Sample Data Analysis Workflows from Alcator C-Mod, DIII-D, and NSTX

M. Greenwald, D. Hillis, A. Hubbard, J. Hughes, S. Kaye, R. Maingi (co-leader), G. McKee, D. Thomas, M. Van Zeeland, M. Walker (co-leader)

U.S. Burning Plasma Organization working group:
Modes of Participation in ITER

June 24, 2013

Sample Data Analysis Workflows from Alcator C-Mod, DIII-D, and NSTX Executive Summary

The U.S. Burning Plasma Organization (BPO) has formed a working group "Modes of Participation in ITER". One main purpose of this group is to work with the IO and provide information on modes of operation and analysis in present day devices, to contribute to the formulation of the strategy for ITER experimental operation procedures and support systems. To this end, the BPO group has engaged in discussions with IO employees, A. Winter and S. Pinches, who are tasked with formulating a proposal for ITER experimentation strategy.

As a first step, the BPO group has documented sample between-pulse analyses done in Alcator C-Mod, DIII-D, and NSTX. These advanced analysis tasks have been used to guide the conduct of specific experiments in these devices. An experiment on these devices typically represents a sequence of plasma discharges (often in a single experimental session) aimed at investigating various aspects of a particular plasma phenomenon. Between-pulse analysis allows some modification or refinement of the run plan based on what has been learned from previous pulses. For operation on ITER, this could translate to the adapted selection of pulse schedules from a set of previously validated set of pulse schedules with the aim of making the most effective use of run time. Moreover if multiple experiments are performed in a single pulse, each sub-experiment will require a set of validated pulse schedules, and the entire pulse will need validation to make sure one experiment isn't affected by changes to another.

The workflows for these tasks are described in the subsequent pages, including a cursory evaluation of the implied requirements for ITER. The sample tasks described below include gyrokinetic analysis (C-Mod), energetic particle instability analysis (DIII-D), and transport and confinement analysis (NSTX). While these tasks are presently carried out between pulses in US devices, processing of certain analysis elements that underpin more in-depth analysis concurrent with the ITER discharge will be needed to insure timely completion between ITER pulses.

The common elements in the workflow analysis from these three devices include:

1. Some level of automated data analysis and required data availability
    a. Magnetics data, from which equilibria are reconstructed as a launching point for further analysis, is presently acquired rapidly enough to allow reconstruction of the equilibrium state during multiple time slices.
    b. Plasma data used as input for subsequent calculations (mostly profile data) is required as inputs to multiple analysis tasks; a more accurate equilibrium calculation actually uses plasma profile data, which is typically available a few minutes after the end of the pulse in present devices.
2. Some level of interactive tasks
    a. Review of data to disregard obviously questionable inputs and to validate data for higher level analysis.
    b. Semi-automated creation of scripts to run in-depth analysis

3. There is a hierarchy of calculations: increasingly constrained equilibrium calculations, e.g. magnetics-only, including kinetic + diamagnetic data, including MSE, etc. are used for different purposes; many of these are presently done between pulses.
4. Analyzed data is written, along with meta-data for provenance, to a central archive accessible by all team members who have been granted access to data, typically through project-wide data usage agreements.

These translate to the following requirements for ITER:

1. Rapid access to "critical" data for between shot inspection and evaluation of run progress
2. Rapid access to allow important automated analysis tasks to begin as quickly as possible – ideally concurrent with the plasma discharge, but minimally to complete soon after each shot. Most notable in this regard is the calculation of the magnetic equilibrium.
3. Reserved cycles on local machines for analysis tasks.
4. Ability to declare and distribute 'events' that signal that needed data is available, allowing the workflow to proceed and be monitored; this is important for tasks with both automated and interactive components.
5. Availability of a central repository where analyzed data and corresponding meta-data can be stored and accessed by all authorized team members.

Comparison of the three tasks also highlights certain flexibility needed:

1. The ability to specify the order of priority for diagnostic signals for specific experiments; e.g. fluctuation experiments will need fluctuation data to come in quickly, while other experiments may not need this.
2. The ability to communicate results of analysis that could guide the experiment execution to the session leader and diagnosticians.
3. The ability to modify the configuration or control of the device between successive pulses of an experiment.

Although not discussed in the sample analysis workflows below, higher level functionality is also desirable and would improve overall coordination and productivity. These capabilities exist piece meal in existing US devices, but will likely need to be developed in an integrated fashion for ITER. Additional specific recommendations will follow in subsequent documents. Examples of the required tools include:

1. Run information database: This information includes descriptions of each day's experiments with links to experimental proposals, run plans, run summaries, logbook entries and data summaries.
2. Data quality system: Infrastructure that allows users to assign and record data quality metrics.
3. Data analysis/review request system: A system that allows researchers to request data analysis or review from other members of the team and to track the status and results of these requests.
4. Data provenance system: Infrastructure that supports the documentation and annotation of raw and processed data through the entire analysis chain.

Sample workflow for data analysis from Alcator C-Mod
Between-shot linear gyrokinetic analysis using GYRO code
M. Greenwald

Task Description

Data from a broad set of diagnostics is marshaled, analyzed and reviewed to prepare inputs for a linear gyrokinetic code (GYRO), running on HPC hardware.

Physics Motivation

When available between shots, the results of linear gyrokinetic calculations are used to help guide machine operations. The relevant experiments require operation at conditions specified relative to stability boundaries (e.g. the boundary when Ion Temperature Gradient (ITG) fluctuations/transport overtake Trapped Electron Mode (TEM) fluctuations/transport) – which would otherwise be computed after the run. By carrying out the analysis between shots, we are able to gather the required data with the fewest number of shots, selecting from a previously prepared menu of planned discharges. Without this capability, we would be operating somewhat blindly and would have to more exhaustively sample the operating space to ensure that we obtained the data needed. Since manual analysis and review are part of the workflow, this approach presents requirements for data availability and coordination among the analysis team.

Workflow

The workflow can be broken roughly into 3 phases. In the 1st phase, automated processes acquire, store data, then carry out low level analysis tasks (steps 1-3). In the 2nd, interactive tools are used for higher level analysis and manual data review (steps 4-7), finally in the 3rd phase, input files for are assembled and the stability jobs are dispatched to an HPC system (steps 8-12).

1. Automated data acquisition & data storage
    a. Required data set includes magnetics, density and temperature profiles from multiple diagnostics, various spectroscopic measurements
2. Automated conversion to physical units, application of calibration factors
3. Automated post-shot analysis
    a. Equilibrium (EFIT – run with several different input options)
    b. Diagnostics: Thomson, Edge Cyclotron Emission (ECE), Motional Stark Effect (MSE), Impurity assessment via effective charge (Z_{eff}), Charge eXchange Recombination (CXR), Xray Ion Crystal Spectrometer (XICS), etc.
    c. Calculation of statistical error bars
4. Set data retrieval parameters for GYRO runs
    a. Select which data to use (which EFIT "tree", which diagnostics for Ti, velocities, etc.)
    b. Set radius and time range of interest

c. Set radial grid for mapping/interpolation

5. Get "raw" EFIT data
6. Process EFIT data
    a. Compute magnetic flux surface shape profiles
    b. Smooth in time as required
    c. Display and review results
    d. Or compute kinetic or MSE constrained EFIT equilibrium if desired and repeat a-c; these provide more a realistic representation of the equilibrium, as they use more plasma data as constrains
7. Retrieve, fit and map profile data for ne, Te, Ti, V, q, etc.
    a. Typically requires interactive analysis and/or expert review
8. Load scalar parameters required by code
9. Load raw and mapped profiles
10. Set GYRO run parameters (krange, nk)
11. Invoke script to run code
    a. Read data
    b. Write GYRO input file for normalization run
    c. Write batch file for normalization run
    d. Submit GYRO job (write output file)
    e. Read/parse output file
    f. Create k-array data
    g. Write set of input files, batch files
    h. Move files to directory structure
    i. Submit each "directory" (run GYRO)
    j. Check queue for job status
    k. Read files
    l. Write data to archive (MDSplus tree)
    m. Compute derived quantities, write to archive (MDSplus tree)
    n. Clean-up
12. Invoke procedures to display data

Implied Requirement

- Need for some level of self-documenting automatic data processing that integrates more than one plant system

    - (Note: scope might be extrapolated based on C-Mod total, between-shot, automated analysis jobs for all purposes = 165)
- Need to have raw and automatically processed data available to end users as quickly as possible - ideally during the ITER pulse duration
    - This might require only some part of the total data set, though this is hard to determine in advance
- Users need fast access to data – processing servers may need to be physically close to data archive
- May require direct data access from HPC machines
- May require reserved cycles on HPC for these between-shot analysis tasks
- User must be able to write data back into shared archive
- Need some sort of "events" or notification system to allow users to know when raw or processed data is available
    - User applications (including HPC tasks) must be able to set these events/notifications for subsequent dependent tasks
- Need to track/document status and provenance (with full chain of dependencies) of all automated and manual analysis tasks
    - User applications must be able to write this data to archive

Sample workflow for between shot analysis of energetic particle (EP) relevant diagnostic data, using instability studies as an example

M.A. Van Zeeland, W.W. Heidbrink, G.R. McKee, D.M. Thomas

Task Description

Data from several fluctuation (ECE, BES, CO2 interferometers, magnetics), EP (Fast Ion Loss Detectors, BILD (Beam Ion Loss Detector), Ion Cyclotron Emission), and equilibrium related diagnostics are acquired and analyzed between shots for DIII-D EP experiments. Fluctuation diagnostics in particular are processed primarily through Fourier analysis.

Physics Motivation

For DIII-D EP experiments, rapid between-shot access to information about the particular instability under investigation is essential. Measurements of eigenmode stability, frequency, toroidal mode number, radial structure/localization and induced EP transport are all used to inform actions for subsequent discharges. These actions can specifically target a physics aspect of the instability itself, for example, changing neutral beam heating power to alter mode drive, or be more diagnostic oriented, such as radially shifting the BES array to be better centered on the instability location.

Workflow

1. Automated data acquisition and data storage
2. Select time and frequency range of interest
3. From local user terminal, trigger calculation of windowed spectra for several fluctuation diagnostics
    a. Retrieve data
    b. Divide into overlapping time windows
    c. Fourier analyze
    d. Store resultant surface plot as .eps
    e. Repeat for:
        i. Crosspower spectrogram of vertical and radial CO2 interferometer data
        ii. Crosspower spectrogram of adjacent Electron Cyclotron Emission (ECE) channels across array
        iii. Crosspower spectrogram of successive Beam Emission Spectroscopy (BES) channels (if appropriate beam was used)
        iv. Autopower spectrogram of large bandwidth FILD Photomultiplier channels
4. Automated post-shot analysis
    a. Equilibrium (EFIT with and without Motional Stark Effect – MSE data)
    b. Diagnostics: Thomson, ECE, MSE, Charge Exchange Recombination - CER (simple analysis between shots, detailed analysis overnight)
    c. Rapid spline fits (ZIPFIT) to kinetic profiles created

    d. Toroidal Alfven Eigenmode (TAE) Frequency calculated and stored using q_0, q_{min}, q_{95} values from EFIT and line-averaged density –used to help identify observed instabilities

5. If Alfven continuum analysis desired
    a. MSE based EFITS are used in combination with ZIPFIT density profiles to create input file for single toroidal mode number on EFIT timebase
    b. Calculation is triggered on cluster
6. Expert review of spectrogram data to inform next steps
    a. Initial determination of mode stability, frequency range, and rough amplitude made from CO2 Interferometer spectrogram
    b. Frequency and spectral behavior (i.e. steady, chirping, sweeping) combined with calculated TAE frequency and/or continua used to help identify instability
    c. Mode localization determined from ECE and BES data
    d. Assessment of induced EP loss determined from spectra of FILD channels
7. If satisfied with ECE/ECEI/BES array location relative to instability, continue with physics investigation, otherwise adjust BES steering and/or make minor toroidal field adjustment for improved ECE/ECEI location
8. If physics investigation includes ECH probing of instability and between shot adjustment of ECH deposition location relative to instability
    a. Launch EFIT viewing program
    b. Load present gyrotron setup
    c. Adjust gyrotron steering angles
    d. Using EFIT, electron temperature, and electron density profiles create input file for ECH deposition code, TORAY
    e. Run TORAY to obtain ECH deposition profile
    f. Iterate until desired steering is obtained
    g. Communicate desired steering angles for next discharge to ECH group

Implied Requirement

1. Immediate access to fluctuation data. At minimum, data from diagnostics that probe large regions of plasma (e.g. central interferometer chord) should be available first to give overall picture of instabilities.
2. Variable diagnostic information, e.g. geometry, frequency, mapping between variable names and locations, etc. for key fluctuation diagnostics
3. Fast access to equilibrium, density, and temperature profile information.
4. Fast access to neutral beam and other heating waveforms
5. Ability to communicate desired changes to responsible diagnosticians

Sample workflow for data analysis from NSTX
Between-shot TRANSP analysis
S.M. Kaye

Task Description
Data from a broad set of diagnostics is prepared as input for between- (or among-) shots cross-magnetic field transport TRANSP analysis (BEAST)

Physics Motivation
When available between shots, the TRANSP can be used to help guide machine operations. TRANSP is used not only as a means of diagnostic and data validation, but also as a guide to understanding confinement and transport trends of various scans during experiments. Results will indicate if the data is good enough for more comprehensive post-run analysis, and can provide a guide for whether discharges need to be rerun, or the run plan modified during operation. Since manual analysis and review are part of the workflow, this approach presents requirements for data availability and coordination among analysis team.

Workflow
The workflow can be broken several phases, and different options for run preparation are offered.

1. Automated data acquisition & data storage
    a. Required data set includes magnetics and other 1D time evolving quantities (neutron emission, diamagnetic signal, voltage, plasma current, etc). density, rotation, temperature and radiation profiles from multiple diagnostics
2. Automated conversion to physical units, application of calibration factors
3. Automated post-shot analysis
    a. Equilibrium (EFIT)
    b. Diagnostics: Thomson, Charge Exchange Recombination Spectroscopy (CHERS), Motional Stark Eeffect data (MSE, if available), Bolometer, etc.
    c. Calculation of statistical error bars
4. Set up TRANSP run using "ELVIS" graphical user interface
    a. Specify EFIT version for run (magnetics-only EFITs for between-shots TRANSP, EFIT with profile information for "among-shots" runs)
    b. Set data source for Zeff
    c. Set time range of calculation (time slice or full evolution; the latter will not complete between shots)
    d. Process data for run setup (run ELVIS interface)
        i. "Scrunch" EFIT data
        ii. Retrieve 1D and profile data, automatically smooth/deglitch data
        iii. User examination of input 1D, 2D data if using ELVIS interface
        iv. Enter comments for run
    e. Submit run directly through ELVIS interface

5. Set up TRANSP run using python script, review input and submit run
    a. Run python script specifying EFIT version, other parameters
        i. Automatically "scrunch" EFIT data, retrieve/smooth/deglitch input 1D and 2D data, write UFILES,
        ii. Graphical user interface allows for setting time range of run
        iii. Prepare and write namelist
    b. Examine input data using trdat, res-mooth/re-deglitch if necessary
    c. Preprocess data using tr_start; enter comments for run
    d. Submit run using tr_send
6. Results of running by both methods written into BEAST MDSplus tree
7. Email notification when process complete
8. Run ELVIS or personal script, or other routine (RPLOT) to examine results

Implied Requirement

- Need for reserved and dedicated CPUs on linux cluster for BEAST processes (serial runs)
- Need to have raw and automatically processed data (equilibrium, profiles, in 3a, 3b) available to end users as quickly as possible
- Users need fast access to data – processing servers may need to be physically close to data archive
- Need "events" or notification system to allow automated scripts to know when raw or processed data is available.
    - Preparation scripts "wait" until data is available