GA–A23745

STATUS OF THE LINUX PC CLUSTER FOR BETWEEN-PULSE DATA ANALYSES AT DIII–D

by
Q. PENG, R.J. GROEBNER, L.L. LAO, J. SCHACHTER,
D.P. SCHISSEL, and M.R. WADE

AUGUST 2001

DISCLAIMER

GA–A23745

STATUS OF THE LINUX PC CLUSTER FOR BETWEEN-PULSE DATA ANALYSES AT DIII–D

by
Q. PENG, R.J. GROEBNER, L.L. LAO, J. SCHACHTER,
D.P. SCHISSEL, and M.R. WADE†

This is a preprint of a paper to be presented at the 3rd IAEA Technical Committee Meeting on Control, Data Acquisition, and Remote PaRticipation for Fusion Research, Padova, Italy, July 16–19, 2001 and to be published in *Fusion Engineering and Design.*

†Oak Ridge National Laboratory

Work supported by
the U.S. Department of Energy under
Contracts DE-AC03-99ER54463 and DE-AC05-00OR22725

GENERAL ATOMICS PROJECT 30033
AUGUST 2001

ABSTRACT

Some analyses that survey experimental data are carried out at a sparse sample rate between pulses during tokamak operation and/or completed as a batch job overnight because the complete analysis on a single fast workstation cannot fit in the narrow time window between two pulses. Scientists therefore miss the opportunity to use these results to guide experiments quickly. With a dedicated Beowulf type cluster at a cost less than that of a workstation, these analyses can be accomplished between pulses and the analyzed data made available for the research team during the tokamak operation. A Linux PC cluster comprised of 12 processors was installed at DIII–D National Fusion Facility in CY00 and expanded to 24 processors in CY01 to automatically perform between-pulse magnetic equilibrium reconstructions using the EFIT code written in Fortran, CER analyses using CERQUICK code written in IDL and full profile fitting analyses (n_e, T_e, T_i, V_r, Z_{eff}) using IDL code ZIPFIT. This paper reports the current status of the system and discusses some problems and concerns raised during the implementation and expansion of the system.

1. INTRODUCTION

The DIII–D [1] tokamak operates in a pulsed mode producing plasmas of 5 to 10 s duration every 10 to 15 minutes, with 25 to 35 pulses per operating day. Some of the analyses that survey experimental data take on the order of ten minutes or more to complete on a single fast workstation. Since they do not fit in the narrow window between pulses, they are typically carried out at a sparse sample rate between-pulse and/or completed as a batch job overnight. Scientists therefore miss the opportunity to use these results to guide experiments quickly.

Supercomputer and its recent low cost version Beowulf [2,3] are traditionally used for modeling and very detailed and time-consuming analyses in fusion community. With affordability and applicability, many experimental data analyses can also benefit from similar systems. A small Beowulf type cluster can be had at a fraction of the cost of a fast workstation. A quick look at the analyses reveals that many of these are computed over time dimension or spatial dimension or both, often independently, and can easily be parallelized or distributed over these dimensions. The execution time can then be dramatically reduced to fit the analyses at full resolution between two pulses.

Based on this observation and preliminary testing, a 12-processor Linux PC cluster [4] was installed at DIII–D early last year to perform between-pulse magnetic equilibrium reconstructions using EFIT code written in Fortran [5]. Typically two sets of EFITs are performed: one using magnetic data only and the other with Motion Stark Effect (MSE) as well. Thi new system produces equilibria eight times faster than a previous distributed system.

The availability of EFITs earlier in a pulse cycle opens up possibilities for other analyses and is one step closer to between-pulse Kinetic EFIT analyses and/or ONETWO power balance analyses. An automatic CER [6] analysis code CERQUICK written in IDL and Fortran was ported to Linux and adapted to the cluster. It provides the profiles of ion temperature (T_i), rotation speed (V_r) and impurity density (Z_{eff}) while the electron density (n_e) and temperature (T_e) profiles are made available by Thomson scattering and CO_2 diagnostics shortly after a pulse ends. The analysis at full resolution was previously performed overnight. It can now be completed between pulses.

With EFITs, complete profiles of n_e, T_e, T_i, V_r, Z_{eff} available, the natural next step is profile fitting analysis. A detailed profile analysis is usually done by experts for the selected times of a pulse and can be very time consuming. But an automated profile fitting code that surveys the full time series and produces reasonable result is available and again it was previously performed on a sparse sample rate between pulses. The code is ZIPFIT and written in IDL. It fits profiles using EFIT data, Thomson data and the results from CERQUICK analysis. Based on the fitted profiles,

it calculates thermal neutron source density and thermal pressure. All results are loaded to MDSplus [7], a data storage system for analyzed data.

With 12 processors, CERQUICK is scheduled after both sets of EFIT although the two are independent. With EFITs and CERQUICK, the 12-processor cluster nearly reached its capacity. In order to perform profile fitting analysis, the cluster needs to be expanded.

2. THE EXPANDED CLUSTER AND BETWEEN-PULSE PROFILE FITTING ANALYSIS

For CY01 DIII–D operation, the size of the cluster was doubled to a total of 24 processors. The additions are 6 dual 800 MHz Pentium III Coppermine processors while the old ones are 6 dual 500 MHz Pentium III Katmai processors. Each processor has 256 MB memory as in the old ones. The expansion in hardware cost about $14K. Both CERQUICK and ZIPFIT require Interactive Data Language (IDL) [8], a commercial software by RSI. It was decided to install IDL on the new 12 processors only, after balancing the cost and performance. Any analysis that requires IDL will run off this part of the cluster, while the remaining modes continue to serve EFITs. With part of the cluster freed up later in a pulse cycle, more types of EFITs can be scheduled. Researchers also have the luxury to improve the accuracy of the MSE EFITs by increasing the number of iterations.

The CERQUICK analysis cannot be started until CER data become available, about three minutes after a pulse ends. To improve the accuracy of physics calculations, CERQUICK has to wait for one more minute for the calculation of the rotation reference to complete. The CPUs are not all wasted however. The first set of EFIT that uses only magnetic data is executed on all 24 processors. This reduces its time to availability by nearly a minute. A minute may not sound substantial, but it can have cascading effects. The MSE EFIT starts after the magnetic EFIT completes to avoid competition for the CPU time. The electron profile part of the ZIPFIT only needs magnetic EFIT data and Thomson scattering. This part can be started as soon as magnetic EFIT data are available and before CERUICK analyses while the remaining is scheduled after CERQUICK completes (Fig. 1).

CERQUICK is distributed along spatial dimension. Each process reads CER data through a library and Thomson data from MDSplus and writes results back independently, one channel at a time. ZIPFIT fits profiles on the spatial dimension therefore has to be distributed along time dimension. Because of the way the data are organized in MDSplus, the results of each fitted profile in ZIPFIT need to be collected and loaded once for the whole time sequence. Since IDL does not have the capability of parallel processing, the distribution and control are done with shell scripts. Each process retrieves EFIT and CERQUICK data from MDSplus. The results are written in files temporarily. Once the analysis is completed, the master process collects the results and writes them into MDSplus. The ZIPFIT analysis contains several stages, including and in the order of electron profiles, ion temperature and toroidal rotation profiles, impurity density profile and finally thermal neutron density and thermal pressure calculations. They are such ordered mostly by data dependency and partially by priority. The impurity density

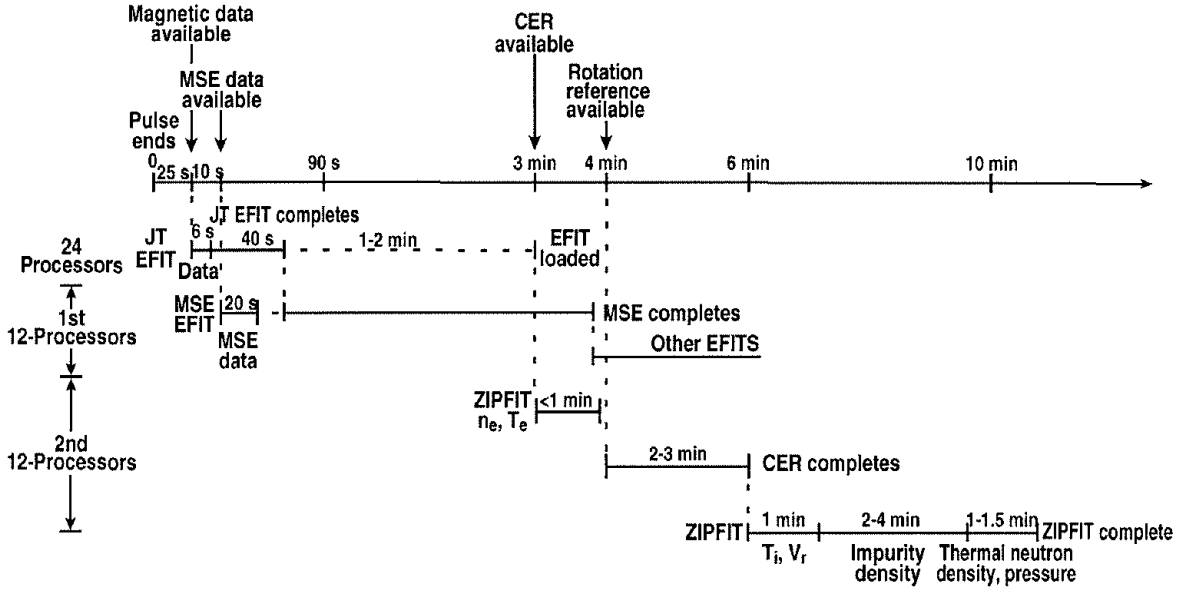

Fig. 1. EFIT, CERQUICK and ZIPFIT analyses start automatically and complete between pulses.

calculation uses a Fortran library that would require large effort to port to Linux, it is therefore left running on an HP workstation.

The entire ZIPFIT analysis in its current shape can last five to seven minutes. We can not afford to wait for all the analyses to complete and then load them altogether although that would be the most efficient in terms of the loads on network and MDSplus server. Therefore, each profile is analyzed and results loaded before next profile proceeds. This simplistic approach requires entering and leaving the IDL sessions at each stage, which results in several reconnections to the MDSplus server and multiple retrieval of the same data, primarily the magnetic EFIT data. The sudden increase of simultaneous connections and data flow puts an extra load on the already heavily loaded MDSplus server and the network during its busiest time. This not only slows down the process itself but also has a large impact on others. Large part of it cannot be helped without infrastructure changes because that is the result of the increased computational capability: there are lots more data to go around. But there is room for improvement. The first area to attack is the multiple fetching of EFIT data from MDSplus server. The solution is data caching. Again, a relatively simple approach was taken. A temporary MDSplus server is installed on the master node of the cluster. The magnetic EFIT data files are duplicated for each pulse. Each ZIPFIT process then retrieves its EFIT data from the local server. Multiple data retrievals are not eliminated but made faster since the cluster has its own local network. It also takes some load off the main network and MDSplus server that everyone shares. About one minute is saved.

The EFIT, CERQUICK and ZIPFIT analyses are automatically triggered by MDSplus event when necessary data are available. Their results are loaded to MDSplus for central storage and unified accessibility. Their progresses are monitored by a data analysis monitoring (DAM) system. It issues warnings when analyzed data are not available at the expected time. A time line of the three analyses and related events is shown in Fig. 1.

3. DISCUSSION

The major problem arisen in the implementation of ZIPFIT is the incompatibility of IDL and parallel processing. ZIPFIT has to make several connections to the MDSplus server and repeated data retrievals because of the staged process and data loading. The inefficiency is very noticeable. The ideal solution will be an MPI type library callable from IDL, which is not a minor task in itself. Before that can become a reality, sub-optimal solutions, presumably easier and faster to implement, have to be found.

For the type of the analyses performed by far, the primary concern is data distribution and collection. A temporary MDSplus server local to the cluster is used to cache EFIT for ZIPFIT. One step further in this scheme will be making the local server a relay point between the main MDSplus server and the cluster to eliminate the need for multiple and simultaneous connections to the main server. The complication in the case of ZIPFIT would be that other than EFIT data, it reads from and writes to an atomic part of the MDSplus data files. The same data file is updated by others at the same time. Synchronization can be challenge.

Since the initiation of a connection to the server is expensive, another option is to write a wrapper for the IDL code so that all the connections remain alive during the entire process. The shell scripts communicate with the wrappers rather than the IDL codes directly. Other options may involve modifications to MDSplus system.

The expanded STAR contains processors of two speeds. The load balancing can potentially be a problem. In the present system, only the first set of EFIT runs across the entire cluster. Afterwards, the cluster is effectively split into two smaller ones. Since EFIT runs fast, the benefit of load balancing would be very small comparing to if not cancelled by the complexity it would introduce. Another benefit of splitting the cluster is to process different analyses simultaneously when there is no data dependency. This is especially valuable when making partial results available as soon as possible is more important than completing the whole analysis quickly.

Regarding the size of the cluster, as far as computation is concerned, it can almost be said: the larger the better. For between-pulse analyses, however, there are many limiting factors. First of all, the analyses are performed upon experimental data. Nothing can be done before the raw data become available. In addition, there are other data dependencies. Secondly, the fact that an analysis is performed between pulses indicates that it completes in at most several minutes. Improvement in the computation time diminishes quickly if the non-parallelizable part cannot be improved. If one is not careful, this part can be made worse for it often involves network and data servers. More rapid analysis puts more demand on these shared resources. This leads to the

last but not the least important point: if the analyzed data cannot be served quickly to the research team, all is in vain. Overall, it is a cost performance analysis.

REFERENCE

[1] J.L. Luxon, et al., Proc. 11th European Conference on Controlled Fusion and Plasma Physics, 5–9 September 1983, Aachen, Federal Republic of Germany (European Physical Society, Aachen, 1983).

[2] Thomas L. Sterling, et al., "How to Build a Beowulf," The MIT Press.

[3] W. Gropp, et al., "A high-performance, portable implementation of the MPI message passing interface standard," Parallel Computing, Vol. **22**, 6 (1996) 789-828.

[4] Q. Peng, et al., "A Linux Cluster for Between-Pulse Magnetic Equilibrium Reconstructions and Other Processor Bound Analyses," to be published in Rev. Sci. Instrum.

[5] L.L. Lao, et al., Nucl. Fusion **25** (1985) 1611.

[6] P. Gohil, et al., "The Charge Exchange Recombination Diagnostic System on the DIII–D Tokamak," Proc. 14th IEEE/NPSS Symposium on Fusion Engineering, (San Diego, California, September 30–October 3, 1991) Vol. II, (1992) 1199.

[7] J.A. Stillerman, et al., "MDSplus data acquisition system," Rev. Sci. Instrum. **68** (1997) 939-942.

[8] Research Systems, Inc. http://www.rsinc.com.

ACKNOWLEDGMENT

Work supported by U.S. Department of Energy under Contracts DE-AC03-99ER54463 and DE-AC05-00OR22725.