

# Coping with the Ultrascale Tsunami of Scientific Data

## Summary

*Ultrascale computing will revolutionize the way science is conducted. With this promise, however, comes a problem – the massive quantities of data produced by ultrascale simulation. Those data must be stored, analyzed, mined, and moved to researchers around the world to extract knowledge – to understand the science. We must not cripple this burgeoning capability with inadequate storage, data management, data flow, or network technology.*

In this “New Millennium of Ultrascale Computing for Scientific Discovery,” *whoever finds the best means for harnessing and controlling the exquisite complexity of Nature will control the science and the technology.* For the first time highly complex applications capable of including all of the relevant physical, chemical, and biological processes at the necessary level of detail can use ultrascale computation to simulate their dynamics in a variety of new ways thought impossible just a few years ago.

Ultrascale computing is changing the face of scientific exploration and discovery. Along with new opportunities come new challenges. The Earth Simulator class computers are on the Moore’s Law trajectory of data generation at an exponential rate. Simulations in astrophysics, climate modeling, computational biology, fusion, high energy physics, combustion modeling, materials, and other scientific applications, will routinely produce data sets that are hundreds of terabytes or even petabytes in size. For example, coupled-climate models already generate terabytes of data per run. Current U.S. models use 140 kilometers per dimension of a cell, and higher resolution using an ultrascale computer will greatly enhance the accuracy of the models and lead to new discoveries of climate behavior. Even a modest goal of improving the resolution to 70 kilometers/dimension, while also improving altitude and time resolution, will generate sixteen times as much data. Similarly, it is estimated that the High Energy and Nuclear Physics community will have several centers by 2006 that will generate 10 petabytes of data per year. That is, in four years there will be 10 times more data than currently exists in this

scientific domain. Thus, the metaphor of a tsunami of scientific data is clearly apt.

The thrilling promise of ultrascale computing will be hampered without adequate storage, file systems, and software to manage, move and understand the ultrascale data effectively. The challenge is not only in managing data within the machine but also in transferring the data to archives to make space for the next ultrascale simulation runs. Moreover, these data need to be read out of storage systems and flow to the processing environment to feed further processing and analysis. Thus, a well-balanced ultrascale facility depends on a smooth flow of data and must include efficient transfer of data to appropriately large storage and file systems and the technology to efficiently search and mine the data during subsequent analyses. To be effective and efficient, we must maintain close coupling between scientific data management and other technologies such as data flow, data analysis, visualization, and networking.

There are several properties of scientific data that demand new computer science approaches to scientific data management. First, while processing power doubles each eighteen months, the quantity of data stored online quadruples in the same period<sup>1</sup>. Increasingly, we will be unable to analyze raw data; we must develop efficient and advanced data extraction and analysis capabilities so that only relevant summarized subsets are presented to the scientist for evaluation.

Second, we must not allow external storage system limitations to hinder ultrascale

---

<sup>1</sup> J. W. Toigo, Avoiding a Data Crunch, *Scientific American*, May 2000.

computing. As data are generated in simulations, they must be moved off the machine to avoid saturating machine resources. Current technology, using parallel transfers to external disk or tape systems, may be inadequate by factors of five to fifty; new technologies or very large aggregations of storage hardware will be necessary to sustain computations. The storage system must not only be fast, it must also achieve new levels of reliability, availability, and manageability. Combining such a reliable storage system with intelligent hardware, hardware that can process streaming data as it is retrieved, for instance, could help reduce the quantities of data to be moved to analysis resources.

Third, massive scientific data sets are essentially immovable. Although raw network capacity is increasing rapidly, data quantities are growing faster. We are already finding it difficult to move terabytes of data to analysis programs. The extreme increase in data quantity resulting from ultrascale computing will be unmanageable, requiring that data extraction, subsetting, and summarization activities take place prior to network transmission. Further, in order to move the reduced sets of data we must have middleware managing intermediate storage systems in the network and mechanisms for scheduling and running analysis jobs at locations remote from the customer. In those circumstances when large amounts of data must to be moved to the scientist's site, reliable and robust data movement over wide-area networks is required. These problems are currently being considered in the framework of data Grids, but ultrascale simulation data quantities and interdependencies present much more severe challenges than those faced by the Grid community.

The enormous quantities of data generated by ultrascale simulations will hold the answers to fundamental questions about the nature of the universe. However, the answers will be subtly hidden in the raw data. The key challenge in the years ahead is a dynamic analysis of this raw

data leading to comprehensive understanding of how individual properties of a physical phenomenon are related. For example, biological cellular systems are very complex. The relationship among genes, proteins, cell structure and organism function are complex and not necessarily linear. For instance, a small change in a gene can ruin the efficiency of a protein and have a profound effect on the organism. Extracting features describing and quantifying these relationships from raw biological data is vital to the understanding of cellular systems. As we study cellular systems via ultrascale simulation, it will be necessary to monitor progress in real time to be sure the simulation is proceeding correctly and to steer subsequent phases.

Discovering such new knowledge will require novel approaches to data management, data analysis, and visualization. Advances in these areas will also enable iteratively validating ultrascale simulations with experimental data, resulting in more accurate computational models. As a result, these technologies will bring revolutionary and unconventional solutions to some of our most pressing and expensive challenges in health, energy, environment, and national security.

The DOE Office of Science supports a broad spectrum of activities that are embarking on deployment of a computational capability that will sustain U.S. leadership in scientific computing. Scientific data management is an essential component of these activities. The "New Millennium of Ultrascale Computing for Scientific Discovery" will depend on both computing power for more complex simulations, and data management power to support subsequent analysis and data understanding.

**For further information on this subject contact:**  
Dr. John van Rosendale, Program Manager  
Office of Advanced Scientific Computing Research  
Phone: 301-903-3127  
JohnVR@er.doe.gov