Exa-DM: Enabling Scientific Discovery in Exascale Simulations

https://computation.llnl.gov/casc/StarSapphire/ExaDM.html

PI: Chandrika Kamath
Lawrence Livermore National Laboratory
kamath2@llnl.gov

Co-PI: George Karypis
University of Minnesota
karypis@cs.umn.edu

Extreme-scale systems are enabling the simulation of increasingly complex phenomena, with the output of these simulations being analyzed in different ways to gain deeper insights into the phenomena being modeled. Analysis that is motivated by scientific discovery is particularly challenging as we may not have a precise notion of what we are looking for in the data. In the absence of a mathematical definition that can be converted into an algorithm to extract information from the data, a typical solution is to use an iterative and interactive approach. Starting with a tentative definition, we extract the quantities of interest, have the domain scientists validate the results, and then iterate, refining the algorithms until the results are satisfactory from the domain and the data analysis viewpoints. This process often involves re-running the simulation to collect and write out additional data or generate data with a different set of input parameters.

As an example, consider the detection and tracking of coherent structures, a problem that occurs in many domains ranging from combustion to fusion and materials science. Figure 1 shows the electrostatic potential from a fusion simulation, generated by the GSEP SciDAC Center (http://phoenix.ps.uci.edu/gsep). We are interested in understanding the dynamics of the coherent structures as they shrink, grow, merge, and split over time. After a

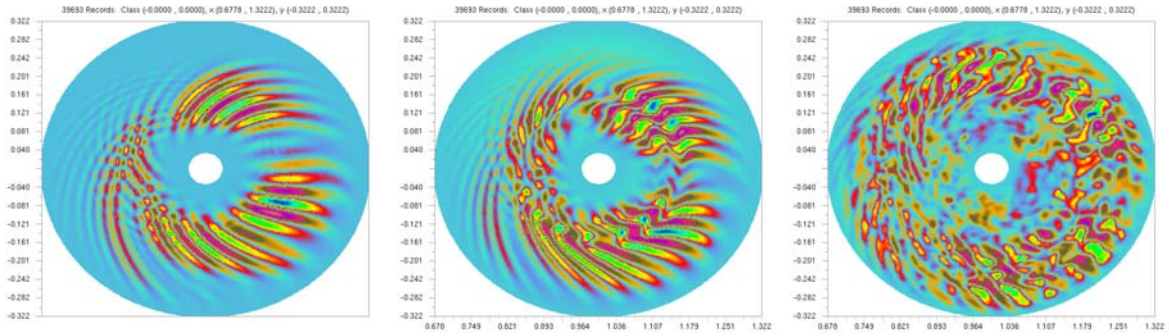

Figure 1: Coherent structures in electrostatic potential. From left to right: time steps 1500, 2000, 2500. Grid points with similar colors have similar values; otherwise, the colors have no significance.

preliminary analysis using a small data set, the simulation was re-run to write out all the variables at each time step to determine if we could use them to design more robust and efficient algorithms for the extraction of the structures in the electrostatic potential. The physicists chose to exploit this opportunity to analyze the structures in the ion heat flux variable which was less well understood. This analysis appeared to indicate the presence of structures with negative flux, which was unexpected. This fact was then confirmed by running a larger simulation, which in turn prompted the physicists to run a different simulation and this time, all the variables were output. As the structures with negative flux have persisted, other variables are being analyzed to gain further insights into these structures and their effect on the performance of burning plasmas. Interestingly, for each type of simulation, and each variable, we had to use a different algorithm to extract the structures of interest.

While this iterative process of scientific discovery through the analysis of massive, complex data sets generated on peta-scale systems is difficult enough by itself, we face an even greater challenge as we prepare for exascale systems. Substantial changes are expected in the architecture of these systems; in particular, the I/O and storage systems are unlikely to provide the required capabilities at the exascale and it may not be possible to write out all

the data from a simulation for analysis later on. A proposed solution is to move all the analysis "in situ", and write out only the results of the analysis, which tend to be much smaller in size. Unfortunately, as shown in our example, this idea is in direct conflict with the process of scientific discovery, which often involves addressing questions which were not even formulated when the simulation was run.

Our goal is to address this conflict by identifying a middle ground where we reduce the amount of data being output while ensuring that we write out enough data to support the process of scientific discovery. We propose to accomplish this in two ways. First, we are exploring techniques to move in-situ as much of the analysis as possible, and second, we are using data mining techniques to intelligently identify a reduced amount of data to be output so the analysis can still be done off line. We are conducting our research in the context of the problem of detection and tracking of coherent structures. Specifically, our work focuses on the following four areas:

- ***Exascale implementations of known algorithms:*** Using well-understood, threshold-based techniques, we are redesigning our algorithms to address the small memory size, the limited memory bandwidth, and the high communication and data movement costs of the exascale systems.
- ***Automated detection of coherent structures:*** We are exploiting the similarity between coherent structures (a group of spatially-close points behaving in a similar manner) and clustering techniques from data mining. Our focus is on graph-based methods for multi-variate data that write out a reduced representation of the structures which are then further analyzed after the simulation has completed.
- ***General reduced representations of the data:*** Considering the problem more generally, we will investigate compressed sensing techniques for their suitability to reduce the data output while enabling near exact reconstruction off-line.
- ***Enhancement of a data exploration tool:*** We are also enhancing an existing tool (see Figure 2) so it can process subsets of larger data sets as well as the reduced representations of the data. This tool has proven invaluable in the initial exploration of the data as well as in the verification of analysis results.

We will use data from the GSEP SciDAC and combustion simulations modeling chemically reacting flows in our work. Both data sets are currently at the petascale. Our domain collaborators are Prof. Zhihong Lin from UC Irvine for fusion and Prof. Sean Garrick from University of Minnesota for combustion.

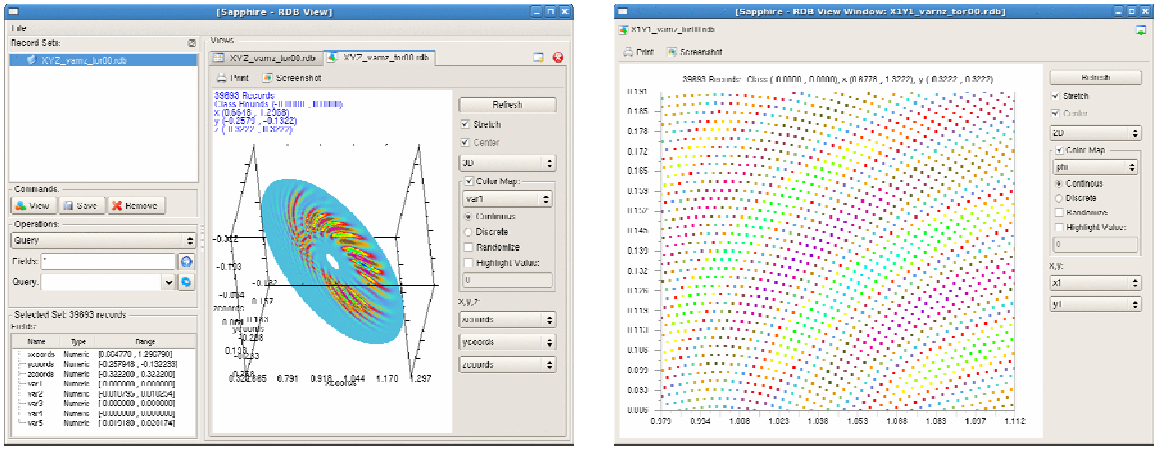

Figure 2: (Left) A screenshot of the data exploration tool showing the electrostatic variable on a poloidal plan in 3-D. The tool allows one to easily view different variables in a multi-variate data set, make simple queries on the data, zoom-in and rotate the figure, etc. (Right) A zoomed-in view of the structures in a 2-D plane. Each point in the grid is displayed and assigned a color based on its value.

LLNL-MI-469913 This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.