Mass Storage Systems

For articles on related subjects see CHANNEL; DATA MINING; DATA WAREHOUSING; DATABASE; FIBER OPTICS; FILE SERVER; HARD DISK; MEMORY, AUXILIARY; VIRTUAL MEMORY.

A *mass storage system* or MSS is a collection of software, computing, input/output, and data storage components that jointly automate the archiving, storage, management, and retrieval of very large quantities of digital information. A typical high-end mass storage system may store from hundreds of *terabytes* (10^{12} bytes) to *petabytes* (10^{15} bytes) of data contained in millions of files. The MSS provides access to those files to client computer systems ranging from desktop workstations to supercomputers at speeds from megabits per second over a local area network to gigabits per second over high-speed I/O channels. Mass storage systems are not automated backup systems, although they may be used as a component of such systems. Mass storage systems are not automated tape management systems, although some early ones evolved from such systems, and most incorporate sophisticated tape management algorithms. Mass storage systems are not distributed file systems, although users may indirectly access an MSS through front-end distributed file system servers. Examples of early mass storage systems are the Common File System or CFS developed at Los Alamos National Labs, Unitree from Lawrence Livermore National Labs, and the National Center for Atmospheric Research (NCAR) MSS, each of which were developed in the mid-1980s.

There are many examples of the need for mass storage systems in science and commerce. A study done by Harvard Medical School and others estimates that a medium-sized metropolitan medical institution will generate two terabytes a year of multimedia patient information, the bulk of which will be digitized x-ray images. The NASA Earth Observing Satellite, slated to launch in 1998, is expected to generate eight petabytes of data by the year 2007, at a peak rate of three terabytes a day, or the equivalent of six 660 megabyte CD-ROMs every minute. The U.S. Department of Energy Accelerated Strategic Computing Initiative (ASCI), responsible for computer simulations of the nuclear stockpile, will require hundreds of petabytes of storage. In each case, not only is a prodigious amount of data being generated, but it must be archived for as long as decades, it must be made easily available for computer processing, and it must be stored as economically as possible. NCAR has found that for every billion floating point operations performed by its climate simulations, about half a megabyte of new data is stored in the MSS. At this rate, a teraflop (10^{12} floating point operations per second) supercomputer would generate 500 megabytes of data per second for the MSS. NCAR has described a linear relationship between computer power and data generation for their applications. However, speed of computation increases exponentially, doubling roughly every two years as new processor technology is introduced, which results in a proportional doubling of the rate at which new data is generated.

Mass storage systems place no interpretation on the data that they store, and so are said to store *bitfiles* or uninterpreted strings of bits. To achieve a balance of performance and economy, a mass storage system uses a *multi-level storage hierarchy*. For this reason, mass storage systems are sometimes referred to as *hierarchical storage managers* or HSMs. Bitfiles that are active are cached on disks. The total capacity of an MSS disk farm may be hundreds of gigabytes in size, and aggregate transfer rates for disk arrays can be hundreds of megabits per second. The expense of direct access storage calls for the bulk of the data archive to be stored on a more economical medium, such as magnetic tape cartridges or optical disks. Larger or less frequently accessed bitfiles are cached in robotic libraries, each of which may contain thousands of tape cartridges or optical disks, and hold a total of hundreds of terabytes. Access time for the robot to fetch a tape or optical disk, *mount* it on a read/write drive, and locate the correct data is typically a few seconds to several minutes, and transfer rates can be a hundred megabits per second or faster. In the largest of archives, the majority of data may be stored offline on tape cartridges shelved in racks, and a human operator responds to requests from the MSS to mount an offline tape manually.

When the MSS determines that a bitfile is being accessed frequently, the bitfile may be *staged* or copied upwards in the storage hierarchy, from the offline archive into the tape library or onto the disks. As the disks or the library fill to capacity, the MSS identifies those bitfiles that are less frequently used and schedules them to be *migrated* or copied downward in the storage hierarchy to offline tapes, and *scrubbed* or *purged* from the disks or library. An entire bitfile, or portions of a bitfile, may exist in several places in the storage hierarchy simultaneously in order to speed up access. A bitfile may be migrated from disk to tape, but not scrubbed until the disk space is needed. The physical location of a file is transparent to the user except in the additional delay caused when a requested bitfile is not cached on disk and must be retrieved from tape. This is somewhat analogous to paging in a virtual memory system, and the MSS can be thought of as extending the memory hierarchy to very low speed, very high capacity (when compared to main memory) storage. Because all or part of the same bitfile may reside simultaneously on different storage devices or different distributed servers, the MSS must deal with cache coherency and concurrency issues to insure data integrity.

A local area network or LAN may be used to move data between an MSS and its clients. Commercial mass storage systems like Unitree, now a product of Unitree Software Inc., use internet protocols such as File Transfer Protocol (FTP) or Network File System (NFS), allowing the MSS to appear to be a terabyte-size file server. However, LAN bandwidth is inadequate for high-performance MSS clients such as supercomputers or massively parallel processors. Also, internet protocols are typically inefficient for moving large amounts of data quickly because most are limited by response time rather than by network capacity; increasing the network bandwidth by a factor of ten may yield only fractional improvement in data throughput. The *separation of control and data paths* in the client-MSS connection is one approach to overcoming the LAN bottleneck. Control messages between the MSS and its clients are still exchanged over the LAN, which is the *control path*. The data transfer occurs over the *data path*, high-speed I/O channels using *lightweight* I/O protocols which have lower overhead and better scalability as bandwidth increases than traditional LAN protocols because they are simpler or depend upon intelligent I/O devices for much of their processing.

To prevent the MSS itself from becoming a bottleneck, it may not participate directly in the data transfer. The data moves directly between the client system and the storage device which are both attached to a common *data fabric* or *storage area network* (SAN). The term "fabric" is used to distinguish it from a more typical local area network (although internet protocols may also run over the same physical network) and as a nod to the fact that data fabrics are frequently woven from (optical) fiber. Data fabrics usually employ switch-based I/O channel interconnection technologies such as High Performance Parallel Interface or HIPPI (at 800 megabits per second), or Fibre Channel Standard (at a gigabit per second). Either may be implemented over optical fiber or copper. The attachment of an intelligent storage device such as a disk array or tape library directly onto the data fabric as an active participant is referred to as *network attached storage*. The control of the storage devices by the MSS is called *third-party transfer*, since the controlling entity is at neither end of the data transfer. Network attached storage and the separation of the control and data paths was first introduced in 1985 in the NCAR MSS. It has since been accepted as a standard architecture for scalable mass storage systems, and incorporated in high-end commercial systems such as the High Performance Storage System or HPSS, jointly developed by IBM, the U.S. Department of Energy's National Storage Laboratory and others.

Associated with each bitfile in the mass storage system is *metadata*, for example the name of the file, who owns it, when it was created and last used, who is allowed to access it, and where it is located in the storage hierarchy. Metadata may also include information of interest to the user, such as what version of a simulation or which scientific instrument created the bitfile, or what types of post-processing has been applied to its contents. Losing metadata is as catastrophic as losing the data itself, since data without context is of little value. The metadata itself may be a valuable corporate or scientific resource. Database and data mining tools provided with the MSS for users to efficiently manage and manipulate metadata are becoming increasingly important as the amount of data stored per user increases exponentially. In the mid-1990s, Sequoia 2000, a research collaboration between the University of California and Digital Equipment Corporation, applied object-oriented database techniques to managing both metadata and data in mass storage systems.

The longevity of all digital information, both data and metadata, is a critical issue in the selection of storage technology for a mass storage system. Of concern is not only the physical lifetime of the media, for example how long a tape cartridge or optical disk will remain readable, but also the lifetime of the hardware and software technology used to create and access them. A digital magnetic tape may still be readable ten years after it is written, but the technology used to read it may not be commercially available after five years. Obsolete storage media such as punch cards and eight-inch floppy diskettes are similar to player piano rolls: the technology to read and write them is rare or non-existent. An MSS may have a hundred thousand or more offline tape cartridges holding petabytes of data. Since the time to mount and copy a single tape cartridge may be several minutes, the time required to copy an entire archive to a new medium is measured in *drive-years*. At the current rate of technological obsolescence, it is possible that a petabyte archive cannot be copied to a new medium before that medium itself is obsolete. The choice of storage technology in a mass storage system is often conservative, using not the fastest or the most dense storage media but rather the most reliable and the most likely still to have spare parts available a decade after its deployment.

Since 1990, the IEEE Storage System Standard Working Group (SSSWG, pronounced "sissy-wig") has been developing a *mass storage system reference model*. Version 5 of the model, which has been renamed the *Open Storage Systems Interconnection* or OSSI model, was released in 1994 for public review. The model describes seven components common to mass storage systems. The *Application Environment Profile* specifies the environmental software interfaces required by open storage system services. The *Object Identifier* defines the format and generation of object identifiers that uniquely identify each object (for example, each bitfile or physical volume) within the mass storage system. The *Physical Volume Library* defines the services that manage removable physical volumes (for example, tape cartridges) and optimize their use on read/write drives. The *Physical Volume Repository* defines the services that store removable physical volumes and selectively mount them onto drives. The *Data Mover* describes how the data fabric is used to transfer data. *Storage System Management* specifies a framework to consistently and portably monitor and control MSS resources and to implement site-specific storage management policies. *Virtual Storage Services* describes how portions of persistent storage are mapped into a single virtual storage image (for example, physical files on several tape cartridges, possibly in different libraries, into an image of a single bitfile). The version 5 OSSI model establishes a standard nomenclature for discussing mass storage systems. It is the goal of the Working Group to develop the model further to include standard programmatic interfaces for each component so that a mass storage system could be built from interchangeable compliant software and hardware from different vendors.

References

1990. Levy, E., and Silberschatz, A., "Distributed File Systems: Concepts and Examples", *ACM Computing Surveys*, 22.4, Dec. 1990

1994. IEEE Storage Systems Standards Working Group, Reference Model for Open Storage Systems Interconnection, Mass Storage System Reference Model Version 5, IEEE Project 1244, Sep. 1994

1995. Cole, J. and Jones, M., "The IEEE Storage System Standard Working Group Overview and Status", *Proc. 14th IEEE Symp. Mass Storage Systems*, IEEE Comp. Soc. Press, Monterey CA, Sep. 1995

1995. National Research Council, *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation’s Scientific Information Resources*, National Academy Press, 1995

1995. Rothenberg, J., "Ensuring the Longevity of Digital Documents", *Scientific American*, 272.1, Jan. 1995

1995. Watson, R., and Coyne, R., "The Parallel I/O Architecture of the High Performance Storage System (HPSS)", *Proc. 14th IEEE Symp. Mass Storage Systems*,

IEEE Comp. Soc. Press, Monterey CA, Sep. 1995

J. L. Sloan