A DISCUSSION ON DATA ACQUISITION, DATA MANAGEMENT AND REMOTE PARTICIPATION FOR ITER

T.W. Fredian,[1] M.J. Greenwald,[1] D.C. McCune,[2] D.P. Schissel,[3] J. Stillerman[1]

[1]Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139
[2]Princeton Plasma Physics Laboratory, P.O. Box 451, Princeton, New Jersey 08544
[3]General Atomics, P.O. Box 85608, San Diego, California 92186-5608

EXECUTIVE SUMMARY

This paper outlines a vision for data systems and remote collaboration for the ITER project. Outlined in this paper is an approach for creating the software infrastructure to satisfy the project's requirements and to maximize the value to the U.S. of ITER participatifon. It also emphasizes that the experience and accomplishments of the U.S. fusion program in these areas and our established links to those conducting relevant computer science research places our community at the forefront to carry out these tasks. We expect that software created for ITER will expand the boundaries of such technology and will likely be applicable to a broad range of scientific disciplines.

While it is clearly too early to choose an architecture or specific technologies for a data system that will not be deployed for over a decade, generic characteristics of such a system are now understood. Chief among these are the integration of all data (raw, processed simulations) and metadata into a coherent structure, allowing a common and powerful set of tools to work across the broadest range of applications, and the creation of a working environment for off-site personnel that is every bit as productive and engaging as when they are physically in the control room.

A two-phase solution is proposed where Phase I is an initial research and development effort to gather requirements and construct prototypes for design testing. This activity would lead to a comprehensive requirements document, a proposal for the data system architecture, and an implementation plan with a schedule consistent with ITER operations. Phase II would entail the implementation and deployment for the ITER community. Given the complexity of the problem, Phase I should begin early in the ITER construction phase to insure adequate time to develop a satisfactory solution. The U.S. fusion program has significant experience with all areas of data acquisition, data management, and remote participation. The MDSplus data acquisition and management system is presently used on over 30 experiments worldwide, becoming a *de facto* standard that greatly facilitates data sharing and collaborations across institutions. Most recently, the U.S. is creating a pilot collaboratory

project to prototype design solutions for large-scale collaborative activities in magnetic fusion research.

We, therefore, propose that the U.S. take primary responsibility for defining and implementing software for data acquisition, data management, and remote participation for the ITER experiment.

INTRODUCTION

The next major step planned for the worldwide Fusion Energy Sciences program is the construction, outside of the United States, of the International Thermonuclear Experimental Reactor (ITER). This $5B class device is expected to be operational by the middle of the next decade and would produce fusion power at the level of an industrial power plant. The importance and cost of this device requires that it operate at the highest possible level of scientific productivity. In this sense, it is useful to think of ITER as the largest and most expensive scientific instrument ever built for fusion research. It is the assertion of the authors, that for experiments as complex as those carried out in this field, scientific productivity is inextricably linked to the capability and usability of their data and computing systems. Such an effective infrastructure is required both for the success of the entire ITER project and will maximize the value of ITER to the U.S. program as well. Thus, careful consideration must be given to choices for architecture and technologies when designing a system that is so crucial to the overall success of the ITER project. Most importantly, the systems must be designed to meet the needs of the thousands of scientists and engineers who will use them.

The ITER device will be a unique collaboration for the fusion program, involving very large numbers of scientists from many different countries. And unlike large experimental collaborations in other fields, such as high-energy or nuclear physics which operate essentially in a “batch” mode, fusion experiments put a premium on near real-time interactions with data and among members of the team. It is reasonable to assume that not all members of the experimental team will be on-site for all experiments. In fact, it is probably desirable and practical to carry out a significant amount of the scientific work remotely. Effective international collaboration on this scale is a technically demanding problem since it requires the presentation of a working environment to off-site personnel for experimental operations that is every bit as productive and engaging as when they are physically in the control room. The technologies developed by this project will push the frontiers of data acquisition, data management, and remote participation and will be significant to a broad range of other scientific research disciplines.

ISSUES AND REQUIREMENTS FOR ITER DATA SYSTEMS AND REMOTE PARTICIPATION

Given the rapid rate of change in information technologies and the long interval before ITER operation, it is clearly premature to define in any detail, the implementation or design of its data software systems. However it is not too early to discuss some general capabilities that these systems must have. Data systems capabilities should include:

- A coherent view of data that is available through simple interfaces and tools.
- The integration of all data including raw, processed, simulations. This allows a common set of tools to work across the broadest range of applications. Artificial distinctions between data structures made on the basis of their origin lead to redundant efforts and impede scientific progress.
- Support for all needed data types and structures.
- The storage of all calibrations, geometry, set up information and analysis assumptions, etc., giving users a complete view of all data.
- Metadata (data about the data) for every data item. This would document, for example, where the data came from, when it was written, who was responsible for it as well as basic information on the data type, size, structure, etc., creating a coherent self-descriptive structure.
- The capability of being browsed.
- The capability of being queried. It should be possible to locate data based on its content as well as on its name, shot number, time, etc.
- Extendibility and flexibility to support an experiment that will operate for many years.
- The capabilities for integration of data acquisition, analysis, and visualization tasks.
- An easy learning path while maintaining powerful capabilities for experienced users.

Remote participation capabilities should include:

- The ability to fully and securely access the entire ITER data system by off-site collaborators.
- The ability to fully and securely access the entire analysis tool set for ITER including visualization codes, data analysis codes, and complex simulation codes.

- The ability to seamlessly communicate with multiple off-site locations including shared video and hands-free integrated audio.
- The ability to easily share complex scientific visualizations amongst remote participants and conduct an interactive discussion of the results.

With the scope of its mission and its increased pulse length, ITER will generate significantly more data than the current generation of experiments (which collect on the order of 1 Gbyte per pulse). Given the rapid growth in computing, communications and storage technologies, this increase in data volume is unlikely to exceed the raw technical capabilities of computer systems that will be available at that time. However, the creation of more data per pulse will challenge our ability to analyze and assimilate all of the data. Enhanced visualization tools will be required that allow this increasing data volume to be effectively used for decision making by the experimental team and to advance the science. Latency issues, associated with the movement of large quantities of data across intercontinental distances, will also likely come into play and require innovative solutions.

The computational challenge will be to perform more and more complex data analysis between plasma pulses. Improvements in plasma diagnostic techniques have made direct comparisons between experimental results and theoretical models a more common and more productive activity in the fusion program. For example the development of diagnostic instruments that can measure profiles of the electric and magnetic fields and make observations of the two dimensional structure of turbulent fluctuations has greatly improved the basic understanding of the mechanisms controlling plasma confinement. Today complete time-histories of the plasma magnetic structure including the effects of measured pressure and current profiles are available between pulses by using parallel processing on Linux Beowulf clusters. Five years ago, only selected times were analyzed between pulses with the entire time-history completed overnight. For ITER, more complex plasma simulations running on thousands of parallel nodes producing significant amounts of data will likely be performed between pulses and the results distributed to the entire team. Today, these comparisons are done over a period of days or weeks after experimental operations have concluded when it is far too late to adjust or optimize experimental conditions.

The magnetic fusion community's requirement for more efficient collaboration is well known and was identified in a review by the National Research Council [1]. More recently, the Integrated Simulation and Optimization of Fusion Systems report [2] concluded that to successfully model the entire burning plasma, a rich collaborative infrastructure will be required so that the geographically separated theoretical scientists can work together to create a unified software environment. The goal for ITER should be to present a working environment to off-site personnel for experimental operations that is every bit as productive and engaging as when they are physically in the control room. Effective remote participation in experiments on the scale envisioned will demand significantly enhanced and coordinated

resource sharing and problem solving in a dynamic, multi-institutional international environment. Key adjustments can be made to hardware/software controls only after vast amount of data has been assimilated in near real-time. Successful operation in this mode will require the movement of large quantities of data between pulses to computational clusters, to data servers, and to visualization tools used by an experimental and theoretical team distributed across the world and the sharing of remote visualizations and decision making back into the control room (Fig. 1).

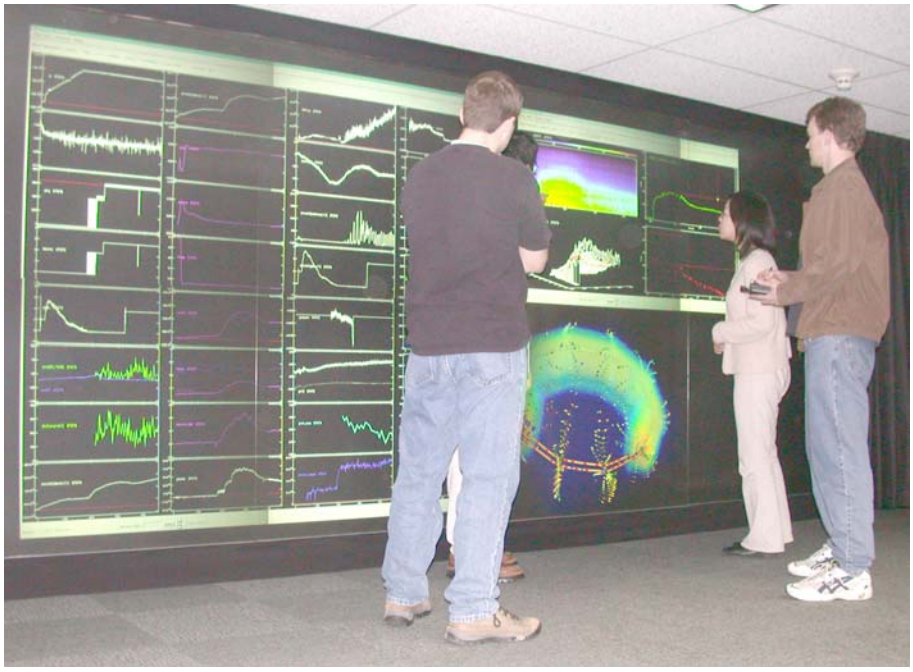

Fig. 1. Large tiled display walls like the one pictured at General Atomics might prove valuable in the ITER control room for collaborative group discussions to support real-time distributed decision making.

POSSIBLE SCOPE FOR U.S. PARTICIPATION

We propose that the U.S. takes primary responsibility for defining and implementing software for data acquisition, data management, and remote participation for ITER (Fig. 2). In the context of rapidly changing information technologies, this task would need to proceed in phases to avoid premature selection of technologies or architectures and to take best advantage of hardware platforms available during ITER operation. The initial research and development effort, Phase I, that is required to define a solution needs to be started early. Waiting too long into the ITER construction phase to start Phase I runs the risk of having inadequate time to develop a satisfactory solution.

For Phase I, we envision two parallel but closely related activities. The first would be to gather requirements from potential users based on their current plans and experience and

with reasonable extrapolations to ITER. (The most significant departure from current experience will likely be the extension to very long pulse lengths.) Those working on data intensive tasks, particularly diagnostics and analysis will be most helpful in this regard, but it will be important to get input from scientists and engineers from across the community. A survey of the approach and methodologies used by other large scientific projects would also be carried out. The second activity would be the construction of a series of prototypes to test concepts for the data system design and remote collaboration. The prototypes would be tested on actual or simulated experiments providing an opportunity to get early feedback from users. Experience with the prototypes would drive refinements in the design or changes to the underlying architecture. The cycle would continue until the design converged. Clearly, the two activities would be tightly coupled since the ultimate criteria is that the systems provided must facilitate the science and meet the anticipated needs of the users. These activities would converge leading to a comprehensive requirements document, a proposal for the data system architecture, and an implementation plan with a schedule consistent with ITER operations.

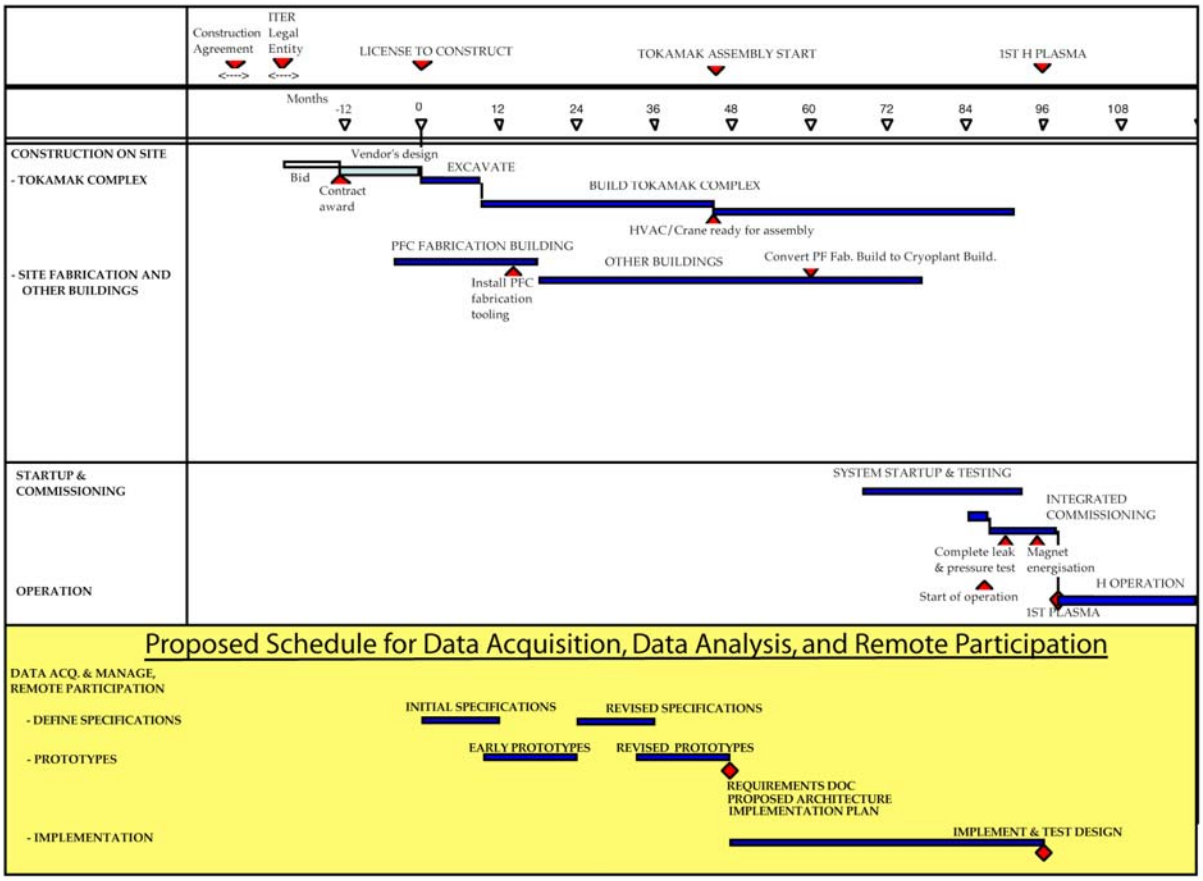

Fig. 2. The creation of an advanced data acquisition, data analysis, and remote participation infrastructure in time to support first plasma will require early planning and prototyping to insure the proper design architecture.

Phase II would involve implementation and deployment for the ITER community as well as continued testing to insure the design is working as specified. A larger team of software engineers with a broad scope of skills would need to be assembled and managed during this phase. A critical component of this activity will be the coordination with other development and construction activities especially those groups responsible for diagnostic systems, data analysis and machine control. The activity must also be integrated with IT infrastructure procurement and deployment, particularly the computer and communications systems. Finally, support must be given for the installation of local facilities at off-site locations to insure smooth integration into the entire ITER data system.

THE U.S. CAPABILITY TO CARRY OUT THIS PLAN

The U.S. fusion program has a proven track record in the areas of data acquisition, management, and remote participation. For example, MDSplus, developed jointly by MIT, LANL, and the IGI in Padua, Italy, is by far the most widely used (Fig. 3) data system in the international fusion program [3]. Based on a client/server model, MDSplus provides a hierarchical, self-descriptive structure for simple and complex data types [4,5]; the majority of the internals of the MDSplus system were designed and implemented in the U.S. Currently it is installed and used in a variety of ways by about 30 experiments. It is deployed as a complete data acquisition and analysis systems for C-Mod (MIT); RFX (IGI, Padua); TCV (EPFL, Switzerland); NSTX (PPPL); Heliac (ANU, Australia); MST (U. Wisconsin); HIT (U. Washington); CHS (NIFS, Japan); and LDX (MIT). It is used to store processed data for DIII–D, for the collaborative data archives assembled by the ITPA, and for the inputs and outputs of several widely used codes including EFIT, TRANSP, NIMROD and GS2. JET and ASDEX-Upgrade are using MDSplus as a remote interface to existing data stores and KSTAR has adopted it as a data acquisition engine for data stored in other formats. The result is a *de facto* standard that greatly facilitates data sharing and collaborations across institutions.

Historically, efforts to improve collaboration within the U.S. fusion community have included sharing of resources and co-development of tools mostly carried out on an *ad hoc* basis. The community has considerable experience in placing remote collaboration tools into the hands of real users [11]. The ability to remotely view operations and to control selected instrumentation and analysis tasks was demonstrated as early as 1992 [6]. Full remote operation of an entire tokamak experiment was tested in 1996 [7,8]. Today's experiments invariably involve a mix of local and remote researchers. Additionally, the U.S. fusion scientists have a solid working relationship with our European colleagues in the area of remote collaboration and remote participation. We would expect that if the U.S. put forth a solution for the ITER infrastructure, we would partner with some of these European colleagues. Decades of experience combined with a multi-national team should make for a powerful proposal.

Beginning in late 2001, the USDOE SciDAC initiative [9] funded the three-year National Fusion Collaboratory Project (NFC) [10,12]. This project builds on the past collaborative work performed within the U.S. fusion community and adds the component of computer science research done within the USDOE Office of Science, Office of Advanced Scientific Computer Research. The NFC is a pilot project, but nevertheless has as its overriding goal to enhance the scientific productivity of magnetic fusion research. The overall objective is to allow scientists at remote sites to participate as fully in experiments and computational activities as if they were working at a common site. This goal is being achieved by creating and deploying collaborative software tools. In its first year, the NFC has deployed a fusion computational and data grid as well as new and innovative collaborative visualization capabilities.

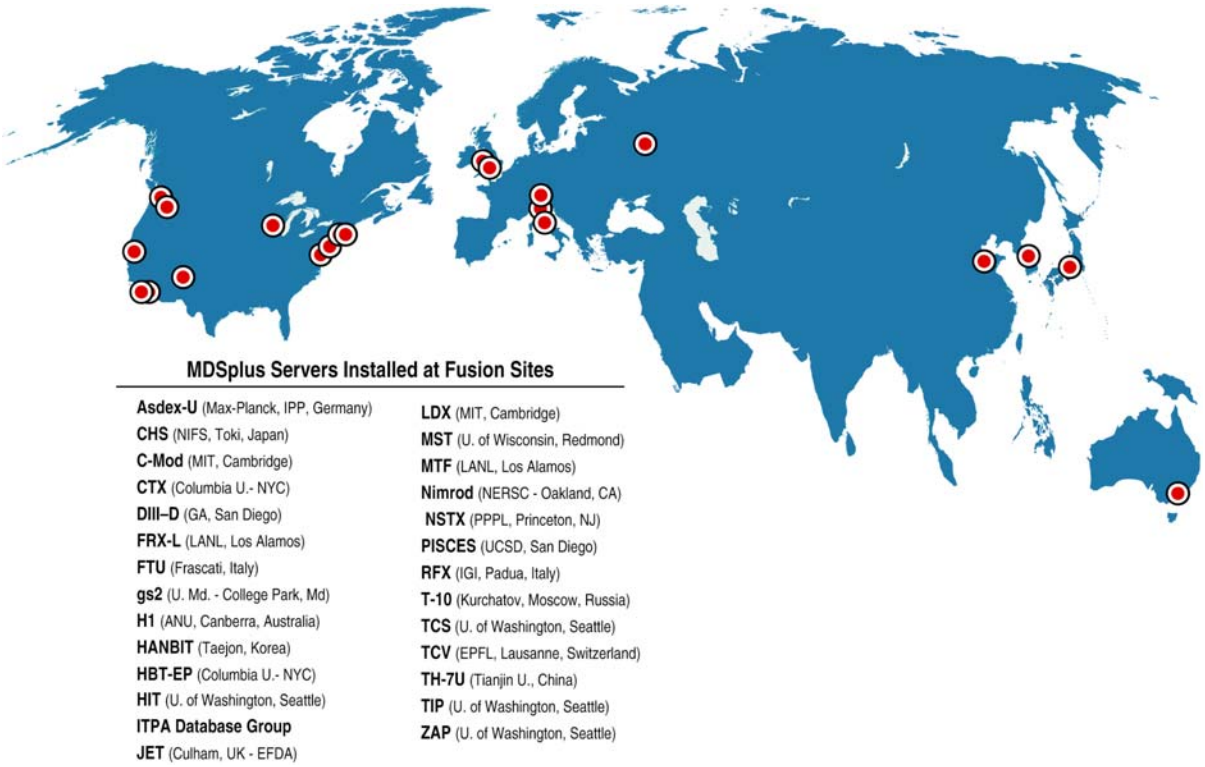

Fig. 3. The MDSplus data system is being used worldwide in magnetic fusion research and its adoption has greatly facilitated data sharing and collaborations across institutions.

CONCLUSION

The success of the ITER project requires a capable and integrated solution to data acquisition, data management, and remote participation. Due to the complexity of this problem and the importance of reaching a satisfactory solution, design work and testing needs to start early in the ITER construction phase. The experience and accomplishments of the U.S. fusion program in these areas and our established links to those conducting relevant

computer science research places our community at the forefront to carry out these tasks. Therefore, we propose that the U.S. takes primary responsibility for defining, testing, and implementing software for data acquisition, data management, and remote participation for the ITER experiment.

REFERENCES

[1] National Research Council, "An Assessment of the Department of Energy's Office of Fusion Energy Sciences Program," National Academy Press (2000).

[2] http://real.krellinst.org/isofs/.

[3] T.W. Fredian, J. Stillerman, "MDSplus, Current Developments and Future Directions," Fusion Eng. Design **60** (2002) 229.

[4] J.A. Stillerman, T.W. Fredian, K.A. Klare, G. Manduchi, Rev. Sci. Instrum. **68** (1997) 939.

[5] J. Stillerman, T.W. Fredian, "The MDSplus Data Acquisition System, Current Status and Future Directions," Fusion Eng. Design **43** (1999) 301.

[6] R. Fonck, et al., "Remote Operation of the TFTR BES Experiment From an Off-Site Location," Rev. Sci. Instrum. **63** (1992) 4803.

[7] S. Horne, M. Greenwald, T. Fredian, I. Hutchinson, B. LaBombard, J. Stillerman, Y. Takase, S. Wolfe, T. Casper, D. Butner, W. Meyer, and J. Moller, "Remote Control of Alcator C–Mod from LLNL," Fusion Technology **32** (1997) 52.

[8] B.B. McHarg, T.A. Casper, S. Davis, D. Greenwood, "Tools for Remote Collaboration on the DIII–D National Fusion Facility," Fusion Eng. Design **43** (1999) 343.

[9] http://www.osti.gov/scidac/.

[10] K. Keahey, T. Fredian, Q. Peng, D. Schissel, M. Thompson, I. Foster, M. Greenwald, D. McCune, "Computational Grids in action: the National Fusion Collaboratory," Future Generations Computer Systems **18** (2002) 1005.

[11] T. Fredian, J. Stillerman, "MDSplus Remote Collaboration Support — Internet and World Wide Web," Fusion Eng. Design **43** (1999) 327.

[12] http://www.fusiongrid.org/.

CONTACT INFORMATION

- Tom Fredian, MIT Plasma Science and Fusion Center, twf@psfc.mit.edu.
- Martin Greenwald, MIT Plasma Science and Fusion Center, g@psfc.mit.edu.

- Doug McCune, Princeton Plasma Physics Laboratory, dmccune@pppl.gov.
- David Schissel, General Atomics, schissel@fusion.gat.com.
- Josh Stillerman, MIT Plasma Science and Fusion Center, jas@psfc.mit.edu.