189
8-18-81

(2)

B6554

DR-2957

GA-A16272
UC-20f

MASTER

DATA MANAGEMENT IN A FUSION ENERGY RESEARCH EXPERIMENT

by
A. GLAD, D. DROBNIS, and B. McHARG

Prepared under
Contract DE-AT03-76ET51011
for the San Francisco Operations Office
Department of Energy

DATE PUBLISHED: JULY 1981

GENERAL ATOMIC COMPANY

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Printed in the United States of America
Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road
Springfield, Virginia 22161
Price Code: Printed Copy A02; Microfiche A01

**GA-A16272
UC-20f**

DATA MANAGEMENT IN A FUSION ENERGY RESEARCH EXPERIMENT

**by
A. GLAD, D. DROBNIS, and B. McHARG**

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

**Prepared under
Contract DE-AT03-76ET51011
for the San Francisco Operations Office
Department of Energy**

**GENERAL ATOMIC PROJECT 3344
DATE PUBLISHED: JULY 1981**

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

GENERAL ATOMIC COMPANY

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

CONTENTS

ABSTRACT	1
1. INTRODUCTION	2
2. FUSION RESEARCH AND THE DOUBLET III EXPERIMENTAL PROGRAM	4
3. THE DOUBLET III DATA ACQUISITION COMPUTER SYSTEM	8
4. MANAGEMENT OF DOUBLET III DATA	14
5. PLANNED EXPANSION AND INCREASE OF FUSION DATA.	22
6. ALTERNATIVES TO DATA REDUCTION	25
7. STANDARDIZING DATA	27
8. CONCLUSION	28
REFERENCES	30

FIGURES

1. Schematic of Doublet III	6
2. NMFE computer network.	7
3. Host minicomputer/microprocessor block diagram	9
4. Stages of data acquisition	11
5. Data file structure.	16

DATA MANAGEMENT
IN A
FUSION ENERGY RESEARCH EXPERIMENT*

A. Glad, D. Drobnis, and B. McHarg
General Atomic Company
P.O. Box 81608
San Diego, California 92138
U.S.A.

ABSTRACT

Present-day fusion research requires extensive support for the large amount of scientific data generated, bringing about three distinct problems computer systems must solve:

- 1) The processing of large amounts of data in very small time frames
- 2) The archiving, analyzing and managing of the entire data output for the project's lifetime
- 3) The standardization of data for the exchange of information between laboratories

The computer system supporting General Atomic's Doublet III tokamak, a project funded by the United States Department of Energy, is the first to encounter and address these problems through a system-wide data base structure.

1. INTRODUCTION

The purpose of this paper is threefold. First, to describe the current methods for managing large masses of data generated in the present-day fusion energy research efforts. Second, to identify the problems associated with handling the large data structures effectively and efficiently without overwhelming the physicists seeking experimental results. Third, to seek solutions to the unique problems associated with fusion data bases from members of other disciplines who may have insights and contributions for the fusion community. Fusion is the joining, or fusing, of two light atoms -- unlike fission, which is the splitting of a heavy atom. When fusion occurs, mass is converted to energy. It is this energy, in the form of heat or charged particles, that can be used to generate electricity or to provide power for process heat applications.

The early fusion experiments did not utilize computer aided acquisition and analysis of data for some sound reasons. Most of the research was channeled into the technology necessary to extract information from the fusion device itself and the experimentalists were satisfied with oscilloscope traces of analog signals as the means of data acquisition, and photographs of the displays as the means of archiving and analyzing the information. The lack of experienced computer professionals in the fusion community and the somewhat primitive, albeit costly, technology made the acquisition and analysis less attractive to experimentalists who became satisfied analyzing from photographs. Finally many experiments were funded in parallel, each given a goal resulting in small amounts of experimental data.

As the fusion community made significant advances and larger amounts of information became available, the old methods of analysis became obsolete and were replaced by computer-controlled data acquisition and analysis. The computer has become an integral part of

each experiment from early fusion devices, which generated approximately 100 kilobytes of data each day to the current devices which generate over 100 megabytes of data each day to future devices which will generate on the order of gigabytes of data each day.

The increasing amounts of scientific data generated by the experiments have created three problems which must be solved if the fusion program is to result in the creation of electrical power from the fusion reaction:

- 1) The acquisition, processing and archiving of the large mass of scientific data generated in the limited time frame of an experiment. This includes a summary output to the operational physicists used as a guide for intelligent experimental direction. The summary provides near real-time analysis, indicating whether all sensing devices are operational and whether the device is behaving as defined by the theoretical models and planned goals of the day's operation.
- 2) The archiving of all the scientific data generated during the life of the experiment. This historical archiving provides the physicist with the ability to analyze old data with new theories and methods and to make comparisons of a sequence of experiments in order to devise or evaluate an experimental plan.
- 3) The standardization of the data generated by all fusion laboratories, thus facilitating the exchange of data between laboratories. Because of the rising costs of new fusion technology, fewer devices are being built, causing an increase in demands for improved data acquisition and analysis for those which are built.

General Atomic's Doublet III tokamak fusion experiment, funded by the U.S. Department of Energy (DOE), has identified and addressed these problems utilizing a system-wide data base structure spanning multiple levels of computer processing.

2. FUSION RESEARCH AND THE DOUBLET III EXPERIMENTAL PROGRAM

The primary goal of fusion research is to harness in a controlled manner the thermonuclear processes occurring in the sun to produce energy for electricity or other purposes. In nuclear fusion, light elements, if heated and confined under conditions approximating those of the sun, will fuse together creating heavier elements and in the process converting some mass into energy. This is in contrast to nuclear fission where heavy elements are split, forming lighter elements and releasing energy.

The most likely reaction candidate for a first time fusion reactor is deuterium-tritium (D-T), these being two heavier isotopes of hydrogen. In the fusion process the nuclei of deuterium and tritium fuse to form a helium-4 nucleus and a neutron, and release 17.6 MeV of energy. This D-T reaction has a large reaction rate peaking around a temperature of 10 keV (100,000,000° C) for the reacting species, a temperature much lower than for other fusion reactions.

At 10 keV the D-T fuel exists as charged particles called a plasma and is confined by magnetic fields used to keep the plasma away from material walls. The leading candidate for a fusion reactor is the tokamak confinement concept which uses a doughnut shaped device with large magnetic fields to confine and shape the plasma.

The Doublet III tokamak experiment at General Atomic, Fig. 1, is one of the major fusion research projects in the United States. The goal of Doublet III is to study the properties of a high temperature hydrogen plasma approaching reactor-like conditions. Numerous individual plasma diagnostics are attached to the tokamak device to diagnose plasma heating, shaping, and confinement and to determine plasma properties such as temperature and density. The term "plasma diagnostic" will be used throughout the paper to define instrumentation used to extract plasma properties and to measure plasma characteristics. Doublet III currently has a plasma duration time, or "shot," lasting about one second. This can be repeated about once every five minutes.

Fusion research at General Atomic as well as at other laboratories generates large amounts of experimental data. To facilitate access to and analysis of this data, and to promote communication throughout the fusion community, the National Magnetic Fusion Energy Network (NMFE) was created, linking the major fusion labs to a central large scale computing facility. This facility, illustrated in Fig. 2, uses supercomputers such as the CDC-7600 and CRAY-1 for extensive and complex data analysis. General Atomic interfaces to this system through its User Service Center (USC) DECSYSTEM-10 computer, which provides access to the NMFE central computers. The USC also provides data archival, retrieval and analysis of the Doublet III experimental data, and supports other fusion research unrelated to Doublet III.

Fig. 1. Schematic of Doublet III

3. THE DOUBLET III DATA ACQUISITION COMPUTER SYSTEM

The Doublet III data acquisition computer system operates at three levels: minicomputer, microprocessor and USC DECsystem-10.

3.1. Minicomputer

The interface between the Doublet III experiment and the DECsystem-10 is via the data acquisition computer. This is a centralized minicomputer with numerous satellite microprocessor systems (Ref. 1). Acquisition and analysis of data is distributed between the minicomputer and microprocessor systems. The microprocessors remove some of the workload from the minicomputer and permit independent development and testing of a plasma diagnostic. Doublet III currently generates approximately 2 1/2 megabytes of data in each five minute time frame, termed a shot cycle. The host minicomputer system accomplishes three primary objectives: 1) acquiring all data for the experiment, including raw data from hardware devices and raw or processed data from the microprocessor systems; 2) archiving all data for permanent storage and retrieval; and 3) providing limited analysis for all data, thereby giving experimentalists critical information about the experiment as well as indicating the correct functioning of various diagnostics.

Figure 3 illustrates the hardware configuration of the data acquisition system for Doublet III. The host computer is a ModComp CLASSIC with 512K 16-bit words of memory. Four moving head disks provide storage for the operating system, user source and object programs, data base files, terminal scratch files, experimental data files, and auxiliary data analysis files. A high speed bulk core memory contains a portion of the acquisition data base requiring frequent and rapid access. Data archiving is accomplished on a 6250 bpi tape drive, with an 800 bpi drive for miscellaneous uses and as a

Fig. 3. Host minicomputer/microprocessor block diagram

back-up for archiving. A communications interface connects the minicomputer to graphics terminals, to some plasma diagnostics microprocessor systems, and a high speed synchronous link to the DECsystem-10. The data acquisition computer is linked to a control system computer (Ref. 2) via an inter-CPU link as well as by priority interrupts initiating different phases of data acquisition.

Most data is acquired through a multiplexed system of standardized modules known as CAMAC. This standard is in common use at nuclear and other laboratories in many countries, and has been adopted by virtually all fusion research projects. Data is collected by modules such as waveform digitizers, which capture waveforms as a function of time, and dual port memory modules shared with the microprocessor systems.

The software of the Doublet III data acquisition system consists of global tables in memory which characterize the data acquisition system, data base files on disk which characterize the data to be acquired, data files on disk which receive the experimental data, and the centralized tasks which accomplish the acquisition, archiving, and processing of data (Ref. 3). Figure 4 illustrates the stages of data acquisition.

Stage 0 occurs with reboot of the computer system. Tasks are executed to initialize the system to a state ready for immediate data acquisition, without human intervention or setup.

Stage 1 is a time prior to tokamak operation for acquisition data base generation and interactive modification. The acquisition data base, only a small subset of the data base discussed in this paper, contains information on the overall status of the acquisition system, each experiment, sequencing modes for acquisition and processing

Fig. 4. Stages of data acquisition

experiments, machine parameters and calculated experimental results, information characterizing individual signals, hardware waveform digitizer and analog-to-digital convertor structure, and summary data output formats.

Stage 2 is the time-dependent stage in which data is acquired and archived. This stage divides the shot cycle into a series of sequential synchronous phases permitting the orderly sequencing of important events in the shot cycle. The acquisition is sequenced to acquire the most important data immediately after a tokamak shot while data used for off-line analysis is input later after analysis of critical parameters is complete. An asynchronous phase occurs any time during the shot cycle and is normally used for the microprocessor acquisition.

All acquisition is performed by a central task that uses different overlays for acquiring different types of data. At the start of a shot, a data file manager task deletes the oldest data file and creates new files for the next shot. Approximately 20 shots are resident at the minicomputer level. After data is acquired, an archiving task reads data from disk and writes it to tape. Throughout this stage, a task manager controls the activation of various tasks executed during the course of data acquisition to insure that analysis commences as soon as the data has been acquired.

Stage 3 is a time of data examination. Some tasks are activated during data acquisition to automatically analyze certain data each shot and provide summary output while other tasks are executed at the discretion of the user. Data examination may influence the next tokamak shot, thus providing feedback as indicated in Fig. 4.

3.2. Satellite Microprocessor Systems

The microprocessor computer systems remove some acquisition and pre-processing work load from the host, providing output soon after a shot is complete. In addition, each plasma diagnostic can be developed, tested, and if necessary operated as an independent system without relying on the host.

The microprocessor hardware configurations, although different for each application, have a number of common features. Each microprocessor is a CAMAC packaged LSI-11 processor with 32K bytes memory and floppy disks with 200K bytes of storage. Each micro contains a hardware floating point chip to speed data analysis and a graphics terminal used to provide graphical output each shot. All microprocessors interface to the minicomputer through CAMAC dual port memories.

The primary software functions of the microprocessors are standard and include: 1) providing control of the experiment and acquiring all of the diagnostic data, 2) archiving the data to the host minicomputer and locally on the floppy disk, and 3) providing initial data reduction, analysis and output.

The microprocessor software systems are designed to operate in an unattended mode requiring operator intervention only to change full floppies and for physicists to examine samples of data output to verify correct operation of the plasma diagnostic.

3.3. DECsystem-10 Computer

The purpose of the General Atomic USC DECsystem-10 computer is threefold. First, the USC provides access to the NMFE large scale central computers (CDC-7600 and Cray-1) for both the theoretical and

experimental investigators at General Atomic requiring such resources. Such large scale requirements arise from the great mathematical complexity of the analysis of the plasma physics problems of interest.

Secondly, the USC provides data archival, retrieval, and analysis facilities for the Doublet III experiment. Use of the DECsystem-10 for this removes unfavorable interaction of later analysis with data collection and allows use of the proven DECsystem-10 file handling and time-sharing software. The problems of data retrieval and analysis are more concerned with the central processing power of the computer system. A planned upgrade of the USC to dual KL-10 processors is intended to satisfy this requirement so that the analysis of Doublet III data may proceed at a pace determined by the progress of the experiment.

Thirdly, the USC is used for local computing unrelated to the Doublet III program. This includes theoretical plasma physics investigations, and engineering and physics calculations for other U.S. DOE Development and Technology tasks.

4. MANAGEMENT OF DOUBLET III DATA

The structure and methods of managing the Doublet III diagnostic data follow a plan dating from the system's initial design. It was estimated that 2-4 Megabytes of data would be generated each shot to be distributed over three levels of computer processing: microprocessor, minicomputer and DECsystem-10 computer. The system was designed to fulfill the following requirements.

- 1) All data generated during a shot of the Doublet III machine will be acquired and archived in a central location.

- 2) All data archived should be in an unprocessed form eliminating the destruction and loss of data due to invalid processing at the acquisition stage.
- 3) The accessing of data by users should not require knowledge of internal data base structures.
- 4) The methods of accessing and processing data should be transferrable for all levels of processing with protocols to access data consistent throughout.

4.1. Centralized Minicomputer Data Structure

The acquired data on the centralized minicomputer is organized into a series of files. Each file contains all the information for a specific shot and diagnostic experiment. The shot numbers for the tokamak device are unique and are generated in sequential ascending order. Within the file, whose structure is illustrated in Fig. 5, there is a directory defining the names of all signals acquired for each plasma diagnostic, the status of the acquisition, and in what phase of acquisition they were acquired. For some plasma diagnostics, data is acquired under the same name for calibration or baseline information prior to a shot and for the actual data after a shot. Finally, a pointer in the directory defines the start of the actual data located in the data file.

The data located in the data file has its own structure, which is described by a header appended to the data at the time of acquisition. The header contains all information related to the signal and necessary to process the data. This header simplifies the accessing of data by applications programs. The system services need only search the directory, locate the point, ensure no fatal errors occurred during acquisition, read in and return to the user both the data and header.

Fig. 5. Data file structure

A brief explanation of the acquisition data base files will clarify the need for a header appended to the data.

The Point Parameters Definition (PPD) files consist of an entry for each signal with information necessary to acquire, process and archive the data. Information includes name, type, description, experiment, engineering units, CAMAC address, processing limits, conversion factors, archiving format, etc. The data acquisition program utilizes this information when processing the points.

The Digitizer Definition (TDD) files contain all the information related to the waveform digitizers and their corresponding integrators. Information contained in the digitizer file includes memory size, external switch settings, sampling frequency, gain factors, and offsets. Each PPD record of a digitizer signal is linked to a record in the TDD file. Only modifications to the links are necessary to reflect the change in the hardware configuration.

The Scan Control (SCAD) file defines all the signals which are to be processed for a given experiment/phase. The information contained in this file is a list of signal names with processing status. This file is used by the data acquisition task to determine which PPD files to process for a given experiment phase and is necessary because a signal can be processed in more than one phase.

The Experiment Control Table (ECT) contains all the information for each experiment on the system. This file is divided into three parts: a global part (experiment information) containing information such as experiment name, current data file sizes, terminal information for graphical output, task file assignment, and experiment status; a data file part containing the list and status of all data files for the experiment which are currently stored locally; a task file part containing the list of all the tasks which should be run each phase on behalf of the experiment. The data acquisition task checks experiment status and sets the data initialized and acquisition complete for the experiment.

The Phase Control Table (PCT) contains a list of experiments to process for each phase and the order in which they should be processed. This table is utilized by the data acquisition task to sequence the experiments for a given phase.

The Diagnostics Information Table (DIA) contains general information both about the shot and phases of acquisition. The shot information includes current shot number, last shot data files created, last shot for complete data acquisition, and shot status. The phase portion includes the current phase being processed, next phase to be processed and phase status. This table is utilized by the data acquisition task for phase sequencing and current shot number.

The Reduced Data Table (RDATA) consists of the summary of all critical parameters of the shot and all calculated values. This information is used to summarize and identify the shot. The data acquisition task inserts both control and diagnostic values into the table and each experimental analysis task inserts its results.

With the information about the signal scattered over several files, accessing these files to process the data is unacceptable, making the data file header necessary. As an example, consider an analog waveform digitized over the one-second plasma duration into 4096 digital samples. Contained in the TDD file is information specifying the gain, offset and drift of that hardware channel used to convert a digital number to an input voltage and the addressing information for the hardware sensor. Contained in the PPD file is information defining the sequence of commands necessary to acquire the data, and an inherent number used to convert the voltage input of the channel to actual physical units.

Header information for various signals must take into account several differences. First, all hardware sensors are not the same and the information contained in the header would be different for the different sensors. Second, the initial header for a signal type can change as more information about the signal becomes available. Third,

certain information is standard for all points. For these reasons, the structure of the header was divided into two parts, a fixed part containing all information common to all signals and a variable part containing specific information for the signal type. By defining standards, the managing of different header types is done in the system services accessing the data.

The fixed header is assigned a point format index (PFI) while the variable headers have data format index (DFI) numbers. These numbers specify the format of the header and data. As the system developed, more information became available and both header parts changed. The system services were designed to handle old and new PFI/DFI types thus making access to old and new data transparent, eliminating the user's need to know a specific shot number when the PFI/DFI changed.

The fixed header contains information used to identify and verify data integrity including signal name, description, experiment, phase of acquisition, shot number, date and time, plasma type, and engineering unit. The variable part contains information such as calibration numbers, conversion factors, number of data words following, format of the data (integer, ASCII, floating point, etc.) and width of data elements.

For all signal data acquired directly by the minicomputer, both headers are built by the centralized data acquisition tasks. For all signal data acquired from the microprocessor, the minicomputer builds the fixed header with the variable header built by the microprocessor and appended to the data.

Finally, the services to access the data were designed with defaults to simplify access by a user. Items which identify the data such as the shot number, phase of acquisition, or plasma diagnostic, are defaulted. For most real-time analysis, experimentalists are examining the current shot to get information about experimental results, and so the shot number is defaulted to the current shot. Since most data examined at this level is shot data, the phase is

defaulted to the plasma data acquisition phase. Since every signal name is unique, it is redundant for a user to specify the plasma diagnostic. Therefore, a physicist accessing data needs to know only the signal name and physical meaning of the data to perform analysis.

4.2. Microprocessor Acquisition

The microprocessor systems acquire and process data related to a specific experiment. The amount of data acquired by a microprocessor may range from 16K 16-bit words to 400-500K words generated by the latest state-of-the-art image processing systems. In most cases, the raw data is too large and costly to transmit directly to the minicomputer system, thus requiring each microprocessor to do preliminary data reduction and processing. This reduction helps in the management of data at the minicomputer and DECsystem-10 level processing. At the current time, most microprocessors transmit approximately 16K words of data.

The sequence of microprocessor operation is to acquire raw data, do processing and data reduction, append the variable header to the data, and transmit to the minicomputer. In addition, the data is normally archived locally on floppy disk. This floppy disk archiving is done both to preserve the data if the minicomputer is down, thus allowing transmission at a later time, and to preserve the data in its rawest possible form. The data written to the floppy is structured as sequential data files with a header containing all of the physics parameters necessary for processing. The structure on the floppy disk varies among microprocessor systems since each system is dedicated to its own plasma diagnostic. The applications tasks on the microprocessor systems are transferrable to other computing levels because these tasks operate off the reduced data transferred to the minicomputer system and are written in high level languages (e.g., FORTRAN).

4.3. DECsystem-10 Computing Network

Finally, the USC DECsystem-10 computing network functions to archive all data for all the plasma shots for all experiments and to provide access and analysis to any subset of this data. The minicomputer transmits the data and directory files for each experiment. The DECsystem-10 combines all experimental data for a specific shot into one large data file and a corresponding directory file. Once the data is processed into a single data file, an entry is made in a master file index (MFI), the files are backed up to tape and purged from the disk. The only shots remaining on disk are those specifically requested by users. This is necessary because the DECsystem-10 has the capability of maintaining only 250 shots of data disk resident from the experiment which generates about 100 shots per day. The MFI keeps track of whether or not a shot has been processed and, if so, whether it is currently resident on disk or what tape number must be mounted to restore the data to disk. When a data file is restored to disk, a least-recently-accessed algorithm is used to delete a data set.

The conversion of the minicomputer data to DECsystem-10 formatted tapes takes into account the necessary transition from 16-bit to 36-bit word sizes. The net result is two back-up tapes, the original minicomputer tape and the DECsystem-10 formatted tape. A request for data not on disk is automatically translated to an operator mount request at the DECsystem-10. The DECsystem-10 uses its formatted tape to restore the data, eliminating reprocessing.

The file structure for a shot on the DECsystem-10 is structured similar to the minicomputer directory and data file. The DECsystem-10 has two files, a directory containing a list of all signal names acquired for each shot and pointer into the data file where the data image is stored. The directory also contains the status of the data indicating errors in acquisition or processing.

All of the tasks that access the data on the DECsystem-10 use the same high level FORTRAN interface routine calling sequence that is available on the minicomputer system.

4.4. Accomplishment of Objectives

The objectives of the initial system design have been met for the current system. All of the plasma data is acquired by the centralized minicomputer system, combined with and archived to the DECsystem-10 which maintains the data for all experiments and shots. All the diagnostic data archived is in the rawest form possible. Data acquired by microprocessors, however, is pre-processed for the purposes of transmitting more meaningful data and reducing the amount archived (e.g., image processing reduction).

The experimentalist needn't know internal data base structures to access data. The interface routines interpret the format specifications in the data, and return meaningful information.

The incompatibilities between the many levels of processing used by experimentalists were reduced as much as possible. The data base structure and access routines are patterned the same on all levels. Most generalized applications tools and user generated applications tasks are transferrable between machines.

5. PLANNED EXPANSION AND INCREASE OF FUSION DATA

Several general trends are apparent among fusion research projects as a class, and at the Doublet III project in particular. These include a trend by users of the acquired data toward increasingly complex analysis in real time and in an off-line mode; a proliferation in the number and complexity of the techniques used for plasma diagnostics; and an increase in the pulse times of operating

fusion devices resulting in a linear increase in the amount of data to be acquired. Future development of systems for fusion data bases must consider these trends.

5.1. Doublet III System Growth

The demand upon the Doublet III system has grown from the time of its initial operation in 1978 from about 500 kilobytes of data per shot to over two megabytes at present. The minicomputer which acquires and archives data has been upgraded with faster CPU and more disks, and is linked to microprocessors which control and preprocess more sophisticated plasma diagnostics. A hardwired link between the minicomputer and DECsystem-10 is being implemented for on-line near real-time archiving of the data.

Continued addition of plasma instrumentation to Doublet III, improvements in the desired quick-look analysis, and lengthening in the effective plasma duration time are expected to produce, over the next two years an additional two-fold increase in the amount of data acquired and archived per shot. This will act to increase the load upon both the data base and computational parts of the system. In order to meet these needs, the system will be further decentralized, and the computing power and data storage at each level increased.

The microprocessors now used with plasma instruments are the LSI-11/23 type. The minicomputer is tentatively planned for upgrade to a dual CPU shared memory configuration with dual-ported disk storage. Data from the current shot will be held in shared memory, eliminating contention for the most commonly accessed data. This will also help to decouple the archiving and summary or quick look processing functions.

Similarly the DECSys-10 is being upgraded to a dual KL-10 CPU configuration, with increased disc capacity, to accomodate more users interacting with more data and performing more complex analysis. Plans for upgrading the MFE computer center are also under study by the DOE.

5.2. New Fusion Project Designs

The upgrades to Doublet III's data base capability are in line with the trend in data system designs at the next generation of fusion devices. The Tokamak Fusion Test Reactor (TFTR), under construction at the Princeton Plasma Physics Laboratory, includes a data collection and quick-look processing system which utilizes a network of six large 32-bit minicomputers, with eight similar but smaller subsystem computers (Ref. 4). A similar concept is used for the Mirror Fusion Test Facility (MFTF), under construction at the DOE's Lawrence Livermore Laboratory to test a somewhat different technique of plasma confinement. Here, a network of nine 32-bit minicomputers perform the acquisition, control, and display functions, while up to 55 PDP-11/03 computers provide for front-end control and acquisition of data from the entire experiment, including all plasma instrumentation (Ref. 5). At MFTF, all data base management is performed at a central node in the network and information supplied to and from it as needed elsewhere in the network. The designers of both these systems feel, on the other hand, that the storage and computational facilities in their local networks are sufficiently great that an intimate connection to the MFE computing network is not necessary. Other large fusion projects elsewhere in the world include the Joint European Torus (JET), under construction at Culham Laboratory in the United Kingdom, and JT-60 in Japan. JET will employ a network of 18 minicomputers (Ref. 6); JT-60 similarly employs a large multi-computer network for control and data acquisition (Ref. 7).

All of these projects expect to operate with plasma durations of from several seconds to perhaps as long as 25-30 seconds, with a repetition rate similar to Doublet III's five minutes. Initial expectations are typically for about 10 megabytes of data per shot. If Doublet III's experience of an order of magnitude increase in data quantity over the life of the project is borne out, these projects will each be generating by the middle of the decade roughly a thousand megabytes of data during each hour of operation.

6. ALTERNATIVES TO DATA REDUCTION

There are some alternatives which could attempt to reduce the ever-increasing data base which contains all the information and data generated by an experiment.

The first alternative is to reduce the amount of physics data by utilizing the hardware capabilities to the utmost. Consider the following example of digitized analog waveforms, which comprise over half of the data. These waveforms are currently sampled continuously at the highest frequency necessary for examination. As the plasma time increases, the waveforms will generate a corresponding increase in data for each signal. From current experience, only a portion of the signal at a critical point in the plasma needs the high frequency sampling with the remainder of the signal being sampled at a lower frequency. If variable sampling frequencies were used over a shot, then as the plasma time increased, the waveform data increases would not increase linearly. This would better utilize the existing hardware and create a reduction in cost for acquiring and archiving the diagnostic data. It would, however, add complexity to the services accessing the data.

A second alternative is to eliminate irrelevant data from the archived data base. Most signals are hardware controlled to capture a signal for a set time frame. If the plasma duration does not meet the setpoints in the hardware, the data collected after the plasma has terminated could be discarded. In addition, many shots are aborted through hardware malfunctions resulting in no plasma. These shots could be marked by the minicomputer system and discarded at the DECsystem-10 level. Unfortunately, care must be taken not to discard valid data by mistake. This is why the initial design objective for all raw data to be archived regardless of the results of the shot was made. Highly skilled physicists are presently needed to determine if data is valid.

But no matter how well the above alternatives are implemented, the amount of data will increase. It is apparent that studies must be made to discover effective algorithms to reduce the data base. There is a somewhat representative data base structure existing on the Doublet III system (approximately 18,000 shots) which could be studied to determine algorithms which most effectively reduce or compress the data. The algorithms must have the following objectives:

- 1) Not changing the data but only representing it in a compressed or reduced form.
- 2) The compression must be worth the effort, reducing data by a factor of over two.
- 3) The compression and uncompression must be efficient and must not provide heavy computing burdens and long time delays to users.

7. STANDARDIZING DATA

It will become increasingly important to standardize the data base management and methods for accessing fusion experimental data, thereby making the most efficient use of future devices and personnel in the fusion community. The cost of financing a device is growing at such a tremendous rate that the building of new devices is limited. Countries are combining their efforts internally (i.e., FED in the USA) and internationally (i.e., JET in Europe). Doublet III now receives support from the Japanese Atomic Energy Research Institute (JAERI), who share the experimental results. This means that the data generated at these devices must be available to all fusion research laboratories because experimental results will be compared between devices in an attempt to determine which experimental approach is more feasible for power generation.

At the Doublet III facility, physics data is available on the NMFE network for any laboratory to access and analyze but only the General Atomic and JAERI physicists directly involved in the research have done so.

There are two generalized system-wide methods designed to access and process the data on the Doublet III. First is a high level FORTRAN access routine which operates on the minicomputer and DECsystem-10. By knowing the signal name and shot number, the user can access any portion of that data set. Once the data is accessed, the user still needs collaboration with the physicist responsible for the data to determine its physical meaning.

With data access standardized, certain general purpose processing programs were developed: a general purpose graphics program PLOTTER is used to perform rudimentary manipulation of the waveform digitizer data; a program to extract data for a magnetohydrodynamic (MHD) analysis, MDAT; a program performing complex calculations by emulating a vectorized HP-45 calculator, ODDBALL, originally developed at Oak

Ridge National Laboratory (ORNL) was adapted to Doublet III data; and an FFT program used to process sophisticated waveform data. These general purpose programs along with a majority of the analysis programs related to the specific experiment utilize the common interface to the data base and operate on all levels of computer processing (Ref. 8).

These standards of diagnostic data created at General Atomic make an initial attempt to simplify the processing by common routines and general purpose programs. By adapting the current analysis tasks to all levels of computer processing, experimentalists can make use of these levels without learning a new human interface. The sharing of these routines will be an important step to solving the future problems of the ever-increasing data base since no single laboratory will be able to staff enough qualified personnel to analyze the enormous amount of data to be generated.

8. CONCLUSION

Until now, fusion technology has ignored the problems of the increasing size of its physics data base. Previous conferences on fusion technology did not devote time to details. A large percentage of papers described the overall systems without delving into the detailed aspects of data base management. Concerns now are that progress is not being made in managing the fusion data base.

It is the authors' opinion that future fusion technology conferences should allocate time in the form of workshops, discussion sessions and expanded papers to address data base technology with experts in data base technology from other disciplines invited to

speak. The different laboratories should exchange information on a regular basis and be organized by the funding hierarchy (i.e., DOE) to develop standards for data management.

It is hoped that this paper has conveyed the current status of fusion research data base management, and, most importantly, the problems facing the fusion community in the future due to the ever-increasing demand for more data.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Pam Henline and Glen Shephard for their technical consultations and to Peggy Ainsworth for her efforts and patience in getting the final draft completed.

*Work supported by the U. S. Department of Energy, Contract No. DE-AT03-76ET51011.

REFERENCES

1. A. Glad, et al., "Aquisition, Archiving and Analysis of Doublet III Diagnostic Data on a Distributed Computer System," GA Report GA-A16048.
2. D. Drobnis, et al., "Computer Control of Tokamak Operation at Doublet III," GA Report GA-A15661.
3. B. McHarg, "A Guide to the Data Acquisition Software of the Doublet III Computer System," GA Report GA-A16134.
4. TFTR Final Design Report, Chapter 9, "Central Instrumentation Control and Data Acquisition (CICADA) System," PPPL 1475, PH-R007, August 1978.
5. W. G. Labiak, "Software for the Local Control and Instrumentation System for MFTF," Proceedings of the Eighth Symposium on Engineering Problems of Fusion Research, IEEE, 1979, Vol. 4, pp. 1991-1994.
6. F. Bombi, et al., "The Control and Data Acquisition System of JET - Status Report," Proceedings of the Eleventh Symposium on Fusion Technology, Culham Laboratory, 1980.
7. K. Kawasaki, et al., "Control and Monitoring System of JT-60 Machine I," Proceedings of the Eighth Symposium on Engineering Problems of Fusion Research, Ibid., pp. 1978-1981.
8. P. Henline, "Computer Aided Analysis of Doublet III Data," GA Report GA-A16149.