

Proposal Cover Sheet Interdisciplinary Science Program

Project Title **An e-Science Grid for Indiana University**

Brief Summary An information network will be built at Indiana University for large data and computationally intensive applications in several sciences, using advanced data grid technologies. With national and international collaborations in physics, bioinformatics, geology, and computer science, this will provide scientists access to local and globally distributed computing resources.

Requested Budget \$1,000,000 Duration 4 years, starting July 2002

Principal Investigator Robert W. Gardner
(primary contact)

Institution Indiana University

Department Department of Physics

Address 701 East Third St, Bloomington, Indiana 47405

Phone 812 855 0236 Fax 812 855 0440

Email rwg@indiana.edu

Co-Principal Investigator Donald G. Gilbert
gilbertd@bio.indiana.edu

Department Center for Genomics and Bioinformatics, Indiana University

Co-Principal Investigator Gary L. Pavlis
pavlis@geology.indiana.edu

Department Department of Geological Sciences, Indiana University

Co-Principal Investigator Dennis Gannon
gannon@cs.indiana.edu

Department Department of Computer Science, Indiana University

1 Objective

Many of the grand challenges of contemporary scientific research involve solving problems at scales orders of magnitude larger than anything previously attempted. With increasing frequency these problems are being addressed by very large, geographically distributed teams of people that need to collaborate and share vast amounts of data and computational resources. The total volume of data that is on-line for scientific research is several orders of magnitude larger than the rest of the World Wide Web taken together. In terms of the computing power needed to solve today's large problems, it is often the case that combining the computational resources of many organizations and individuals may solve them. Building a distributed information infrastructure that allows scientific collaborations to share data and computational resources in secure manner over a wide area is a very difficult task. Such a system, known as a "data grid", is currently a very active area of research in Computer Science¹, and substantial investments are being made in the United States, Europe, and Asia to deploy international data grid infrastructures. Two large, collaborative initiatives in the U.S. have formed to build large-scale data grids for particle physics, gravitational physics and astronomy². CERN, the European Laboratory for Particle Physics in Geneva, Switzerland, is leading the Data Grid Project³, a collaboration of particle physicists, earth scientists, and biologists building a grid infrastructure to enable next generation scientific exploration in these disciplines. In addition, several national data grid initiatives in EU member states have been launched in the past year⁴.

The objective of this proposal is to develop and deploy a community-based, internationally integrated "e-Science Grid" infrastructure needed to advance science in three areas: Genomics and Bioinformatics, High Energy Physics and Earth Seismology. The common thread that unites these disciplines is the need for data grid technology to manage computational and data resources of distributed collaborations. Physics and Computer Science grid initiatives are more advanced, and their technology can be transferred to the similar emerging needs in bioinformatics and seismology, with a broader goal of transferring these to several disciplines. Interdisciplinary distributed data and computing tasks can be shared, and lead to solving problems that span sciences, in a cost-effective way. Advances in each discipline will require integration of heterogeneous computing, human, and data storage resources distributed locally and internationally. At Indiana University, we participate in several projects advancing grid technology for these fields, including distribution of bioinformatics databases, the Grid Physics Network for particle physics, and the collection and analysis of earthquake waves to image the earth's deep geologic structure. We also participate in a growing community developing grid standards and protocols (the Global Grid Forum) and the development of high performance networks for High Energy and Nuclear Physics.

This initiative exploits investments made by Indiana University in several key areas of information technology, including: advanced networking (Internet2/Abilene, TransPAC, Global Network Operations Center), high performance computing, and massive data storage. Most importantly, we expect the advances in e-Science grid infrastructure at Indiana University to provide a practical, transferable model that benefits a wide range of institutions and scientific disciplines (such as astronomy and astrophysics, chemistry, biology, biophysics), medicine, business, and the arts (for example, digital libraries).

2 Science Requirements

In each section below we provide an overview of the computational and data access requirements from each of the scientific areas.

2.1 Genomics and Bioinformatics Research

Needs and opportunities for Internet grid distribution of bioinformatic data and computing are high and immediate. Human Genome data, comprising 100s of gigabytes that are continuously improved, full genome computing tasks, biochemical and protein structure predictions, and proteomics data all challenge current computing resources. Grid improvements here are of economic importance to biomedical, pharmaceutical, agribusiness, and related industries. To empower bio-scientists to answer questions using these large data sets, automation of data and compute resource distribution are essential. A goal of this project is improving data exchange among science centers within Indiana University and with global partners. The US and Asia-Pacific bioinformatics data distribution project (Bio-Mirror⁵), European Union's molecular biology network, national and regional bioinformatics centers are currently investigating grid technologies for this. As exemplified by the Globus package⁶, grid methods are a good match to bioinformatics needs in reliance on standard protocols for data distribution, secure and authenticated resource use, and replica management. Bioinformatics software requires improvements to take advantage of these. IUBio and Bio-Mirror projects at Indiana University have a decade of experience in bio-data distribution, and can help lead internationally in deploying grid methods.

2.2 High Energy Physics

Indiana University physicists are collaborating with 2000 physicists and engineers from 150 institutes in more than 40 countries to construct the ATLAS detector at the Large Hadron Collider (LHC), at CERN, Geneva, Switzerland. The ATLAS detector will explore a new realm of physics up to the TeV energy scale, including the Higgs particles thought to be responsible for mass generation in the Standard Model of Particle Physics, supersymmetry, and evidence for extra dimensions of space-time. Within the first year of LHC operation, the project will store, access, process and analyze 10 Petabytes of data, using of order 200 Teraflops of fully utilized compute resources situated at the Tier-N centers in a global Grid hierarchy. The LHC data volume is expected to increase rapidly, reaching 1 million Terabytes and several Petaflops of computing power by approximately 2015. LHC physicists need to seamlessly access their experimental data and results, independent of location and storage medium, in order to focus on the exploration for the new physics signals rather than the complexities of worldwide data management. ATLAS is in the process of implementing object-oriented software frameworks, database systems, and middleware components to support the seamless access to results at a single site. ATLAS will rely on the GriPhyN Virtual Data Toolkit, grid security and information infrastructures, to provide global views of the data and worldwide rapid access to results, as the foundations of their scientific data exploration. Leading up to operation of the experiment will be a series of Monte Carlo "data challenges" of increasing scale and complexity designed to simulate access and analysis of LHC data.

2.3 Seismology Imaging

Since the birth of digital computers seismic imaging has been a field that pushed the limits of computation and data storage. The reason is that seismic imaging is the fundamental tool for ex-

ploring for oil and gas deposits. Recently the concepts of seismic imaging used in oil and gas exploration have begun to be extended to imaging larger scale earth structure. New instrumentation made possible by the IRIS Consortium⁷ have made it possible to image earth structure at unprecedented scales. This has led to a major new earth science initiative called Earthscope⁸ that includes a continental scale seismic array called USArray. USArray will produce ten times the volume of data of any previous passive array experiment and will require new analysis tools to realize its full potential. Grid computing is a clear direction to make this happen. The geophysics group at Indiana University is a leader in the development of seismic imaging technology with large scale, passive seismic arrays. The methods they are developing for direct imaging of seismic data from distant earthquakes are well suited to grid computing.

3 Proposed e-Science Grid Development

3.1 Community Grid Infrastructure

A community e-Science grid infrastructure, integrated with internationally available grid resources, will be created. Several key grid technologies will be deployed. *Workflow management*: define and implement community-level architecture for distributed scheduling and resource management for large numbers of local (campus) and external users. Co-allocation of resources, optimization of execution based on data location, access to mass storage systems, computation and network resources. *Grid monitoring*: specify, develop, integrate and test tools and infrastructure to enable end-user and administrator access to status and error information in a grid environment. *Security and authorization*: deploy systems to permit local intra-grid, and wide area access to data and computing resources. *Data access*: deploy, test, support tools that permit management and sharing of petabyte-scale information volumes in high-throughput production-quality grid environments. Grid software technology choices: Use of present and next generation grid software toolkits of Globus, Condor⁹; GriPhyN Virtual Data Grid Toolkits¹⁰, to be released annually.

3.2 Bioinformatics specific development

In conjunction with IUBio Archive, Bio-Mirror¹¹ and international partners, apply grid methods to daily distribute data sets among regional and global computing centers, and to employ these data in shared computing for these high usage applications: BLAST¹² sequence search; SRS¹³ sequence retrieval system; EMBOSS¹⁴ sequence analysis package. Integrate these community grid services into a new Bio-Grid architecture, to be linked with future national and international genomics and bioinformatics data grids.

3.3 High energy physics specific development

The Physics Department, together with University Information Technology Services, is building a Tier 2 regional data center for the ATLAS and GriPhyN/iVDGL collaborations. Apply ATLAS grid instantiation methods to other disciplines, building discipline specific, exportable “layers” in the community e-Science grid.

3.4 Seismic imaging specific development

Develop and define grid architecture for grid computing in seismic imaging. Seismic processing involves application of a series of algorithms to well-defined subsets of the combined data volume, and is well suited to grid computing. No application of this kind currently exists in the

field, however, and we will develop a leadership role through the proposed project. This development will unquestionably be of great interest to the oil and gas industry as well, and we would propose to develop partnerships with one or more major oil companies and/or software developers for that industry.

4 Outline Budget

Funds of \$1,000,000 are requested, including a 15% institutional allowance (\$150,000), to be allocated as follows:

4.1 Personnel

A request is made for \$800,000 to employ 3 to 4 postdoctoral or experienced science information technologists for four years, at \$50,000/yr to \$70,000/yr in salaries and fringe benefits. Salaries are required to support the technical personnel required to evaluate, test, and deploy advanced grid software technologies. Three to four personnel are needed to be effective in spanning the discipline range of physics, bioinformatics, geology, and information technology, and be effective in the technology transfer among disciplines that is needed. A major focus will be on integration of grid toolkits with application specific software. Therefore, experienced researchers with significant computing and information technology skills and with knowledge of bioinformatics, physics and geosciences disciplines will be required to communicate requirements and use-cases, perform large-scale tests, and debug software. These personnel will apply existing and evolving grid methods to build a generic infrastructure in conjunction with integrating discipline-specific applications and needs into this infrastructure.

4.2 Hardware

The project will utilize existing hardware resources at Indiana University, from the University Information Technology division, Physics, Bioinformatics, Geology, and Computer Science groups, to a large extent. Requested is \$50,000 to provide support a central grid software server facility and dedicated “grid nodes” for support, evaluation, and testing of grid software.

4.3 Duration

This project will need a four-year duration to adequately test and deploy project plans. A shorter duration would not allow for significant deployment of grid tools or integration with science applications. A longer duration would reduce the number of needed personnel, and limit the range of applications to be included in this project.

¹ “The Grid, Blueprint for a New Computing Infrastructure”, Foster & Kesselman, Eds, Morgan Kaufmann Publishers, Inc., 1999

² The Particle Physics Data Grid (<http://www.ppdg.net/>); The Grid Physics Network (<http://www.griphyn.org/>)

³ The European Union Data Grid Project, <http://www.eu-datagrid.org/>

⁴ The Dutch Grid, <http://vlabwww.nikhef.nl/>; The French Grid (<http://grid-france.in2p3.fr/>); Particle Physics Grid in the UK (<http://www.gridpp.ac.uk/>); E-Science Grid Initiative for UK physicists and astronomers: (<http://www.pparc.ac.uk/nw/sr2000.asp>).

⁵ <http://www.bio-mirror.net/>

⁶ “Grid Information Services for Distributed Resource Sharing”, K. Czajkowski, S. Fitzgerald, I. Foster, C. Kesselman, Proceedings of the Tenth IEEE International Symposium on High-Performance Distributed Computing (HPDC-10), IEEE Press, August 2001. Also <http://www.globus.org/datagrid/>

⁷ <http://www.iris.edu/>

⁸ <http://www.earthscope.org/>

⁹ <http://www.cs.wisc.edu/condor/>

¹⁰ The Grid Physics Network, <http://www.griphyn.org/>

¹¹ Gilbert, D., 1993. "The Global Library", *Trends in Biochem. Sci.* 18: 107-108; See also <http://iubio.bio.indiana.edu/>

¹² Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." *J. Mol. Biol.* 215:403-410. See also <http://www.ncbi.nlm.nih.gov/BLAST/>

¹³ Etzold, T and P. Argos, 1993. SRS – an indexing and retrieval tool for flat file data libraries. *Comp. Appl. Bio-sci.*, 9: 49-58. See also <http://srs.ebi.ac.uk/>

¹⁴ <http://emboss.org/>