**A grand challenge for DOE collaboratories research:
Creating a digital laboratory system for the 21st Century**

Comments to: Ian Foster <foster@anl.gov>

The Office of Science’s $5B/year budget supports 26,000 investigators at 300 academic institutions and all DOE laboratories; 27,000 researchers use its scientific user facilities. These investigators and facility users are all information workers, consuming and producing large quantities of digital data. The performance of these information workers must be of the greatest concern to DOE and the nation. Yet despite the central role of digital data in DOE research, the methods used to manage this data and to support the information and collaboration processes that underpin DOE research are often incredibly primitive. We must do far better if we are to address 21st Century energy challenges in a timely and effective manner. Improving on the current state is a worthy grand challenge for DOE that, if successfully pursued, has the potential to transform DOE research and facilities.

**Current state: Primitive collaboration systems hinder discovery**

Lest the reader think that we exaggerate when we describe current information processes as primitive, consider the following synopsis of the current state: The journal article (a 17th Century innovation) is a primary information exchange method. The paper notebook (1st Century?) is widely used to document research. The workstation file system or portable drive (modern analogs of the filing cabinet, 19th Century) is a primary data storage method, and due to poor indexing much data stored on those devices is inaccessible to even its original producer. Information sharing is clumsy at best, due to idiosyncratic schema, semantics, and data sharing conventions. Most computational results cannot be reproduced, due to their reliance on manually created and inadequately documented computational pipelines. Email (1970s), telephone (1870s), and airplane (1920s) remain the most widely used collaboration tools. Despite much progress in distributed authentication and authorization mechanisms, security concerns remain a frequent obstacle to collaboration, sometimes shutting down inter-laboratory communications for weeks at a time. And much modeling and simulation is performed using spreadsheets and proprietary packages, rather than the far more powerful methods developed by DOE and other computational groups.

These primitive methods for research and productivity severely constrain researcher productivity and the pace of discovery and innovation.

**Desired future state: A digital laboratory system that accelerates discovery**

Let us imagine what a fully digital laboratory system, designed to accelerate discovery to 21st Century speeds, would look like. In this system:

* All data, code, and documents system-wide would be accessible, discoverable, reusable, reproducible, and computable over.
* Those same information products would be linked by a distributed knowledge base that permits automated navigation of content and connections.
* Advanced software and computational processes would be available on-demand and used routinely by every researcher.
* Collaboration would occur within spaces that people want to use even when they are not collaborating.
* Intrinsic and proactive security mechanisms would encourage rather than discourage collaboration, while protecting against attacks.
* These capabilities would be as intuitive, flexible, and collaborative as the best modern consumer software. Imagine if research data and software were as easily accessible as movies from Netflix and applications from the Apple App store …

Such a system would transform DOE research in many ways. For example, DOE researchers could rapidly design efficient, cheap, and manufacturable catalysts. Today, catalyst design is largely an artisanal process. Prior data is obtained via manual literature review. Expert judgment suggests experimental strategies. Supercomputer simulations provide occasional input. And there is little or no sharing of simulation results. In contrast, a fully digital laboratory system would permit a rational design process, in which automated review of literature, experimental data, and simulation data suggests design strategies; user-guided heuristics guide exploration via large-scale ensemble computations; automated experiments evaluate promising candidates; experimental results inform further refinement; and publications are backed by full documentation of all relevant data, code, and computations.

We could also make far more effective use of next-generation instruments at DOE light sources. Today, such experiments are still largely a cottage industry. A researcher travels to the light source for an experiment. Data is collected, transferred to a USB drive, and carried back to their home institution, where reconstruction and evaluation are performed by a grad student. If problems are detected, the researcher must wait months to repeat the process. In contrast, a fully digital laboratory system would permit high-throughput experiments in which the researcher performs an experiment remotely; data is collected at 100+ MB/sec; real-time reconstruction, analysis, and quality assurance processes permit interactive control of experiments; coupled simulations are used to guide experiments; and both experimental data and a rich set of derived products are instantly available to a worldwide community.