Astronomical Data Analysis Software and Systems VI
ASP Conference Series, Vol. 125, 1997
Gareth Hunt and H. E. Payne, eds.

The LASCO Data Archive

D. Wang,[1] R. A. Howard, S. E. Paswaters,[1] A. E. Esfandiari,[1] and N. Rich[1]

Solar Physics Branch, Naval Research Lab (NRL), 4555 Overlook Ave SW, Washington DC 20375

Abstract. The data archive for the Large Angle Spectrometric Coronagraph (LASCO) and the Extreme Ultraviolet Imaging Telescope (EIT) is designed to contain 1 B of image data in an easy to use CD-R based archive. This paper discusses the planning, implementation and cost considerations of designing the archive. The problem of getting data into the archive and distributing data to Co-I institutes is also discussed.

1. Introduction

The Large Angle Spectrometric Coronagraph Observatory (LASCO) and the Extreme Ultraviolet Imaging Telescope (EIT) are instruments aboard the Solar Heliospheric Observatory (SOHO). SOHO was launched by ESA/NASA on December 2, 1995. The three coronagraphs of LASCO and the EIT telescope produce the equivalent of 100 1024×1024 16-bit images of the Sun and the solar corona out to 30 solar radii each day. Since first light on Jan 2, 1996 approximately 40,000 images of the corona of the sun have been taken. The LASCO data archive is responsible for storing all of the data and distributing all of LASCO data.

2. Planning and Implementation

Planning for the LASCO Data Archive started in 1994 with the purchase of a Sybase database, development of database tables and understanding what the database could and could not do. By 1995 we were ready to make hardware purchases. Since LASCO is a joint effort between the Naval Research Lab and three major European co-investigator institutions, the problem of sharing the data was an important consideration. Each institute wanted a copy of the data in a timely fashion. Magnetic tape was clearly the cheapest solution for distribution but difficult to work with since the data would have to be staged on and off of magnetic disk to be useful. The amount of data predicted was clearly beyond what could be budgeted for magnetic disk at each institute (in early 1995 magnetic disk was approximately $1/MB). Optical disk was a possibility but had uncertain support for all the various computers and operating systems used by our co-investigators.

[1] Interferometrics Inc., 14120 Parke Long Court—Suite 103, Chantilly VA 20151

The recordable CD-ROM (CD-R) was the ideal solution. The format was usable by almost any computer. The disks are archival with an estimated 100 year lifetime with proper storage. The amount of data was convenient with each CD-R containing several days of data and easy to browse. The media was inexpensive with blank media costing only $0.01/MB which is important when five copies of each CD-R must be made (one for each of four institutes and one for archival storage). The cost of 1 TB is only about $10,000; an amount that is affordable. The only disadvantages are that CD-Rs are much slower to access than magnetic disks and require several hours to assemble a CD-ROM image and write the CD-Rs. CD-Rs have become very popular. The price of CD-R recorders has dropped from $2,500 in 1995 to less than $1,000 at present. CD-Rs are so popular that the media manufacturers were overwhelmed with demand earlier this year and there were long back order times.

Large CD-ROM jukeboxes meant that all of our data could be on-line. Solar scientists often wish to conduct studies that span images spread over a solar rotation (27 days) or even several rotations. Having some data off-line makes assembling datasets more difficult and harder to manage. The 500 disk jukebox we chose provides 325 GB of storage. Using smaller jukeboxes would offer faster access times under heavy usage because the ratio of CD-ROMs to drives would be better, but it also would have cost significantly more money for the same storage capacity. The jukebox format was favored over a carousel both because of its capacity and because a carousel rotates all of its CD-ROMs each time a CD-ROM is loaded or unloaded whereas a jukebox handles only a single CD-ROM with each load or unload operation. Although good statistics for the lifetime of CD-ROMs in carousels versus jukeboxes do not exist, it is intuitive that less motion is better for the life of the CD-R. The total cost of the CD-ROM jukebox, software and 500 CD-Rs to fill it was about $0.07/MB in 1995. Even though the cost of magnetic disk has dropped to $0.20/MB in 1996, CD-Rs other advantages still make it a good choice for use as an on-line storage medium.

3. Archive Operations

The LASCO ground support system at Goddard Space Flight Center receives data in three ways. During times when the NASA Deep Space Network (DSN) is in contact, the data packets are handled in realtime. Since the contact period is typically 8 hours, the remaining 16 hours of non-contact time are covered by an on-board tape recorder which is dumped into packet files when contact is re-established. The final method used to input data is a CD-ROM from the Deep Space Network some weeks after the date of observation. This CD-ROM is produced after DSN has applied all of its error correction and packet recovery techniques. The raw packets are processed by the data reduction pipeline which takes the raw telemetry and produces raw image files. The raw image files are decompressed, rotated so that solar north is up and made into FITS files. The FITS files are then made available to observers at Goddard. The entire process is automated and the FITS files are usable and displayed at the LASCO Experiment Operations Facility in about a minute. The raw image files and packet files are also processed at NRL and stored on magnetic disk on the NRL data server. This provides redundant data reduction capability in case some malfunction occurs at NRL or GSFC. The DSN CD-ROMs containing corrected

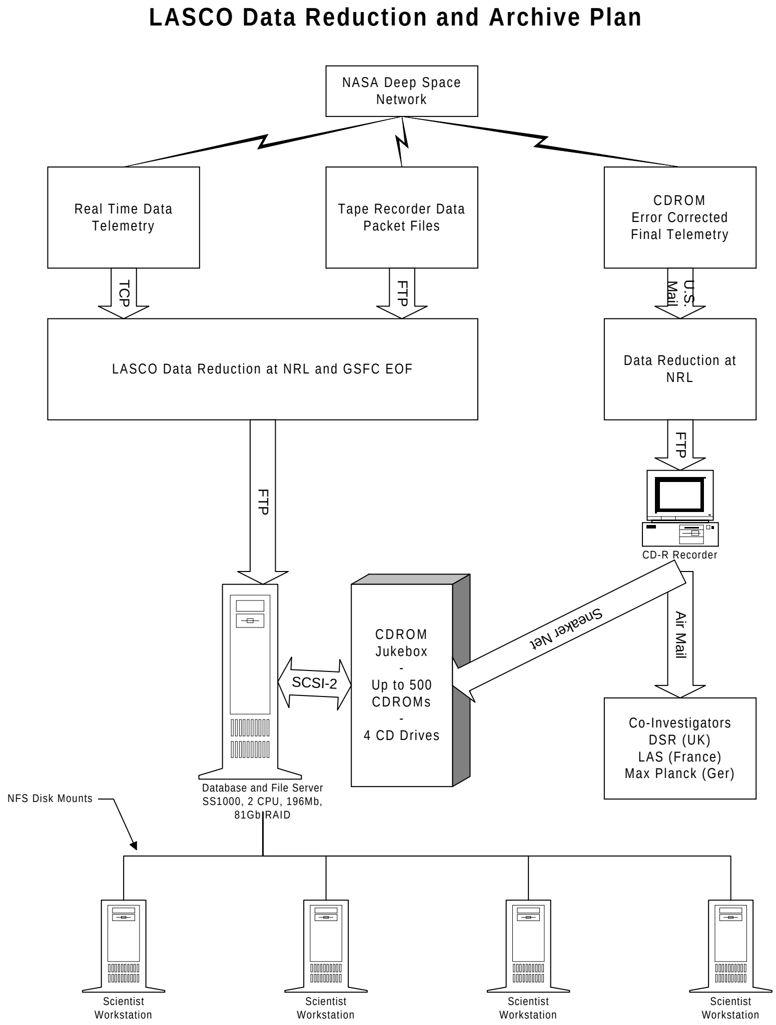
LASCO Data Reduction and Archive Plan
NASA Deep Space Network
Real Time Data Telemetry
Tape Recorder Data Packet Files
CDROM Error Corrected Final Telemetry
TCP
FTP
U.S. Mail
LASCO Data Reduction at NRL and GSFC EOF
Data Reduction at NRL
FTP
FTP
CD-R Recorder
CDROM Jukebox - Up to 500 CDROMs - 4 CD Drives
Sneaker Net
Air Mail
SCSI-2
Co-Investigators DSR (UK) LAS (France) Max Planck (Ger)
Database and File Server SS1000, 2 CPU, 196Mb, 81Gb RAID
NFS Disk Mounts
Scientist Workstation
Scientist Workstation
Scientist Workstation
Scientist Workstation

Figure 1. Schematic of the LASCO data archive and data flow.

data are generally processed at NRL as they arrive. This final data is put onto CD-Rs and then into the CD-ROM jukebox. The data flow is shown in Figure 1.

The data server is arranged so that as data is moved from magnetic disk to the CD-ROM jukebox the directory path does not change, so a scientist accessing data via NFS doesn't need to know whether the data is on CD-R or on magnetic disk. The jukebox software maintains a disk cache which stores the directory of each CD-ROM and the first kilobytes of each file making it possible to browse the directories and for file manager software to identify the type of each file (document, image, JPEG image etc.) without mounting the CD-ROM.

The data reduction process also generates a text file list of images taken for our WWW server and a set of SQL commands for updating our database. The text file is commonly used to monitor the data taking in realtime. The

SQL commands update the image information in the database and also store a JPEG browse image. The database and browse images can be accessed through our WWW page[2] using software written by the European Southern Observatory and generally updated within a day or two of the date of observation. Although this could be done as a realtime process, batch processing is preferred because of occasional network outages between our Sybase database server at NRL and the LASCO EOF at Goddard. The database can also be accessed during data analysis by using IDL routines and C routines written at NRL (Esfandiari et al. 1997).

4. Future Plans

Immediate plans for the archive are a network upgrade to Fast Ethernet in November 1996. Future plans include the possibility of adding more CPUs to our data server as the amount of data grows and another CD-ROM jukebox when we outgrow our present jukebox. One nice feature of jukeboxes is that the access time does not increase as jukeboxes are added since the ratio of CD-ROM drives to CD-ROMs can be kept constant. If the new Digital Video Disk (DVD) technology becomes successful it may not even be necessary to buy another jukebox since we could just replace the drives in our present jukebox with new DVD drives. DVD promises 4.7 GB of storage on a CD-ROM sized disk with readers which are backward compatible with CD-ROM and faster access times than CD-ROM. Double sided DVD disks offer over 8 GB on a single disk. Recordable DVD disks are perhaps a year or two in the future but are expected to cost little more than the present CD-Rs.

References

Esfandiari, A. E., Paswaters, S. E., Wang, D., & Howard, R. A. 1997, this volume, 353

[2] http://lasco-www.nrl.navy.mil/lasco.html