STORAGE SYSTEMS AND TECHNOLOGIES

Jai Menon
IBM Academy of Technology, San Jose, CA, USA

Clodoaldo Barrera
IBM Systems Group, San Jose, CA, USA

Keywords: Storage Systems, Hard disk drives, Tape Drives, Optical Drives, RAID, Storage Area Networks, Backup, Archive, Network Attached Storage, Copy Services, Disk Caching, Fiber Channel, Storage Switch, Storage Controllers, Disk subsystems, Information Lifecycle Management, NFS, CIFS, Storage Class Memories, Phase Change Memory, Flash Memory, SCSI, Caching, Non-Volatile Memory, Remote Copy, Storage Virtualization, Data De-duplication, File Systems, Archival Storage, Storage Software, Holographic Storage, Storage Class Memory, Long-Term data preservation

Contents

Summary

Our world is increasingly becoming a data-centric world. Corporations differentiate themselves from their competitors based on what data they have and how they are able to leverage that data for competitive advantage.

The amount of data in the world continues to grow by leaps and bounds. According to International Technology Group, in 1996, the average corporation had 5 TB of data and storage expenditures represented 10% of the IT budget. In 2007, the average corporation had almost 225 TB of data and storage represented 22% of the IT budget.

Data continues to be stored primarily on disk drives, which were invented more than 50 years ago. The amount of data one can store per square inch of the surface area of a disk drive, also known as areal density, grew at between 60-100% CGR between 1990 and 2004. Since then, areal density has been increasing at a more modest rate of about 25-25% CGR. As it becomes harder and harder to pack more bits per square inch on a disk drive, researchers are looking at alternative storage technologies. We will briefly describe some of these alternative technologies that may emerge in the next 5-10 years. IT departments of enterprises and small and medium businesses use disk based storage systems to store their online data. Storage systems combine 10s to 100s of disk drives, plus a storage controller which caches the data to improve access performance and provides other advanced functions such as RAID. We will describe the functionality of a state-of-the-art storage system. We will also describe emerging new functionality in storage systems – such as enhanced forms of RAID, storage virtualization, and data encryption.

The nature of data stored in storage systems is beginning to change in some fundamental ways. This shift will have a profound impact on storage system design. The requirement to store traditional transactional data is being supplemented by the necessity to store other kinds of information for long periods of time. In 2005, 2,700 petabytes (PB) of storage was allocated worldwide for information that required long-term retention, and this amount is expected to grow to an estimated 27,200 petabytes by 2010. This chapter will review the requirements for long-term storage of data and will describe state-of-the-art storage systems that are emerging to deal with these new requirements.

1. Introduction

In this chapter we will begin by describing market trends that have an impact on the storage business. Then, we will describe the underlying storage technologies and how they are changing as the market needs are evolving.

The amount of data in the world continues to explode. IDC [19] predicts that the amount of data created and copied will grow from 281 Exabytes in 2007 to 1800 Exabytes in 2011, an annual growth rate of 60% (1024 MB = 1 GigaByte (GB); 1024 GB = 1 TeraByte (TB); 1024 TB = 1 PetaByte (PB); 1024 PB = 1 ExaByte (EB)) . 80% of the data is unstructured, generated by email, documents, images and videos. The traditional structured data used in large enterprises and stored in relational databases is now only

20% of total data.

Enterprises are storing more data and for longer periods than before for two main reasons – first, the large number of new rules and regulations in the US and internationally that companies have to adhere to that require corporate data preservation, and second, the availability of cheap Mips and new analysis techniques that provide corporations with new ways to analyze their business data for competitive advantage.

The rest of this chapter is organized as follows. We begin by covering the various forms of storage devices that store data, including the most popular and ubiquitous form of storage device which is the disk drive. Next, we describe storage systems which combine 10s to 100s of disk drives, plus additional intelligence. Customers typically buy storage systems to store their data. Storage systems can retrieve and store customer data and can hide most forms of storage device failures from the applications. They also support storage functions such as caching and RAID. Over the years, storage systems have become more functional and we will describe a current state-of-the-art storage system as well as what we can expect in the next few years. Following that, we will describe how computers connect to storage systems, through what are called storage networks. Finally, we discuss storage software. Storage software is necessary to manage the storage systems and storage networks in a data center.

2. Storage Devices

2.1. Storage Device Industry Overview

Development and production of storage devices is now a distinct industry segment occupied by companies with high volume manufacturing capability. Storage devices are built to a collection of industry standards for form factor, power, system interface, etc., and sold as OEM products to computer and storage systems integrators. Hundreds of millions of devices per year find their way into laptops, desktops, other mobile devices such as cameras, and computer data center equipment. We describe the most common of these devices, and make observations about the system requirements these devices seek to satisfy.

2.2. Hard Disk Drives

The Hard Disk Drive (HDD) is the principle non-volatile storage medium for practically every computer system in use today. The HDD serves as a secondary layer in the memory hierarchy, beneath the solid state Random Access Memory (RAM). They offer information storage that is much cheaper (about 100 times cheaper) than RAM, but with much longer access times (1000 times slower or more).

The basic design of HDDs uses one or more platters or disks coated with magnetic material stacked on a spindle turned by an electric motor. For each platter, a moving arm positions a read/write head over a particular track, a narrow ring of data around the circumference of the platter. The head flies over the surface of the platter as the platter turns below it, and senses magnetic domains encoded on the surface. Bits are recorded on a track using one of two basic schemes – longitudinal recording and perpendicular

recording. In longitudinal recording, the magnetic orientation of the data bits is aligned horizontally, as its name indicates, parallel to the surface of the platter. By contrast, in perpendicular recording, the magnetic orientation of the data bits is aligned vertically, perpendicular to the platter or disk. When the magnetic bits become too tiny, ambient temperature can reverse their orientations and data can get lost. Perpendicular recording allows smaller physical bits that are still stable at room temperature.

The response time of a disk drive is determined by the time to interpret a command, plus the time required to move the read/write head to the correct track (a 'seek' operation), plus the time for the disk to rotate to the correct record on the track, plus the data transfer time. Typical I/O operations take 1-10ms to perform. The technology to write and sense magnetic domains on rotating media has been one of the most dramatic success stories of the IT industry. Since its first introduction by IBM in 1956, magnetic recording technology has improved by a factor of 10**8 (100 million), allowing an iPod to store 10,000 times more data than the original HDD (the IBM RAMAC), which was the size of a refrigerator.

HDDs are manufactured in high volume, with over 400M units produced worldwide in 2007. There are four major product types:

- Desktop drives – used in desktop computer systems. These drives use the 3.5" form factor, and spin platters at 5400, 7200, or 10K RPM. The system interface is most commonly the IDE, ATA, or SATA interface. In 2006, desktop drives accounted for 40% of worldwide drive production. These drives are increasingly used in external storage systems attached to servers or storage networks.
- Laptop drives – used in notebook computers and some high-end portable music players. These drives accounted for 20% of worldwide production in 2006. Laptop drives use the IDE, ATA or SATA interfaces. The portable applications are sensitive to power and weight, so these drives are smaller and spin slower – a 2.5" form factor and 5400 RPM.
- Consumer drives – portable drives used in music players, GPS guidance systems, cameras, or other mobile applications. These are the smallest drives, with form factors as small as .8 inches. They accounted for 20% of production in 2006.
- Enterprise drives – the highest performance and highest capacity drives used as embedded drives in servers or in external storage systems that are attached to multiple servers through network interfaces. These drives use the 3.5" form factor, and currently range from 72GB up to 1TB in capacity, and spin at 10K and 15K RPM. They use the SCSI, SAS, and FC-AL system interfaces.

Disk areal density (bits/sq in) has improved historically at about 35%/year, with future improvements in the range of 25-35% in the coming years. During the years of 1992-2002 areal density improved at a rate of nearly 100% per year, with similar rates of improvement in the cost per capacity or $/GB (dollars per Gigabyte).

Improvements in other attributes of device behavior have been much slower. Most seriously, increases in performance measured by read operations/second and write operations/second have seriously lagged, with increases in the range of 5-8% per year.

As a result, the rate of operations/second per Gigabyte for an enterprise HDD has worsened by 100X in the past 10 years and Enterprise system workloads that require a high rate of small record operations have begun to see storage as a performance bottleneck.

The cost per gigabyte (GB) of disk drives will continue to decrease, but at a reduced rate. Instead of drive capacity doubling every 12 to 18 months, it will now double in 24 to 36 months. The 2007 high volume price estimates are \$1.00-2.00/GB for enterprise disks and \$0.30-0.60/GB for desktop disks. The recent trend, which is expected to continue indefinitely, is for the cost per gigabyte to decline at approximately 40% per year (CAGR).

The form factor of disk drives is also evolving. Currently, a transition is taking place from 3.5″ to 2.5″ drives. This transition is being driven by a number of factors, including the proliferation of laptops and the need of the enterprise system for higher storage performance in a smaller space. It is likely that the next form factor transition will take place after 2015. This transition will be from 2½″ to 1.8″ disks. If current areal density and packaging trends continue, then, in 2020, the likely capacity of disk drives will be: ~20 terabytes (TB) for a 3.5″ drive, ~10TB for a 2.5″ drive and ~5TB for a 1.8″ drive.

The power P supplied to a disk drive is given by: $P = C + M + S$. In this equation, C is the power supplied to the control logic of the disk drive. M is the power supplied to the motor to spin the disk drive. $M \sim d^{4.6} \times r^{2.8}$, where d is the diameter of the disk, and r is the rotational speed. S is the power supplied when the heads are moved to a new track. It is dissipated *only* when the heads are moving. Only S varies during normal system operation. A common approximation rule suggests that each of these draws about one third of the total disk power. To lower power consumed, one can either use a smaller disk form factor or one can shut the HDD down completely when not in use. However, because it takes such a long time to power-up a disk drive (~20 seconds), the latter is only practical for disk-based archival systems [1]. As power becomes the central issue for data centers, this power constraint will become a significant drawback for using disks as a storage medium.

2.3. Digital Tape Drives

Digital tape has been in use slightly longer than HDDs, and forms the third layer of the classic information storage hierarchy. Tape has several key attributes that keep it in use in data center operations:

- It is the cheapest large scale storage medium (\$/Gbyte), and has maintained a 10X advantage over disk for several decades.
- The media cartridge can be removed from the tape drive or tape player, allowing the media to be stored in dense libraries with no power usage.
- Media cartridges can be removed & sent to offsite vaults or exchanged between data centers.

Digital tape has historically been used for backup data and for archival storage.   The

drawbacks of tape use are:

- Slow access to data stored on tape. A cartridge must be mounted on a drive (player), and the tape must be sequentially searched for the data desired.
- Tapes are fragile, and can be damaged by environmental conditions, or mechanical stress. Tapes written on one drive may be unreadable on another drive.
- Tape archives require special software to create indices, since it is not searchable.

Modern data center tape operations use automated libraries with robot arms to pick cartridges from shelves and place them in drives. This automation eliminates human error, and minimizes exposure of tapes to hazards. Libraries come in a variety of capacity points, from 10's of TB (Terabytes) up to 10's of PB (Petabytes) in a single library.

Worldwide spending for datacenter tape has been flat or slightly declining in recent years. Tape for small computers (laptops, desktops) is in significant decline, as these users rely more on disk backup and network backup approaches. In the high-end market, the decline of tape for backup is more than offset by their use in archives, which will grow very large in some industries. Medical records and film and music archives need to be preserved 'forever', and continue to grow as more new content is added.

There have been a variety of tape formats historically, with different cartridges, tape widths and lengths, and data formats. As the use of tape becomes more limited to data center backup and archive applications, there has been consolidation around fewer formats. The Linear Tape Open (LTO) format, now on its fourth generation, appears to be positioned as the surviving midrange format, together with the half inch enterprise formats from IBM and STK.

2.4. Optical Storage

Optical storage has a well established role as a distribution media for video, music, games, and software. These high volume media are non-writeable, and therefore play a limited role in computer systems.

Optical storage is any storage method in which data is written and read with a laser. Typically, data is written to optical media, such as CDs and DVDs. Optical drives of all kinds operate on the same principle of detecting variations in the optical properties of the media surface. CD and DVD drives detect changes in the light intensity, and magneto-optic drives detect changes in the light polarization. All optical storage systems work with reflected light. For several years, proponents have spoken of optical storage as a near-future replacement for both hard drives in personal computers and tape backup in mass storage. Optical media is more durable than tape and less vulnerable to environmental conditions. On the other hand, it tends to be slower than typical hard drive speeds, and to offer lower storage capacities. A number of new optical formats, such as Blu-ray, use a blue laser to dramatically increase capacities.

Write once or re-writeable optical media have a role in the computer storage hierarchy, mostly in low-end or personal systems. Optical storage has higher performance and easier retrieval attributes than low-end tape for small office and home systems, and has become the preferred media for those users who want to create their own archives of data, downloads, pictures, etc. Slower write performance, and lower capacities than magnetic drives limit the value of optical storage in larger computing environments.

2.5. Emerging Storage Technologies

Attempts have been made for years to replace spinning magnetic technology as the second layer in the storage hierarchy, with technologies such as holographic storage, magnetic bubbles and battery protected memory.

2.5.1. Holographic Storage

In **holographic data storage** [18], an entire page of information is stored at once as an optical interference pattern within a thick, photosensitive optical material. This is done by intersecting two coherent laser beams within the storage material. The first, called the object beam, contains the information to be stored; the second, called the reference beam, is designed to be simple to reproduce. The resulting optical interference pattern causes chemical and/or physical changes in the photosensitive medium: A replica of the interference pattern is stored as a change in the absorption, refractive index, or thickness of the photosensitive medium. When the stored interference grating is illuminated with one of the two waves that was used during recording, the other wave is reconstructed. Illuminating the stored grating with the reference wave reconstructs the object wave, and vice versa.

A large number of these interference gratings or patterns can be superimposed in the same thick piece of media and can be accessed independently. Any particular data page can then be read out independently by illuminating the stored gratings with the reference wave that was used to store that page. The theoretical limits for the storage density of this technique are around tens of terabits per cubic centimeter.

In addition to high storage density, holographic data storage promises fast access times, because the laser beams can be moved rapidly without inertia, unlike the actuators in disk drives. With the inherent parallelism of its page-wise storage and retrieval, a very large compound data rate can be reached by having a large number of relatively slow, and therefore low-cost, parallel channels.

A rather unique feature of holographic data storage is associative retrieval: Imprinting a *partial* or search data pattern on the object beam and illuminating the stored holograms reconstructs all of the reference beams that were used to store data. The intensity that is diffracted by each of the stored interference gratings into the corresponding reconstructed reference beam is proportional to the similarity between the search pattern and the content of that particular data page. By determining, for example, which reference beam has the highest intensity and then reading the corresponding data page with this reference beam, the closest match to the search pattern can be found without initially knowing its address.

Because of all of these advantages and capabilities, holographic storage has provided glimpses of becoming an intriguing alternative to conventional data storage techniques for three decades. It is still unclear whether the promise will be realized in practice. In 2005, InPhase technologies showed a working prototype at the National Association of Broadcasters convention in Las Vegas, and in 2006, they published a chapter reporting an achievement of a holographic storage device storing data at 500 Mbits/sq. inch.

2.5.2. Flash Storage

There has been a lot of recent interest in **Flash** memory, which is solid state (meaning it has no moving parts) and non-volatile (meaning the contents are not lost when power is turned off). The interest in Flash stems from its pervasive use in consumer electronics applications such as cameras and iPods, and because Flash prices have fallen from $600 per MB in 1987 to $0.01 per MB in 2007 (a factor of 60,000 in 20 years). While it is still significantly more expensive than hard disk drives, there are cases where it is being used instead of magnetic disk drives because of its much higher performance (25 μs access time versus 5-10 ms for disk), its lower power consumption, or its better reliability (no moving parts). For example, we are already seeing many laptops that only use Flash drives instead of magnetic drives.

Flash drives will need to overcome three problems before we expect to see them used instead of magnetic drives in any significant way. First, the cost must continue to drop. By 2010 we expect to see Flash drive prices come within a factor of 10 of disk drive prices. This is through use of MLC (multi-level cell) technology which allows for storing multiple bits per cell. Second, Flash controllers must be developed that overcome the write endurance problem of Flash which is the limitation that a particular cell in the Flash memory can only be written 10^4 to 10^5 times before it becomes unusable. Finally, Flash controllers must be able to effectively hide the poor write performance of Flash caused by the need to erase a location before it can be rewritten. The first generation of Flash controllers which attack these problems are becoming available. So, by 2010-2011, we expect to see more pervasive use of Flash drives, primarily in applications where performance is important. However, Flash will never completely replace disk drives.

-
-
-

TO ACCESS ALL THE **24 PAGES** OF THIS CHAPTER,
Visit: http://www.eolss.net/Eolss-sampleAllChapter.aspx

Bibliography

[1] D. Colarelli, D. Grunwald, and M. Neufeld, "The Case for Massive Arrays of Idle Disk (MAID)," *Usenix FAST Conference*, Monterey, California (2002); see *http://www.usenix.org/events/fast02/wips/colarelli.pdf.*

[2] P. Arnett and J. Chang, "Non-volatile diode Cross-Point Memory Array," United States Patent, No.

3838405 (1973).

[3] Special Issue on "3D Integration," *IEEE Design&Test of Computers*, 22, No. 6 (2005).

[4] B. Eitan and A. Roy, "Multilevel Flash Cells and Their Trade-Offs," *IEDM Technical Digest* , 8-11 Dec. (1996), pp. 169-172.

[5] Gopalakrishnan, K.; Shenoy, R.S.; Rettner, C.T.; King, R.S.; Zhang, Y.; Kurdi, B.; Bozano, L.D.; Welser, J.J.; Rothwell, M.E.; Jurich, M.; Sanchez, M.I.; Hernandez, M.; Rice, P.M.; Risk, W.P.; Wickramasinghe, H.K., "The micro to nano addressing block (MNAB)" Electron Devices Meeting, 2005. *IEDM Technical Digest*, 5-7 Dec. (2005) pps. 471-474.

[6] S. Hudgens and B. Johnson, "Overview of Phase-Change Chalcogenide Nonvolatile Memory Technology," *MRS Bulletin* 29 No. 11, 829-832 (2004).

[7] Patterson, D. A., Gibson, G., and Katz, R. H. 1988. A case for redundant arrays of inexpensive

disks (RAID). In Proceedings of the ACM SIGMOD International Conference on Management of Data (Chicago, IL). 109–116.

[8] Chen, P. M., Lee, E., Gibson, G., Katz, R., and Patterson, D. 1994. RAID: High-performance, reliable secondary storage. ACM Computing Surveys 26, 2 (June), 145–185.

[9] Blaum, M., Brady, J., Bruck, J., and Menon, J. 1995. EVENODD: an efficient scheme for tolerating double disk failures in RAID architectures. IEEE Trans. Computer. 44, 2 (Feb.), 192–202.

[10] Corbett, P., English, R., Goel, A., Grcanac, T., Kleiman, S., Leong, J., and Sankar, S. 2004. Row-diagonal parity for double disk failure correction. In Proceedings of the 3rd USENIX Conference on File and Storage Technologies (FAST) (San Francisco, CA). 1–14.

[11] A. Dholakia, E. Eleftheriou, X.–Y. Hu, I. Iliadis, J. Menon, K.K. Rao, "A New Intra-Disk Redundancy Scheme for High-Reliability RAID Storage Systems in the Presence of Unrecoverable Errors," IBM Research Report, RZ 3689, June 2007.

[12] J. McKnight, "ESG Research Report: Digital Archiving: End-User Survey and Market Forecast 2006-2010," Enterprise Strategy Group (January 2006).

[13] J. Menon and Cortney, "Architecture of a Fault-Tolerant Cached RAID Controller", Proceedings of 20th International Symposium on Computer Architecture, April 1993.

[14] M. Farley, "Building Storage Networks, Second Edition", Osborne Publishing 2001.

[15] J. Hufferd. "iSCSI: The Universal Storage Connection", Addison-Wesley Publishing Company 2003.

[16] S. Rabinovici-Cohen, M. E. Factor, D. Naor, L. Ramati, P. Reshef, S. Ronen, J. Satran, and D. L. Giaretta, "Preservation DataStores: New storage paradigm for preservation environments," *IBM Journal of Research and Development*, Vol. 52, Number 4/5, 2008.

[17] *ISO Standard 14721:2003,* "Space Data and Information Transfer Systems—A Reference Model for an Open Archival Information System (OAIS)," International Organization for Standardization, 2003.

[18] J. Ashley, M.-P. Bernal, G. W. Burr, H. Coufal, H. Guenther, J. A. Hoffnagle, C. M. Jefferson, B. Marcus, R. M. Macfarlane, R. M. Shelby, and G. T. Sincerbox, "Holographic Data Storage," *IBM Journal of Research and Development*, Vol. 44, Number 3, 2000.

[19] John F. Gantz, Christopher Chute, Alex Manfrediz, Stephen Minton, David Reinsel, Wolfgnag Schlichting, and Anna Toncheva, "The Diverse and Exploding Digital Universe," *IDC White Paper*, March 2008

[20] "Digital Archiving: End-User Survey and Market Forecast 2006-2010", *ESG Research*, January 2006

Biographical Sketches

Dr. Jai Menon is responsible for shaping IBM's technical strategy and identifying emerging technologies critical to the future of IBM as chair of the IBM Academy of Technology - an elite body of worldwide leaders whose mission is to provide technical direction and leadership for IBM. In addition, Jai serves as the global leader of IBM's University Collaboration programs and works with academia to establish

innovative research and build 21st century skills in support of the innovation economy.

Jai began his career with IBM Research in 1982. He directed IBM's world-wide research in Storage Systems and his leadership helped establish IBM Almaden as a world-class center of competence in storage research. Research projects that Jai lead or initiated contributed to the creation of IBM's flagship Shark Storage Server and the industry's leading storage virtualization product. In 2003, Jai joined the Systems Division as CTO for Storage and Systems Software, where he was responsible for setting the strategy and direction of IBM's storage and systems software products.

As an IBM Fellow, Jai has achieved the highest technical position within IBM. He is an IEEE Fellow and a Member of IBM's Academy of Technology. He is also an IBM Master Inventor. Over the course of his career, Jai has been awarded 52 U.S. Patents and has been the recipient of many IBM technical awards and external honors. For example, in 2002, he received the IEEE Wallace McDowell Award and in 2004 received the Distinguished Alumnus Award from the College of Engineering at Ohio State University. In 2006, he became a Distinguished Alumnus of IIT, Madras and received the IEEE Reynold B. Johnson Information Systems Award. In addition, Jai has published 31 papers and 47 technical reports and served as a contributing author to three books.

Jai received a Bachelor of Technology in Electrical Engineering from the Indian Institute of Technology in 1977 and earned M.S. and Ph.D degrees in Computer Science from Ohio State University in 1978 and 1981.

Clodoaldo Barrera is a Distinguished Engineer and the Chief Technical Strategist for IBM Systems Storage in San Jose California. His responsibilities include development strategy for IBM's disk and tape subsystems, storage management software, and SAN and NAS solutions. Prior to his current position, Mr. Barrera was the development executive for IBM's serial storage product development, and for future storage systems development. Before joining IBM's Storage Division, Mr. Barrera was a programmer and a development manager in printer product development in San Jose, Tucson Arizona, and Boulder Colorado. He has also served for two years on the IBM Corporate Development Staff in Armonk, New York.

Mr. Barrera is a founding member of the Storage Networking Industry Association, and served as a Director and Secretary of the Board. He has also served as a board member of the Fibre Channel Association. Mr. Barrera is a graduate of Stanford University, and holds a Bachelors Degree in Mathematics and a Masters Degree in Electrical Engineering. He is also a Senior Member of the IEEE Computer Society and of the ACM.