GA–A23737

DIII–D DATA MANAGEMENT

by
B.B. McHARG, JR., J.R. BURUSS, J. FREEMAN, C.T. PARKER,
J. SCHACHTER, and D.P. SCHISSEL

AUGUST 2001

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.

GA–A23737

DIII–D DATA MANAGEMENT

by
B.B. McHARG, JR., J.R. BURUSS, J. FREEMAN, C.T. PARKER,
J. SCHACHTER, and D.P. SCHISSEL

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.*

Work supported by
the U.S. Department of Energy under
Contract No. DE-AC03-99ER54463

GENERAL ATOMICS PROJECT 30033
AUGUST 2001

ABSTRACT

The DIII–D tokamak at the DIII–D National Fusion Facility routinely acquires ~500 Megabytes of raw data per pulse of the experiment through a centralized data management system. It is expected that in FY01, nearly one Terabyte of data will be acquired. In addition there are several diagnostics, which are not part of the centralized system, which acquire hundreds of megabytes of raw data per pulse. There is also a growing suite of codes running between pulses that produce analyzed data, which add ~10 Megabytes per pulse with total disk usage of about 100 Gigabytes. A relational database system has been introduced which further adds to the overall data load. In recent years there has been an order of magnitude increase in magnetic disk space devoted to raw data and a Hierarchical Storage Management system (HSM) was implemented to allow 7×24 unattended access to raw data. The management of all of the data is a significant and growing challenge as the quantities of both raw and analyzed data are expected to continue to increase in the future. This paper will examine the experiences of the approaches that have been taken in management of the data and plans for the continued growth of the data quantity.

1. INTRODUCTION

DIII–D is a large tokamak fusion research experiment at the DIII–D National Fusion Facility operated by General Atomics and funded by the United States Department of Energy. The mission of the DIII–D program is to establish the scientific basis for the optimization of the tokamak approach to fusion energy production.

DIII–D is a pulsed experiment with plasma discharges (called "shots") of 5–10 s duration, 30–40 discharges per day, and operations for 15–18 weeks per year. Depending on the mode of operation, a discharge has generated as much as 513 Megabytes of raw data. Through the history of the experiment since 1986, over 4 Terabytes of raw data have been generated as illustrated by the accumulated raw data in Fig. 1. Since data is compressed, the actual storage space required is about half this size. The amount of raw data is expected to continue to increase further. A general description of the DIII–D computing environment can be found in Ref. [1].

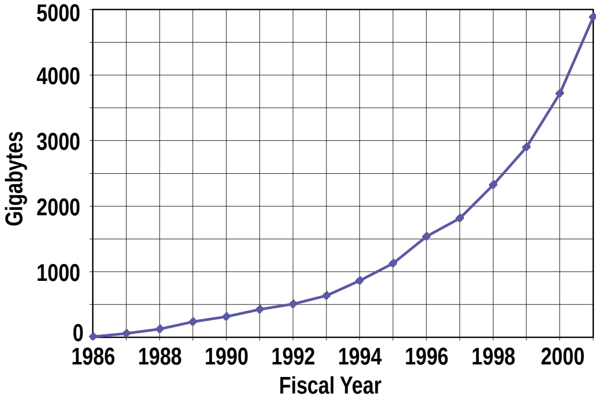

Fig. 1. Cumulative raw data acquired by the DIII–D experiment.

Besides the raw data there is an increasing amount of analyzed data generated, and more recently, selected highlights are being put into a relational database. The analyzed data currently occupies about 100 Gigabytes, with about 10 Megabytes being added per shot. This quantity continues to increase as advances in research requires more analysis between discharges, the results of which are stored as analyzed data. There are also new diagnostics, such as the infrared TV (IRTV) data acquired from digital cameras, which alone generate so much raw data that they do not presently fit into the centralized data management scheme.

The management of the storage of all this data is very challenging as data must be acquired, archived, made available to users, and restored upon request. In recent years improvements to storage management have involved increasing the amount of magnetic disk space, introducing a hierarchical storage management system, and the evolution of the three types of data into separate server systems. This paper will discuss the ways in which the various types of data are managed, the experiences of the approaches that have been taken in management of the data, and future issues and plans for data management.

2. RAW DATA

Raw data is acquired by a large number of diagnostics. All of these diagnostics first acquire and write data onto local disks, data is compressed, and then transferred to a central raw data shot server. Throughout this process, the data signals are always available for access via a PTDATA data access routine, and access is available over the wide area network [2].

By 1997, during times of high data production by the experiment, or when large numbers of requests were made for older data, the data management system had inadequate performance. Restores of data required an operator to mount archive tapes, and the operators were not staffed 24 hours/day, 7days/week. Frequently the residence time of a shot file on disk was limited to 24 hours because of the demand.

To alleviate the poor performance, a number of improvements were made. A separate shot server computer was purchased with initially a 100 Gigabyte RAID disk array, an increase of 2.5 times over previous data storage allocations. In addition an HP optical jukebox of ~600 Gigabytes and an ATL DLT tape robot with an ultimate capacity of ~3.3 Terabytes was purchased. Hierarchical Storage Management (HSM) software was purchased from Veritas to manage the migration from magnetic disk to optical media to tape media and back. This new system went into service in early 1998. It required over a year to get all older raw data restored and into the HSM system after which it was accessible 24 hours/day, 7 days/week. In early 2000 the RAID array was doubled to 200 Gigabytes and in late 2000 an additional 400 Gigabytes of RAID disk were added. This provides magnetic disk space for ~3000 complete, compressed shots. However since the shot is made up of ~20 files, not all of which need to be on disk, the space currently provides for ~14000 partial, compressed shots.

The HSM system achieved the goal of 7×24 unattended access to data, but at the expense of a substantial amount of time devoted to system management of the software and the development of routines to manage the HSM software efficiently. Figure 2 is a block diagram illustrating the life cycle of a shot data file through the HSM system. There are three functional parts to this system as indicated by the three dashed rectangles in the figure: HSM data management which encompasses the online availability of all data, HSM archiving management which involves the permanent archiving of the data, and pool management which is a separate large data area that is not directly part of the HSM system. These parts are discussed in the following paragraphs.

For HSM data management (upper and right part of Fig. 2), initially two functionally identical HSM volumes for raw data were created, /d3data0 and /d3data1 (large solid rectangle in the figure). Only one was actually needed, but issues related to the RAID disk array led to the necessity of two. As new data enters the system it is written to the HSM volumes. If a file has not

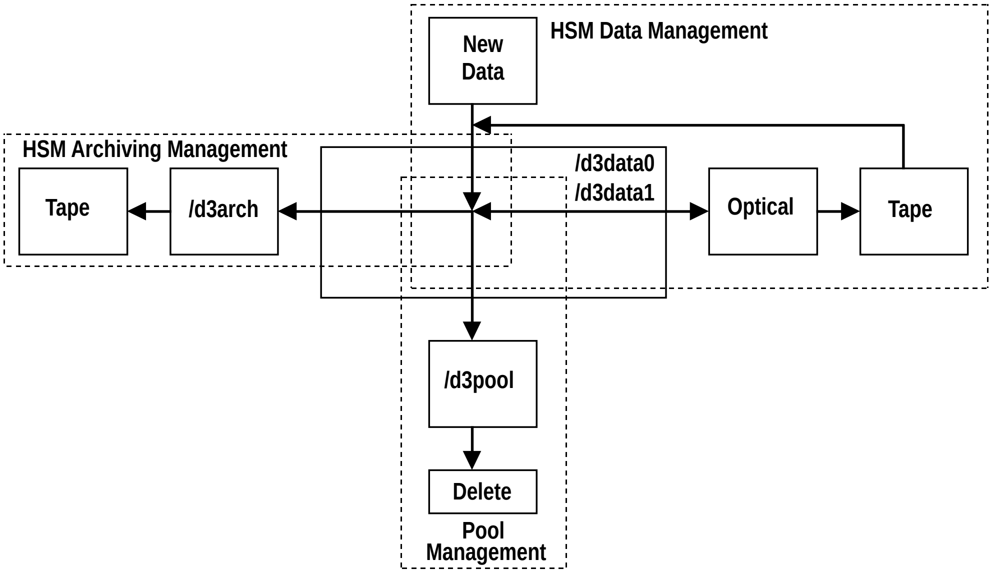

Fig. 2. The life cycle of a shot file in the HSM system.

been accessed after 14 days then it becomes eligible for migration from disk to optical media. The HSM software will automatically migrate files if the HSM data volume becomes too full. The restore of data from optical media happens automatically when that data file is accessed. During times of very heavy restore activity or new data generation, the 14 days has had to be lowered to accommodate the influx of data.

Data is typically on optical media for about 3 months and if not accessed becomes eligible for migration to tape, but this procedure must be run manually and can take a long time depending on when the last migration was run. After the migration, a process of consolidating optical media must be performed. Restore from tape happens automatically when the data file is accessed, but can take a highly variable amount of time. When data is restored from tape, it becomes eligible for removal from disk after 14 days and would then reside on tape again. However the assumption is made that if data was restored, it has some importance and might be looked at more frequently. Thus the file is renewed, that is its lifecycle is started over just as if it were new data.

When another 400 GB of magnetic disk space was added in late 2000, it was decided not to incorporate it into the actual HSM system partly because of the many difficulties that had occurred with that software. To provide more control over the data area, the new space was designed as a large pool volume separate from the HSM volumes. This is the Pool management (lower part of Fig. 2). As the HSM volumes fill, older data is copied to the pool volume, /d3pool, and the corresponding data on the HSM is forced to migrate to optical media. As the pool volume fills, the files with the oldest access date are deleted, but they still exist on secondary media in the real HSM system. Scripts manage all of these procedures. This additional disk space has resulted in dramatic improvements in the overall efficiency of the raw data system because more data is on disk and there are fewer restore requests. Data exists in the pool for 3–4 months.

It was found that it would be convenient for the archiving of data to be through the HSM system and this is the HSMarchiving management (left part of Fig. 2). A small disk volume was created, /d3arch, which is the HSMarchiving volume. As data is brought into the regular HSM system, it is copied by scripts to the archiving volume. Once a day, unarchived files are automatically copied to two DLT tapes and then removed from /d3arch. One copy is kept in the computer center with the shot server. The other copy is kept offsite for disaster protection. This uses the same tape robot as the HSM, but uses different tapes.

If the HSM software worked perfectly, then what has been described would require little system management. However the software is very “fragile”. In the initial use of the software, there was some type of error condition almost every day, and usually it was a condition that had not happened before. Recovering from errors was often very time consuming. Examples of these error conditions included “stuck” media and corruption of the HSM database. The HSM software and its reliability have improved over time. Many bugs have been fixed, though some still remain.

3. ANALYZED DATA

Meaningful scientific interpretation of the data requires analysis of the raw data and these analysis results need to be stored for later access. Researchers from different institutions or even within the same institution often structured their analyzed data differently and stored them in different locations. A convenient and uniform method of access and location was clearly needed.

Beginning in 1997–1998 a new methodology was created whereby analyzed data was put into MDSplus [3]. This method offered a convenient way to organize the analyzed data, to compress the data, and to provide a uniform location and uniform access method. A server system was implemented called the MDSplus server. This is a tru64 UNIX Alpha Server with 300 Gigabytes of RAID disk. Only about 100 Gigabytes is currently filled with analyzed data, but that is growing at the rate of about 10 Megabytes per shot. At the time of installation it was felt that disk prices were sufficiently low and would continue to drop so that all analyzed data could be maintained online. This does appear to be the case for the foreseeable future [4].

Access to this data is via MDSplus routines. MDSplus also has wide area access. Data is written to the MDSplus server from various diagnostics or analysis codes on many different computers, in particular by between shot analysis for MHD equilibrium (EFIT) as well as analysis of other diagnostic data [5].

The MDSplus data system contains the capability of multiple analysis versions and updating or replacing files. Because of the changing nature of MDSplus files, they are backed up on a regular basis. This is done as a differential incremental backup whereby the incremental includes all changed files since the last full backup rather than just changed files since the last incremental. Should the disks fail, this method provides for faster restoration of the data. Once the differential incremental becomes a significant fraction of the size of the full backup, then a full backup is done again. This procedure is somewhat reasonable now, but as this repository grows in size, it will become more difficult to continue this method because the size of the full backup will lead to a large increase in backup time.

4. RELATIONAL DATABASE DATA

The newest part of the triad of data types is the relational database. Selected analysis highlights and summary discharge quantities are placed into a Microsoft SQL server 7.0 relational database for quick searching to find associations or shots with certain characteristics [6]. The relational database runs on a Windows NT PC SQL server. Access to this data is primarily through IDL or through the Web. The amount of data is presently only about 6 Gigabytes and thus storage and backup are only minor issues at this time. The relational database disk is mirrored, and is also backed up as a database backup. As with the analyzed data, all of the relational database data is kept on disk and is accessible over the wide area network.

5. UNMANAGED RAW DATA

Recently there have been new diagnostics such as the IRTV and Lithium Beam, which by themselves can produce hundreds of Megabytes of digitized data each shot. These are too large to directly incorporate into the existing data system at the current time because of the large individual file sizes that would have to be transferred over the network and archived. This data is acquired on a local system, stored locally in MDSplus, and is only examined on that system and not accessed over the network. The raw data is analyzed on the local system and then the analyzed results are written to the MDSplus server for wider access.

6. FUTURE DATA MANAGEMENT

One of the important assumptions in implementing the HSM system was that much data would never get accessed and would migrate to tape and stay there forever. Unfortunately, this assumption has not been found to be correct. In fact a substantial amount of older data is routinely accessed. In addition there is a growing desire to survey thousands of shots at a time, thus treating the raw data as a searchable database. These surveys can take an unacceptably long amount of time if data is on tape and can also interfere with other users that are only requesting a small amount of data. In 1997 it would have been prohibitively expensive to put all data on magnetic disk. However the cost of disk storage keeps dropping dramatically and magnetic disk has a better cost to performance ratio than optical media. The current goal is to get all data onto magnetic disk within the next two years. If disk prices keep dropping as they have, then this is a realistic goal.

One significant idea is to create a bulk data system in which the tape and optical parts of the HSM system are effectively replaced with inexpensive disks. What is being explored is to utilize the existing system (HSM, archiving, pool) to add another component, bulk data storage, which would lie between the /d3pool and delete blocks in Fig. 2. As data aged on the faster "pool" disks, it would migrate to the bulk storage rather than be deleted. Data would be retrieved from the bulk storage if it were accessed, or deleted from the bulk storage if it became too full. What is undecided at this point is whether to keep the HSM system for a fallback capability once all data is on disk.

As the analyzed data repository continues to grow, its backup could become an important issue. It may be necessary to define some data as unchangeable so that it does not need to continually be backed up. Because the relational database data is such a small component of the overall data management, it is not expected to have an impact. Finally it would be desirable to include unmanaged raw data within the normal data system, and this capability will be explored.

REFERENCES

[1] B.B. McHarg Jr., The DIII–D Computing Environment: Characteristics and Recent Changes, Fusion Eng. Des. **48** (2000) 77.

[2] B.B. McHarg Jr., “Access to DIII–D Data Located in Multiple Files and Multiple Locations,” Proceedings of the Fifteenth Symposium on Fusion Engineering, 1993, p. 123.

[3] J.A. Stillerman, T.W. Fredian, K.A. Klare, G. Manduchi, “MDSplus Data Acquisition System,” Rev. Sci. Instrum. **68** (1) (1997) 939.

[4] J. Schachter, Q. Peng, D. P. Schissel, Data Analysis Software Tools for Enhanced Collaboration at the DIII–D National Fusion Facility,” Fusion Eng. Des. **48** (2000) 91.

[5] Q. Peng, et al., “Status of the Linux PC Cluster for Between-Pulse Data Analyses at DIII–D,” these proceedings.

[6] J. Burruss, “Enhanced DIII–D Data Management Through a Relational Database,” presented at the 42nd Annual Meeting of the Division of Plasma Physics of the American Physical Society, October 23–27, 2000, Quebec City, Canada.

ACKNOWLEDGMENT

Work supported by US Department of Energy under Contract No. DE-AC03-99ER54463.