Appeared in ACM Symposium on Applied Computing, ACM SAC, San Jose Feb.-Mar. 1997

A BRIEF SURVEY OF TERTIARY STORAGE SYSTEMS AND RESEARCH *

S. Prabhakar D. Agrawal A. El Abbadi A. Singh
Department of Computer Science
University of California
Santa Barbara, CA 93106.
{sunilp, agrawal, amr, ambuj}@cs.ucsb.edu

Keywords: Tapes, Digital Libraries, Mass Storage

Abstract

This report summarizes current state of the art in tertiary storage systems. We also summarize the current technologies and research efforts to integrate tertiary storage in operating systems, databases and advanced applications.

1 Introduction

With the recent improvements in network and processor speeds, several data intensive applications have become much more feasible than ever before. These applications are characterized by very large computational and storage requirements. In the present commercial setting and most likely in the near future, the only practical solution for storing such enormous amounts of data is tertiary storage. Although tertiary storage, in particular magnetic tapes, has been used solely for archiving or backup purposes, the exploding storage requirements and the high cost of secondary storage are forcing computer architects and designers to re-evaluate the role of tertiary storage. In this paper we present some of the more recent research activities that study tertiary storage and their integration into computer systems.

There have been three major directions in which research on tertiary storage has been pursued. The first is research on tertiary storage systems from the operating system point of view. These include studies on I/O scheduling, file systems and data striping. The second is the investigation of issues involved in the integration of tertiary storage systems directly into database management systems (DBMS). Traditional database management systems are unaware of tertiary storage devices and do not optimize for them. This can lead to significant performance degradation if data resides on tertiary storage. The third area of research has developed out of existing applications that are compelled to use tertiary storage due to their large data storage needs. These include scientific applications that manipulate terabytes of multidimensional array data and digital libraries. We summarize the initial experience from all these efforts in this paper.

Due to space limitations, detailed descriptions have been omitted in this report. A more detailed version of the paper is available in [9].

2 Tertiary Devices - Current Technology

The most common tertiary storage devices are magnetic tapes, optical disks and magneto-optical disks. Tapes have been the traditional devices for archiving

*Work partially supported by a research grant from NSF/ARPA/NASA IRI9411330, and from NSF CDA9421978 and by a research gift from NEC Japan.

large amounts of data whereas optical devices such as CD-ROM and WORM are more recent tertiary devices. Tapes offer the highest storage densities of all current storage media, however they are sequential access devices and most allow append-only updates. Optical disks are similar to magnetic disks except that the recording medium is not magnetic. As compared to magnetic disks, optical disks are slower, have less capacity and are more expensive. Their advantages over magnetic disks are that they are removable and are not as susceptible to head crashes. Like magnetic disks, optical disks are random access devices with similar latencies. Magneto-optic disks use both magnetic and optical technologies. Magneto-optic disks are faster and less expensive than read/write optical disks and have been more successful commercially. In order to provide automated access to these media, robotic changers are often employed to load and unload tapes or disks from the drives.

The transfer rates of magnetic tape devices vary from about 250 KB/s to 32 MB/s. Tapes vary in capacity from as low as 1.3 GB to 165 GB. The main advantage of tapes over disks is the cost of storage. They are two orders of magnitude cheaper than magnetic disks and almost an order of magnitude cheaper than optical disks. The major disadvantage of tapes is that they are sequential access devices whereas disks (magnetic, optical or magneto-optical) are random access devices. The low seek rate of tapes renders them highly unsuitable for random access workloads. Tapes also suffer from high wear of the drive heads and tape.

CD-ROMs have a standard capacity of 600MB and data rates that are multiples of 153.6 KB/s. The average latency is 150-200ms. Other optical disks have capacities between 600MB and 15GB, with transfer rates between 0.6MB/s and 2.7MB/s. Magneto-optical disks have capacities ranging from 120MB to 2.6GB and transfer rates between 512KB/s and 3.37MB/s. The average seek time is about 28 to 40ms. Both optical and magneto-optical disks have greater reliability and storage life than magnetic disks, but their write rates are lower than their read rates.

While removable media result in cheaper storage, they require cartridge switching. Because cartridge switching takes place at mechanical speeds, switch times are of the order of several seconds or higher. This results in two or three orders of magnitude worse performance as compared to magnetic disks with latencies of the order of tens of milliseconds. Thus in order to obtain acceptable performance from tertiary devices it is important to reduce media switches.

3 Current Research Directions

With the changing role of tertiary storage devices from being used solely for backup to holding "on-line" data, better operating system support is needed. Hillyer and Silberschatz have developed various strategies for reordering batched I/O requests for single serpentine tapes [5]. At the file level, several file systems for managing tertiary storage have been proposed, based upon file-level access through FTP [7]. More recently, log structured file systems have been developed that take advantage of the append-only and sequential limitations of tapes [6, 3]. Drapeau and Katz have studied striping in the context of large tape libraries in the presence of concurrent random I/O [2]. They show that in order for striping to be effective in a concurrent environment, it is necessary to have an adequate number of readers. Golubchik and Muntz [4] have studied striping using a more general open system model with multiple sizes of requests within a single run and various stripe widths.

Commercial database systems are optimized for performance with primary and secondary memory. However, relational database operations such as joins can perform poorly if data is stored on tertiary storage [11]. Sarawagi and Stonebraker have investigated optimizations of 2-way joins of relations which are both tape resident [11]. Techniques for reordering the data access to reduce the amount of switching are described. These techniques result in about two orders of magnitude savings in the number of switches and fetches. Myllymaki and Livny have studied join operations where one relation resides on secondary storage and the other on tertiary storage [8]. The benefits of executing disk and tape I/O in parallel have been investigated. The authors observe that the operations of disk and tape access can be overlapped to reduce the total execution time. Sarawagi and Stonebraker have also investigated the architecture of database management systems that incorporate tertiary devices directly [10]. The authors argue in favor of a central *Scheduler* that has knowledge of the currently pending queries, the contents (and semantics) of the disk cache and the state of the tertiary memory.

Applications with very large data storage requirements need to use tertiary storage to hold active data. Prime examples of such applications are digital libraries and scientific applications that generate terabytes of data at a regular rate. Researchers working on such applications have focused on performance improvements tailored for the applications where data are accessed in small multidimensional blocks which require retrieval from widely separated locations in storage resulting in

large seeks and media switches. Two independent studies [12, 1] have suggested that such data should be stored in a manner that facilitates retrieval for specific access patterns. Both techniques require knowledge of the user access patterns and recommend data duplication to handle conflicting optimization requirements.

4 Concluding Remarks

The storage requirements of data intensive applications cannot be met by secondary storage due to its high cost and low storage density. Tertiary storage has traditionally been relegated to the role of storing archival or backup data. However, with its low cost and high storage density, tertiary storage, in particular magnetic tape technology, is the only reasonable solution to the large scale storage requirements. Tape technology however, is sequential in nature and is therefore ill suited for applications that require random access. Optical and magneto-optical disks do not suffer from this problem but current optical disks have low transfer rates. Even though it is highly likely that in the future, optical technology will overcome these limitations and become the technology of choice for tertiary storage, the current need for large scale storage can only be filled by magnetic tapes. Hence in the near future, tapes will be used for storing data that is accessed randomly. The use of tapes in random access applications results in poor performance. Research efforts to overcome this limitation have been made but the gap is still large. Due to the variability of the various characteristics of tertiary systems, it is important that solutions take into account the parameters of the system for which the solution is designed. Thus in contrast to solutions for magnetic disks, it is not obvious that optimizations that work for one type of tertiary technology will also work for others. Much work needs to be done to overcome the problems associated with integrating tape storage.

References

[1] L. T. Chen, R. Drach, M. Keating, S. Louise, D. Rotem, and A. Shoshani. Efficient organization and access of multi-dimensional datasets on tertiary storage systems. In *Information Systems*, volume 20, pages 155–83. Elsevier Science, 1995.

[2] A. L. Drapeau and R. H. Katz. Striping in large tape libraries. In *Proc. of Supercomputing*, pages 378–387, Portland, Oregon, 1993. ACM.

[3] D. A. Ford and J. Myllymaki. A log-structured organization for tertiary storage. In *Proceedings of the Twelfth International Conference on Data Engineering*, pages 20–7, New Orleans, Louisiana, 1996.

[4] L. Golubchik and R. Muntz. Analysis of striping techniques in robotic storage libraries. In *Proceedings of the Fourteenth IEEE Symposium on Mass Storage Systems*, pages 225–38, Monterey, CA, 1995.

[5] B. K. Hillyer and A. Silberschatz. Random I/O scheduling in online tertiary storage. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, Canada, 1996.

[6] J. Kohl, M. Stonebraker, and C. Staelin. HighLight: a file system for tertiary storage. In *Proceedings of the Twelfth IEEE Symposium on Mass Storage Systems*, pages 157–61, Monetery, CA, 1993.

[7] F. McClain. DataTree and UniTree: Software for file and storage management. In *IEEE*, pages 126–8, 1990.

[8] J. Myllymaki and M. Livny. Disk-tape joins: Synchronizing disk and tape access. In *Joint International Conference on Measurement and Modeling of ComputerSystems. SIGMETRICS '95/PERFORMANCE '95*, pages 279–90, Ottawa, Canada, 1995.

[9] S. Prabhakar, D. Agrawal, A. El Abbadi, and A. Singh. Tertiary storage: Current status and future trends. Technical Report TRCS96-21, Dept. of Computer Science, Univ. of Calilfornia, Santa Barbara, 1996. http://www.cs.ucsb.edu/TRs/TRCS96-21.ps.

[10] S. Sarawagi. Query processing in tertiary memory databases. In *Proc. of the 21st Int. Conf. on Very Large Data Bases*, pages 585–596, San Francisco, California, 1995. Morgan Kaufmann.

[11] S. Sarawagi and M. Stonebraker. Single query opimization for tertiary memory. Technical Report s2k-94-45, Computer Science Div. U.C. Berkeley, December 1993.

[12] S. Sarawagi and M. Stonebraker. Efficient organization of large multidimensional arrays. In *IEEE Int. Conf. on Data Engineering*, pages 328–336, Houston, TX, USA, Feb. 1994. IEEE Comput. Soc. Press.