Workflow Case Study: Fusion Experimental Data Processing Workflow

Jong Choi (ORNL), Tahsin Kurc (SUNY Stonybrook), Scott Klasky (ORNL)

1 Background

Fusion experiments provide critical information to validate and refine simulations that model complex physical processes in the fusion reactor as well as to test and postulate hypotheses. Monitoring, predicting, and mitigating instabilities are critical components of Fusion experiments. Unstable high-energy plasmas can cause serious damage to the reactor chamber, costing hundreds of millions of dollars to repair or substantial loss in productivity.

2 Network and Data Architecture

Local and wide-area networks. Datasets are stored in files on file systems. Streaming data through wide-area networks; streaming experimental data in near-real-time in order to support remote analysis.

3 Collaborators

Korea Superconducting Tokamak Advanced Research (KSTAR), a fusion experiment facility located in Korea, Joint European Torus (JET) in UK, Princeton Plasma Physics Lab (PPPL), LBNL, and ORNL

4 Instruments and Facilities

Present: JET and KSTAR are the current fusion experiment facilities in UK and Korea, respectively. Currently, JET, the world's largest magnetic confinement plasma physics experiment in the UK, is collecting 60 GB of diagnostic data per pulse. An imaging system, called Electron Cyclotron Emission Imaging (ECEI), in KSTAR alone generates 10-100 GB of images per pulse. Mostly post and batch-based data/image analysis is performed locally.

Next 2-5 years: Due to the continued advancement in sensor technologies, we expect 2x-5x increases in data volume in the next 5 years. The rapid imaging system development will contribute on the data explosion. We expect the rate and spatial coverage will be 2x-4x faster and wider in the next 5 years, leading 10x-100x increased data volumes. Researchers need to perform near real-time analysis without restrictions on data locality. Stream-based analysis and workflows through wide area networks need to be supported.

Beyond 5 years: ITER, the next generation fusion facility being built in France, is going to start its initial plasma experiments in 2020. We expect 300-3,000 second pulses, which is 10x-100x longer than current ones produced in JET and KSTAR. Not only near real-time local/remote analysis, but also on-line feedback workflows over wide area networks will take an important role in ITER.

5 Process of Science

Present: Fusion experiments provide critical information to validate and refine simulations that model complex physical processes in the fusion reactor as well as to test and postulate hypotheses. Recent advances in sensors and imaging systems, such as sub-microsecond data acquisition capabilities and extremely fast 2D/3D imaging, allow researchers to capture very large volumes of data at high rates for monitoring and diagnostic purposes as well as post-experiment analyses. However, currently most data and image analysis is performed locally after experiments.

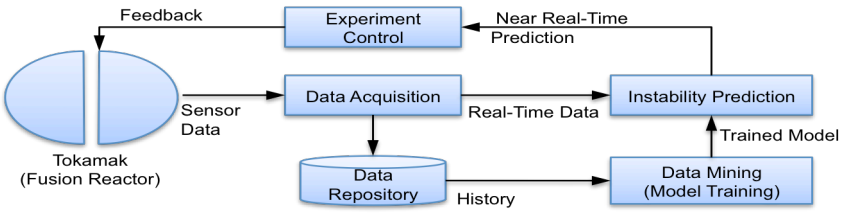

Figure 1. Fusion instability monitoring and mitigation workflow.

Next 2-5 years: The volume, velocity, and variety (data elements from thousands of sensors) of data will make it extremely challenging for researchers to analyze the data only using computational resources at experiment facilities. Researchers need ability to compose and execute workflows spanning local resources and remote large-scale high performance computing facilities. Moreover, near-real-time (NRT) analysis and decision-making is of paramount importance in fusion experiments. Monitoring, predicting, and mitigating instabilities during an experiment need strong NRT analysis capabilities. Unstable high-energy plasmas can cause serious damage to the reactor chamber, costing hundreds of millions of dollars to repair or substantial loss in productivity. A workflow to monitor, predict, and mitigate instabilities is being considered

(Figure 1). This workflow is a multi-level workflow in that each box consists of one or more sub-workflows. Figure 2 shows an example workflow for analyzing 2D imaging data as part of analysis workflows for instability prediction during experiment run using a previously trained model.

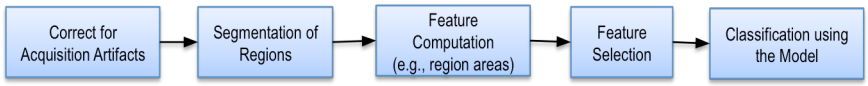

Figure 2. Workflow for analysis of 2D image data as one of ensemble of workflows for instability prediction during experiment run.

To facilitate more efficient experimental work in fusion science, analysis workflows and underlying middleware infrastructure to execute them on local and remote resources should be able to handle thousands of streams of multi-dimensional sensor data within near-real time analysis constraints.

We have been researching and developing systems to support various data challenges in fusion science for the next 2-5 years, which involves the development of ICEE framework to support science workflows execution over the wide area network (WAN). ICEE is developed to support near-real-time streaming of experiment data to and from an experiment site and remote computing resource facilities. We focus on how we execute remote workflows over WAN with NRT requirement.

Beyond 5 years: We anticipate that fusion researches will have more remote workflows scenarios and require strong NRT supports in order to collaborate with remote scientists and exchange live feedbacks. Streaming data thorough WAN will be an important technical element in managing and executing remote workflows.

6 Remote Science Activities

Remote science activities in Fusion experiments can be divided into a few categories. During the run of an experiment, collaborators at multiple sites will want to monitor the experiments and apply analyses to mitigate problems that may arise from instabilities. Between experiments, collaborators may analyze experimental results to evaluate hypotheses as well as design new experiments.

7 Software Infrastructure

Present: A variety of software systems and methods, mostly developed and maintained by research groups in house, are used locally. There is strong need to develop software and tools for stream data processing and large scale data management.

Next 2-5 years: We expect variety of stream-based signal processing and data mining methods need to be integrated in the fusion data processing workflows. Strong NRT support is also necessary. In order to keep up with high-speed data generations, intensive researches will need to be performed on data management technology for the next-generation infrastructure, such as indexing, compression, and feature detection. Hardware and network development needs to be aligned with software development for NRT support.

Beyond 5 years: We expect the complexity of software and workflow system will be highly increased. Efficient software and network infrastructures need to be developed.

8 Outstanding Issues

The volume, velocity, and variety (data elements from thousands of sensors) of data make it extremely challenging for researchers to analyze the data only using computational resources at experiment facilities. Researchers need ability to compose and execute workflows spanning local resources and remote large-scale high performance computing facilities. Moreover, near-real-time (NRT) analysis and decision-making is of paramount importance in fusion experiments.

[1] J. Farthing, T. Budd, A. Capel, N. Cook, A. Edwards, R. Felton, F. Griph, and E. Jones. Data management at jet with a look forward to iter. In International Conference on Accelerator and Large Experimental Physics Control Systems, 2006

[2] G. Yun, W. Lee, M. Choi, J. Kim, H. Park, C. Domier, B. Tobias, T. Liang, X. Kong, N. Luhmann Jr, et al. Development of kstar ece imaging system for measurement of temperature fluctuations and edge density fluctuationsa). Review of Scientific Instruments, 81(10):10D930, 2010