT's approach to knowledge acquisition has been evaluated successfully in diverse domains with users such as Army officers, project assistants, and non-CS students. Building on the ideas in EXPECT, we recently developed KANAL, a prototype tool for checking the correctness of manually developed process models and plans by exploiting Interdependency Models [58].

The main research problems to be addressed by this tool are: (1) how to enable users who are not experts in programming or knowledge engineering to formulate computational pathways that are ready for the execution environment; (2) how to hide from users the details of the models of the software components and their constraints, while enabling them to understand enough to decide which alternative simulation models to use and; (3) how to design plan-authoring tools that take into account scheduling, execution failures, and replanning issues brought up by the Grid execution environment.

Although many issues related to software modeling, reuse, and configuration may arise during this work, the tool that we are proposing to build will be highly focused on earthquake applications. Similarly focused research projects for configuring software have been successful for applications such as image processing, engineering design, and planning interplanetary science missions [59,60,61,62]. We believe that the knowledge-rich community models will enable us to develop more sophisticated and capable environments that will make the pathway-assembly process accessible to end-users.

C.3.e. Integration Benefits

The synergistic collaborations among different computer science disciplines will enable advances in each of them that would not be possible without these collaborations:

- Knowledge-rich repositories will be grounded in large-scale, real-world data sources. This will enable a layered approach to building knowledge bases, starting from simpler representations such as database schemas and IDL specs going all the way up to more expressive descriptions in first-order logic. These repositories will be available to other researchers in knowledge representation and reasoning, databases, etc., and will facilitate much needed synergies among these communities.
- Grid technologies will have access to deep knowledge about program components and access to powerful planning techniques not typically used for scheduling. This integration has the potential to result in the construction of more robust and efficient Grid applications than are possible using current techniques.
- Knowledge-based collection management systems will have access to more powerful knowledge representations and inference engines than are currently being used. Furthermore, the tight coupling between knowledge-based collection management and the knowledge representations used for planning has the potential to increase the utility of managed collections by broadening the types of queries that can be posed to the collection management systems.

In addition to integration across information technology efforts, there must also be a tight coupling between information technology activities and the earthquake science. To help ensure that this takes place, we have structured the effort so that each major institution has participants in both information technology and earthquake science. We have also included funds in the budget to support an Annual SCEC/IT Workshop to bring larger groups within each field together on an annual basis and smaller tutorial workshops to cross-train students in computer science and geoscience.

C.4. Milestones and Schedule

The SCEC Community Modeling Environment will be developed by collaborative efforts between the SCEC institutions and ISI, SDSC, IRIS, and USGS. The research and development tasks will be supported by teams consisting nominally of supervisory staff, one of more post docs, and one of more graduate students. The teams will be constructed to ensure the linkage of expertise in information technology and geoscience, with supervisory staff in geoscience provided by SCEC, USGS, and IRIS, and supervisory staff in computer science provided by ISI and SDSC.

Geophysics activities will be supervised by primarily by senior scientists from the Scripps Institution of Oceanography, Carnegie Mellon University, University of California at Santa Barbara, University of Southern California, and San Diego State University. However, we expect many other geoscientists to be involved through their participation in SCEC research, which will be tightly coordinated with this project. Activities associated with knowledge-based collection management will be centered at the San Diego Supercomputer Center, while Grid, knowledge-base and knowledge-acquisition activities will be centered at the Information Sciences Institute. Integration activities will be the responsibility of all participants, but they will be coordinated by the P.I. and the

SCEC Information Architect, who will act as Project Manager.

Each year of this five-year project will have an overarching goal and a series of specific objectives, listed below. For the purpose of indicating the division of task responsibilities, the efforts of USGS personnel, who will participate at no cost to this project (and are thus not in our budget), are not listed separately, but rather combined as part of SCEC.

Year 1: Constructing Basic Building Blocks

Workshop to assess end-user needs for SCEC community models (SCEC/IRIS/SDSC/ISI).
Ontology design and knowledge servers for Pathway 1 (SCEC/ISI).
Workshop to structure initial SCEC community model components (SCEC/IRIS/SDSC/ISI).
Extension of FISSURES data model to geodetic and geologic data sets (SCEC/IRIS).
Deployment of initial Grid testbed with support for SCEC virtual organization (ISI/SDSC).
Pathway assembly tool for Pathway 1 templates (ISI/SCEC/SDSC).
Execution framework for Pathway 2 simulations (ISI/SCEC).
Collection management including recording execution environment and results (SDSC).
Delivery of tools and prototypes to SCEC community (SCEC/ISI/SDSC).

Year 2: Integration and Prototype System

Initial integration of Grid, pathway assembly, collection management, and knowledge representation, focused on Pathway 1 specification and execution (SCEC/ISI/SDSC).
End-to-end demonstration of integrated prototype for Pathway 1 (SCEC/ISI/SDSC).
Workshops to demonstrate SCEC community model components to end-users (SCEC/IRIS/SDSC/ISI).
Code verification and ontology design for Pathway 2 (SCEC/IRIS/ISI).
Basic end-user interface for instantiation of simple pathways (ISI/SCEC).
Knowledge base and collection integration through PowerLoom/Topic Map translation (ISI/SDSC).
Delivery of tools and prototypes to SCEC community (SCEC/ISI/SDSC).

Year 3 : Initial Technology Transition and Complex Pathways

Incorporation of resource constraints into Pathway assembly planning (ISI).
Incorporation of Pathway 2 into end-to-end demonstration (SCEC/ISI/SDSC).
Workshops to evaluate SCEC community model components by end-users (SCEC/IRIS/SDSC/ISI).
Mediators to existing seismic collections (SDSC).
Characterize of consistency between pathway resource constraints and simulation collection attributes (SDSC).
Unified structural representation for Southern California (SCEC/ISI/SDSC).
End-user interfaces for multi-pathway instantiation (ISI/SCEC).
Design of simulation comparison tools for validating pathway results (SDSC).
Delivery of tools and prototypes to SCEC and earthquake-engineering communities (SCEC/ISI/SDSC).

Year 4: Catalog Integration and Advanced Planning

Investigate use of replanning to handle resource failure and unavailability (ISI).
Workshops to plan enhancements to the SCEC community model components and extensions to EarthScope/USArray applications (SCEC/IRIS/SDSC/ISI).
Mediators to existing seismic collections and simulation collections (SDSC).
Analysis of consistency between pathway resource constraints and simulation collection attributes (SDSC).
Code verification and ontology design for Pathway-4 inverse problems, including waveform inversion for 3D structure (SCEC/IRIS/ISI).
Development of simulation comparison tools for validating pathway results (SDSC).
Delivery of tools and prototypes to SCEC and IRIS communities (SCEC/IRIS/ISI/SDSC).

Year 5: Evaluation, Catalog Integration, and Transfer

Integration of pathway assembly and execution planning (ISI).
Workshops to promote use of the SCEC community model components by national geoscience community (SCEC/IRIS/SDSC/ISI).
Extension of unified structural representation to other geographic regions (IRIS/SCEC).
Mediators to researcher simulation collections for publishing output (SDSC).
Validation of consistency between pathway resource constraints and simulation collection attributes (SDSC).
Implementation of unified structural representation for the SCEC community (SCEC/ISI/SDSC).
Characterization and analysis of accuracy of pathway models (SDSC).
Technology transition to EarthScope (IRIS/SCEC).
Delivery of tools and prototypes to national geoscience and earthquake-engineering communities (SCEC/IRIS/ISI/SDSC).

C.5. Knowledge Transfer, Education, and Outreach

This project will sponsor an Education and Outreach (E&O) program with four primary goals: (1) to transfer the technology developed under this project to end-users of earthquake information; (2) to cross-educate advanced students and postdocs in the fields of geoscience and computer-science; (3) to make the general public aware of the benefits of applying advanced information technology to the problems of earthquake risk; and (4) to use public interest in earthquake information to attract beginning students into geoscience and computer science. Because SCEC services a Southern California population of 20 million with an Hispanic plurality, a specific objective of our E&O program will be to engage young Hispanic Americans in the intellectual challenges of earthquake information technology.

The SCEC/IT Partnership comprises organizations with powerful, nationally recognized E&O programs, and we will use these organizational resources to achieve our E&O goals. SCEC maintains a large, very successful E&O program to transfer the knowledge gained by SCEC researchers to end-user communities, to educate students at all levels, and to engage the general public in earthquake-related issues. SCEC also coordinates its activities with its partners in earthquake engineering and risk management, including the Consortium of Universities for Research in Earthquake Engineering (CUREE) and the Pacific Earthquake Engineering Research (PEER) Center. IRIS's E&O program is designed to enhance seismology and Earth Science education in K-12 schools, colleges and universities, and in adult education at the national level. The UCSD NPACI program maintains an E&O program to apply advanced technology in support of human resource development, increase participation of underrepresented groups, and promote national programs in education.

The technology-transfer goal will be achieved primarily through the SCEC E&O program. We have specifically budgeted two tutorial workshops per year to make the current results of this project available to the end-user communities and to get their feedback on our development activities. The locations of these workshops will be distributed at various SCEC institutions to ensure participation by end-users throughout California.

This project will provide an excellent means for the cross-training of students and postdocs in geoscience and computer science. A structure based on research teams comprising representatives of both the IT and the geoscience communities (see Management Plan) will ensure this cross-training for young scientists directly involved in the project, and our annual meetings and tutorial workshops will extend this educational endeavor to a much larger group. We will also use the SCEC Community Modeling Environment as a case study in two USC courses, CS 541 (Artificial Intelligence Planning) and CS598 (Knowledge Based Systems/Knowledge Representation and Reasoning). Dr. Y. Gil, a senior scientist on this project, is the faculty member currently responsible for both courses. The courses will be revamped by using the framework described in this proposal as a practical sandbox for course projects, and their structure will be revised so that they can taken by both geophysics and computer science students.

To expose less advanced students from various disciplines to the excitement of this research, we have included funds for 5 undergraduates as part of the USC budget. These students will participate in the SCEC Summer Interns Program, which matches students from around the country with advisors who supervise summer research projects. We will use these funds specifically to recruit students from computer science and related disciplines, and we will specifically seek out and recruit students who are Hispanic American or from other underrepresented groups.

The output of the proposed project will include many products of general interest that can be used to improve public understanding of earthquakes (e.g., 4D visualizations of earthquake and ground-motion simulations). These products will find extensive use in SCEC and IRIS's excellent compendium of curricular and on-line educational materials, such as the Electronic Encyclopedia of Earthquakes (E^3). E^3 is an NSF-sponsored venture involving SCEC, IRIS, and CUREE to develop a digital library collection of earthquake information in partnership with the Digital Library for Earth System Education (DLESE).

The proposed project will produce new methodologies and analysis products of potentially great utility to other technical groups outside of Southern California earthquake science. The inclusion of the USGS and IRIS in the SCEC/IT partnership will facilitate this technology transfer. The USGS will take the lead in applying the information infrastructure developed in this project to real-time operations and post-earthquake emergency response. The USGS will also ensure that the results of this ITR effort are exported to other regions of earthquake risk, as well as to other USGS activities. IRIS will ensure that the methodology developed for the study of Southern California earthquakes can be transported to other regions and used to facilitate the NSF-sponsored EarthScope Project [63].

C.6. Management Plan

The work in this proposal will be managed as an independent research project. The management plan has four key features: (1) Project leadership will be the responsibility of a Steering Committee comprising the P.I., the three Co-P.I.'s,, and representatives of IRIS and the USGS. The P.I. and Steering Committee will report directly to the NSF. (2) Project management will be the responsibility of an IT Architect, who will be jointly supported by this project and SCEC. (3) Project organization and execution will leverage on the existing strengths of SCEC, which maintains associations with many representatives of the end-user communities. (4) Project structure will focus on interdisciplinary teams comprising geoscientists and computer scientists. Research teams at SCEC, ISI, and SDSC will generally involve a senior scientist, one or more postdocs, and one or more graduate students.

C.6.a. SCEC Structure and Daily Administrative Management

The management of this project will be built on the existing management structure of SCEC. SCEC will graduate from the NSF Science & Technology Center (STC) program in January, 2002. A proposal has been submitted to NSF/EAR and the USGS for a 5-year continuation of the Center, and the agencies are expected to return a decision on this proposal by May, 2001. The University of Southern California (USC) will continue as the SCEC managing institution, with T. Jordan, the Principal Investigator of this proposal, as the new Center Director. A diagram showing the structure of SCEC and its relationship to the proposed project is given in **Figure 7**.

SCEC is an institutionally-based organization governed by a Board of Directors representing its core institutions. The fourteen core institutions enrolled in the Center renewal proposal are Caltech, Columbia, Harvard, MIT, San Diego State, Stanford, USGS Golden, USGS Menlo Park, USGS Pasadena, UCLA, UCSD, UCSB, University of Nevada, and USC. The Center Director acts as Chair of the Board and the representative of the managing institution (USC). The Center Director is also the Chief Executive Officer of the Center and bears ultimate responsibility for the Center's programs and budget. The Center has an External Advisory Council that serves as an experienced advisory body to the Board of Directors, which comprises a diverse membership representing all aspects of Center activities. Knowledge transfer, education, and public outreach are managed by a Vice-Director for Communication, Education and Outreach (CEO), who supervises a staff of CEO specialists.

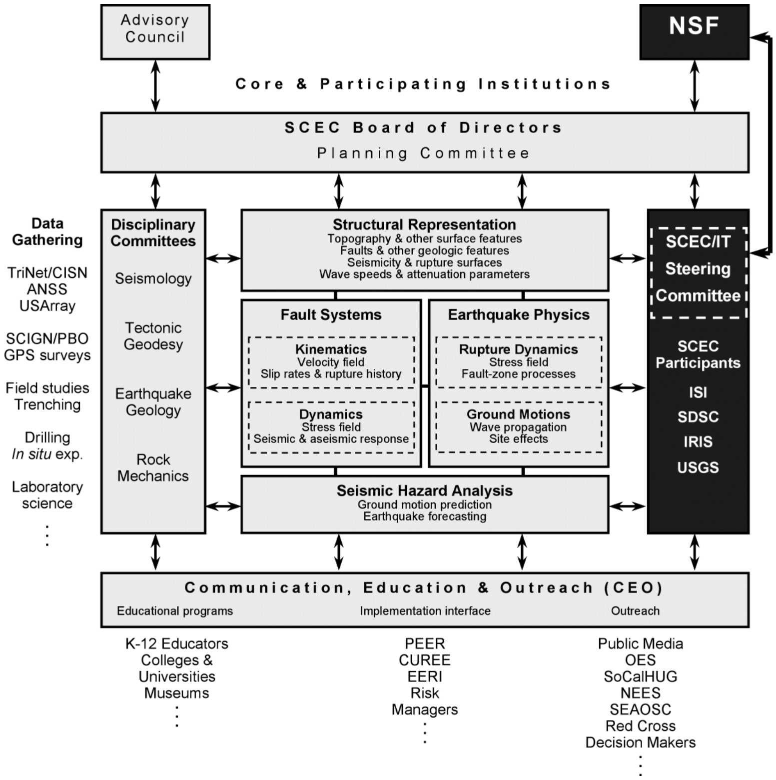

Figure 7. Diagram showing the relationship of the SCEC/IT Partnership (black box) to other activities in the SCEC matrix (gray boxes). The Partnership will be managed by a Steering Committee that will be responsible for project planning and will report directly to NSF. Within SCEC, four focus groups (central boxes) will organize the interdisciplinary research needed for the SCEC Community Modeling Environment. Knowledge transfer, education, and outreach will be coordinated through SCEC's CEO office.

The P.I. will be assisted in setting up the subcontracts and other project administrative matters (such as reporting requirements to NSF) by the Vice-Director for Administration of SCEC, John McRaney. Mr. McRaney has been with SCEC since its inception in 1991 and has a national reputation for his administrative skills. The SCEC administrative staff will assist the P.I. in organizing the tutorial workshops and the SCEC/IT annual meeting.

C.6.b. Project Management

The SCEC/IT Partnership will be managed by a SCEC/IT Steering Committee, which will oversee all aspects of this project. Members of the Steering Committee will include the P.I., acting as chair, and three Co-P.I.'s (C. Kesselman, R. Moore, and J.-B. Minster), representatives from IRIS (T. Ahern) and the USGS (N. Field), and the SCEC Information Technology Architect (to be named). The Steering Committee will be responsible for planning all Partnership activities and reporting to NSF on the results of the proposed project. The IT Architect will act as the Project Manager; s/he will coordinate project activities on a day-to-day basis and will be responsible for managing the SCEC Community Modeling Environment.

C.6.c. Project Advisory Council

We will establish a Project Advisory Council to serve as an experienced advisory body to the Steering Committee. The Advisory Council will comprise both geoscience and computer science representatives and will have a diverse membership representing basic and applied research, related technical disciplines, formal and informal education, and outreach. The Council will report to the Steering Committee and will be drawn from academia, government, and the private sector. The Council will meet once per year at the time of the annual project meeting to review programs and plans and prepare a report for the P.I. Council members will be kept informed of project activities and will be invited to participate in all appropriate functions and activities. Summaries of Council reports will be made available to NSF.

C.6.d. Team Approach

The practical implementation of individual tasks, whenever it makes sense to do so, will be accomplished through teams with representatives of both the IT and the geoscience communities. A team will minimally consist of, but not be restricted to, two researchers (graduate students or postdocs) working under the supervision of a senior staff member. Some problem will require creating ephemeral teams, others will create more persistent, perhaps larger, teams. Annual management reviews and workshops will help define and assemble the team in the most flexible and effective way.

For example, the development of distributed data collection support might require a team consisting of a seismology post-doc very familiar with the various SCEC data collection, working with one or more IT graduate students, under the guidance of an IT research staffer for a year. On the other hand, the federation of and access to inhomogeneous and distributed SCEC data collections (seismic, strong-motion, GPS, geologic) might best be handled by a smaller team working for several years.

This team-based approach to the development of the SCEC Community Model Environment will be an evolving process, wherein each community in turn might take the lead on a particular task at hand. In order for this process to converge to a satisfactory product, frequent interactions and assessments will be necessary. The collaboratories enabled by the Grid environment, punctuated by periodic face-to-face meetings and workshops are the basic tools that we propose to apply. Each of the major participating institutions is already equipped with the necessary infrastructure.

C.7. Results of Prior NSF Research

Thomas H. Jordan and J. Bernard Minster. These geophysicists have been involved in a number of NSF-sponsored projects. THJ is P.I. of NSF Grant EAR-0049042 ("Western Deep Levels Gold Mine, South Africa, as a Natural Laboratory for Studying Rock Deformation Processes"), and JBM has participated in NSF Grant ATM-9873133 ("KDI: Structure Preserving Algorithms and Model Reduction in the Natural Sciences"). However, the most relevant project for this proposal is the NSF Cooperative Agreement for funding of the Southern California Earthquake Center [EAR-8920136]. Minster is the current Science Director of SCEC, and Jordan is SCEC board member for the University of Southern California. Jordan will assume the directorship of the Center on January 1, 2002.

Since 1991, SCEC has been the primary organization in Southern California for coordinating earthquake research. Among the most significant scientific accomplishments attained by scientists within this extended collaboration are the following:

- *Seismic hazard science:* Synthesis of seismic, geologic, and geodetic data to estimate earthquake potential. Procedure for balancing seismic and tectonic moment rates, including allowance for blind thrusts and off-fault earthquakes. Recognition of hazard-estimate sensitivity to magnitude distribution and non-Poissonian recurrence.
- *Los Angeles Basin hazard:* Fundamental reformulation of tectonics and seismic hazard of the L.A. Basin, including recognition of blind thrusts and the potential for very large (magnitude 7+) earthquakes.
- *Strong ground motion:* Improved understanding of how sedimentary basins influence earthquake ground motion—focusing effects in Santa Monica, sediment nonlinearity from the Northridge earthquake, and basin-depth/edge effects—and the effects of surface deposits on ground shaking. Matching of low-frequency ground-motion amplitudes from Northridge using three-dimensional wave-propagation simulations.
- *Landers earthquake:* Joint inversion of multiple data sets to determine rupture history of a large earthquake. Demonstration of rupture propagation by dislocation pulse, rather than expanding crack. Detailed observations and physical modeling of post-seismic relaxation and the effects of fault segmentation during large-scale rupture. Use of a dynamical rupture model to satisfy observed ground motions.
- *Fault systems:* Paleoseismic fieldwork, geodetic observations, and integrative studies showing clustering of large earthquakes, prolonged earthquake interactions, large cascading ruptures, and general consistency with the historical record.
- *Deformation map:* Development of a detailed crustal deformation map, combining all available geodetic data, and use of this map to investigate tectonic loading of faults and post-earthquake response. Recognition from geodetic data of rapid strain accumulation in the eastern Ventura basin prior to the Northridge earthquake.
- *Evolution of stresses and slip deficits:* Modeling of stress evolution due to earthquakes, tectonic motions on faults, and stress relaxation. Demonstration that some earthquake sequences are consistent with triggering by stress interactions. Recognition of seismic slip deficits on faults in the Ventura and Los Angeles basins.
- *Los Angeles Basin structure from LARSE:* Discovery of a mid-crustal reflector (possible detachment zone) under the San Gabriel Mountains, apparent offset of the crust-mantle transition under the San Andreas Fault, and the displacement of the crustal root north of the topographic maximum. Revision of the depths of the San Gabriel and Los Angeles sedimentary basins.
- *Fault-zone waves:* Demonstration of the existence of waves trapped by fault-zone low-velocity waveguides and use of these waves in determining fault-zone properties and observing the fault-zone healing after earthquakes.
- *3D seismic velocity model:* Development of a 3D velocity model that includes geologic constraints, sedimentary basins, tomographic background velocities, a detailed geotechnical surface layer, and topography on the crust-mantle boundary. Demonstration that this model is consistent with independent gravity measurements.

These and other scientific results have been published in more than 500 scientific articles and special publications. The results have been synthesized into a "Master Model" of probabilistic seismic hazard in the Los Angeles region through a series of integrative reports:

- **Phase I**: Future Seismic Hazards in Southern California, Implications of the 1992 Landers Earthquake Sequence.
- **Phase II**: Seismic Hazards in Southern California: Probable Earthquakes, 1994 to 2024.
- **Phase III**: Accounting for Site Effects in Probabilistic Seismic Hazard Analyses of Southern California.

- **Phase IV**: RELM: Regional Earthquake Likelihood Models.

SCEC has organized and obtained funding for major new facilities in Southern California. The 250-station Southern California Integrated GPS Network (SCIGN) is the world's second-largest (behind Japan), making continuous, densely spaced geodetic measurements of strain accumulation and release in the L.A. Basin and surrounding regions. The Southern California Earthquake Data Center (SCEDC) is the primary data repository and distribution center for seismic networks in the region. The Portable Broadband Instrument Center (PBIC) provides high-performance seismic instrumentation for field experiments and post-seismic response.

Along with these facilities, SCEC has constructed a new infrastructure that allows researchers to share data, instruments, expertise, and effort. It has developed on-line data archives for all available seismic records, geodetic data, and satellite radar images for Southern California, and established the first on-line, web-based relational database for retrieving strong-motion data. It coordinated the field observations and science analysis following the 1992 Landers, 1994 Northridge, and 1999 Hector Mine earthquakes, and provided much of the expertise for post-event response to the damaging 1999 Turkey earthquakes. SCEC's most enduring accomplishment may be the demonstration that an effective, organized collaboration among disciplines is the best way to make progress in understanding earthquakes and communicating this understanding to others.

Carl Kesselman (CCR-8899615) (NSF-ASC 96-19020 Center for Research in Parallel Computation). Funding provided under this NSF funded science and technology center was used to conduct research in parallel programming languages and the Nexus runtime system. This work let to the development of the development of the Globus toolkit which provides mechanisms for communication, resource management, security, data access, and information in high-performance distributed environments. Globus is used extensively within the National Technology Grid being established by the two NSF PACIs as well as by many NSF-supported research projects. (NSF ASC 96-1920 National Partnership for Advanced Computational Infrastructure) Funding from this PACI center was used to deploy Globus in production environments, and develop advanced Grid applications. Kesselman is also part of the CGrADS project (NSF-EIA 9975020), which is investigating the construction of Software development environments for Grid applications. Work in GrADS has focused on the development of so called virtual organization tools for structuring Grid based execution environments for program execution. Finally, as senior personal on the GriPhyn Project (NSF ITR-0086044), Kesselman is leading research in the use of "virtual data" as a basic usage paradigm for data-intensive science.

Reagan Moore. The development of advanced systems for data, information, and knowledge management has been conducted under multiple NSF supported projects. The primary project is the NSF National Partnership for Advanced Computational Infrastructure, and supplements to the NPACI project from NARA for persistent archives and NASA for an Information Power Grid. Ongoing NSF projects include the Digital Library Initiative Phase 2, the National Science, Mathematics, Engineering, and Technology Education Digital Library, and the Grid Physics Network. Across the multiple projects, SDSC has focused on development of common software infrastructure for creating digital libraries, data grids, and persistent archives. The approach is based upon the concepts of logical representations for digital objects (global name space), logical representations for storage systems (uniform access semantics across file systems, archives, and databases), and logical representations for information repositories (automation of SQL generation, export of attributes as XML DTDs, and support for extensible schema). The resulting data and information management system is the SDSC Storage Resource Broker. The software supports collections distributed across network connected archives, file systems, and databases. Applications of the SRB include replication of the 2-Micron All Sky Survey (5 million images, 10 TBs) using containers, integration of collections into a virtual data grid as part of GriPhyN, and federation of neuroscience brain image collections. The SRB software is unique in that it provides the interoperability mechanisms needed to federate access to heterogeneous storage resources, as well as the management of digital objects that are distributed across the storage resources. A persistent archive is created by providing the ability to migrate digital objects from old storage technologies to new technologies, and collections from old databases to new databases. Reports on the capabilities of the system have been published in conferences, government workshops, and in project final reports.

REFERENCES:

---

1. A. Cornell, Engineering seismic risk analysis, *Bull. Seismol. Soc. Am.*, *58*, 1583-1606, 1968; L. Reiter, *Earthquake Hazard Analysis: Issues and Insights*, Columbia University Press, New York, 254 pp., 1990.
2. For a summary of seismic attenuation and site-response relationships, see the collection of papers introduced by N. A. Abrahamson & K. M. Shedlock, *Seismol. Res. Lett.*, *68*, 9-23, 1997.
3. E. H. Field, Modified ground-motion attenuation relationship for Southern California that accounts for detailed site classification and a basin-depth effect, *Bull. Seismol. Soc. Am.*, *90*, S209-S221, 2000.
4. E. H. Field and the SCEC Phase III Working Group, Accounting for site effects in probabilistic seismic hazard analyses of Southern California: Overview of the SCEC Phase III report, *Bull. Seismol. Soc. Am.*, *90*, S1-S31, 2000.
5. For a comprehensive discussion of the SCEC scientific framework, see http://www.scec.org/scec2.
6. See discussion of Computational Seismology on p. 15 of the President's Information Technology Advisory Committee (PITAC), *Information Technology for the 21st Century: A Bold Investment in America's Future*, White House, Washington, D.C., 35 pp., 1999 (http://www.ccic.gov/it2/initiative.pdf).
7. D. Wald, V. Quitoriano, T. Heaton, H. Kanamori, C. W. Scrivner & C. B. Worden, *Earthquake Spectra*, *15*, 537-556, 1999; http://www.trinet.org/shake.
8. *HAZUS®99 Estimated Annualized Earthquake Losses for the United States*, Federal Emergency Management Agency Report 366, Washington, D.C., September, 2000, 32 pp (http://www.fema.gov/pdf/FEMA366.pdf).
9. A. Frankel, C. Mueller, T. Barnhard, D. Perkins, E. V. Leyendecker, N. Dickman, S. Hanson, and M. Hopper, National seismic hazard maps: documentation, *U. S. Geol. Surv. Open-File Rept. 96-352*, 1996.
10. S. N. Ward, *J. Geophys. Res.*, *101*, 22,393-22,418; *Bull. Seismol. Soc. Am.*, *90*, 370-386, 2000; J. B. Rundle, W. Klein, K.F. Tiampo & S. Gross, in *Geocomplexity and the Physics of Earthquakes*, ed. J. B. Rundle, D. L. Turcotte and W. Klein, American Geophysical Union Monograph 120, Washington, DC, 127-146, 2000.
11. S. Peyrat, K.B. Olsen, and R. Madariaga, Dynamic modeling of the 1992 Landers earthquake, *J. Geophys. Res.*, in press, 2001.
12. K. Olsen, Site amplification in the Los Angeles Basin from three-dimensional modeling of ground motion, *Bull. Seismol. Soc. Am., 90*, S77-S94, 2000
13. http://www.scec.org/research/RELM/
14. Sanders, M., and P. Crotwell, FISSURES Framework for the IRIS Consortium, Incorporated Research Institutions for Seismology, Washington, D.C., January, 2001.
15. Report of SCEC/PEER Working Group on Validation of 3D Wave Propagation Codes, in preparation, 2001.

16 B.L. Humphreys and D.A Lindberg. The Unified Medical Language System project: a distributed experiment in improving access to biomedical information. *Medinfo* 1992;7(Pt 2):1496-500.

17 McCray A.T. Representing biomedical knowledge in the UMLS semantic network. In: Broering NC, editor. *High-performance medical libraries: advances in information management for the virtual era.* Westport (CT): Meckler; 1993. p. 31-44.

18 Gene Ontology (GO) Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25-29. 2000.

19 Lenat, D. B. Cyc: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM* 38, no. 11, November 1995.

20 R. Fikes, A. Farquhar; Large-Scale Repositories of Highly Expressive Reusable Knowledge; *IEEE Intelligent Systems*, Vol. 14, No. 2, March/April 1999.

21 N. Guarino. Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human and Computer Studies,* special issue on *The Role of Formal Ontology in the Information Technology,* Vol 43 no. 5/6, 1995

22 Gruber, T. R Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies* 43(5,6):907—928, 1995.

23 Moriarty, D. Determining effective military decisive points through knowledge-rich case-based reasoning. In *Proceedings of the Thirteenth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*. Springer Verlag. 2000

24 Chalupsky, H.; Gil, Y.; Knoblock, C.; Lerman, K.; Oh, J.; Pynadath, D.; Russ, T.; and Tambe, M. Electric Elves: Applying agent technology to support human organizations. In *Proceedings of the Thirteenth Innovative Applications of Artificial Intelligence Conference (IAAI-01).* Menlo Park: AAAI Press. To appear. 2001

---

25 Chalupsky, H. OntoMorph: a translation system for symbolic knowledge. In Cohn, A.; Giunchiglia, F.; and Selman, B., eds., Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR2000). San Francisco, CA: Morgan Kaufmann, 2000

26 *Information Processing—SGML Applications—Topic Navigation Maps*, ISO/IEC FCD 13250**,** 1999.

27 McGuinness, D., and Wright, J.. An industrial-strength description logic-based configurator platform. *IEEE Intelligent Systems* 13(4):69—77, 1998

28 Sabin, D., and Weigel,. Product configuration frameworks -- a survey. *IEEE Intelligent Systems* 13(4):42—49, 1998

29. I. Foster & C. Kesselman, *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, San Francisco, 677 pp., 1999.

30 I. Foster, C. Kesselman & S. Tuecke, The Anatomy of the Grid: Enabling Scalable Virtual Organizations, *International Journal of Supercomputer Applications*, to appear**.**

31 I. Foster, C. Kesselman, et al., Design and Deployment of a National-Scale Authentication, *Computer*, IEEE, 33(12):60-66 2000.

32 I. Foster, C. Kesselman, G. Tsudik, S. Tuecke, A Security Architecture for Computational Grids, *Proc. 5th ACM Conference on Computer and Communications Security Conference*, pg. 83-92, 1998.

33 S. Fitzgerald, I. Foster, C. Kesselman, G. von Laszewski, W. Smith, S. Tuecke A Directory Service for Configuring High-Performance Distributed Computation,, *Proc. 6th IEEE Symp. on High-Performance* Distributed Computing, pg. 365-375, 1997.

34 K. Czajkowski I. Foster, N. Karonis, C. Kesselman, S. Martin, W. Smith, S. Tuecke**,** A Resource Management Architecture for Metacomputing Systems,. *Proc. IPPS/SPDP '98 Workshop on Job Scheduling Strategies for Parallel Processing,*1998.

35 W. Allcock, A. Chervenak, I. Foster, C. Kesselman, C. Salisbury, S. Tuecke, The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets, *Journal of Network and Computer Applications*, to appear

36 I. Foster, A. Roy, V. Sander, L. Winkler **,** End-to-End Quality of Service for High-End Applications**,** *IEEE Journal on Selected Areas in Communications Special Issue on QoS in the Internet, 1999.*

37 I. Foster, C. Kesselman **,** The Globus Project: A Status Report**.**, *Proc. IPPS/SPDP '98 Heterogeneous Computing Workshop*, pg. 4-18, 1998.

38. S. Smallen, W. Cirne, J. Frey, F. Berman, R. Wolski, M. Su, C. Kesselman, S. Young, Combining Workstations and Supercomputers to Support Grid Applications: The Parallel Tomography Experience, *Proc. 9th Heterogeneous Computing Workshop*, to appear.

39. K. Czajkowski, I. Foster, C. Kesselman, Resource Co-allocation in Computational Grids, *Proc. of the 8th IEEE Symp. on High Performance Distributed Computing,* 1999.

40 *H*. Casanova, G. Obertelli, F. Berman and R. Wolski,The AppLeS Parameter Sweep Template: User-Level Middleware for the Grid, *Proceedings of the Super Computing Conference (SC'2000).*

41. R. Moore, C. Baru, P. Bourne, M. Ellisman, S. Karin, A. Rajasekar, S. Young, Information Based Computing, *Proceedings of the Workshop on Research Directions for the Next Generation Internet*, May, 1997.

42. R. Moore et al., D-Lib Magazine, Collection-Based Persistent Digital Archives, March 2000, Volume 6 Number 3 [Part 1], D-Lib Magazine, April 2000, Volume 6 Number 4 [Part 2].

43. C. Baru, Managing Very Large Scientific Data Collections, poster session, *5th International Conference on High Performance Computing HiPC'98*, Dec. 17-20, 1998, Chennai, India.

44. C. Baru, R. Moore, A. Rajasekar & M. Wan, The SDSC Storage Resource Broker, *Proc. CASCON'98 Conference*, Nov.30-Dec.3, 1998, Toronto, Canada.

45. A. Gupta, B. Ludäscher, M. E. Martone, Knowledge-Based Integration of Neuroscience Data Sources , 1*2th Intl. Conference on Scientific and Statistical Database Management (SSDBM)*, Berlin, Germany, IEEE Computer Society, July, 2000.

46 J. Allen, J. Hendler, and A. Tate. Readings in Planning. Morgan Kaufmann, 1990.

47 D. E. Wilkins. Practical Planning: Extending the Classical AI Planning Paradigm, Morgan Kaufmann, 1988.

48 Y. Gil. Plan Representation and Reasoning with Description Logics. In The Description Logic Handbook, F. Baader, D. McGuinness, D. Nardi, and P. Patel-Schneider (Eds). Cambridge University Press. To appear.

49 D. Fensel, V. Benjamins, E. Motta, & B. Wielinga. UPML: A framework for knowledge system reuse. In Proceedings of the International Joint Conference on AI (IJCAI-99), 1999.

50 F. Tissot & M. Gruninger, NIST Process Specification Language. Technical Report, NIST, 1999.

---

51 Smith, Becker, An ontology for constructing scheduling systems , in *AAAI Spring Symposium on Ontological Engineering*, Stanford, CA, 1997.

52 de Barros, L. N., Hendler, J., and Benjamins, V. Par-KAP: A Knowledge Acquisition Tool for Building Practical Planning Systems . *In Proc. Fifteenth International Joint Conference on Artificial Intelligence* . Nagoya, Japan, 1997.

53 Advisable Planning Systems , in *Advanced Planning Technology*, Tate, A., ed., AAAI Press, 1996.

54 J. Blythe, J. Kim, S. Ramachandran, and Y. Gil. An Integrated Environment for Knowledge Acquisition. Proceedings of the *2001 International Conference on Intelligent User Interfaces (IUI-2001)*, Santa Fe, New Mexico, January 2001.

55 Y. Gil and E. Melz. Explicit Representations of Problem-Solving Strategies to Support Knowledge Acquisition. *Proceedings of the Thirteen National Conference on Artificial Intelligence (AAAI-96)*, Portland, OR, August 4-8, 1996.

56 J. Kim, and Y. Gil. Acquiring Problem-Solving Knowledge from End Users: Putting Interdependency Models to the Test. *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-2000)*, Austin, TX, July 30-August 3, 2000.

57 A. Valente, T. Russ, R. MacGregor, and W. Swartout. Building, Using and Reusing an Ontology of Air Campaign Planning. *IEEE Intelligent Systems, special issue on Ontologies*, 14(1), January 1999.

58 J. Kim and Y. Gil. Knowledge Analysis on Process Models. To appear in *Proceedings of the International Joint Conference on AI (IJCAI-2001),* 2001.

59 A. Lansky and L. Getoor. Practical Planning in COLLAGE. *Planning and Learning: On to Real Applications: Papers from the 1994 AAAI Fall Symposium*, pp. 98-104. Y. Gil and M. Veloso (Eds.). AAAI Press, Menlo Park, California, 1994.

60 M. Stickel, R. Waldinger, M. Lowry, T. Pressburger, and Ian Unerwood. Deductive Composition from Astronomical Software Libraries. Technical Report, SRI International and NASA Ames Research Center, 1993.

61 F. Fisher, S. Chien, E. Lo & R. Greeley, Using Artificial Intelligence Planning to Automate SAR Image Processing for Scientific Data Analysis, *Proceedings of the National Conference on Innovative Applications of Artificial Intelligence (AAAI/IAAI-98)*, 1998

62 J. Runkel, W. Birmingham & A. Balkany, Solving VT by reuse, *International Journal of Human-Computer Studies*, **44**, 1996

63. EarthScope is an NSF initiative to employ new technologies for synoptic observation of the active tectonics and structure of the North American continent.(http://www.earthscope.org).