

## **Gleaning insight from scientific simulation data**

Nagiza F. Samatova\*, Oak Ridge National Laboratory  
George Ostrouchov, Oak Ridge National Laboratory

### **Summary**

*Terascale computing has enabled simulations of complex natural phenomena, on a scale not possible just a few years ago. With this opportunity, comes a new problem – the massive quantities of data produced by these simulations. However, the answers to fundamental questions about the nature of the universe still remain hidden in these data. The goal of this work is to provide a data-understanding infrastructure to help simulation scientists perform a dynamic analysis of these raw data to extract knowledge.*

Terascale computing has enabled advanced simulations that probe deeply into natural phenomena, on a scale not possible just a few years ago. However, terascale simulations in astrophysics, climate modeling, computational biology, and other scientific applications produce so much data that answers to fundamental questions about the nature of the universe are hidden — needles in a huge haystack.

The next, immensely challenging step is to analyze these data so we can identify important properties of a phenomenon and understand how they are related. To do so we need a *data-understanding infrastructure* that will let scientists employ a wide range of analysis and visualization tools without having to deal with the intricacies of accessing and moving data. This infrastructure must consist of a fully integrated suite of software tools and algorithms for data storage, transfer, analysis, and visualization. The DOE SciDAC program has provided an unprecedented opportunity to establish a collaborative multi-disciplinary team to tackle this challenge resulting in the development and deployment of a prototype

of this infrastructure, called ASPECT<sup>§</sup>, as part of the Scientific Data Management (SDM) ISIC center<sup>¶</sup>.

The ASPECT activity has advanced the state of the art in the following key areas.

**Data Reduction.** An adaptive data reduction methodology developed at ORNL in collaboration with SciDAC Terascale Supernova Initiative (TSI) provides 30-fold compression of 3D data with 99% accuracy (total variability). It utilizes a new block-based Principle Component Analysis technique. Data fields adaptively compressed 15 to 200 times over the course of a simulation retain their full visual impact.

**Feature Extraction.** When faced with huge amounts of information, it is easier to understand if you can determine the most important features, a task made tougher when phenomena include many descriptors. Elements of a supernova simulation have seven parameters – 3 space dimensions, 2 radiation directions, color and time. In collaboration with the SciDAC TSI project ORNL researchers have introduced a set of

---

<sup>§</sup> <http://www.scidac.org/SDM/ASPECT/>

<sup>¶</sup> <http://sdm.lbl.gov>

---

\* 865-241-4351, samatovan@ornl.gov

techniques to reduce the entire simulation to a concise set of two-dimensional views that capture its salient features. These techniques have been used to explain the stability of standing, spherical accretion shocks that arise in a 2D simulation of core-collapse supernovae and provide methodology for 2D and 3D simulation comparisons.

Using similar techniques on data from a 120-year transient CO<sub>2</sub> climate simulation led to the discovery that global warming affects winter temperatures the most and summer temperatures the least. This discovery involved a collaboration of ORNL mathematicians, SciDAC CCDCSM.

**Advanced Data Mining Algorithms.** Most data mining algorithms break down on data sets beyond 10 gigabytes, as they may require years on terascale computers to perform typical analyses. ORNL researchers have developed a set of algorithms to analyze datasets that are located at multiple sites and deployed those tools within ASPECT. The tools have been used on data-intensive applications including astrophysics, climate simulations, and biological databases.

**Scientific Visualization.** Visualization is often the only effective way to glean insight from raw simulation data or data analysis results. In collaboration with various visualization and application groups including SciDAC TSI, DOE ASCI TeraScale Browser (TSB), and DOE ASCI and Kitware Inc. ParaView, we have established visualization architecture within ASPECT. It utilizes TSI's application driven visualization pipelines and ParaView's parallel rendering engine to meet the requirements of various target applications such as climate and astrophysics.

**Data Management and Networking.** The ASPECT architecture is designed with end-to-end performance in mind. Bottlenecks in the I/O path or in data transmission are being addressed by various

technologies. Collaboration with the SciDAC SDM parallel disk access project provides high-performance disk I/O. UIC's Sabul network protocol improves network performance by an order of magnitude. Finally, high performance hardware and software established in the Probe project provide an ideal testbed for data storage and access as well as ASPECT's deployment and experimentation. Our demonstration of the value of integrating these technologies as applied to the analysis and visualization of TSI supernova data was a success at SC2003.

**Assembling the Pieces.** ASPECT now has a coherent and easy-to-use data analysis and visualization infrastructure. Its modular architecture includes a number of key loosely coupled components: (i) a GUI front-end to a rich set of statistical data analysis algorithms developed by the R project, high performance distributed and streamline data mining algorithms and a rich set of visualization features; (ii) a set of analysis servers that communicate with the GUI and one another; (iii) a transparent, unifying parallel I/O interface to NetCDF and HDF5 scientific data formats; and (iv) a parallel rendering engine built on top of ParaView. Data reduction and data analysis algorithms are implemented as pluggable, dynamically loadable modules on each ASPECT server. XML-based interfaces to these modules permit customization of the server and the GUI.

Thus, ASPECT, a prototype data-understanding infrastructure, brings a potential to utilize terascale computing for more practical exploration of the massive datasets generated by scientific simulations.

**For further information on this subject contact:**

Dr. John van Rosendale, Program Manager  
Mathematical, Information, and Computational  
Sciences Division

Office of Advanced Scientific Computing Research  
Phone: 301-903-3127

[JohnVR@er.doe.gov](mailto:JohnVR@er.doe.gov)

