Lawrence Berkeley National Laboratory

In-situ Analysis: Challenges and Opportunities

Position Paper at the DOE Exascale Research Conference
Portland, Oregon, April 16–18, 2012

Gunther H. Weber, Peer-Timo Bremer

The expected I/O bandwidth increase for future exascale platforms is only a factor of $10 - 30$ compared to the factor change of 500 for system peak performance. This difference will increase the already existing divide between I/O and system performance, and make it increasingly difficult, if not impossible, to write all necessary data to disk for further analysis. One solution to this problem is performing visualization and analysis in situ, concurrently with the simulation. DOE is already anticipating this trend as evidenced by recent grant solicitations that call for the use of in situ analysis to gain access to all data produced in a simulation and for the purpose of data compression.

Topological data analysis methods provide means to represent data compactly while supporting a wide range of data analysis based on this representation. This type of analysis also supports flexible feature definitions, and will make it possible to control the amount of data written to disk based on feature analysis. In situ topological data analysis poses both challenges and opportunities. One particular challenge is that in situ analysis needs to run on the same machine with the simulation. Thus, it is not possible to have different machine characteristics for simulation and analysis/visualization. Current architecture decisions are based mainly on the behavior and needs of simulations. Achieving the full potential of in situ processing will require that the different characteristics of data analysis are taken into account when making architecture and programming model decisions. For example, in analysis reduction operations (to a single data structure describing a compact result, or a single operation) are much more prevalent than in simulations. Furthermore, graphs play a more important role in topological and other data analysis methods, and operations on graphs are more sensitive to the effects of deep memory hierarchies. For DOE investments into situ analysis to pay off completely, future architectures need to take data analysis specific needs more into account.

The transition to exascale architectures also offers new opportunities for in situ data analysis. Fault tolerance and resiliency make it more difficult for simulations to estimate and balance workloads. Simulation code design teams need to develop new models for load estimation and make significant changes to simulations. Since this transition is inevitable, this change provides a unique opportunity for simulation code development and data analysis teams to develop these load estimation measures and balance techniques jointly. At the same time, it is possible to use this transition to factor in analysis costs in the simulation code and make steps toward a joint load estimate. This need will make it more feasible to integrate more elaborate analysis techniques into a simulation.

Acknowledgements

This work was supported by the Director, Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Department of Energy under Contract Nos. DE-AC02-05CH11231 (Lawrence Berkeley National Laboratory), DE-AC52-07NA27344 (Lawrence Livermore National Laboratory) and DE-FC02-06ER25781 (University of Utah) and the use of resources of the National Energy Research Scientific Computing Center (NERSC).

Disclaimer