

The DOE Science Grid

Principal Investigators: William E. Johnston, Lawrence Berkeley National Laboratory, Ray Bair, Pacific Northwest National Laboratory, Ian Foster, Argonne National Laboratory, Al Geist, Oak Ridge National Laboratory, William Kramer, National Energy Research Scientific Computing Center. Science Grid Working Group: Keith Jackson, LBNL, Tony Genovese, ESnet, Mike Helm, ESnet, Von Welch, ANL, Steve Chan, NERSC, Kasidit Chanchio, ORNL, Scott Studham, PNNL

Summary

DOE's large-scale science projects involve many collaborators at multiple institutions. This leading edge of science depends critically on an infrastructure that supports widely distributed computing and data resources. The DOE Science Grid is being developed and deployed across the DOE complex to provide persistent Grid services to advanced scientific applications and problem solving frameworks. By reducing barriers to the use of remote resources, it is making significant contributions to SciDAC and deploying the cyber infrastructure required for the next generation of science.

A recent workshop¹ solicited the opinion of several scientific disciplines as to how the cyber infrastructure aspects of the process of doing their science needed to change in order to facilitate significant scientific advances. Several general observations came out of this.

The first, and perhaps most significant, observation is that a lot of DOE science is already, or is rapidly becoming, an inherently distributed endeavor involving many collaborators that are frequently multi-institutional. Rapidly increasing data and computing requirements must be addressed with resources that are often more widely distributed than the collaborators. As scientific instruments become more automated and complex (and therefore more expensive) they are frequently used as shared facilities with remote users. For somewhat different reasons this trend is true of numerical simulation as well². Thus, leading edge science depends critically on a cyber infrastructure that supports the process of distributed science.

A second observation is that all the science areas need both high-speed networks and advanced middleware to smoothly couple, manage, and access widely distributed, high-performance computing and storage systems. The many medium-scale and desktop systems

of the scientific collaborations, high data-rate instruments, and massive data archives must also be easily integrated into this environment.

The goal of an integrated, advanced cyber infrastructure – commonly known as “The Grid” – is to deliver:

- Computing capacity adequate for tasks, provided at the time the task is needed by the science;
- Data capacity sufficient for the science task provided independent of location, and in a transparently managed, global name space;
- Communication capacity sufficient to support all of the aforementioned is provided transparently to both systems and users, and;
- Software services supporting a rich environment that lets scientists focus on the science simulation and analysis aspects of software and problem solving systems, rather than on the details of managing the underlying computing, data, and communication resources.

The goal of the DOE Science Grid project is to provide this advanced cyber infrastructure as persistent, scalable, community standards based³, Grid services to support DOE's large-scale science projects. Grid services provide

¹High Performance Network Planning Workshop, doecollaboratory.pnl.gov/meetings/hpnpw. DOE Office of Science. August, 2002.

²“The Computing and Data Grid Approach: Infrastructure for Distributed Science Applications.” W. E. Johnston, Computing and Informatics, Special Issue on Grid Computing, winter 2002. www.itg.lbl.gov/~wej/Grids.

³The Global Grid Forum is an IETF-like international standards organization that represents the community effort to define a common approach to Grids. See www.gridforum.org.

security, resource discovery, resource access, monitoring, data access, tools for integrating with Web portals, and so forth, to advanced scientific applications and problem solving frameworks. These services reduce barriers to the use of remote resources and facilitate large-scale collaboration. Thus, we are making significant contributions to SciDAC-wide software standards and resources, and addressing the infrastructure for the next generation of science process.

The project is integrating activities in deployment, research and development, and application outreach that allow us to develop and refine the Grid tools and their deployment and support. The DOE Science Grid is focusing on identifying and resolving scalability issues so that the Grid can support large-scale science collaborations. Close cooperation with a variety of application projects is ensuring relevance to SciDAC goals and enabling innovative approaches to scientific computing via secure remote access

to online facilities, distance collaboration, shared petabyte datasets, and large-scale distributed computation.

Major accomplishments to date include:

- Construction of a Grid across five major DOE facilities with an initial complement of computing and data resources;
- Integration of NERSC's production, large-scale storage systems into the Grid;
- Design and deployment of a Grid security infrastructure that is facilitating collaboration between US and European High Energy Physics Projects, and within the US Magnetic Fusion community. This infrastructure provides a global, policy based method of identifying and authenticating users, which leads to a "single sign-on" so that any system on the Grid can accept a uniform user identity for authorization. This work is currently used by the SciDAC Particle Physics Data Grid, Earth Systems Grid, and Fusion Grid projects.

- A resource monitoring and debugging infrastructure that facilitates managing this widely distributed system and the building of high performance distributed science applications;
- Establishing development and deployment partnerships with several key vendors;
- Use of the Grid infrastructure by applications from several disciplines – computational chemistry, ground water transport, climate modeling, bioinformatics, etc.

These are important steps in developing and deploying a realistic scale Grid environment that supports advanced Grid services for DOE science.

For more information visit doesciencegrid.org or contact wejohnston@lbl.gov.

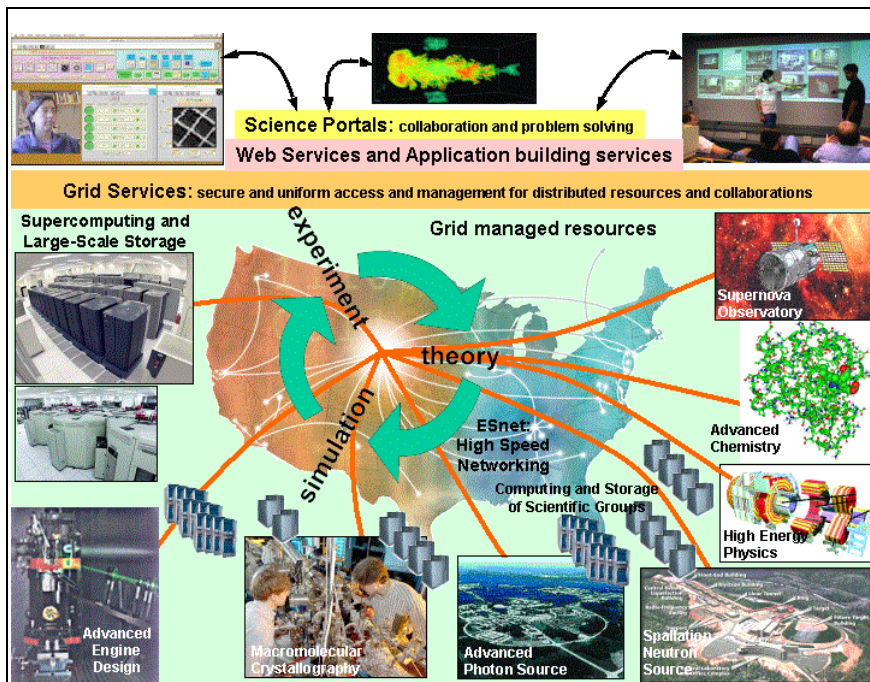


Figure 1. Integrated, Advanced Cyber-Infrastructure – Very High-speed Networks, High Performance Computing, and Grid middleware – Enables Advanced Science: A Vision for the U. S. Dept. of Energy, Office of Science

- Enable the collaborative and interactive use of the next generation of massive data producing scientific instruments
- Facilitate large-scale scientific collaborations that integrate the Federal Labs and Universities