

Addressing the Challenges of Large-Scale Data

Virtually all fields of scientific endeavor base hypothesis testing methodologies on some form of data analysis. Scientific disciplines vary in how they produce data (via observation or simulation), in how they manage data (storage, retrieval, archiving, indexing, summaries, sharing across the science team), and in how they analyze data and communicate results. It is widely agreed that one of the primary bottlenecks in modern science is managing and discovering knowledge in light of the tsunami of data resulting from increasing computational capacity and the increasing fidelity of scientific observational instruments. Further, as data become too large to move, we are evolving towards a model where data-intensive services are centrally located.

Scientific discovery will increasingly be based on the creation, maintenance, and analysis of exa- to zettabyte data archives and collections. This tsunami of data will need to be stored, accessed, analyzed, processed, shared, and understood. Techniques and technology in data analysis, visualization, analytics, networking, and collaboration tools will be essential in all data-rich scientific applications.

To meet these challenges NERSC is developing a diverse set of activities that form the basis of the “NERSC Data” effort, including, but not limited to: community-oriented data repositories; browsing, exploration, and analysis capabilities that operate on the centrally located community repositories; and providing and maintaining the centrally located hardware and software infrastructure that enables these capabilities. NERSC envisions hosting community data repositories with integrated information management and analytics tools and services, as well as new exa-scale storage technologies.

In response to the needs of the NERSC user community, which consists of a diverse set of science stakeholders that span both observational and computational sciences, one key element of our long-term strategy includes an emphasis on integrated, production-quality

Biography in Brief



Dr. Horst Simon is Associate Laboratory Director at Lawrence Berkeley National Laboratory for Computing Sciences, Division Director for the Computational Research

Division, and Adjunct Professor of Computer Science at the University of California—Berkeley.

His research interests are in the development of sparse matrix algorithms, algorithms for large-scale eigenvalue problems, and domain decomposition algorithms. His recursive spectral bisection algorithm is a breakthrough in parallel algorithms, honored with the 1988 Gordon Bell Prize. He has served as a senior manager for Silicon Graphics, the Computer Sciences Corporation, Boeing Computer Services, and has been a member of the faculty at the State University of New York. He is currently a member of the advisory boards of more than five research organizations located throughout the world and is a member of many journal editorial boards and one of four editors of the twice-yearly “TOP500” list of the world’s most powerful computing systems.

data management and analytics with sufficient resources to meet science needs. The status quo—relying solely on “center balance” (NERSC’s hardware infrastructure other than computational systems) and HPSS to address the massive volumes of data on the horizon—does not allow the NERSC Center to fulfill its mission to its stakeholders or DOE. Furthermore, the Office of Science does not yet have a comprehensive program to solve these problems.

Profound Impact

The potential impact to DOE in long-term cost savings and scientific opportunity is profound. For example, the Earth Systems Grid (ESG) currently has about 2,300 registered users and 140 TB of data. The ESG has expressed the desire to migrate their data operations from a workstation-class machine to a scalable platform that is maintained by professionals; this move would potentially free up ESG resources currently committed to system administration for use on other tasks central to its mission. Provid-

ing this kind of data platform to ESG and other imminent projects (such as the Large Hadron Collider, ITER, the Joint Dark Energy Mission/Supernova Acceleration Probe, Planck, the SciDAC Computational Astrophysics Consortium, and the Joint Genome Institute) is a challenge that the DOE Science community must address now. Failure to act decisively may cost DOE potentially tens of millions of dollars in duplicated effort when scientific staff set up and administer their own clusters for doing community-based data management and analysis.

By the end of the next decade we will see an exponential increase in experimental and simulation data. At NERSC we are setting plans in motion to be prepared. ●

Contributor: Dr. Horst Simon, Associate Laboratory Director at Lawrence Berkeley National Laboratory for Computing Sciences, Division Director for the Computational Research Division, and Adjunct Professor of Computer Science at the University of California—Berkeley