Scientific knowledge discovery in data-intensive, large-scale fusion simulations

R.M. Churchill[1],*
with input from C.S. Chang[1], S. Ku[1], S. Klasky[2], J. Choi[2], R. Hager[1], D. Stotler[1], J. Lin[1], S. Janhunen[1]
1 Princeton Plasma Physics Laboratory, Princeton, NJ, USA
2 Oak Ridge National Laboratory, Oak Ridge, TN, USA
*email: rchurchi@pppl.gov

Where is the knowledge we have lost in information? -T.S. Eliott

Introduction

This paper focuses on a central question: how can we derive knowledge from data generated by large-scale simulations? Simulations today are often hypothesis-driven, carried out with specific goals which justifies reduced physics or data subsets. With whole device modelling, which will include multi-physics and span several spatio-temporal scales, there is an opportunity, even a need, to focus additionally on data-driven discovery, which encompasses a broad range of techniques, all aimed at enabling researchers to effectively explore and learn from these large data sets. If Fusion Energy Sciences is to successfully execute whole device modeling, including difficult regions such as the boundary edge plasma and material interface, we must utilize and refine data-driven discovery tools to aid fusion researchers in extracting as much physics and understanding as possible from these simulations.

Simulation Data

The fundamental issue is that simulation data already today is becoming large enough that we can't store it all (due mainly to disk I/O speeds, not available storage), yet we don't know a priori all that is of interest to look at in the vast amounts of data. This makes it more difficult for researchers to gain the insights and actionable intelligence sought. This data problem will only become more acute as exascale simulations become available in the near future (<10 years).

XGC1 Example

A current example will help illustrate the issue. XGC1 is a large-scale, highly parallelized, gyrokinetic PIC code which uses billions of particles to simulate tokamak devices, focusing on the physics in the edge (pedestal + SOL). To simulate a DIII-D size machine, approximately 25 billion particles are used, generating almost 4TB of data per time step, which is roughly every 30 seconds on the Titan supercomputer. Clearly the particle dataset must be reduced to stay within filesystem I/O limits (about 50 GB/s at best).

One data reduction method is simply calculating fluid moments (density, temperature, velocity, etc.), generating a much smaller dataset (about 1 GB / time step). However, the full

particle distribution function is often desired in the the edge region due to significant non-Maxwellian features that are often present. This data totals a much more manageable 20GB per time step, for a total of 50TB during the simulation.

These data sizes increase by a factor of 10 when moving to larger scale machines such as ITER.

Focus Areas

Here I list general areas for FES researchers to focus on to enable working with and gaining knowledge from large simulation data sets. This is by no means comprehensive, but rather a sampling of algorithms and tools which could be of use.

In-situ or streaming analysis: This is an enabling tool to carry out data analysis before data is written out to disk. The idea is to have either dedicated nodes for data analysis, or a separate staging area where data analysis can be performed. Current tool examples include the ADIOS framework.

Feature Extraction/Pattern Recognition: Automatically finding anomalies, or specific user-defined features in the data sets, helping researchers discover patterns or features to further investigate. If done in-situ, this can trigger additional data analysis or data subset storage actions on data that would normally be discarded due to I/O constraints. For example, in-situ analysis which detects a “blob” in the edge region could then trigger additional analysis on particle trajectories in that region, or save high resolution particle distribution functions in that region for post-processing, to better understand blob generation and propagation mechanisms.

Dimensionality Reduction: Algorithms which can reduce higher dimensionality data to a more fundamental lower dimensional space. This could be useful for example to visualize structure in the higher dimensional particle phase space.

Unsupervised Machine Learning connected to Simulation Database: Creating a database of reduced dimensionality or feature extracted data, then comparing to a current running simulation using algorithms such as k-means clustering to detect novel features in the current simulation. Not only can this guide researchers to new physics discoveries, but also serve as an indicator of “code health”.

Synthetic diagnostics: Observables must be generated to connect to experiment and validate simulations. This data analysis may particularly benefit from in-situ analysis when kinetic simulations are run, as the kinetic effects would otherwise be lost with the reduced dataset being stored. An example where this can have a significant effect could be, for example, Langmuir probes in the scrape-off layer.

Luckily we are able to partner with researchers from ASCR, which has and will increasingly dedicate resources to developing tools and algorithms to work with large data sets generated by scientific simulations (see as a main example the SciDAC SDAV Institute) . However, I am strongly advocating here that FES researchers involved in fusion simulations play an active role in the development and utilization of such algorithms and tools. These tools only become truly useful when combined with the intuition and insight gained from having science domain-specific knowledge.