Accelerating Scientific Discovery in Experimental Science (ASDES)

A Workshop to help Define a SciDAC-2 Program

Organizers: Ian Foster (Argonne), Vicky White (Fermilab)

Table of Contents

1 Introduction

Experimental research in science and engineering, and the design, construction, and operation of experimental facilities, are at the heart of DOE's mission. 40% of DOE's FY06 Basic Energy Sciences budget request is devoted to experimental facilities, which support more than 18,000 users. A substantial fraction of DOE scientists and engineers are engaged in experimental research. Furthermore, experimental science and facilities continue to increase in importance: for example, DOE's nanoscience centers have major experimental components.

Experimental science is getting increasingly challenging and expensive. More complex problems, more sensitive instruments, and growing and more interdisciplinary user communities all place pressure on both experimental scientists and facility operations. These pressures relate partly to increased quantities of data: for example, operators and users at light sources and nanoscience centers, and in environmental monitoring projects such as ARM, who used to process megabytes of data, must now access, manage, and analyze terabytes, while the high energy physics community will need a 10^{18}-byte archive by 2012 for data from the four major large hadron collider (LHC) experiments. However, the underlying challenges also encompass the need to design and operate complex experimental apparatus, facilities, and protocols and to collect, manage, access, analyze, exchange, and discuss data that is not only large in volume but also heterogeneous, distributed, and diverse in its origins and purpose.

We believe that it is now timely to define a new program aimed at applying and enhancing information technology to improve the effectiveness of DOE experimental research and facilities. Much as the current SciDAC program has empowered theorists by automating (via numerical simulation) important steps in the process of exploring the implications of theory, so a comparable program can empower experimentalists by automating important steps in the experimental process, from the operation of experimental facilities to the collection and analysis of experimental data. Furthermore, the state of the art in information technology is such that significant progress in these areas seems quite possible at reasonable cost.

To this end, we are convening a workshop aimed at defining the major problems that stand in the way of the DOE experimental community and in identifying opportunities for the application of advanced information technology to those problems. The FY2005 PITAC "Blue Book" report formulated 16 new illustrative grand challenges, including "Knowledge Environments for Science and Engineering," "Collaborative Intelligence: Integrating Humans with Intelligent Technologies," "Generating Insights from Information at Your Fingertips," and "Managing Knowledge-Intensive Organizations in Dynamic Environments"—all directly relevant to the concerns of DOE experimental science.

Each of the offices of the Office of Science spends a great deal of money each year on the application of information technology to their particular experimental science—in effect, on grappling with many of the "grand challenges" listed above. A major goal of the workshop will be to identify R&D activities that can complement and advance that existing work. We expect these activities to adopt the highly successful multidisciplinary SciDAC approach, in which discipline scientists and computer scientists work together to achieve significant gains in efficacy and scientific potential, either by pioneering entirely new approaches or by bringing to bear best-practices and experiences from one discipline to other disciplines.

2 Candidate Work Areas

To start discussion, we outline five broad areas in which we believe major advances in capability can be achieved via focused research, development, and deployment.

2.1 Information Management for Scientific Experiments

The different users and experiments hosted by a major DOE facility such as a light source or nanoscience center will generate, on any one day, many thousands of data items of different types, formats, and sizes. The management and analysis of that data is today largely a manual process, resulting in a tremendous amount of effort (easily hundreds of thousands of person hours per year) being spent storing, finding, converting, and analyzing data—as well as many missed opportunities for progress due to mislaid data and analyses that are never attempted due to their inherent complexity.

We hypothesize that these difficulties can be overcome via the creation of a scientific information management system (SIMS) capable of linking the many different data formats and sources that exist in a typical experimental environment. With such a system, any data can be discovered and accessed in any format, regardless of how it was originally generated and stored. Powerful workflows can then be defined, applied, archived, modified, and reused. Furthermore, while the creation of such systems certainly raises technical challenges that should be addressed, the state of the art in both commercial and open source technology is such that it is likely possible to create workable solutions at reasonable cost.

The SciDAC ASDES program represents a unique opportunity to perform the research, development, and prototype deployments needed to achieve these advances. The following are examples of potential projects within this area:

- Creation and evaluation of a scientific information management system supporting content and workflow management across DOE's Nanoscale Science Research Centers.
- Creation and evaluation of a scientific information management system integrating diverse environmental data sources across DOE laboratories, including both observational data (e.g., ARM) and simulation data (e.g., Earth System Grid, PCMDI).
- R&D on technologies for automating the discovery and mediation of data schemas.

- R&D on technologies for creating, discovering, and reusing scientific workflows on data maintained in a SIMS, and for tracking the provenance of results created by such workflows.

2.2 Telepresence and Telecollaboration for Experimental Facilities

Few of the 18,000 users of DOE experimental facilities are collocated with the facilities that they use. Thus, they must either travel at great expense and inconvenience, or alternatively work with reduced effectiveness from a remote location.

Telepresence technologies allow remote users to participate effectively in remote experiments. Telecollaboration technologies allow distributed communities of users and facility operators to collaborate on the design and implementation of experiments and the analysis of experimental results. In both cases, the state of the art (which owes much to earlier DOE research) is such that large-scale deployments are now possible, although further R&D is also required.

The following are examples of potential projects within this area:

- Creation of remote control rooms for national and international fusion and high energy physics experiments.
- Deployment of telepresence technologies at light source beamlines and nanoscience facilities to enable remote participation in experiment design and execution, with the goals of enhancing both scientific productivity and providing educational opportunities.
- R&D aimed at extending Access Grid technology to link seamlessly with scientific information management systems.

2.3 Linking Simulation and Analysis with Experiment

The “collect-then-analyze-offline” cycle used with most instruments today can be limiting if experimental success is sensitive to experimental configuration and parameter settings. Experimental efficiency and effectiveness can be improved tremendously by coupling with online data processing (perhaps including simulation) for quality control or to provide guidance concerning parameter selection for the ongoing experiment.

The following are examples of projects that could be supported within this area:

- Linking of light source beamlines with analysis capabilities for human-in-the-loop steering of experiments, for example small-angle-scattering experiments.
- Between-shot computation in fusion experiments.
- Selection and classification of data for permanent recording (triggering) in high energy physics experiments

2.4 Data Grids for Large-Scale and Collaborative Data Analysis

An increasing number of science disciplines are finding it useful to create and operate shared, sometime global, data sharing and analysis infrastructures that allow distributed communities to share storage, computing, and data. The Open Science Grid (OSG) is one example (certainly not the only example) of such an infrastructure that is being used by DOE scientists. Originally created as Grid2003 to support high energy and nuclear physics experiments, OSG now supports a range of research projects, including bioinformatics, computational chemistry, environmental science, and computer science. Many other discipline-specific grid infrastructures exist in the U.S. including, but not limited to, NEESGrid, BIRN, Earth System Grid, National Fusion Collaboratory and Collaboratory for Multi-Scale Chemical Science.

The following are examples of potential projects within this area:

- Interoperability and sharing of both best practices and resources between grid infrastructures
- Expansion of OSG to include all major DOE Office of Science research laboratories, and its operation in support of Office of Science projects.
- Service-oriented architectures for evolution and expansion of services available to Grid users
- R&D focused on monitoring and troubleshooting a large distributed system such as OSG.

2.5 Security Technologies for Experimental Science

The competing demands of operational security at DOE laboratories ("if it moves, shut it down") and collaborative science ("if it's useful, provide remote access") are on a collision course. The resolution of these competing demands requires new approaches to system security, by which ….

The following are examples of potential projects within this area:

- Improved intrusion detection technologies.
- Automated system management technologies for detecting and correcting vulnerabilities across DOE machines.
- Virtual organization (VO)-oriented technologies that enable facilities to provide access and services just to selected remote users.

3 Workshop Structure

We propose a two-day workshop to be held in the Chicago area in July 2005. The goal of the workshop will be to develop recommendations concerning the goals and structure of a major (~$30M/yr) DOE program aimed at empowering experimental science. More specifically, it will be charged with producing a ~3 page description of such a program plus a ~10 page workshop report with further information.

Due to limited time, the meeting will be tasked with developing plans in a specific set of areas, for which we will provide, as much as possible, preliminary writeups.

- Day 1: (Full-day - long)
    - Present and review proposed program scope and set of program areas (see an initial set of ideas below); solicit ideas for additional program areas from participants.
    - Breakouts to develop both program scope and program areas.
    - Plenary to take stock, discuss and help prioritize
- Day 2 : Half-day: (to 2.0 pm say)
    - Breakouts to develop concise recommendations (2 pages)
    - Plenary presentations of breakout results (15 minutes each)
- Day 2 : Half-day:
    - Smaller group to complete development of document.