ExTENCI: Extending Science Through Enhanced National Cyberinfrastructure

University of Florida (PI)
Pittsburgh Supercomputing Center (co-PI)
University of Chicago (co-PI)
Clemson University
Louisiana State University
Purdue University
University of Wisconsin, Madison
Fermi National Accelerator Laboratory
Brookhaven National Laboratory
Florida State University (supporting partner)
Florida International University (supporting partner)

Table of contents

Project Summary

As the conduct of scientific and engineering research and the provisioning of cyberinfrastructure (CI) to support that research evolve and mature, new opportunities for leverage and technology sharing and co-development to provide increasing levels of support and services to the computationally-enabled research community are emerging. TeraGrid and Open Science Grid (OSG) have been working over the past year to identify such opportunities, and in this proposal (ExTENCI: Extending science Through Enhanced National CyberInfrastructure) we describe how we can jointly advance the support and capabilities provided to a set of key research activities that represent pathfinders in their respective communities in harnessing CI resources and services.

ExTENCI will work with several representative research applications for which a limited but sustained CI effort will significantly increase the science they deliver. These applications span a range from large collaborative science to small research groups and include earthquake engineering, biology, protein structure, physics, and environmental studies. Our team will work with specialists from these research areas to exploit enhancements we will make in four technology areas: (1) *Workflow and client tools* that permit an application to exploit either TeraGrid and OSG resources, to use both simultaneously or to utilize new resources (e.g., clouds); (2) *Distributed file systems* operating across wide area networks that simplify access to and delivery of data; (3) *Virtual machine technologies* that can hide the complexity of application environments and allow them to run in developing environments such as clouds; (4) *New job submission paradigms* that utilize distributed grid resources more efficiently.

ExTENCI will enable the Southern California Earthquake Center (SCEC) to advance their hazard prediction curves by providing extended workflow capabilities and a new distributed storage solution, to support data transfer and to allow integrated use of OSG and TeraGrid resources appropriate for each part of the workflow. The extended workflow capabilities will also permit protein 3-D structures to be determined for significantly longer amino acid sequences. The distributed storage mechanism will also significantly improve access by U.S. universities to Large Hadron Collider data and will simplify the sharing of simulated data by institutions participating in the Lattice Quantum Chromodynamics (LQCD) project. Additionally, we will provide improved Virtual Machine (VM) and job submission capabilities to extend the range of CI resources available to applications as diverse as experiments at the Relativistic Heavy Ion Collider (RHIC) and oil reservoir simulations dependent on the Ensemble Kalman Filter algorithm.

TeraGrid and OSG management strongly endorse this effort because it provides immediate benefit to both projects by strengthening and extending the capabilities of their cyberinfrastructures and, through the continuing interactions it enables between their team members, fosters opportunities for future collaborative efforts. They agree that the extra resources made available to ExTENCI, and the close integration of its deliverables with those of OSG and TeraGrid, will enhance – not distract from – their core mission of delivering cyberinfrastructure resources to their respective research communities.

Intellectual merit: ExTENCI will provide and then use cyberinfrastructure enhancements in the form of new tools in the four areas described above, building on long-term investments in their development to enable or simplify science and engineering results on a broad array of CI resources.

Broad Impact: The activities in this proposal will provide direct scientific benefits to several research areas by significantly expanding the scale of the CI resources they can exploit. This impact will extend to other communities as well. Possibly the strongest impact will result from showing the strengths of OSG and TeraGrid working more closely together to provide solutions of general utility towards a comprehensive national cyberinfrastructure. It will help in better understanding the strengths of the two projects and foster better concrete collaboration for the future, and will enable future new and/or expanded partnerships with other international, national and regional cyberinfrastructures.

Project Description

1. Introduction

1.1 ExTENCI goals

As the conduct of scientific and engineering research and the provisioning of cyberinfrastructure (CI) to support that research evolve and mature, new opportunities for leverage and technology sharing and co-development to provide increasing levels of support and services to the computationally-enabled research community are emerging. TeraGrid[1] and Open Science Grid[2] (OSG) have been working over the past year to identify such opportunities and in this proposal (ExTENCI: Extending science Through Enhanced National CyberInfrastructure) describe how we can jointly advance the support and capabilities provided to a set of key research activities that represent pathfinders in their respective communities in harnessing the CI resources and services available to them.

ExTENCI will enable the Southern California Earthquake Center (SCEC) to advance their hazard prediction curves by providing extended workflow capabilities and a new distributed storage solution, to support data transfer and to allow integrated use of OSG and TeraGrid resources appropriate for each part of the workflow. The extended workflow capabilities will also permit protein 3-D structures to be determined for significantly longer amino acid sequences. The distributed storage mechanism will also significantly improve access by U.S. universities to Large Hadron Collider data and will simplify the sharing of simulated data by institutions participating in the Lattice Quantum Chromodynamics (LQCD) project. Additionally, we will provide improved Virtual Machine (VM) and job submission capabilities to extend the range of CI resources available to applications as diverse as experiments at the Relativistic Heavy Ion Collider (RHIC) and oil reservoir simulations dependent on the Ensemble Kalman Filter algorithm.

1.2 Background and recent activities

TeraGrid and OSG operate general-purpose, distributed national cyberinfrastructures (CIs) that serve multiple scientific and engineering communities. During the past year, they jointly provided the U.S. research community with more than a billion hours of computing and several petabytes of online storage, and facilitated more than 1200 scientific publications that benefited from the computational resources and support provided. To date, however, OSG and TeraGrid operate essentially disjoint CIs and have deployed technologies and conducted operations in relative isolation from one another. The lack of common tools, the human cost required to take advantage of advanced CI and other social factors have led to most disciplines using resources from only one CI or another. Aside from the opportunity cost, the situation leads to inefficient resource use.

The projects have over time come to recognize the similarities of the disciplines we serve as well as the challenges we face. For the past two years, we have obtained our grid middleware (Condor, Globus, GridFTP, etc.) from the integrated packages maintained by the Virtual Data Toolkit[3] team. We have also interacted much more closely for the past 18 months, taking part in one another's agency reviews, participating jointly at external meetings, co-authoring white papers and, in summer 2009, developing a joint document[4] laying out a set of agreed principles for collaboration. The opening statement of that document summarizes the shared goals of our collaboration:

> *TeraGrid (TG) and Open Science Grid (OSG) jointly subscribe to the vision of a common long-term US National Cyberinfrastructure (CI) that is based on a federation of CIs. We strongly believe that establishing a coherent framework across all cyberinfrastructure programs will significantly enhance the effectiveness of our national investments. Together we plan to work on forming such a common view of a "national federation" of cyberinfrastructures, including relationships and interaction points. We also agreed that we will use the joint activities as a "laboratory" for defining, building and maintaining inter-CI interfaces. What we learn from this experience will help us guide other CIs and hopefully lead to an effective model for a national CI community. The activities described here are guided by this longer-term vision and commitment.*

Thus, one of our shared principles is a joint commitment to contribute to National Cyberinfrastructure in the US and to work together on the principles and architecture that would drive it. One specific activity has already started, with project representatives attending one another's management meetings (OSG Council meeting and TeraGrid Quarterly Meeting). Other meetings have been scheduled to overcome barriers resulting from the fundamentally different ways in which the organizations are organized in order to provide access to resources (e.g., OSG's Virtual Organizations, TeraGrid's allocations process).

1.3 Summary of ExTENCI goals

ExTENCI extends this collaboration in a focused TeraGrid – OSG effort that will significantly improve the ability of research groups to exploit advanced CI resources. Its unifying goal is to allow researchers to do better science and engineering by

- *lowering the barriers* that currently limit their effective use of CI;
- *providing tools* that expand their ability to exploit large-scale CI resources;
- *extending the range* of CI resources beyond what they can currently utilize.

We have selected several representative applications for which a limited but sustained CI effort can significantly increase the science they deliver. These applications span a range from large collaborative science to small research groups and include earthquake engineering, biology, protein structure, physics and environmental studies. The applications vary in their sensitivity to specific CI technologies, but taken together they can achieve significant gains if enhancements are made in four areas:

(1) *Workflow and client tools* that permit an application to exploit either TeraGrid and OSG resources, to use both simultaneously or to utilize new resources (e.g., clouds);

(2) *Distributed file systems* operating across wide area networks that simplify access to data;

(3) *Virtual machine* technologies that can hide the complexity of certain application environments and allow them to run in developing environments such as clouds;

(4) *New job submission paradigms* that utilize distributed grid resources more efficiently.

A modest but sustained joint effort in these areas, as we describe below, can provide significant benefits to computationally- and/or data-intensive applications.

1.4 Alignment of ExTENCI goals with OSG and TeraGrid

The ExTENCI effort will make improvements in several CI technologies and substantially increase the range of CI resources that scientific and engineering communities can utilize. TeraGrid and OSG management strongly endorse this effort because it provides immediate benefit to both projects by strengthening and extending the capabilities of their cyberinfrastructures and, through the continuing interactions it enables between their team members, fosters opportunities for future collaborative efforts. They agree that the extra resources made available to ExTENCI, and the close integration of its deliverables with those of OSG and TeraGrid, will enhance – not distract from – their core mission of delivering cyberinfrastructure resources to their respective research communities.

2. Applications to be enabled

The applications that are the focus of this proposal are described in this section.

Southern California Earthquake Center (SCEC)[5]: SCEC is a community of more than 400 scientists from over 54 research organizations that conducts geophysical research in order to develop a physics-based understanding of earthquake processes and to reduce the hazard from earthquakes in the Southern California region. One element of SCEC is the CyberSHAke project[6], which generates probabilistic seismic hazard analysis (PSHA). These are effectively sets of simulated seismograms for the response of a point on the earth to a very large set (~660,000) of potential ruptures (earthquakes). For each seismogram, a peak ground motion can be calculated. This can then be used to generate a probabilistic peak ground motion over the set of potential ruptures. These research calculations produce a probabilistic estimate of seismic hazard at a particular site. These calculations need to be repeated for each site of interest. SCEC wants to produces PHSA calculations for 2000 sites in 2010.

A site calculation requires use of both parallel and serial codes during three main processing steps: 1. Generate ~660,000 earthquake descriptions 2. Run forward wave propagation simulations for a given volume. 3. Extract strain green tensor data for the ruptures that affect a given site, generate synthetic seismograms for those ruptures, and calculate peak ground motion from the synthetics seismograms. Step 1 is done once for all sites, and is done at USC. Step 2 involves a small workflow with two 400-core parallel jobs (~16 hours run-time) and .a small number of quick single processor tasks. Step 3 is a large workflow with ~7000 single-processor tasks to extract the strain green functions, ~410,000 single-processor tasks to generate the seismograms, and ~410,000 single processor tasks to calculate the ground motions. The run-times for Step 3 jobs (in some case, bundles of jobs) range from one to sixty minutes.

Steps 2 and 3 are currently run on the TeraGrid as workflows that are submitted to standard queues. Step 3 uses Condor Glide-ins. We will continue to run Step 2 on TeraGrid, and move Step 3 to OSG. Using both CIs will allow more science to be done in the same amount of time with the same amount of human effort. This is the simplest use of both CIs; it does not require new technologies, but does require that CIs work together seamlessly, and this application will be used to ensure that they do work together for an actual science problem. Specifically, we will extract a test version of the code from the working code in the SCEC production system, so not to interfere with SCEC production runs, but will still run both workflows on the TG. We will then port the Step 3 workflow to the OSG, stage the initial large data set to OSG (generated in step 1, does not change from one PHSA curve to another), stage the smaller data that the Step 2 workflow produces to OSG (uniquely for each PHSA curve), and determine how to move the overall output from Step 3 on OSG into a test database (standing in for the SCEC production database). Other issues that will be addressed include authentication, authorization, and accounting. Once the CIs work together, we will examine ways to improve the running application, such as using Lustre-WAN, and potentially examining the use of overlay job scheduling on OSG.

Protein structure prediction pipeline: Computational procedures to predict 3D protein structure are critically important for a vast number of applications in the study of biological function[7]. We will implement a novel hybrid structure prediction pipeline that integrates two complementary prediction algorithms (the threaded RAPTOR[8] and the non-homology-based ItFix[7]) and provide them as a valuable community service which leverages the respective and complementary strengths of TeraGrid and Open Science Grid.

The new pipeline will deliver previously unattainable results for critical cases where the number of known homologous proteins is insufficient to predict structure with threading alone. It will be useful in understanding metagenomes where structures are needed, in areas ranging from health to energy production to bioremediation, and in developing drugs against pathogens such as *Staphylococcus aureus* (which have a huge impact on human health).

The RAPTOR stage of a typical prediction request will be run as a single TeraGrid MPI job, sized based on criteria of protein set, sequence lengths, and estimates of likely homologies. This stage of the overall application pipeline can be effectively executed as a pipelined workflow that involves PSIBLAST, PSIPRED, and the core RAPTOR optimization algorithm, as parallel MPI jobs.

The ItFix stage of the application is executed as a much larger set of serial jobs, with a dynamic and less predictable termination criteria, based on multiple rounds of iterative fixing. It will be sized dynamically, reporting incremental results back to the science gateway user as its independent tasks complete. Both large and small rounds can be initiated simultaneously, and the user can evaluate results as they become available and determine the further course of simulation and analysis.

While RAPTOR has been executed with excellent results on both OSG and parallel MPI resources, we will use it here to tackle larger single problems by using both resources. In addition, within the next year, ItFix will be able to use parallel MPI resources to speed up the processing of each individual Monte-Carlo simulated annealing task, by using parallel energy calculation algorithms. (The ItFix code was successfully revised in August 2009 to expedite this parallelization). We will measure and evaluate both RAPTOR and ItFix performance and effectiveness on MPI vs. serial resources and implement flexible, configurable (and if possible, dynamic) balancing between available resources on both grids. This approach will enable best use of parallel MPI vs. serial single-threaded resources for a given run and resource availability situation.

Successive enhancement of the pipeline’s efficiency and parallelism through improved utilization of the complementary resources of TeraGrid and OSG will enable continued increase of the protein sizes that the integrated prediction pipeline can handle, from the current limits of < 200 residues to what we hope will be at least a ten-fold increase in the 2-year course of this effort, to the 2000-residue range. This greatly increases the scientific value to the worldwide community of the proposed science gateway.

This application will seek to leverage the enhanced Lustre-WAN system for community-wide data sharing, the overlay job scheduling mechanism for greatly reducing request latency from the prediction science gateway, and the VM capabilities to make the installation of the entire multi-program application suite for the prediction pipeline much easier (and thereby expand the number of available grid resources and greatly reduce the manual costs of maintaining the prediction software suite).

Ensemble Kalman Filter (EnKF): Ensemble Kalman Filters (EnKF) are recursive filters that can be used to handle large, noisy data; the data are sent through the Kalman filter to obtain the true state of the data. EnKF are the kernel for a wide range of science and engineering applications, ranging from CO2 Sequestration to simple reservoir engineering problems (history matching). Specifically, we will use the EnKF with oil reservoir models. The KF stage is where the data assimilation is performed and comparison of the real measured value from the oil-field (or historical data from Well production logs) with different models. Models are modified so that they are "closer" to the real produced value. After assimilation of all the historical data, a set of models are generated that are representative of what the actual oil-reservoir looks like This approach will be extended to study CO2 Sequestration. The history of a depleted/abandoned oil field is matched to get its properties and inject CO2 based on the newly determined good reservoir description.

Using the EnKF formulation, a scientific problem is solved using “multiple models” (ensemble members). The physical models are represented as ensembles that vary in size from large MPI-style jobs to long-running single processor tasks. Varying parameters sometimes also lead to varying systems of equations and entirely new scenarios, increasing both computational and memory requirements. Each model must converge before the next stage can begin, hence dynamically load-balancing to ensure that all models complete as close to each other as possible is a desired aim. In the general case the number of jobs required varies between stages. For these applications the resource requirement is dynamic and unpredictable. It is difficult to define in advance a static scheduling strategy that will be effective throughout the execution of a complete application. The use of a combined TeraGrid-OSG CI will enable effective resource matching to application requirement, with a concomitant decrease in the time-to-solution for increasingly larger and more realistic physical models.

EnKF-based applications must be able to scale-up (scaling on a single machine) as well as scale-out (scaling to utilize multiple different machines), because the relative importance of the two types of scaling changes depending upon specific physical problem under investigation. Ensuring optimal time-to-solution requires that scaling-up be balanced with the ability to scale-out.

The ability to use both OSG and TeraGrid resources either concurrently or sequentially is important for two reasons: Firstly, long running or exceptionally large models, requiring TG-scale resources, co-exist with models that are often best placed on OSG resources. This motivates concurrent use of OSG and TeraGrid resources. Secondly, the (same) EnKF application is used to simulate a range of physical systems which induce a range of models some of which are most suitable for OSG style resources (1-4 cores), and other appropriate for TeraGrid resources (4-128 cores).[9] As different instances of the problem map better to the OSG or the TeraGrid, but are frequently part of a single overall single solution, it is important that execution be able to toggle between the OSG and TeraGrid seamlessly.

Managing the distribution and coordination of models (tasks) across disparate platforms is currently done at the application level. Because there is currently considerable fluctuation, this application is not trivially amenable to common/existing workflow tools. The EnKF currently uses multiple resources (concurrently) either directly through “spawning” functionality or with overlay job scheduling (currently in a more rudimentary and less usable stage, but conceptually in place). The short ensemble models have been shown to work effectively using VMs[10], thus a hybrid mix of underlying resources lowers the overall time to solution. This will be expanded to include TeraGrid-OSG and cloud resources provided as part of NSF’s Track 2 FutureGrid.[11]

Lattice Quantum Chromodynamics (LQCD[12]): Lattice Quantum Chromodynamics (LQCD) is the numerical study of QCD – the theory describing the dynamics of quarks and gluons. LQCD solves problems in High Energy and Nuclear Physics that are beyond the reach of traditional perturbative methods of quantum field theory. Such calculations when combined with results from accelerator experiments allow physicists to extract the fundamental parameters of the Standard Model of particle physics. Large-scale simulations are carried out using a combination of computing resources including the TeraGrid facilities, the DOE Leadership Computing Facilities (LCF) and dedicated computing systems at BNL, Jefferson Lab and Fermilab which are part of the DOE national computational infrastructure for LQCD. Simulations most efficiently exploit the combined computing resources by generating QCD vacuum gauge configurations on the very high capability computing resources available at the LCFs and the TeraGrid and then distributing the gauge configuration files to other resources for extensive analysis. The analysis step requires significantly more I/O than is required for the generation of configuration files and is better suited to high-capacity dedicated computers. The gauge configurations are typically shared among many scientific collaborations and they are reused in many different physics programs. The availability of Lustre over the wide area – across the TeraGrid and other facilities in use by the LQCD – will enable a consistent implementation for the access and transfer of data between the applications. The use case that will be explored is producing a set of LQCD gauge configurations on TeraGrid facilities, writing a copy of the gauge configurations to a Lustre-WAN file system and making the configurations available for analysis at any of the LQCD facilities in the US that share the WAN file system.

A recent evaluation of storage implementations done jointly by the Fermilab LQCD facility support and central storage groups resulted in Lustre being chosen as the implementation for the locally distributed computing infrastructure.

ATLAS & CMS LHC experiments: The Large Hadron Collider (LHC) experiments ATLAS[13] and CMS[14] will collect prodigious data samples at the world's highest energy collider to look for new particles such as the Higgs and supersymmetric particles as evidence of new fundamental forces of nature, test for the existence of extra spatial dimensions and search for the appearance of new phenomena at the energy frontier. Petabytes of LHC data will be copied to U.S. Tier-1 centers and analyzed by faculty and students at over 100 universities that are part of the US LHC collaborations, with their underlying distributed facility services and software provided by the OSG. An increasing number of faculty from these institutions are planning to cache data locally from the experiments and perform both local and remote analysis, depending on the scale of the application needs. The majority of these sites – the Tier-3s – receive no support from the central US LHC software and computing organizations, and rely on very limited effort to deploy and support their computing infrastructures, which generally lack the storage and sophisticated data management tools of the Tier-2 sites. Successful evaluation and support for the TeraGrid developments in Wide Area Lustre will lower the support costs and increase the usability of the distributed resources both locally and remotely. The experiments plan to run their existing workflows on the enhanced infrastructure possible through the production deployment of this wide area file caching system to centrally host the applications, the calibration and non-event data needed for analysis, while maintaining participation in the large scale data distribution and access services of the experiment. This ability to directly access information at Tier-2 institutions will significantly improve scientific productivity by decreasing the physicist time required to develop and test analysis codes and permitting access to subsamples without moving an entire dataset to a remote institution.

STAR nuclear physics experiment: Over the next decade, the STAR experiment[15] will carry out a broad experimental program in both heavy ion and spin physics, using the high-energy, high luminosity beams at the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Lab.   The heavy ion program is entering a phase of precision determination of the properties of the Quark Gluon Plasma (QGP), a new form of matter produced at RHIC. In spin physics, RHIC and STAR will provide the world's best constraints on the contribution of gluon polarization to the spin of the proton, exploit transverse spin phenomena to investigate transversity and orbital angular momentum, measure the polarization of sea quarks and antiquarks, and study the onset of gluon saturation.

STAR's data processing framework allows processing real and simulated data, complex simulation based on mixed events as well as user analysis. This versatile framework has been deployed at several sites around the world and was recently packaged into a Xen virtual machine for use of the Amazon.com virtual appliance using a Nimbus interface for job submission. The presence of VM capabilities on TeraGrid and OSG alike will allow STAR to not only greatly enhance its data processing opportunities – resources not specifically allocated to STAR could run a STAR image with a fully tested environment and address the last minute peak demands problem STAR has faced over the years – but also, simplify its software provisioning approach to Tier-2 centers at nearly no local software maintenance cost. Furthermore, the Software as a Service (SaaS) approach would provide the means to preserve and evolve software and allow full reproducibility of past datasets. The Lustre-WAN system will be leveraged for data sharing to and between Tier-2 centers as well as aggregating sparse storage resources over the wide network of centers. We envision such capabilities playing a central role in making a vast amount of simulated data sets easily available across our community.

Broader Impact: The activities in this proposal will benefit not only the communities involved but provide broader impact in showing the strengths of OSG and TeraGrid working more closely together to provide solutions of general utility towards a comprehensive national cyberinfrastructure. It will help in better understanding the strengths of the two projects and foster better concrete collaboration for the future, and will enable future new and/or expanded partnerships with other national and regional cyberinfrastructures.

There are other applications waiting in the wings that will be able to use the combined CIs and new technologies developed here, including a biology group at Purdue University that uses the electron cryo-microscopy, or cryo-EM, techniques to study viruses at a resolution fine enough to capture the 3-D structure of the virus capsid, or protein shell, for the first time, employing both serial and parallel computation[16]. For each virus, tens of thousands of CPUs are needed for high throughput computation; cycles available from the integration of the two CIs will significantly improve their productivity. The VM technologies developed and deployed in this proposal will improve user experience in many ways, e.g., expanding the choices of resources for long running applications that do not checkpoint (for instance, the GAUSSIAN computational chemistry program).

3. Summary of OSG and TeraGrid

3.1 Open Science Grid

The Open Science Grid mission is to promote discovery and collaboration in data-intensive research by providing a computing facility and services that integrate distributed, reliable and shared resources to support computation at all scales.

Open Science Grid:

- Is a Consortium of software, service and resource providers and researchers, from universities, national laboratories and computing centers across the U.S., who together operate a national, distributed cyberinfrastructure with funding provided by NSF and DOE augmented by Consortium members. OSG Consortium members' independently owned and managed resources make up the distributed facility, agreements between them provide the glue for it, their requirements drive its evolution, and they contribute their effort to make it happen.
- Offers participating research communities low-threshold access to more resources than they could afford individually, via dedicated, scheduled and opportunistic alternatives.
- Provides common middleware provided by the Virtual Data Toolkit - packaged, tested and supported collections of software for installation on participating compute and storage nodes and a client package for end-user researchers. Individual research communities, the “virtual organizations”, add services according to their scientists' needs.
- Works with research communities to help them evaluate their cyberinfrastructure needs and plan their solutions both locally across the campus and as part of national or international efforts. OSG works jointly with partners to create worldwide interoperable systems, for example the World Wide LHC Computing Grid for the upcoming CERN LHC experiments.

- Provides training through hands-on workshops and focused engagement with the community, helping new users to run applications on the infrastructure and resource

3.2 TeraGrid

The TeraGrid is an advanced, nationally distributed, open, user-driven cyberinfrastructure that enables and supports leading-edge scientific discovery and promotes science and technology education. It is comprised of supercomputing, storage, visualization systems, data collections, and science gateways, integrated by software services and high bandwidth networks, coordinated through common policies and operations, and supported by computing and technology experts.

TeraGrid's three-part mission is to support the most advanced computational science in multiple domains, to empower new communities of users, and to provide resources and services that can be extended to a broader cyberinfrastructure. Accomplishing this vision is crucial for the advancement of many areas of scientific discovery, ensuring US scientific leadership, and increasingly, for addressing critical societal issues. TeraGrid achieves its purpose and fulfills its mission through a three-pronged focus:

- **deep**: ensure profound impact for the most experienced users, through provision of the most powerful computational resources and advanced computational expertise;
- **wide**: enable scientific discovery by broader and more diverse communities of researchers and educators who can leverage TeraGrid's high-end resources, portals and science gateways; and
- **open**: facilitate simple integration with the broader cyberinfrastructure through the use of open interfaces, partnerships with other grids, and collaborations with other science research groups delivering and supporting open cyberinfrastructure facilities.

TeraGrid is coordinated through the Grid Infrastructure Group (GIG) at the University of Chicago, working in partnership with eleven Resource Provider sites: Indiana University, the Louisiana Optical Network Initiative, National Center for Supercomputing Applications, the National Institute for Computational Sciences, Oak Ridge National Laboratory, Pittsburgh Supercomputing Center, Purdue University, San Diego Supercomputer Center, Texas Advanced Computing Center, and University of Chicago/Argonne National Laboratory, and the National Center for Atmospheric Research.

4. Major work activities

This section summarizes the major activities. Each of them has its key ideas, personnel, work plans and utilization by applications woven into their narrative.

4.1 Lustre across a wide area network

The goal of deploying distributed Lustre file systems for use across the wide area network is to evaluate the performance, robustness, and capabilities of a generally available "global wide area file system" as an integrating service across TeraGrid and OSG sites for access to software, catalogs, and processing data. Evaluation of such a service is already part of the program of work for TeraGrid for its extension phase from March 2010 through July 2011. The work in EnCITE will leverage and extend this work to testing in the OSG environment to support both existing OSG applications and applications and workflows spanning TeraGrid and OSG resources. The effort to deploy Lustre technologies and support will be led by J. Ray Scott from PSC, with the infrastructure testing being centered at the University of Florida and the initial application integration and testing being at Fermilab and the University of Chicago.

TeraGrid will decide on the particular implementation of Lustre across a wide area network to pursue within the project. The EnCITE work will track this decision and work with TeraGrid in adopting the same solution.

Both US LHC collaborations will benefit from a robust, wide area file system with good performance in reducing the effort needed to support, maintain and use the up to 100 Tier-3 sites in the US, enabling hosting of data, applications and other information at larger (Tier-2) data serving sites. The University of Florida and other local institutions will contribute to the initial deployments and tests as the software and security components of Lustre over the wide area are released by PSC. Fermilab will subsequently perform system integration and performance tests of the Lattice QCD, CMS and ATLAS applications. For CMS the CRAB server, ProdAgent and CMSSW distributions will all be adapted and integrated to the use of Lustre over the wide area as the data storage system. We will work with the local CMS framework development group to configure and tune the “event data block” streaming from the event store to the job execution site in a Lustre-WAN environment. We will work with US ATLAS to test the PACMAN software distribution and integrate with the PANDA framework.

Providing production-quality wide area Lustre is a large undertaking. The Lustre development group at SUN has undertaken to remedy some of the security and performance shortcomings of wide area Lustre. For instance, they are adopting Kerberos authentication as a way to secure servers and validate client access. We find, however, that for data to be shared by people operating on different systems at different sites, each with its own authentication procedure, there are mundane but annoying problems associated with allowing user IDs from the various systems to be authenticated to Lustre servers in a way transparent to the users. PSC is tackling parts of this problem in order to push a secure Lustre file system into the TeraGrid as a wide-area filesystem solution.

There are six work packages for the Lustre-WAN software development and deployment project. They include 1) kerberized client and server deployment, including a test environment; 2) an environment for user identity management; 3) network performance tuning; 4) tests with a public database of results; 5) adaptation, integration, testing and deployment of Lustre WAN with the existing application services and environment. The sixth work package is ongoing project management. The components in these packages are required to ensure a production quality service.

Server Deployment involves getting a core set of Lustre servers (hardware and software), capable of being a wide area file system, up and running at UF. Effort is required to understand the application environment and then to design a file system solution best suited for those applications.

Specific work items are described below:

4.1.1 Security Infrastructure

The first work package (WP1, 7 person months) will conduct the development and deployment of a Kerberos based security infrastructure for Lustre-WAN. A test server environment will be installed at PSC. Many of the services will be run on one server using virtualization, which allows us to simulate multiple server domains and network latencies. PSC has found a very close fit between its virtualized test bed and the production deployment.

Wewill also design and setup the Kerberos infrastructure at UF. As the core site for this project, the authentication practices in production at UF will be key to a smooth roll out of the subsequent phases. We will incorporate Lustre clients that are Kerberos enabled into the file system. Having all servers authenticate to the project's Kerberos realm assures that the servers remain secure, precluding the ability to bring up rogue servers and pretending to be authorized participants of the core filesystem services. We will also create Lustre client installation packages for the Tier 3 sites. These packages will allow staff at those sites to easily install the necessary software for Lustre and Kerberos and will not require any special systems administration expertise at the Tier 3 sites beyond very basic systems administration. PSC has been creating packages like this for the TeraGrid testbed and has had success in helping client sites get on the Lustre WAN. This effort to simplify OSG Lustre-WAN deployments will also benefit the TeraGrid’s expansion of Lustre-WAN beyond testbed sites.

Once the client packages are ready for deployment, clients will be installed at UF, FSU, FIU, PSC, and potentially other partner sites. Assistance for Lustre and WAN Lustre system administration process will also be made available to clients sites. Application adaptation and testing using both the server and client technologies will be done at Fermilab once the initial software infrastructure is ready for this level of deployment.

During this phase, the interoperability of the Kerberos operational environment will be designed and deployed. With this interoperability, OSG application data will start to be able to be shared between the UF Tier 2 site and the test Tier 3 sites. This will give the TeraGrid a good reference point for a wide area file system in a project.

4.1.2 Environment for User Identity Management

The next work package (WP2, 1 person month) will create the software and administrative environment for user identity management in the Lustre-WAN environment. This effort will include user Kerberos principal creation and management, integration with grid software (e.g., Globus), integration with local site accounts, and id mapping across sites. PSC will apply the tools it has been working for streamlining the administration of user identities in a WAN environment. The output will be evaluated by the application and administration contacts for ease of use and integration with the other registration management tools used within the OSG and TeraGrid

4.1.3 Network Path Analysis and Tests

The third work package (WP3, 1 person month) will start in the second month of the client installation phase of WP1. It will involve PSC's expertise in wide area network performance tuning. The network paths between all the partner sites will be analyzed. Network diagnostic and tuning software will be installed at the endpoints of the file system. The Lustre servers and client systems will be tuned as needed for optimal network performance.

Work package four (WP4, 1 person month) will start in the third month of the client installation phase of WP1. It will install and run file systems tests on the Lustre WAN. Fault monitoring and recovery tools will also be deployed in this WP. PSC has been active in network data movement performance testing since the beginning of the TeraGrid. (PSC created the speedpage[17], an automated testing environment for tracking the bi-lateral performance of files moved from one site to another. All results of these tests across time are stored in a database, which is made available to the community for analysis. This technique has recently been expanded to start testing the TeraGrid Lustre WAN. The WANpage[18] show the results of tests run between the TG Lustre WAN sites.) In addition to running test with synthetic loads, application cores from the OSG LHC experiments will be used to determine the performance of the WAN file system. This will allow tuning the file system and associated components, such as the network, for the types of loads that will typical in production.

4.1.4 Application Integration

Work package five (WP5, 1 person month) will start after the file systems test work package (WP4). It will look at other software tools that will be needed to integrate the WAN Lustre file system with the initial applications and work with the application effort to adapt the applications themselves. This will be the point at which the LQCD and CMS application groups will start their program of work.

Tools for network congestion control, end-to-end data verification, quota management, etc. will be evaluated in the OSG application environment. We will work with the OSG software and operations teams on how best to package, distribute and configure these where needed.

While not strictly a part of this project, the team will look at replication services as a tool for spreading data to more than one site. This can be useful for backup as well as higher performance file serving, leading to being able to read different parts of the same file from different sites, hence striping the data transfer across multiple network paths.

The information gathered as a part of both work packages 3, 4, and 5 will provide valuable guidance to the TeraGrid to help that project understand the data requirements and performance metrics of data intensive projects.

4.1.5 Application Performance Testing, Analysis and Reporting

Work package six (WP6) is an on-going effort for project management, documentation, site and application support, and monitoring. Documentation will be provided to tier 2 and tier 3 site administrators to help with any on-going problems that may arise. There will also be a description provided for the operation of the monitoring tools and suggestions for remediation.

CMS will adapt and test its applications in parallel with the work packages on each release of the WAN system. This will be done in collaboration with LQCD extending existing evaluation requirements and criteria. The groups will build on the existing expertise and testing infrastructure to evaluate the performance, robustness and capabilities of WAN Lustre, and to provide an independent assessment of the technologies provided for the LQCD and US CMS application services. This will include: data integrity, security and accessibility; usability, maintainability, ability to troubleshoot and isolate problems; namespace and its performance.

4.1.6 Metrics and Assessment

Within this program of work we will assess and document performance, technology, operations and maintenance requirements for Lustre WAN to be acceptable as a production storage element for US LHC physics. The reports will be reviewed by the collaboration as well as OSG software and production management. We expect this to be ongoing during the lifetime of the project, with iterations between the software development teams at PSC, the first level testers at U Florida and the application integration teams at Fermilab and elsewhere. We will compare, where possible, between observed performance and LHC target goals.[19] We will also compare Lustre results with a possible Hadoop20 implementation.

This solution will then be tested with the SCEC and Protein Structure applications.

4.2 Virtual Machine technologies and clouds

Cloud computing is a rapidly emerging computing paradigm providing computing infrastructure as a service. In high performance computing, clouds are being used by applications with extremely complex software environments (such as the STAR experiment) to provision dynamic clusters that encapsulate its certified codes. Additionally, applications dependent on the operating system, such as a forestry simulation using the code SEARCH that runs on Windows,[21] have also used cloud computing by provisioning machines from a pre-configured virtual machine pool.

Virtualization is one of the key aspects in cloud computing. The last decade has seen extensive research and practices in virtualization of servers, storage and networks. For the scientific computing communities, virtualization provides several key advantages, including ability to package complete application environment that guarantee execution of the code on multiple systems, isolation from the physical layer which provides more flexibility to the system administrators, and advent of new computing paradigms where resource providers become infrastructure providers onto which communities build up their distributed systems.

For this project, we propose to develop and deploy virtualization technologies at TeraGrid and OSG sites to support a number of scientific applications. This work is driven by large projects, including the STAR experiment and CMS (LHC), and TeraGrid applications of individual research groups, including the simulation of the entire human body arterial tree[22,23] to utilize cloud resources across the two CIs. Several groups in OSG and TeraGrid have already prototyped virtualization-based systems and this work will leverage these prototypes to deliver several key results that can help plan a future architecture and computing model of the national cyberinfrastructure. This proposed work focuses on enabling and supporting scientific applications that can benefit significantly from virtualization. The effort will be led by Carol Song at Purdue University and Sebastien Goasguen at Clemson University.

The specific aims are discussed further below.

4.2.1 Virtual machine configuration and deployment at OSG and TeraGrid sites

We propose to deploy VO configured virtual machine at two grid sites, Purdue University and Clemson University. Purdue is a resource provider to both the TeraGrid and OSG and Clemson is a resource provider to OSG.

STAR’s data processing framework has recently been packaged into a Xen virtual machine. The experiment has demonstrated running on Amazon’s EC2 and on the Science Clouds running the Nimbus software.[24] The STAR application is currently running at Clemson University with its own application environment packaged as a virtual machine. We will expand on this prototype and deploy the STAR VM at Purdue to provide a STAR cloud that interoperates across TG and OSG.

Leveraging the CMS Tier-2 effort at Purdue, we will also develop a CMS VM and deploy these VMs at Clemson. We will ensure good coordination with other ongoing CMS efforts in this area. In this prototype Purdue will provision CMS worker nodes on Clemson’s Public cloud. This work demonstrates how a grid site can gain additional resources on the cloud.

To further expand the above efforts, we propose to deploy hypervisors on a large scale and in full interoperability with traditional cluster operations. Several sites (e.g Purdue, Clemson, U of Wisconsin, Madison, UC, FermiLab) have deployed on the order of tens to several hundred hypervisors but nothing close to the overall size of the clusters that are on TG or OSG. We will demonstrate how this can be done by planning and executing the deployment of a large cluster with hypervisors. Clemson is currently working on this using the KVM system and the Palmetto Cluster (#62 in the world) and using Virtual Box on its Windows based campus grids. Purdue will deploy thousands of hypervisors on the campus teaching lab systems and research clusters. This work will enable applications, including STAR and a number of TeraGrid applications, to run at a much large scale than what they are able to do currently, potentially enabling additional resource providers to join the TG/OSG interoperation cloud. We also look to international partners to further expand the collaboration. For instance, INFN and CERN in Europe are currently planning to manage all their compute nodes using virtualization. Goasguen at Clemson has spent the summer at CERN working in the CERN cloud computing prototype which aims at supporting 16,000 VMs. Currently no site in the US is operating or even planning to do so except corporate data centers.

4.2.2 Virtualization to reduce Tier-3 sites support cost

Most of the intellectual capital of the nation is located on campuses of universities. Reaching out to these campuses and building bridges so that their faculty can easily use the national CI is a strategic aspect of enabling science and education all across the US. However, too few campuses have the resources necessary to devote time to building these CI bridges. Virtualization and clouds are one potential answer to this problem, providing pre-packaged middleware systems that can be deployed on campuses border to connect everyone to the national CI. To drive this effort of large potential, we propose to focus on the OSG Tier-3 sites and campus labs. We will gather requirements from and collaborate with the existing OSG and US CMS Tier-3 support organizations.

As stated in Section 2, most of the CMS Tier-3 sites have very limited funding for cyberinfrastructure support. To help CMS Tier-3 sites to connect to the CMS data and application environment, this project will provide pre-configured CMS virtual machines so that a Tier-3 site can access the CMS data and run analysis and other applications without dedicated IT support personnel since VM deployment is being supported by most IT departments today. Two types of CMS VM appliances will be created: a CMS user appliance and a CMS Tier-3 compute appliance. The CMS user appliance is a virtual machine that allows a scientist to connect to CMS resources (data and tools) from his or her own computer (e.g., one that runs Windows). The CMS T3 compute appliance will enable a Tier-3 site to join a cloud resource made available through an OSG Compute Element hosted and managed at Purdue, essentially supporting the expansion of computing resources with minimal IT admin support from the local sites.

To help campus grids and CMS Tier-3 sites to easily manage CI resources, we will also demonstrate how virtualization can help transform Windows resources in useful scientific computing engines by deploying VMware and/or Virtual Box on Windows resources at Purdue and Clemson. This work is already under way and Purdue has demonstrated how VMWare based VMs can be run on Windows lab machines and join a cloud using virtual networks. We will expand on this to build a Windows based cloud across TeraGrid (Purdue) and OSG (Clemson).

4.2.3 Extending access to larger resources through virtualization

Purdue currently runs an experimental cloud resource *Wispy*. Utilizing the Nimbus virtualization software[25,26]. Wispy allows users to submit a virtual machine image to the resources in the cloud in much the same way as submitting jobs into a grid. The user may interact with the virtual machine as it is executing on some hardware in the cloud. With a similar interface to Amazon's EC2, users can compute the same way they do in a commercial cloud. Currently as a TeraGrid resource, Wispy supports research and development for operating system and grid services, as well as support complex batch and interactive jobs that require very specific dependencies and complicated configurations.

Leveraging the work on Wispy, we propose to extend the Nimbus cloud computing solution with "pilot" codes to enable a resource provider to use large Linux clusters and their batch systems as a cloud infrastructure. Both PBS clusters and large Condor pools (using Condor's Virtual Machine Universe) will be connected to Nimbus resources with the developed pilot scheduling software. This would allow other

TeraGrid sites to potentially join the cloud. This work will provide a much larger pool of resources accessible to such applications as STAR and CMS experiments, potentially improving the productivity of research communities. We will also create and test a TeraGrid VM appliance to run on a non-TeraGrid system at Clemson to investigate the potential of smaller sites joining the TeraGrid. The build and deployment process will be documented and shared with TeraGrid and OSG resource providers. A concrete test will include the STAR experiment using the Nimbus EC2-like interface to submit the VO configured VM to run on this cloud. We will coordinate with US-CMS Monte Carlo production coordinator to use the Nimbus EC2-like capability to provide additional resources for time-critical data simulation.

4.3 Overlay job scheduling

The overlay job-scheduling mechanism includes two concepts: submission of a "pilot" job – which is basically a container job without a specific workload – and the dynamic assignment of tasks to the container job once it is active. Conceptually, the overlay job-scheduling abstraction enables the separation of job-scheduling from resource provisioning, and by supporting the delayed binding of a workload task to a specific resource, scheduling overlays provide a flexible approach to resource assignment and task orchestration. Specific implementations of the overlay job-scheduling approach are the well known Glide-in (Condor project) and Pilot-Job (PANDA/ ATLAS) and the more rudimentary SAGA-based BigJob (TeraGrid). The effort will be led by Shantenu Jha at Louisiana State University and Miron Livny at the University of Wisconsin, Madison.

A number of applications utilize overlay scheduling on the TG and OSG separately, as they move beyond their native Grid environments -- either due to Scaling-up of problem instance size (OSG to TG), or Scaling-Out (TG and OSG concurrently) due to problem instance formulation, providing (and supporting) the same abstractions on the extensible infrastructure is critical, whether or not they use this concurrently.

An important difference between the HPC and HTC domains is the nature of the Overlay Job. In the Condor worldview, the Glide-In (Condor's Overlay Job Scheduling mechanism) is just a Condor job that schedules further jobs. In the Overlay Job abstraction developed for large MPI jobs on the TeraGrid (using SAGA-based BigJob), the Overlay Job is itself a MPI job, which then "massages" the specific node-files for the specific scheduler(s).[27] The SAGA-based scheduling overlay is general-purpose being user-level code currently used on the TeraGrid.[28]

The aim of this work item is to extend current existing scheduling overlay approaches to make them (i) general purpose overlay job-scheduling tools on the TeraGrid, (ii) integrate with overlay approaches on the OSG, and, (iii) to provide the conceptual framework for overlay scheduling on the TeraGrid for a range of VOs/Applications and harmonize its usage across the TG and OSG to support general-purpose High-Throughput Parallel Computing (HTPC).

There are several reasons why an application should be able to utilize high-end (TeraGrid) and commodity cluster machines (OSG) concurrently. For example, in the EnKF-based application, there are many instances where for a given stage, the individual ensemble members have a distribution of sizes, e.g., starting from small 1-4 processor simulations all the way up to simulations requiring 32-64 cores. There can be O(100) such ensemble members running at a given instant of time. For largest system sizes attempted, the number of ensemble members can approach approximately 1000. Not surprisingly there are more tasks to finish than can be supported at any given instant of time on any one TeraGrid machine, thus there is a need to (i) utilize multiple TG machines, and (ii) to utilize Overlay scheduling on multiple TeraGrid machines concurrently.

Clearly a smart way of assigning resources is required -- first at the level of HPC vs HTC, and then secondly, coordinating the many tasks that go to the HPC resources. It is desirable to utilize multiple machines, which although very simple (by design) on the HTC resources, is still a major challenge on the TG, i.e., limited Scale-Out. The SAGA Based Overlay Job mechanism has shown how many large MPI (sub-) jobs up to 64 cores each, have been run within a 256 processor Overlay Job on Ranger (65K); each job runs for up to 2 hours. Not surprisingly, up to 500,000 shorter running jobs have been run using Glide-In. The challenge is to merge the two capabilities into one: the large number of jobs that Glide-in supports with the large size of jobs that SAGA-based Overlay Jobs provide. This capability will be used by multiple applications, including but not limited to EnKF-based applications (History Matching, CO2 sequestration).

4.4 Education and Outreach

The ExTENCI project staff will participate in and contribute to existing education and training programs of OSG and TeraGrid. The project will be able to leverage the infrastructure, logistics and events of the existing EOT efforts. OSG and TeraGrid are already working to increase collaboration in training and workforce development, and ExTENCI's efforts can be most effective in adding to these existing efforts. The EOT efforts will be coordinated by the project management support staff.

5. Major Milestones

Task	**+6 months**	**+1 year**	**+18 months**	**+24 months**
SCEC application	Application running on TG and OSG	Application running using Lustre-WAN	Application running using pilot jobs on OSG	Task completed – no milestone this period
Protein Folding application	Initial pipeline; add 2D feedback. Enhanced ItFix energy calculator. 500-aa capacity. CASP8 tests. Enhance result integration. Initial science gateway.	No milestone this period	Performance and accuracy analysis. Result storage database. Q/A checks on predictions. 1000-aa capacity. Validate w/CASP9 results. Process *S. aureus*.	Calibration/interaction tests. 3000-aa capacity. MSA handling. Domain-domain association ("4D structure"). Metagenome processing. Enhanced science gateway.
Ensemble Kalman Filter application	Applications using TG and OSG resource concurrently	Applications using integrated Overlay Job Scheduling, tools and mechanisms on TG and OSG	Scaling-out to support ~10K ensemble members per stage, and ~100 stages per physical problem (i.e. manage 10^6 jobs/tasks across TG & OSG)	Scale-up problem sizes, sophisticated extensions to EnKF to include optimization (thus requiring more ensembles per stage) and reduce noise and statistical sampling errors.
Lattice QCD application	Receive and install deployment release of Lustre WAN server and client tools	Iteratively test Lustre WAN service and clients against HPC storage system requirements.	Develop a production deployment and operation plan for LQCD and make initial production deployment.	Operate initial Lustre WAN deployment. Evaluate system performance, reliability and administration effort required for steady-state operations.
ATLAS & CMS applications	Set up CMS framework and data samples at test sites. Conduct initial performance tests with campus-wide Lustre at UF.	Adapt CMS CMSSW distribution, CMS framework, Phedex, CRAB to interface to initial releases of Lustre WAN.	Performance, reliability and operational testing, with release of integrated software "ready for production".	Develop a production deployment and operation plans for ATLAS and CMS and make initial production deployment.
STAR applications	Be able to submit jobs to TG and OSG using VM at TG/Purdue and one more OSG site.	Deploy & test initial Lustre WAN release on ≥1 Tier2 and interface with job running on TG &OSG. Test concurrent OSG &TG job scheduling.	Consolidation of VM, Lustre approach within TG / OSG. Scalability and regression testing. Prepare plan for integration into production operation.	Phase in deployment to other sites.

Lustre-WAN	Test Environment installed at PSC. Kerberos infrastructure installed at UF, USF, FIU and PSC. Client packages deployed at UF, USF, FIU, and PSC	Client packages deployed at remaining partner sites. Kerberos principal creation & mgmnt procedures in place. Network performance test tools & procedures in place at partner sites	CMS and LQCD application testing complete. Network and file system monitoring and testing packages in place at all sites.	Widespread application adoption.
Virtual Machine technology	STAR application running on VM cloud across Purdue (TS) and Clemson (OSG).	Extend Purdue CMS Tier2 nodes on Clemson cloud (CMS VMs running at Clemson).	CMS User VM Appliance available; Distribute CMS T3 compute appliance to Tier 3 sites. TeraGrid VM appliance available and tested at Clemson.	Completion of Nimbus pilot code to provide EC2 cloud over TG and OSG resources. STAR and TG virus 3D reconstruction application running at large scale.
Overlay Job submission	Prototype of Integrated Overlay scheduling	Integration with EnKF-based applications and others	Deployment, testing on joint infrastructure	Scale-Out and stress tests. Scale number of jobs to support largest production grade simulations.

6. Management plan

The management of these activities will follow the principles given in the Joint Statement of Agreed Upon Principles between OSG and TeraGrid.[4] "*Brief status reports will be provided to management of both team every six weeks. Effort will be given to align with established report schedules. OSG and TG management will convene to discuss progress before TG prepares its quarterly reports.*"

The project PIs will be responsible to NSF for the execution and coordination of all project activities. The ExTENCI project management will include regular interactions with the existing management of OSG (Executive Team) and TeraGrid (Forum) to assure continued alignment of the ExTENCI project activities with the activities of OSG and TeraGrid. The ExTENCI Executive Board (EEB) will consist of the PIs and the lead person of each task. The EEB will meet by phone bimonthly and face-to-face twice a year co-located with the OSG all hands and TeraGrid annual conference. These meetings will be used as internal reviews of progress on all tasks every six months, specifically reviewing the milestones listed in Section 5. As part of the review process, resource allocations might be reconsidered. The EEB will facilitate interactions among the applications and provide a common view of technology integration.

To provide close oversight and alignment with the activities of TeraGrid and OSG, the parent projects will augment the coordination effort funded by this proposal (0.20 FTE). They will help coordinate ExTENCI planning, execution, metrics, lessons learned and status reports. The team will have members from PSC (0.25 FTE, Janet Brown, effort coordination), University of Chicago (0.10 FTE, Jeff Koerner, lessons learned) and OSG (0.1 FTE, James Weichel, metrics), a total of 0.65 FTE including the ExTENCI-funded person (James Weichel, project coordination).

Task	**Task Lead (other participants)**
Project oversight	Paul Avery, Ralph Roskies, Dan Katz
Reports, coordination, metrics, lessons learned	James Weichel (Jeff Koerner, Janet Brown)
Applications	Dimitri Bourilkov (Dan Katz, Ruth Pordes)
Lustre-WAN	J. Ray Scott (Yujun Wu)
Virtual Machine technology	Carol Song (Sebastien Goasguen)
Overlay Job submission	Shantenu Jha (Miron Livny)

The ExTENCI projects will make use of wikis (with pointers to and from the existing OSG and TeraGrid wikis) to publish their plans, progress and discussions. We will jointly plan to submit papers to report on the work for publication in a refereed journal annually.

Any software developed will be publicly available under an open source license agreed upon by OSG and TeraGrid management. On agreement with OSG the software will be made available for inclusion in the Virtual Data Toolkit. On agreement with the TeraGrid, the software will be made available as “capability kits.” In accordance with NSF practice, the software will be built on the NMI Build and Test under the Metronome system.

Finally, the project will be subject to the security oversight and operational security of the OSG and TeraGrid projects, leveraging the expertise of the Security Officers from the two infrastructures.

7. Co-PI’s background and Previous NSF Awards

Paul Avery (U. Florida) is Professor of Physics and PI and Director of iVDGL Project (NSF PHY-122557, $14.4M). Responsible for the overall mission for iVDGL in developing a national Grid laboratory. P.I. and Director of GriPhyN (NSF ITR-0086044, $12M). Co-P.I. of UltraLight (NSF ITR-0427110, $2M) developing an advanced cyberinfrastructure integrating computing, storage and networking and deploying it for flagship applications. Co-PI of DISUN (NSF-05333280, $10M) which is developing a national cyberinfrastructure for the CMS experiment at the LHC. Co-PI of Open Science Grid (NSF-0621704, $30M), which operates a general purpose, national CI serving multiple disciplines.

Ralph Roskies is Professor of Physics at the University of Pittsburgh and a founder and Co-Scientific Director of the Pittsburgh Supercomputing Center (PSC), where he oversees operations, plans its future course, and concerns himself with its scientific impact. PSC has pioneered developments in file systems, heterogeneous computing, parallel algorithms, and scientific visualization and is renowned for outstanding user support. R. Roskies: TeraGrid Resource Partners, NSF, SCI 04-56541, $52M, 8/05-3/10; ETF Grid Infrastructure Group: Providing System Management and Integration for the TeraGrid, University of Chicago (NSF), SCI 05-03697, $4.7K, 8/05-3/10; Terascale Computing System, ACI 03-07136, NSF/ACIR, $63M, 10/00-9/05.

Daniel S. Katz is a Senior Computational Research Scientist at the University of Chicago and at Argonne National Laboratory, and an Affiliate Faculty in the Center for Computation & Technology (CCT) at LSU and an Adjunct Associate Professor in Electrical and Computer Engineering at LSU. He is the TeraGrid GIG Director of Science, using 20+ years of computational science experience to encourage and promote science applications that use multiple TeraGrid sites, and to act as an advocate for TeraGrid users. He has been Co-PI and technical lead of the HPCOPS award (NSF OCI-0710874, $2.5M, 10/07-3/10 HPCOPS: The LONI Grid – Leveraging HPC Resources of the Louisiana Optical Network Initiative for Science and Engineering Research and Education) that brought LONI into the TeraGrid.

References

[1] TeraGrid home page, http://www.teragrid.org/.

[2] Open Science Grid homepage, http://www.opensciencegrid.org/.

[3] Virtual Data Toolkit home page, http://vdt.cs.wisc.edu/.

[4] "Joint Statement of Agreed Upon Principles between OSG and TeraGrid", http://osg-docdb.opensciencegrid.org/cgi-bin/RetrieveFile?docid=882&extension=pdf

[5] Maechling, P., Gupta, V., Gupta, N., Field, E. H., Okaya, D., Jordan, T. H., "Grid Computing in the SCEC Community Modeling Environment", *Seismological Research Letters*, v. 76, pp. 581-587, 2005.

[6] Maechling, P., Deelman, E., Zhao, L., Graves, R., Mehta G., Gupta, N., Mehringer, J., Kesselman, C., Callaghan. S., Okaya, D., Francoeur, H., Gupta, V., Cui, Y., Vahi, K., Jordan, T., Field, E., "SCEC CyberShake Workflows—Automating Probabilistic Seismic Hazard Analysis Calculations", in Workflows for e-Science, pp. 143-163, Springer London, 2007.

[7] DeBartolo, J., Colubri, A., Jha, A. K., Fitzgerald, J. E., Freed, K. F., Sosnick, T.R., "Mimicking the Folding Pathway to Improve Homology-free Protein Structure Prediction," *Proc. Natl. Acad. Sci. U S A*, v. 106(10), pp. 3734-3739, 2009.

[8] Jinbo Xu, Ying Xu, Dongsup Kim and Ming Li, "RAPTOR: Optimal Protein Threading by Linear Programming," Journal of Bioinformatics and Computational Biology, April 2003.

[9] Y el-Khamra and S Jha, "Investigating Autonomic Behaviours in Grid-Based Computational Science Applications," Grids Meet Autonomics (GMAC09), http://www.cct.lsu.edu/~sjha/select_publications/lazarus_gmac09.pdf.

[10] H Kim, Y el-Khamra, S. Jha and Manish Parashar, "Autonomic Approach to Integrated HPC Grid and Cloud Usage", accepted for IEEE eScience 2009, Oxford, Dec 2009.

[11] FutureGrid home page, http://www.futuregrid.org

[12] LQCD home page, http://www.usqcd.org/.

[13] ATLAS collaboration home page, http://atlas.ch/.

[14] CMS collaboration home page, http://cms.cern.ch/

[15] STAR experiment home page, http://www.star.bnl.gov/.

[16] Jiang, W., Baker, M. L., Jakana, J., Weigele, P., King, J., and Chiu, W., "Backbone Structure of the Infectious Epsilon15 Virus Capsid Revealed by Electron Cryomicroscopy," *Nature* v. 451(7182), 2008.

[17] Speedpage home page, http://speedpage.psc.edu.

[18] Wanpage home page, http://wanpage.psc.edu.

[19] LHC storage requirements (CMS here but ATLAS are similar) are discussed in http://indico.cern.ch/materialDisplay.py?materialId=paper&confId=67969.

[20] Hadoop file system home page, http://hadoop.apache.org/.

[21] Zollner, P.A., Roberts, L.J., Gustafson, E.J., H.S. He, Mladenoff, D.J. and V.C. Radeloff. (In Review) "Modeling the influence of forest management alternatives on patterns of ecological succession in Northern Wisconsin, USA," Submitted to *Forest Ecology and Management*.

[22] Steven Manos, Marco Mazzeo, Owain Kenway, Peter V. Coveney, Nicholas T. Karonis, Brian R. Toonen, "Distributed MPI cross-site run performance using MPIg." HPDC pp 229-230, 2008.

[23] Suchuan Dong, Nicholas T. Karonis, George E. Karniadakis, "Grid solutions for biological and physical cross-site simulations on the TeraGrid," IPDPS, 2006.

[24] Virtual Workspaces home page: http://workspace.globus.org

[25] Keahey, K., T. Freeman, "Contextualization: Providing One-Click Virtual Clusters," eScience 2008, Indianapolis, IN. December 2008.

[26] Alex Younts, "Virtualization at Purdue on TeraGrid," presented at the Virtual Technology Workshop at the Open Science Grid All Hands Meeting, March, 2009.

[27] Andre Luckow, http://randomlydistributed.blogspot.com/2009/10/bigjob-saga-based-pilot-job.html.

[28] Andre Luckow, Shantenu Jha, Joohyun Kim, Andre Merzky, "Adaptive Distributed Replica-Exchange Simulations," *Phil. Trans. R. Soc. A*, v. 367(1897), pp. 2595-2606, 28 June 2009.