Visualization and Data Analysis at the Exascale

A White Paper for the National Nuclear Security Administration (NNSA) Accelerated Strategic Computing (ASC) Exascale Environment Planning Process

ASC Leads: James Ahrens[1], David Rogers[2], Becky Springmeyer[3],

ASC Participants: Eric Brugger[3], Patricia Crossno[2], Ming Jiang[3], Cyrus Harrison[3], Laura Monroe[1], Bob Tomlinson[1], Dino Pavlakos[2]

ASCR Liaisons: Hank Childs[4], Scott Klasky[5], Kwan-Liu Ma[6]

Los Alamos National Laboratory[1], Sandia National Laboratories[2], Lawrence Livermore National Laboratory[3], Lawrence Berkeley National Laboratory[4], Oak Ridge National Laboratory[5], University of California at Davis[6]

Scope

The scope of our working group is scientific visualization and data analysis. Scientific visualization refers to the process of transforming scientific simulation and experimental data into images to facilitate visual understanding. Data analysis refers to the process of transforming data into an information-rich form via mathematical or computational algorithms to promote better understanding. We share scope on data management with the Storage group. Data management refers to the process of tracking, organizing and enhancing the use of scientific data. The purpose of our work is to enable scientific discovery and understanding. Visualization and data analysis has a broad scope as an integral part of scientific simulations and experiments; it is also a distinct separate service for scientific discovery, presentation and documentation purposes. Our scope includes an exascale software and hardware infrastructure that effectively supports visualization and data analysis.

Assessment of current effort with the ASC program and community

By the late nineties, it was becoming increasingly difficult to efficiently and effectively visualize the largest datasets with existing tools. This was a significant concern to the scientific simulation community, because large-scale results were being generated and needed analysis. The Advanced Simulation and Computing (ASC) Computational Systems and Software Environment (CSSE) program changed this by supporting the development of multi-platform parallel visualization applications and toolkits. This software suite includes open-source visualization applications, ParaView and VisIt, an open source visualization library, the Visualization Toolkit (VTK), and a commercial package, CEI's EnSight. These solutions use parallel and distributed computing methods to offer visualization, imaging, and rendering algorithms. The result of these efforts significantly changed how large-scale, scientific visualization is carried out. Scientists today use parallel supercomputers and commodity visualization clusters to routinely visualize terascale and petascale sized datasets that could have never been effectively analyzed. The ASC visualization and data analysis solutions have become key elements in the computational science efforts supported by many government programs including Office of Science (OSC) Advanced Scientific Computing Research (ASCR) and Biological and Environmental Research (BER), the National Science Foundation (NSF) and the Department of Defense (DOD).

Data analysis seeks to characterize data using higher-level abstractions that can be used to find associations between data elements. Data analysis approaches recognize anomalies, find correlations, categorize data, make predictions, and assist in decision making. There are a broad range of techniques used in data analysis, including statistical methods, dimensionality reduction techniques such as principal component analysis (PCA), vector space modeling, machine learning, and clustering. The choice of analysis technique depends upon not only the type of data being analyzed, but also the intent of the user. Some methods permit exploration of data with a target of discovery, others assist in hypothesis testing, and some produce descriptive summaries.

Unlike in visualization, there is little packaged scalable software infrastructure for large-scale parallel data analysis. Currently, the Titan Informatics Toolkit and VTK provide open source implementations of parallel statistics (bivariate correlative statistics, bivariate contingency statistics) and parallel canonical correlation analysis. Another open source project, R, provides some parallel support for statistical analysis, though only a subset of R's algorithms are fully parallelized. Commercially, Mathworks and SAS provide parallel support for a subset of their statistical analysis software packages.

Specific technical challenges to get to exascale

Technical challenges to get to exascale in the visualization and data analysis area are intertwined with challenges in other areas of exascale research, including storage, I/O, programming models, and hardware. In particular, visualization and data analysis at the exascale must be tightly coupled with applications development. While there will continue to be a need for the current post-processing capabilities, analysis at the exascale will require co-design with scientists who develop applications in order to enable a completely new class of exascale analysis.

A new in-situ exascale visualization and data analysis approach is needed since the petascale approach of storing a full-range of results for later analysis becomes impossible due to exascale storage technology trends - The rate of performance improvement of rotating storage is not keeping pace with compute. Provisioning additional disks is a possible mitigation strategy, however, power, cost and reliability issues will become a significant issue. This trend suggests that our ability to generate data on future supercomputing architectures will significantly exceed our ability to store this data. In addition, it is becoming clear that data movement from the chip out through memory hierarchies to storage has significant power costs. At the exascale, visualization and analysis will need to occur as the simulation is run; the option of writing full-scale results to storage will not work. The creation of an effective in-situ visualization and analysis infrastructure is an important technical challenge. A related challenge is the design and addition of data-intensive hardware to the exascale supercomputer, such as memory buffers and analysis-enabled storage systems, to support visualization and data analysis.

Exascale simulation results must be distilled with quantifiable data reduction techniques – In order to achieve scientific understanding from massive simulations, visualization and data analysis techniques reduce the amount of data to a smaller understandable representation. Visualization techniques transform data proportional in size to the scale of the platform, (10^{18}) for exascale, to a visual representation shown on a display that typically has (10^{6}) pixels. Today, parallel graphics algorithms are primarily used to achieve this massive data reduction. This approach provides the foundation for our current successes in handling massive data, however, it is a brute force approach that requires significant computing resources to reduce the data and it is difficult to quantify the bias of this approach. Approaches that quantifiably reduce data as it is generated need to be explored.

New exascale-enabled statistical, parametric, multi-physics, multi-scale simulation approaches require corresponding new visualization and data analysis approaches – Simulation scientists are using new approaches to model more complex phenomenon on exascale resources. Parametric studies record how a simulation responds in a parameter space of possibilities. Multi-physics approaches simulate a linked model of different related phenomena such as a linked physics and chemistry simulation. Multi-scale approaches simulate phenomena at different spatial and temporal scales. Understanding and presenting both summarized and highlighted results from multiple sources such as parameter studies, multi-physics or multi-scale simulations are an important technical challenge that needs to be addressed.

Just like exascale simulations, visualization and data analysis approaches will need to run efficiently on exascale platform architectures and take advantage of a very high degree of parallelism – Technical challenges include achieving portability, efficiency and integration flexibility with simulation codes.

Research and Development Opportunities for the Near Term

For the near term, it is important to prototype, test and deploy a variety of approaches to address the identified technical challenges and evaluate the advantages and disadvantages of these solutions. A list of opportunities is presented below. An evaluation metric and linkages to other areas are identified for each of these opportunities to track progress towards exascale.

1. **In-situ visualization and data analysis software infrastructure** - As we progress to exascale we must move to a model of visualization and data analysis that occurs as part of the simulation run to avoid costly data movement. Visualization and analysis results will be generated as part of the simulation and decisions about what data can be saved will be made dynamically as the simulation progresses. This project addresses the infrastructure required for in-situ analysis. Linkages to the application teams are critical since this infrastructure needs to be collaboratively designed. Success metrics include improving time-to-insight, reducing storage costs and improving result quality.

2. **Advanced data reduction techniques including statistical sampling, compression, multi-resolution and science-based feature extraction approaches** - The massive data that is generated while the simulation is executing needs to be prioritized and reduced immediately. Statistical sampling techniques, information theory and compression approaches are techniques that reduce data size and identify elements of interest. Collaborative design of these approaches with applied mathematics groups and applications groups is critical. Success metrics will be tied to appropriate application milestones.

3. **Visualization and data analysis techniques to help understand advanced exascale physics approaches** – Processing collections of simulation results and providing summarized insight from an ensemble of answers is an important opportunity to pursue. Identifying how results from different aspects of a simulation suite relate to each other, that is, how data relates across scales and how data relates across different joint-physics simulations, is a key aspect of this opportunity. Applied mathematics and applications are vital partners. Success metrics will be tied to appropriate application milestones.

4. **Implement core visualization and data-analysis capability using a scalable parallel infrastructure** – Our visualization and data analysis solutions need to work on the exascale supercomputers. We will partner with the programming models, tools and applications groups to produce a scalable code base. Our success will be measured by our readiness for applications as machine delivery milestones are met.

5. **Exascale visualization and data analysis hardware infrastructure** – We have the opportunity to create an efficient data-intensive hardware infrastructure for the exascale platform. Examples include exploring the use of memory buffers for staged analysis and storage, as well as analysis-enabled storage systems. Success metrics include improving time-to-insight, reducing storage costs, and improving result quality.

6. **Knowledge infrastructure** – Tracking and using knowledge about the scientific goals makes visualization and data analysis more effective. Our opportunity is to integrate the knowledge from workflow tracking and data management systems to improve the quality of the visualization and data analysis solutions we create.