Automated Rich-Metadata Management for Integrated Fusion Energy Simulations

Suren Byna[*], Spyros Blanas[+], and John Wu[*]
[*]Lawrence Berkeley National Laboratory
[+]The Ohio State University
E-mail: {SByna@lbl.gov, KWu@lbl.gov, blanas.2@osu.edu}

Topic: F (Data management, analysis, and assimilation)
Oral presentation: Yes

Motivation

Fusion energy simulations play a crucial role in understanding behaviors of plasma, such as turbulence spectrum, transport, and plasma profiles [1]. For example, XGC1 is a full functional (full-f) gyrokinetic particle-in-cell code developed at Princeton Plasma Physics Lab (PPPL) for simulating realistic tokomak geometry including magnetic separatrix and the biased material wall [1][2]. The knowledge gained from analyzing the data produced by these simulations can be used for comparing with experimental observations.

As fusion simulations typically produce massive amounts of data based on the number of particles, writing data for future analysis is impractical. To avoid storing the raw data, *in situ analysis*, where data is analyzed in memory or on staging nodes, has been gaining popularity. Since some analyses require data from multiple simulations, there are plans to run multiple simulations and analysis tasks concurrently on large-scale computing systems. In Figure 1, we show an example of integrated simulation of various common-volume and scale-separated fusion codes [4]. Integrated simulation with *in situ* analysis is challenging because of heterogeneous input parameters of various components. Moreover, the data structures of output are different for various components, which makes *in situ* analysis complicated. For instance, an analysis function should be able to handle data structures related to different processes. Metadata related to the input and output have to be coordinated extensively to perform efficient data movement as well as to conduct *in situ* analysis tasks for reducing the amount of data to be stored.

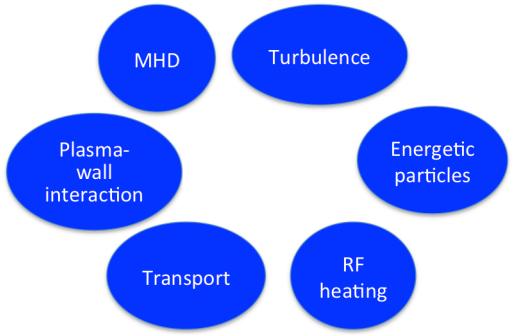

Figure 1. Integrated simulation example of various common-volume, scale-separated fusion codes [4]

Potential Solution Approaches

One way to handle heterogeneity of inputs and outputs is for an analysis code developer to run all the simulation components and extract the metadata of data structures, and then to map them with the analysis data structures. However, this is a laborious and error-prone process even for advanced application developers. Furthermore, this solution may have limited applicability to a specific application.

We propose an automatic rich-metadata management approach that collects metadata of inputs and outputs of simulations and *in situ* analysis codes, stores the metadata efficiently for future use, and maps the data production and consumption patterns of different processes. In addition to tracing the data movement, we propose to profile the performance of data movement so that users of this rich-metadata can compare the obtained performance of their applications with the peak capacity of the systems. In addition, the rich-metadata can include systematic management of information extracted from analyzing the data using *in situ* analysis.

Implementation of the integrated and systematic rich-metadata management system includes three components:

1. *Definition and collection of rich-metadata*: This step requires interfaces for tracing data movement, for analyzing and mapping the data production patterns of simulation components with the data consumption patterns of *in situ* analysis functions, and for defining the extracted information from analysis to be stored. Potential solutions for collecting these different types of metadata is to apply binary instrumentation to dynamically intercept all data movement calls and I/O accesses. Using this approach reduces the burden of recompiling the simulation and analysis codes.
2. *Development of data structures to manage the new types of metadata efficiently*: The metadata we propose here include data production and usage patterns, performance metadata, and information of analyses. Storing these diverse types of next-generation metadata to be used for gaining scientific insights quickly needs to be efficient. The data structures to store these different types of metadata also need to be amenable for searching and for mining the insights efficiently.
3. *Development of methods for mining the new types of metadata*: As the rich-metadata proposed in our solution contains significant amount of data related to data usage patterns, performance data, as well as data related to analysis results, methods for mining the metadata are needed. These methods should consider mapping the data production and consumption patterns of simulation components that could be used for developing or refactoring the in situ analysis codes. Techniques and tools are needed to summarize the performance observations that could be used to set data movement performance expectations that measure the efficiency of integrated simulation and analysis process. Search methods to extract and present the information from analysis results are also required.

More details of our vision for rich metadata management at exascale are available in [3].

Impact

Successful implementation of the automatic rich-metadata management solution will facilitate fusion data analysis application developers with the views of data structures to enable integrated simulation and *in situ* analysis. In addition, scientists will be able to use analysis results for deriving future analysis more effectively. Generic use of the proposed metadata management methods and tools will also be useful for supercomputing facilities to understand the data production and consumption patterns on their systems.

References

[1] S. Ku, C.S. Chang, and P.H. Diamond, "*Full-f gyrokinetic particle simulation of centrally heated global ITG turbulence from magnetic axis to edge pedestal top in a realistic tokamak geometry",* Nuclear Fusion **49**, 115021 (2009)

[2] C.S. Chang, S. Ku, "*Spontaneous rotation sources in a quiescent tokamak edge plasma",* Phys. Plasmas **15**, 062510 (2008)

[3] Spyros Blanas and Suren Byna, "*Towards Exascale Scientific Metadata Management",* in submission, http://arxiv.org/abs/1503.08482

[4] C.-S. Chang, "Fusion Edge Physics and UQ", presentation at the SciDAC-3 Institute QUEST Annual Meeting 2015.