Data Compression Opportunities in EOSDIS

Ben Kobler
EOS Systems Development Office
Mail Code 902.1, NASA GSFC
Greenbelt, MD 20771
(301) 286-3553
(301) 286-3221 (FAX)
bkobler@gsfcmail.nasa.gov

John Berbert
EOS Systems Development Office
Mail Code 902.1, NASA GSFC
Greenbelt, MD 20771
(301) 286-5916
(301) 286-3221 (FAX)
jberbert@gsfcmail.nasa.gov

Abstract. The Earth Observing System Data and Information System (EOSDIS) is described in terms of its data volume, data rate, and data distribution requirements. Opportunities for data compression in EOSDIS are discussed.

1. Introduction

The Earth Observing System Data and Information System (EOSDIS) is being developed by the National Aeronautics and Space Administration (NASA) to be a comprehensive data and information system providing the Earth science research community with easy, affordable, and reliable access to Earth Observing System (EOS) and other appropriate Earth science data. The EOS program, as a part of the Mission to Planet Earth is intended to study global-scale processes that shape and influence the Earth [1, 2, 3]. Beginning in 1998, EOSDIS will archive approximately one terabyte of data per day over a 15 year period [4, 5, 6, 7]. Many opportunities for data compression exist in EOSDIS for alleviating problems due to large data volumes, high bandwidth requirements, and data access requirements.

2. EOSDIS Requirements

There are 5 proposed EOS instruments on the EOS AM-1 spacecraft to be launched in June 1998 and 6 proposed EOS instruments on the EOS PM-1 spacecraft to be launched in December 2000. These instruments will generate data at a rate of 281 gigabytes per day [8]. Other instruments will follow on spacecraft to be flown later. Data from the EOS instruments will be transferred to an EOS Data and Operations System (EDOS), from where data will be batched to an appropriate Distributed Active Archive Center (DAAC), selected with responsibility for further processing. The Product Generation System (PGS) located at the DAACs will generate higher level products (L1 through L4) for storage in the Data Archive and Distribution System (DADS). The data product processing levels are defined as follows:

- L0 Raw instrument data at original resolution, time ordered, with duplicate packets removed
- L1A L0 data, which may have been reformatted or transformed reversibly, located to a coordinate system, and packaged with need ancillary and engineering data
- L1B Radio metrically corrected and calibrated data in physical units at full instrument resolution as acquired
- L2 Retrieved environmental variables (e.g. ocean wave height, soil moisture, ice concentration) at the same location and similar resolution as the L-1 source data
- L3 Data or retrieved environmental variables that have been spatially and/or temporally resampled (i.e., derived from L1 or L2 data products) and may include averaging and compositing
- L4 Model output and/or variables derived from lower level data which are not directly measured by the instruments such as new variables based upon time series of L2 or L3 data

Generation of these higher level data products will expand total data volume by a factor of 3.3, resulting in a total data volume from the AM-1 and PM-1 platforms of approximately 0.9 terabytes per day.

The sustained combined daily rate for data input into EOSDIS from the AM-1 and PM-1 platforms will be 26 megabits per second. The sustained daily rate for data access into and from the DADS will, however, be substantially larger to accommodate, in addition to the initial data processing, subsequent data reprocessing and data distribution to users.

A distributed Information Management System (IMS) will be implemented to provide a common user interface to database management systems at the DAACs, providing the capability to easily construct complex queries to search, locate, select, and order products. The IMS will be sized to accommodate 100,000 users. A load of 100 concurrent IMS sessions will be distributed across the DAACs. Approximately 500 IMS queries per hour can be expected for log-on authorization, directory search, catalog search, inventory search, status checks, browse selection, document search, and ordering services.

EOSDIS will be capable of distributing data via physical media and via communications networks, each at a rate equivalent to approximately 1 terabyte per day. Data requested on physical media will be made available for delivery within 24 hours and data requested over networks will be available to the network within an average of 5 minutes.

3. Data Compression Opportunities

Conventional lossless compression techniques such as Huffman coding, Ziv-Lempel compression, and arithmetic coding have been shown to be very effective at compressing a wide range of data types with compression ratios of approximately 2:1 [9, 10, 11]. The potential cost savings to the EOSDIS data archive facility due to reduction of hardware for data storage is obvious. Perhaps less obvious is also a concomitant reduction in requirements for bandwidth of storage devices. To be most effective, however, compressed data needs to stay in its compressed form as long as possible, so that data is not needlessly decompressed and then re-compressed, and so that the potential savings in network bandwidth are not lost. This requires standardization on a common set of data compression schemes, on associated common data format structures, and on common compression/decompression tool kits that are integrated across all of EOSDIS. For example, callable routines that decompress a block or record at a time, would be essential to PGS, as would routines that decompress data at user workstations.

Lossy compression techniques such as DCT, wavelet transform, and vector quantization [12, 13, 14] can play a significant role in optimizing data access by providing tools for storage and retrieval of display quality browse data. EOSDIS will permit users to browse subsetted, subsampled and/or summarized data sets that are created during routine production processing. These browse data sets will be generated by algorithms provided by scientists. Since some of these browse products are designed for visual display, they may be further compressed by lossy compression techniques that can have significantly higher compression ratios than lossless techniques. Because EOSDIS needs to retrieve and begin to display these browse data sets within one minute, they need to be stored on faster access devices than other data. The associated reduction in bandwidth requirements due to data compression could aid in reducing costs.

More innovative lossy/lossless techniques, such as progressive vector quantization [15], have the potential for allowing browse quality lossy compression, while also allowing lossless restoration of full datasets. Such combined techniques can benefit from the best features of both and can result in reduced total I/O requirements and better compression ratios. To be most useful, these techniques require standardization on a common format structure that allows storage of the browse component on a fast access device and storage of the complementary lossless data component on a slower access device. Unfortunately, however, data compression techniques such as vector quantization are extremely processor intensive, although the decompression phase is much less so. The benefits of reduced I/O and higher compression need to be balanced against the compression cost and the impact of that cost on the PGS design.

Finally, the concept of using very large codebooks to achieve very high compression, both lossless and lossy, although still unproven, has potential for success in extremely large data archives such as those planned in EOSDIS. Fundamental issues need to be investigated that explore the redundancy, and hence compression limit, of these data archives, the stability of the resultant codebooks, and the most effective method for the generation, storage and exchange of those codebooks.

References

[1] Ramapriyan, H. K., "The EOS Data and Information System," Proceedings of the AIAA/NASA Second International Symposium on Space Information Systems, Pasadena, CA, September 1990.

[2] Dozier J. and H. K. Ramapriyan, "Planning for the EOS Data and Information Systems (EOSDIS)," The Science of Global Environmental Change, NASA ASI, 1991.

[3] Taylor, T. D., H. K. Ramapriyan, and J. Dozier, "The Development of the EOS Data and Information System," Proceedings of the AIAA 29th Aerospace Sciences Meeting, Reno, NV, January 1991.

[4] Kobler, B. and J. Berbert, "NASA Earth Observing System Data Information System (EOSDIS)," Eleventh IEEE Symposium on Mass Storage Systems, Monterey, CA, 18-19, October 1991.

[5] Berbert, J. and B. Kobler, "EOSDIS DADS Requirements," NSSDC Conference on Mass Storage Systems and Technologies for Space and Earth Science Applications, Greenbelt, MD, NASA 3165, III-141, July 1991.

[6] Functional and Performance Requirements Specification for the Earth Observing System Data and Information System (EOSDIS) Core System, NASA, Goddard Space Flight Center, July 1991.

[7] EOSDIS Core System Statement of Work, NASA, Goddard Space Flight Center, July 1991.

[8] Lu, Y., "EOS Output Data Products and Input Requirements, Version 2.0," NASA, Goddard Space Flight Center, August 1992.

[9] Huffman, D. A., "A method for the construction of minimum redundancy codes," Proc. IRE, 40, 1098-1101, 1952.

[10] Ziv, J. and A. Lempel, "Compression of Individual Sequences via Variable Rate Coding," IEEE Trans. Information Theory, Vol. IT-24, 530-536, September 1978.

[11] Langdon, G., "An Introduction to Arithmetic Coding," IBM Journal Research Development, Vol. 23, No 2, 135, March 1984.

[12] Clark, R. J., Transform Coding of Images, Academic Press, London, 1985.

[13] Antonini, M., M. Barlaud, P. Mathieu, and I. Daubechies, "Image Coding Using Wavelet Transform," IEEE Transactions on Image Processing, Vol. I, No. 2, April 1992.

[14] Gray, R. M., "Vector quantization," IEEE ASSP Magazine, 1 (2), 4-29, 1984.

[15] Manohar, M. and J. Tilton, "Progressive Vector Quantization of Multispectral Image Data using a Massively Parallel SIMD Machine," Proceedings of the Data Compression Conference, Snowbird, Utah, 181, March 1992.