N89-22334

SPACE DATA MANAGEMENT AT THE NSSDC: APPLICATIONS FOR DATA COMPRESSION

James L. Green
National Space Science Data Center

ABSTRACT

The National Space Science Data Center (NSSDC), which was established in 1966, is the largest archive for processed data from NASA's space and Earth science missions. The NSSDC manages over 120,000 data tapes with over 4,000 data sets. The size of the digital archive is approximately 6,000 gigabytes with all of this data in its original uncompressed form. By 1995 the NSSDC digital archive is expected to more than quadruple in size reaching over 28,000 gigabytes.

The NSSDC is beginning several new thrusts allowing it to better serve the scientific community and keep up with managing the ever increasing volumes of data. These thrusts involve managing larger and larger amounts of information and data online, employing mass storage techniques, and the use of low rate communications networks to move requested data to remote sites in the United States, Europe and Canada. The success of these new thrusts, combined with the tremendous volume of data expected to be archived at the NSSDC, clearly indicates that new and innovative storage and data management solutions must be sought and implemented.

Although not presently used at the NSSDC, data compression techniques may be a very important tool for managing a large fraction or all of the NSSDC archive in the future. Some future applications would consist of compressing online data in order to have more data readily available, compress requested data that must be moved over low rate ground networks, and compress all the digital data in the NSSDC archive for a cost effective backup that would be used only in the

event of a disaster.

INTRODUCTION

The purpose of the NSSDC is to serve as an archive and distribution center for data obtained on NASA space and Earth science flight investigations and to perform a variety of services to enhance the overall scientific return from NASA's initial investment in these missions. The NSSDC usually receives data from NASA principal investigators or directly from NASA projects where facility instruments are being flown. However, the NSSDC also obtains data from other government and international agencies involved in space research.

Although the NSSDC does not currently store the data it manages in its archive in compressed form, data compression may very well be a future requirement. In this paper I will discuss the reasons for considering the use of data compression techniques at the NSSDC by looking at the future requirements for data distribution and the growing size of the data center's archive. I will concentrate on the situation at the NSSDC but it must be recognized that many other institutions, universities and other government agencies are in a similar situation.

CURRENT AND FUTURE NSSDC ARCHIVE

The NSSDC archives and manages both digital and non-digital data. The digital archive is stored on approximately 120,000 magnetic tapes with the volume of over 6,000 gigabytes. There are over 4,000 data sets that are supported with appropriately 250 new data sets being archived per year. The most requested digital data sets are stored at the NSSDC (comprising about 35,000 tapes) with the remainder of the archive stored in the Federal Records Center (FRC) about 20 miles away. In addition to the digital archive, the NSSDC has a photo or film archive of over 2 million feet of film, 41,000 sheets of

microfiche, and 39,000 microfilm roles. The charge for obtaining data from the NSSDC is usually for replacement costs in materials and supplies (example: a blank tape or equivalent is needed for one tapes worth of archived data).

From the time it was established in 1967 until 1985 all requests for NSSDC held data were in the form of letters or phone messages which was consistent with the "offline" management of the data that was employed. Requests for offline archived data typically takes 2 weeks if the data is held locally at the NSSDC. If the needed data is in the FRC, the request will take a month or more to be satisfied. This situation is labor intensive and involves interacting with another federal organization. Currently, the NSSDC must accumulate requests for data stored in the FRC and makes two trips per month to obtain the data.

In 1985 NASA was acquiring approximately 360 gigabytes of data per year. Assuming both currently approved and most likely approved NASA missions, the acquired data volume by 1995 will reach well over 2,400 gigabytes per day(1). This is a staggering rate. Figure 1 shows the data volumes per scientific discipline that have been archived and are expected to be archived at the NSSDC. As discussed above, the NSSDC has currently about 6,000 gigabytes of data. The size of the archive past 1988 is a projection and considers the arrangements being made with the NSSDC and the missions that are currently approved. If this projection holds true, then by 1995, the NSSDC will have over 28,000 gigabytes in its archive.

The physical space that the NSSDC has to manage is nearly full, both locally and at the FRC. From the predicted amount of data to be archived, as shown in Figure 1, the NSSDC must implement mechanisms to store data on higher density media (by a factor of 5 to 10) and/or implement data compression techniques. At the NSSDC, use of data compression techniques as a routine mechanism to pack data on media can only be a viable mechanism when it becomes accepted and is in wide

spread use in the scientific community. This acceptance is occurring (see section, Networking of AVHRR Data), but only very gradually.

ARCHIVE SAFE STORAGE

A national archive needs to have operational plans for insuring that a natural disaster, such as fire, does not permanently destroy irreplaceable data. For a data center, where most of the archive is in digital form, then a copy of the data stored in another location would be the best solution for safe storage of archived digital data.

In the case of the NSSDC, over the last 20 years, it has received data for archiving from investigators at hundreds institutions across the country. In this situation, the remote investigators retain the original data with a copy sent the NSSDC. Due to budget constraints and inadequate resources to provide a complete backup, the NSSDC's disaster recover plan is to request a copy of the data from the original producers. These plans are inadequate as a viable disaster recovery plan since many of the investigators would not be able to reproduce the data that is more than four or five years old for a variety of reasons (ex., inadequate resources, older tapes written at low density formats, etc). In addition, NASA's missions are now moving toward facility instruments where the NSSDC must assume full responsibility for the data being archived since it is coming from a short lived project with resources that are usually just adequate to keep up with the new incoming data with little or no reprocessing possible.

Plans are now being devised at the NSSDC, that once in place, will provide for a complete safe storage as a backup of the NSSDC digital archive. With such large volumes of data to back up, a cost effective solution requires an extremely dense media with a very small cost per megabyte of storage, a high data transfer rate, and adequate data compression schemes (preferably lossless) to further reduce the volume. In this case, even thought scientists are reluctant to

provide NSSDC with compressed data for distribution, there is little argument against data compression techniques being applied in order to provide for a cost effective backup of the entire digital archive.

As operational mass storage software and hardware systems mature, optical tape would be a prime candidate as an archive backup. The data write rates can be very large (100 MB/s) with a $3,000 cost per reel containing 1 TB of data. Drives are estimated in the range of $200,000 apiece. Another possible media is the digital videotape cassettes which costs about $135 per cassette that can hold data up to 125 gigabytes. Recorders/readers are also in the $200,000 range. Once again, these devices are just now in beta testing with commercial units available within a year and little operational software.

ONLINE INFORMATION AND DATA SYSTEMS

There has been an explosion in the use of available communication technology for the movement and access of mission data and information. Many large universities and nearly every NASA center and other government institutions that work with NASA data have relatively high speed local area networks and many wide area network connections.

There are two major wide area NASA networks that are used extensively; SPAN[2] and the NSN. SPAN contains over 2050 nodes in the United States and is internetworked with over 6000 nodes in the U.S., Europe, Canada, and Japan. Like SPAN, NSN is internetworked with other wide area networks such as ARPANET and the NSFNET that can reach many thousands of computers. In general, these wide area networks are of relatively low speed but are serving a tremendously valuable service for the remote users to gain access to space and Earth science computer resources and to fellow researchers all across the country. Although a modest amount of data is transmitted over the wide area networks, it is not real-time (coming directly from a NASA orbiting spacecraft). The bulk of the wide area traffic is of informational purposes such as remote logon and mail.

The NSSDC is responding to an ever increasing number of user requests by putting more of the data and information about the data in its archive online to electronic access. In this way, the NSSDC can "remain open" past the normal working hours allowing scientists and students to "browse" through the online information looking for an important data set. The online data and information systems that are currently operational at the NSSDC are shown in Table 1.

As will be discussed in the next section, the rapid access to selected data through the NSSDC interactive systems is frequently requested. Since it is not known ahead of time what sections of any one data set will be requested, the NSSDC has loaded all the International Ultraviolet Explorer (IUE) data sets online to accommodate user demand. The data is available through the IUE request system. It is important to note that NSSDC manages its archive offline. Storing all the IUE data online was done with full project co-operation and to gain valuable experience with highly requested online data sets.

IUE ONLINE EXPERIENCE

The NSSDC has loaded all the IUE data (in uncompressed form), consisting of over 61,000 unique star images and spectra, in the NASA Space and Earth Science Computer Center's IBM 3850 Mass Store[3]. The Mass Store device is managed by an IBM 3081 system and connected to the NSSDC interactive VAX front ends by a high speed local area network (called SESNET), as shown in Figure 2. An interactive system on the NSSDC VAXs has been created that allows for a remote SPAN user to logon and order IUE data from the electronic Merged Observer Log. Once the exact data segment requested has been identified, the NSSDC request coordinator networks the IUE data from the Mass Store through the local area network and through SPAN to the target computer of the requesting individual. This system became operational in November 1987 and by January 1988 requests were routinely serviced with this system.

In addition, requests for IUE data sent on magnetic tape are easily handled by this system, saving the manual locating of the data. These requests come to the NSSDC from letters, phone calls (not all our users are on computer networks), or electronically.

Figure 3 shows the monthly averages of IUE images requested by individuals (we also send large amounts of IUE data to other archives) from 1979 to 1988. From 1979 to 1987, the only service the NSSDC offered was an offline service resulting in a tape copy of the data being produced and sent to the requestor. The bar graph also shows the monthly number of IUE images in 1988 (averaged over the first four months of that year) requested using the online data requesting and electronic delivery system. As clearly seen, there is a dramatic increase in the amount of IUE data requested in 1988 reaching approximately 350 images/spectra per month. The 1988 requests have come from many scientists at 13 institutions in the United States, Europe, and Canada (locations serviced by SPAN).

Since there has been no new money from NASA Headquarters for increased IUE data analysis, the results clearly show that the tremendous increase in requested data is believed to be due to the convenience this system provides to the user. The following factors are a major part of the user convenience provided by this service:

- Data are loaded to the target system (no tape handling), - Data arrives in the desired format;
- No replacement tape is needed to be sent to the NSSDC (currently the network is a "free" service to the users);
- Rapid turnaround provide the desired data while the scientists are interested.

To be able to use low rate communication networks such as SPAN requires that the size of data requested is relatively small. The IUE example is a good one since the data is managed by the stellar object

observed which in itself forms a small enough data subset that it can be easily networked. The amount of time required to network an observation is 2 to 15 minutes depending on the communication rate and the load on the network.

The IUE example serves to illustrate that to better serve some of the users, faster access to requested data is desirable than what the NSSDC has been doing when the data resided offline on magnetic tape.

Many of the most requested data segments come from very large data sets. Keeping large amounts of data online is expensive. Key questions as to the management of the larger volumes of data being archived and promoted to online status must be addressed within the next few years. But if the IUE example is representative of what users need, then to accommodate large amounts of data online, the NSSDC will have to consider use of data compression techniques.

Even though the NSSDC may not compress the data that resides online there are other uses of data compression/decompression techniques when the use of low rate networks are employed to moved the data to a remote location, as discussed in the next example.

NETWORKING OF COMPRESSED AVHRR DATA

Many universities gain access to the Oceanographic data being collected at the University of Miami using SPAN. The University of Miami routinely networks compressed data from the Advanced Very High Resolution Radiometer (AVHRR) instrument onboard a polar orbiting NOAA spacecraft. The AVHRR data is received in real-time are quickly processed (stripping out the infrared portions), compressed, and transmitted to the University of Rhode Island via SPAN where it is decompressed and remapped into a standard set of projections used for several real-time ship activities such as cruise support and chart generation. These images are also networked to Harvard University for their Gulf Stream predication models. In this example, a Lempel-Ziv

compression algorithm is used.

SUMMARY

If NASA is going to fly future missions which will produce several orders of magnitude more data than in the past, then it needs to continue to develop better techniques and facilities for data management, data storage and data distribution if it has a chance of preserving and providing the continual extraction of science from its archived data in a timely manner. I predict that as data compression techniques become accepted by the scientific community their implementation at the NSSDC and other data centers will be a necessity. Data compression should be considered an important element of data management for very large data bases.

Access is one of the most important aspects necessary for the proper management of very large data bases. It is projected that the NSSDC will be inundated with data, quadrupling its archive by 1995. Although mass store technology has progressed considerable and must be employed in the managing of very large data bases, it appears that data compression will also play a significant role by providing more data online in conjunction with mass storage capability.

In a similar way, the technology of wide area computer networking has grown considerably. If the speeds of wide area communication networks don't keep up with the demand for the electronic transfer of data, then data compression techniques will also be necessary to implement for data that is transferred over wide area networks.

With respect to the above capabilities, we are somewhat dependent on the ability of the data producers to use the new mass storage technology, the wide spread use of very high speed networks and the acceptance and wide spread use of data compression techniques by the scientific community.

However, the use of lossless data compression techniques as part of the implementation of a complete backup of a very large archive is a viable and, perhaps, necessary step for a cost effective way of preserving irreplaceable digital data archive.

REFERENCES

1) Carper, R., J. Dalton, M. Healey, L. Kempster, J. Martin, F. McCaleb, S. Sobieski, and J. Sos, "Mass Storage Systems for Data Transport in the Early Space Station Era 1992-1998." NASA TM 87826, July 1987.

2) Green, J., V. Thomas, B. Lopez-Swafford, and L, Porter, "Introduction to the Space Physics Analysis Network," NSSDC 87-4, January 1987.

3) Perry, C. M., "The National Space Science Data Center and International Ultraviolet Explorer 1978 - Present," To be Published in the 10th Anniversary IUE Conference Proceedings, March 1988.

ACKNOWLEDGEMENTS

I would like to acknowledge valuable discussions with Charleen Perry who provided updated statistics for the NSSDC IUE requests used in this paper. In addition, I would like to gratefully acknowledge Robert Price who presented this paper at the Data Compression Workshop when I was not able to attend.

NSSDC ARCHIVE

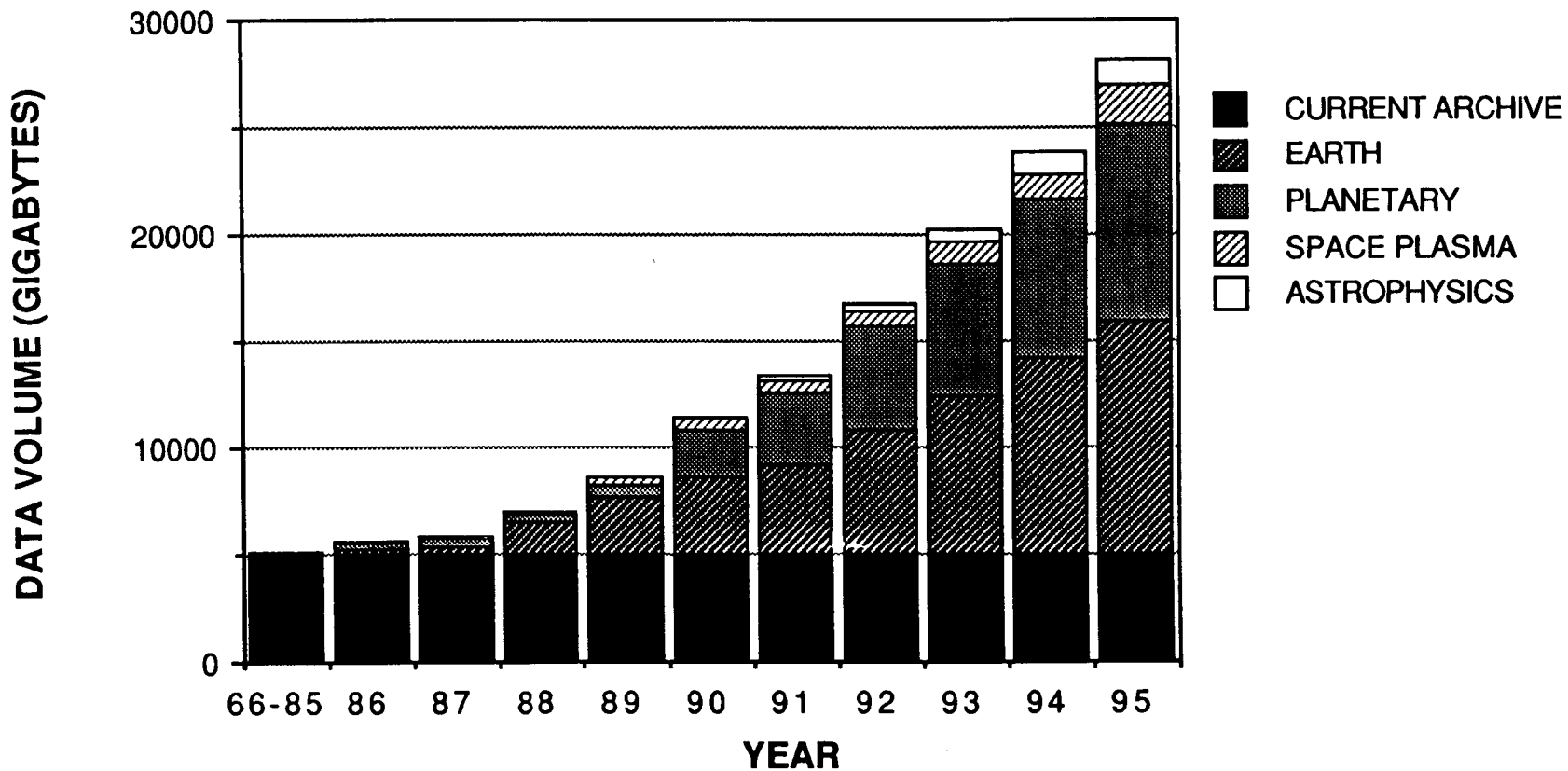

FIGURE 1

MASS STORAGE DATA ACCESS AT GSFC

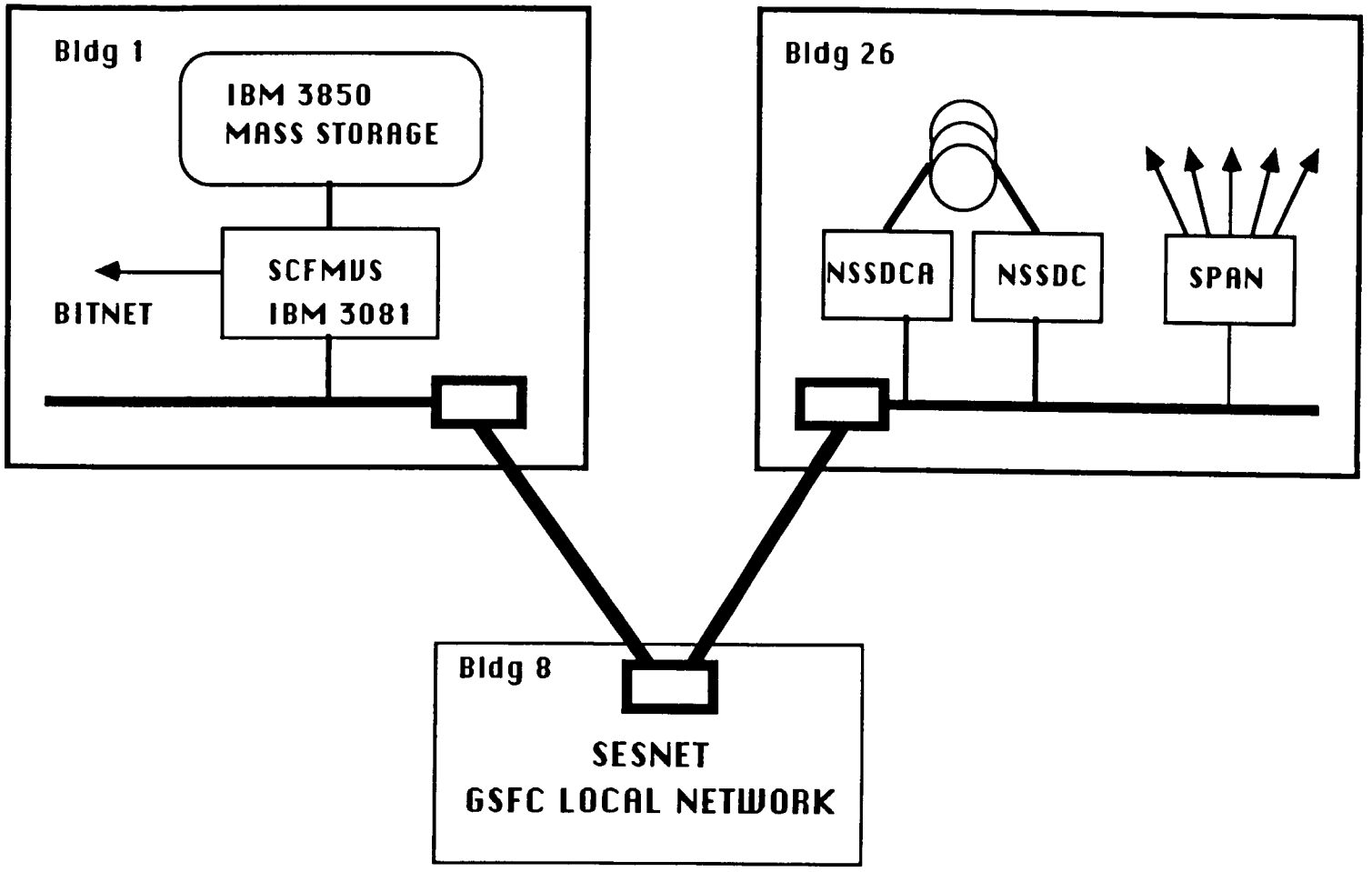

FIGURE 2

IUE IMAGES REQUESTED BY INDIVIDUALS FROM NSSDC ARCHIVE

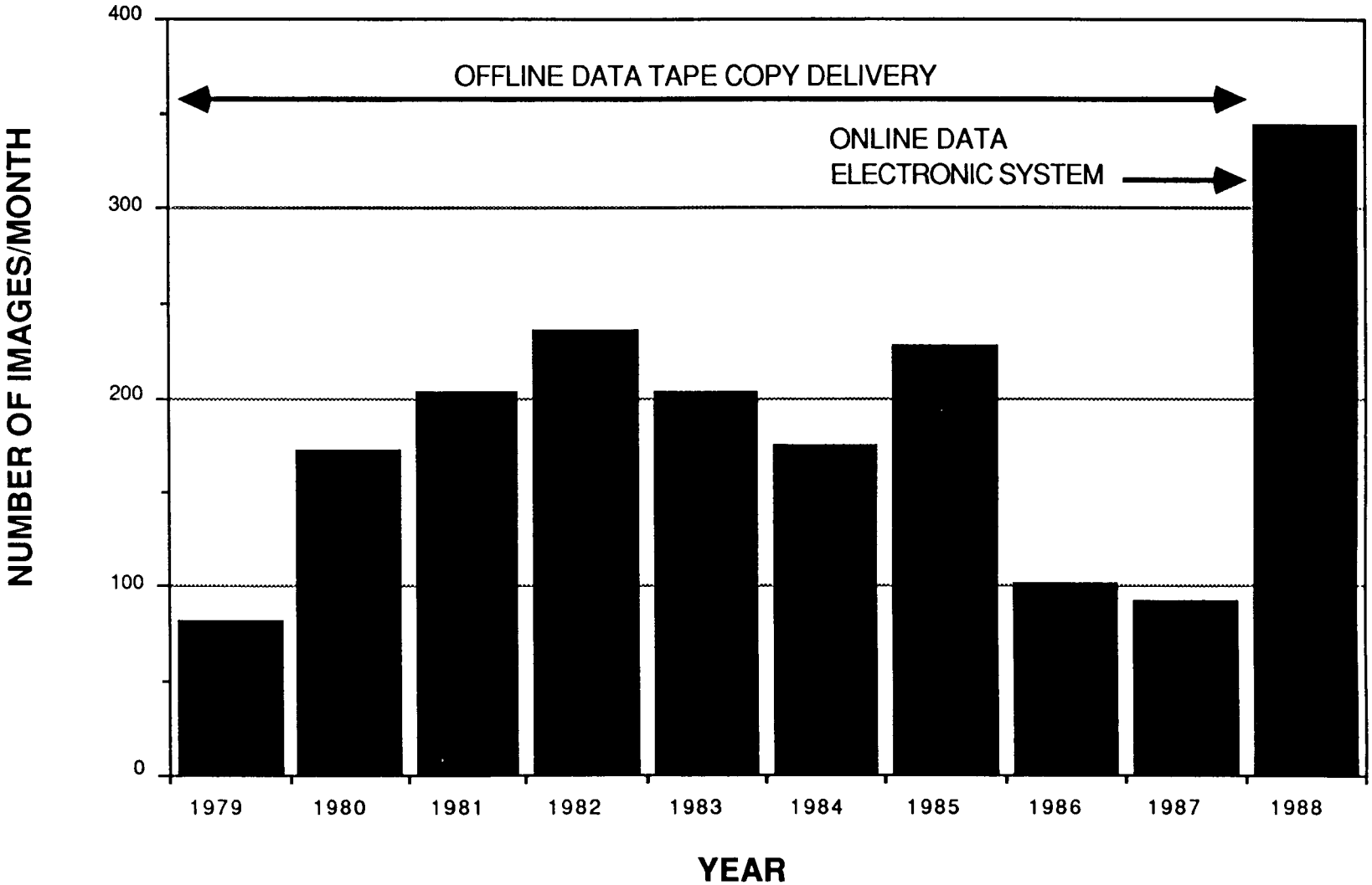

FIGURE 3

TABLE 1

NSSDC ONLINES SYSTEMS

SCIENCE DISCIPLINE	SERVICE	INFORMATION	DATA #
All			
	Master Directory	X	
Astrophysics			
	IUE Requests System	X	X *
	ROSAT Info. Manage. Sys.	X	
	Astronomy Catalog Sys.	X	X
	Starcat with SIMBAD acc.	X	X
Atmospheric Science			
	NASA Climate Data System	X	X
	Ozone TOMS data	X	X
Land Sciences			
	Crustal Dynamics	X	X
	Pilot Land Data Systems	X	
Space Plasma Physics			
	Central Online Data Dir.	X	
	Omni Solar Wind Data Sys.	X	X
	Plasma and Field Models	X	X +
	Coordinated Data Ana. Wk.	X	X
General			
	SPAN-Network Info. Center	X	

NOTES:

Only partial data sets are available

* All available data is online

+ Only software is being distributed

C-2