CHAPTER

16

STORAGE AND ENTERPRISE ARCHIVING

PAUL G. NAGY • THOMAS J. SCHULTZ

A study conducted at the Berkeley School of Information Management and Systems concluded that the world generated more than 5 exabytes (5 billion gigabytes [GB]) of recorded information in 2002. This represents a 30% growth per year for the past 5 years. More information has been generated in the past 3 years alone than in the previous 40,000 years of civilization combined. The entire printed Library of Congress holds 17 million books, or 136 terabytes (TB) of data.

Hospitals have seen a parallel data explosion, especially in the imaging modalities. A single 2000-slice computed tomography (CT) procedure captures 1 GB of information. Advances in molecular imaging, multispectral imaging, and physiologic imaging suggest that the rate of growth in storage requirements will only accelerate in the future. We have clearly entered the era of information overload, and our success now depends on how well we can capture, retrieve, and synthesize this avalanche of information.

TABLE 16.1
Storage Terminology: Powers of 10

Kilo	10^3	1 kilobyte (KB) = 1000 bytes
Mega	10^6	1 megabyte (MB) = 1000 KB
Giga	10^9	1 gigabyte (GB) = 1000 MB
Tera	10^{12}	1 terabyte (TB) = 1000 GB
Peta	10^{15}	1 petabyte (PB) = 1000 TB
Exa	10^{18}	1 exabyte (EB) = 1000 PB
Zetta	10^{21}	1 zettabyte (ZB) = 1000 EB
Yotta	10^{24}	1 yotta (YB) = 1000 ZB
Google	10^{100}	

The storage industry has been providing innovations at a breakneck pace to attempt to manage this data deluge. Expanding storage and the escalating pace of capacity and performance are among the true marvels of the computer age. Historians like to study the storage industry in particular because it is somewhat like studying the evolution of fruit flies, with companies innovating and the market reforming itself every few years to constantly embrace disruptive changes. Gordon Moore, the cofounder of Intel, predicted in 1965 that the number of transistors on a microprocessor would continue to double every 18 months for the same cost. Now known as Moore's law, this observation applies not only to microprocessors but to almost every facet of computing, especially storage. Over the last 50 years, the storage areal density (number of bits that can be squeezed into 1 square inch) has increased by a factor of 17 million at the same time that the vocabulary of storage terminology has expanded to accommodate previously undreamed of capacities (Table 16.1). Widespread agreement on the accuracy of Moore's law has led to an entirely new philosophy of equipment buying: if everything will be better and cheaper next year, then buy only what you need and make sure your system is designed to incorporate new technology that has not yet been invented.

Archiving technology is described as either online, near-line, or offline. Hard drives are online technology, because they are instantly accessible on a moment's notice. Tape and optical media are considered near-line because

they may need to be automatically retrieved from a robotic jukebox and mounted before being accessible. This generally takes 10 to 30 seconds if drives are available in the jukebox.

Offline storage requires manual intervention to load the media off a shelf and into the reader. As storage technology is becoming dramatically less expensive, offline storage should be discouraged. The true cost of offline storage is higher than most users perceive, because it includes the delay in time to retrieve data, the operator's time, and, most important, the much higher probability of losing data by mislabeling, misplacing, or damaging the media through improper storage.

HARD DRIVES

The hard drive was invented at the research laboratories of IBM in 1952. The random access method of accounting control (RAMAC) stored its data in 50 stacked, 2-ft-wide aluminum platters. The platters rotated at 1200 revolutions per minute (rpm), and the system weighed more than 1 ton. At the time, RAMAC was considered a marvel because it could hold 5 megabytes (MB) of data—the storage equivalent of 50,000 punch cards.

DRIVE MECHANICS

A hard drive consists of cylindrical platters, with thousands of single-bit-wide tracks of data (Figure 16.1). The surface of the platter is layered with a ferromagnetic material and polished so that the surface is extremely uniform. To read information off the surface of the hard drive, an electromagnet on a moving arm skims the surface of the platter at a height of less than 0.10 mm. When the head is in read mode, it will induce a current on the head from the magnetic surface of the hard drive and thus be able to read information. When the hard drive is writing to the disk, the head is energized and changes the polarity of the magnetic surface of the platter. When the ferromagnetic material is exposed to an external field, it is permanently magnetized.

The hard drive surface is broken into rings, also known as tracks. Each track on the hard drive is broken into sectors. A platter can have upwards of 30,000 tracks and between 100 and 500 sectors per track. When a computer requests a file from the hard drive, the file is indexed by the track and sectors in which it is located. Sectors can be of various sizes, but, for large

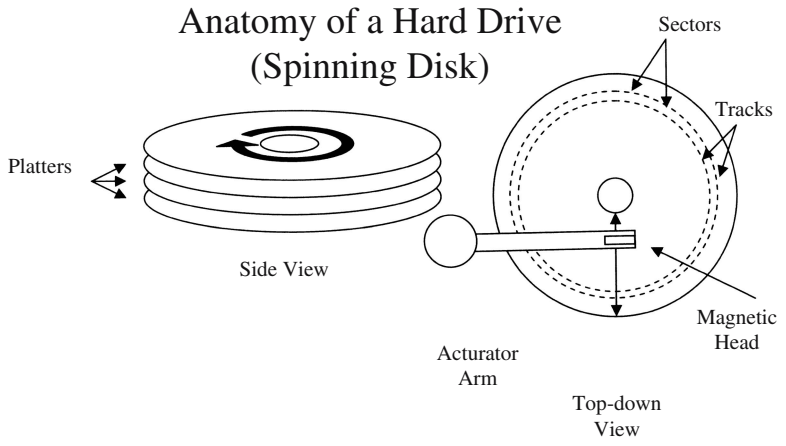


FIGURE 16.1

Terminology and internal mechanisms of a hard drive.

file applications like a picture archiving and communications system (PACS), a typical sector size is 64 kilobytes (KB). The hard drive spins the platter underneath the head and rotates at a rate of 5400 rpm for relatively slow drives and up to 15,000 rpm for high-speed drives. The two key performance metrics for hard drives are defined as the seek time and the data transfer rate.

The seek time is the time that it takes to access the beginning piece of data requested. On average, this is the time it takes for half of a rotation to position the sector requested underneath the head (a lag known as rotational latency). This is dependent on the speed at which the hard drive is spinning. For a 5400-rpm hard drive, the access time would be calculated as:

$$\frac{1 \text{ minute}}{5400 \text{ rotations}} \times 0.5 \text{ rotation} = 5.6 \text{ milliseconds}$$

For a 15,000-rpm hard drive, the access time is significantly reduced:

$$\frac{1 \text{ minute}}{15{,}000 \text{ rotations}} \times 0.5 \text{ rotation} = 2.0 \text{ milliseconds}$$

A high-speed hard drive is ideal for applications such as databases and e-mail servers that can process more than 300 in-and-out requests per second from

a storage system. For storing large image files, however, it is arguable that a high overall transfer and throughput rate is more important than saving a few milliseconds on access time.

Data transfer rate is not quite as simple to calculate because it is dependent on the geometry of the hard drive and the location of the data. Data transfer rates are twice as great around the outer radius of hard drives because the circumference at the outer radius is twice that of the tracks around the inner radius. On average, high-rpm drives transfer data at 40 to 60 MB per second, and slower-rpm drives transfer data at 25 to 40 MB per second.

The highest-capacity hard drive released to date is the 400-GB hard drive by Hitachi (Tokyo, Japan). The 400-GB hard drive has 5 platters and 10 heads that read both sides of the platters. The Hitachi drive rotates at 7200 rpm, has an access time of 8.5 milliseconds, and is rated at an average transfer rate of 45 MB per second.

DISK ARRAYS AND INTERFACES

Hard drive performance is often limited by the way in which the drive is connected to the computer. The hard drive is defined by this connector, called the interface. Many types of interfaces are available, but the principal ones for enterprise storage applications are serial ATA, small computer system interface (SCSI), and Fibre Channel. Parallel ATA, or AT attached, was originally named for the IBM AT computer and is the most common type of hard drive in the commercial PC market.

Serial ATA is an evolved version of parallel ATA that provides enterprise performance and reliability at commercial pricing. ATA drives are included in the majority of the PC market, with a volume of 10 times that of SCSI, which allows them a much lower cost per gigabyte of storage.

SCSI has historically been the hard drive style of choice for enterprise applications. Multiple hard drives can be daisy-chained off a single SCSI connection. Fibre Channel is an outgrowth of SCSI. In fact, it uses the same SCSI command set but transports it over a serial connection, whereas SCSI uses a parallel bus. Fibre Channel is an unfortunate misnomer in that it is a protocol and does not depend on fiber-optic connections at the physical layer. Although Fibre Channel can be run over copper wire, it is usually run over fiber optics because of the extremely fast speed that optical switching provides. Fibre Channel hard drives have the fastest interface and are also the most expensive type of hard drive on the market. Table 16.2 summarizes hard drive interface performance.

TABLE 16.2
Hard Drive Interface Performance

Interface	Theoretical Performance (MB/s)
USB 1.0/1.1	1.5
Ultra ATA/33	33
IEEE 1394/Firewire	50
USB 2.0	60
Ultra ATA/66	66
Ultra ATA/100	100
Ultra ATA/133	133
Serial ATA/150	150
Ultra 160/SCSI	160
Serial ATA II/300	300
Ultra 320/SCSI	320
Fibre Channel	400+

Source: Connolly C. Highpoint 1520 RocketRAID Serial ATA/150 controller. Game PC. July 27, 2002. Available at: www.gamepc.com/labs/view_content.asp?id=hpsataraid&page=2. Accessed September 15, 2004.

RELIABILITY AND CALCULATING MEAN TIME TO FAILURE

Failure of computer components is a given and must be taken into account when designing an archive. Computer components with moving parts, such as the hard drive, have a higher probability of failure than solid-state components, such as the central processing unit and motherboard. Hard drive bearings can seize, heads can crash into platters, or the actuator arm can lock up. Most enterprise hard drives are rated with a mean life-to-failure of more than 1 million hours of operation. This does not mean that the average hard drive will actually last for 1 million hours. This model assumes a 5-year lifespan that is covered in the warranty and that all drives are replaced every 5 years. This is an important difference and means that a migration plan should be in place to prevent going over the warranty. Over the 5 years (43,800

hours) of warranty for the hard drive, there is a 4% failure rate. In other words, for every 22 hard drives running over the 5 years, expect at least 1 hard drive to fail.

REDUNDANT ARRAY OF INEXPENSIVE DISKS

Redundant array of inexpensive disks (RAID) is one of the core concepts of implementing a high-performance, high-availability large storage solution. RAID allows multiple hard drives to work in concert and appear to the computer as a single storage device. Several RAID configurations are available and provide different combinations of redundancy and performance characteristics. To orchestrate the hard drives, a RAID controller is utilized between the hard drives and the computer. The RAID controls the hard drives directly and provides blocks of storage to the operating system.

RAID 0, striping, splits data to 2 hard drives simultaneously. The benefit to striping is that the storage system can be twice as fast at reading and writing data because it utilizes 2 drives. The problem with striping, however, is that there is no data redundancy. If either drive fails, all data on both drives are lost.

RAID 1, mirroring, copies the data to a shadow hard drive. RAID 1 provides complete redundancy in the event that 1 of the hard drives fails but does not enjoy any performance benefits because the user is still reading from a single drive. The other challenge to RAID 1 is effectively the loss of the capacity of the second drive.

RAID 10 is called a nested strategy and combines RAID 1 and RAID 0 levels. Two hard drives are striped, and both of those hard drives have a mirror copy in case of failure. This simple technique provides the performance benefits of 2 drives while allowing for failure. The drawback of RAID 10 is that half the storage is used making a backup of the data.

RAID 3 requires a minimum of 3 hard drives and is a combination of striping and a checksum. As data are split among multiple hard drives, a checksum, or parity bit, is calculated. If any of the drives fails, the information can be reconstructed onto a new hard drive from this checksum. RAID 3 uses a fixed parity drive with synchronized disk rotation and calculates checksums at the bit level. The challenge to parity drives is that the capacity of 1 drive is lost and that writing data to the disk array can be slower because it is calculating the parity bit. The read speed from RAID 3 is quite fast because the data is striped over all the drives in the array.

RAID 4 is a variant of RAID 3. RAID 4 calculates the parity bit on a dedicated drive at the block level instead of at the lower bit level. This

improves the random access performance of the disk array compared with RAID 3.

RAID 5 is a variant of RAID 4 that also calculates the parity at the block level. However, RAID 5 does not use a dedicated drive to store the checksum but, instead, distributes the parity information among all the drives; this improves random access write performance. RAID 5 has relatively poor write performance but excellent read performance that matches well to the clinical requirements of PACS. Performance truly matters with PACS when a user selects a patient name on a list and wants instantly to read large image studies. PACS is considered a read-intensive application in that every study that is written to the system is read 5 to 10 times in its lifetime for diagnosis, relevant priors, enterprise distribution, archiving, and migration.

A new technique not yet in common use is RAID 6. RAID 6 is similar to RAID 5 but calculates 2 parity checks that allow the system to suffer any 2 hard drive failures without compromise. The catch is the loss of capacity of 2 drives on the system.

In clinical practice, RAID 5 is the predominant technique for retaining medical images on storage arrays. RAID 5 is considered the ideal configuration for the PACS industry because it provides good performance through striping across multiple drives and at the same time provides redundancy and can survive losing the capacity of 1 drive in the array. With 1 drive out, however, the system must run in compromised condition in which another drive failure would result in the loss of the entire RAID array. That is why it is always good policy to have an extra hard drive sitting on a RAID array to act as a hot spare. During the event of a failed drive, the system will activate the spare and start rebuilding the spare drive. The time it takes to rebuild the spare hard drive and restore the RAID is called the mean time to recovery (MTTR). This recovery rebuild rate is usually in the range of 20 to 50 GB/hour. For a 200 to 300 GB drive, this exposed time could be several hours. A RAID 5 system is fully operational to external requests during the rebuild time, although its performance will be degraded because it is rebuilding a drive in the background. The probability of another failure during the rebuild time is extremely remote. When a system seems to go against the odds by having more failures than are statistically probable, environmental issues, such as high internal temperature or excessive vibration in the system rack, may be the cause. One common myth is that hard drives and system components fail without any warning signs. A study performed in 1999 tracked the system log of 368 hard drives and showed 4 hard drive failures over an 18-month period. The drives started warning the system between 5 and 186 hours before each failure, generating on average more

than 1000 messages. It is important to have a systems management tool in place and adequately configured so that the administrator can be alerted to the problem and be proactive in finding a solution.

The most common systems management tool is the simple network management protocol (SNMP). This protocol allows health information about a server to be communicated between systems to provide a holistic view of the status of the entire system. SNMP tools can track fans, hard drives, and power supplies for failure and can e-mail administrators in the event of failure. These tools also can routinely monitor the temperature of these components, as well as the drawing power used. Limits can be set for identifying impending problems. The Storage Management Initiative Specification (SMI-S), an open standard developed by the Storage Network Industry Association (SNIA), has been recently introduced. Whereas SNMP is developed around a passive mode for catching error messages, SMI-S is designed to be more active, allowing administrators to control and reconfigure enterprise storage devices.

ROLE OF THE FILE SYSTEM

Now that all your hard drives are sitting in a nice disk drive rack mount array, how does the PACS application connect to and make use of them? Sitting between an application and the low-level block access storage system lies the file system, as shown in Figure 16.2 The file or filing system is the

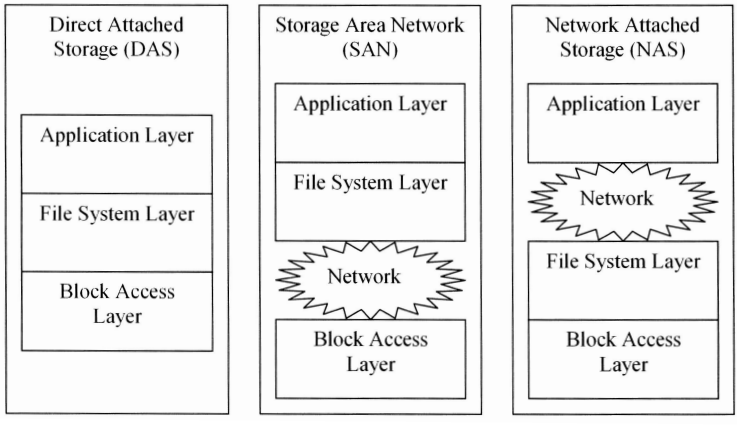


FIGURE 16.2

File system.

card catalog of a storage system and knows where files are stored and presents a hierarchical view that helps users locate them. When searching through the treelike structure of your file system in an application such as Windows Explorer, you are viewing the file system and not actually scanning through the entire hard drive. When you defragment a hard drive, the file system reorganizes the hard drive to put files into contiguous sectors for faster reading. The common file systems for Windows are FAT16, FAT32, and NTFS. NTFS is the newest file system from Microsoft and has the best features for PACS applications. NTFS can address up to 256 TB of data using a 64-KB cluster size and can support up to 16-TB files. FAT32 can address up to only 32 GB, and the largest file it can support is 2 GB.

DIRECT ATTACHED STORAGE, STORAGE AREA NETWORKS, AND NETWORK ATTACHED STORAGE

Direct attached storage (DAS) is using the hard drives that come with the server (Figure 16.3). Storage area networks (SANs) and network attached storage (NAS) are the only 2 methods for scaling storage by inserting a network topology at either the file level (NAS) or the block level (SAN) between the storage system and the application.

DIRECT ATTACHED STORAGE

A 5U server might have the capacity for 10 to 12 drives to be added, typically yielding a capacity of around 2 to 3 TB. This is the least expensive approach for small storage requirements but also provides the least amount of redundancy. If any component fails on the server and stops the application, access to the data will also be lost. Another challenge of DAS is the likelihood of running out of space. A direct attached solution does not scale well.

STORAGE AREA NETWORK

Storage area network refers to a network dedicated to storage access at the block level (Figure 16.3). Networking typically involves Ethernet and

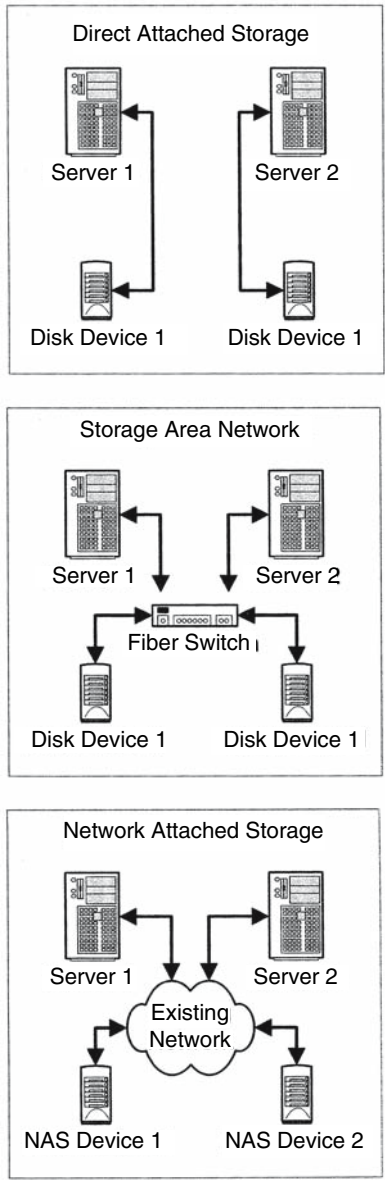


FIGURE 16.3

Simple example of direct attached storage (DAS), storage area network (SAN), and network attached storage (NAS).

TCP/IP standards, but a SAN uses neither. A SAN uses the Fibre Channel communications protocol to directly access storage devices. The file system remains on the computer, but low-level block-level access is routed through an external network of storage devices. These external storage devices can be daisy-chained together into an arbitrary loop or can be driven from a star topology using a SAN switch.

A SAN is perfectly suited for a busy information services department running dozens or even hundreds of servers. A SAN can be used to handle the storage needs of servers running different operating systems, although each system sees only its own blocks of data. Instead of having to monitor and manage the storage needs on each of those computers, storage management can be consolidated into a SAN and managed centrally. A SAN puts the storage in one place to manage and can automatically grant computers more storage on demand when it sees them approaching their storage limits. This provides significant economies in the operating costs for the servers. A SAN can provide logic at the block level to copy data to different storage devices and provide a transparent level of fault tolerance to the computers as well. Another unique advantage of SAN is that it can move and back up data from 1 node of the SAN (such as a RAID array) to another node (such as a tape archive) without drawing any resources from the host computers. This is called a clientless backup.

A SAN is the most frequently used storage method for PACS database redundancy. With fault-tolerant clustering, 2 servers share the same SAN partition, although 1 server sits passive. The 2 servers are linked via a heartbeat connection that monitors the active server and, in the event of failure, alerts the passive server to assume the network name of the server and take over the shared storage partition and database processes.

SAN is the highest performance type of scalable storage architecture because it provides such low-level access—but it is also the most expensive. One reason is that optical connections are more costly than copper ones. In addition, the storage network industry is orders of magnitude smaller than the Ethernet network industry and so cannot provide buyers the benefits of economies of scale. Because it is not used as widely, the Fibre Channel standard is not as plug-and-play as networking standards. Some incompatibilities are often intentional. SAN providers often include proprietary code for the following: performance/value-added features (such as multihost bus adapter [HBA], the card that connects the server to the SAN), unique software for caching read/write requests in fast memory, replication (backup to another location) technology, fault monitoring, and SAN management. A SAN can become an engineering exercise and require complex installation

and ongoing maintenance. Expertise in Fibre Channel is not as common in information technology organizations as networking expertise and is therefore more costly. Figure 16.4 is a PACS example of a multivendor SAN implementation.

To alleviate these problems, an exciting and potentially disruptive technology has been developed: Internet SCSI or iSCSI. Internet SCSI is a communications protocol that sends and receives low-level SCSI

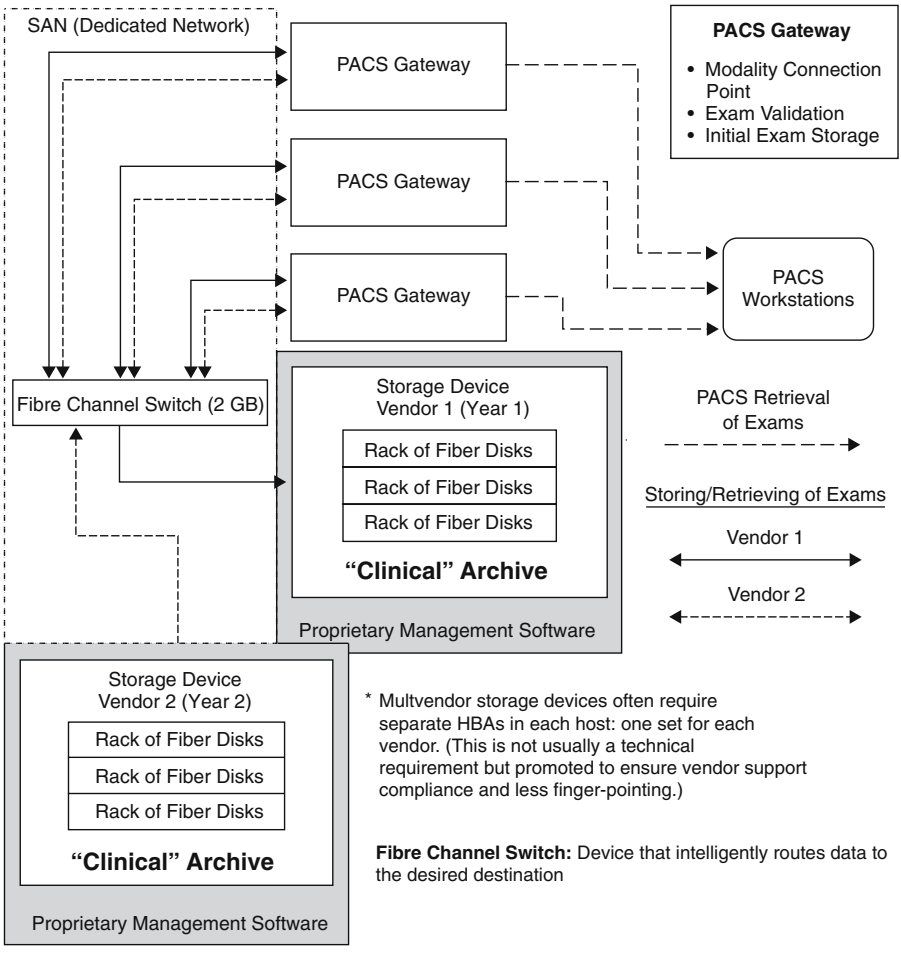


FIGURE 16.4

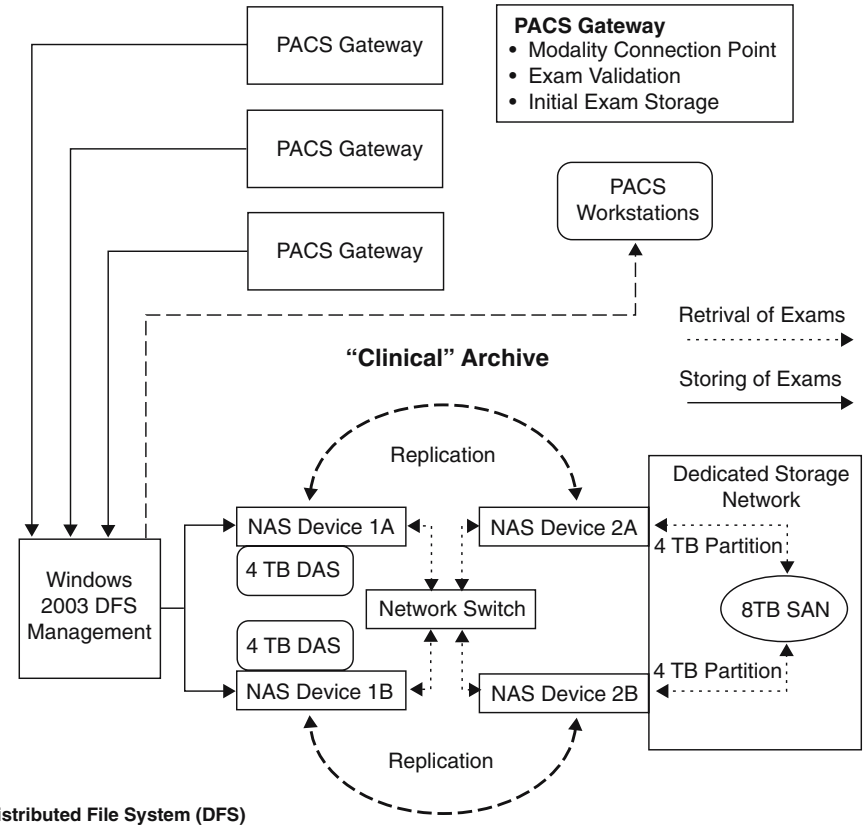
Sample PACS (SAN) implementation.

commands on a copper Category 5 Ethernet cable sitting on top of TCP/IP routing protocols. Internet SCSI can use the same inexpensive cabling as Ethernet networking and even sit on an Ethernet network. Internet SCSI is giving rise to what is being referred to as IP SANs, and it is predicted that this strategy will drastically reduce the cost of SANs installation and support with the introduction of plug-and-play (multivendor) components.

NETWORK ATTACHED STORAGE

The third technique for adding storage is by working at the file level and using NAS (Figure 16.5). NAS servers, also known as filers, are dedicated to providing remote file-level access. PACS applications can access storage sitting on NAS servers using a high-level file-level access protocol called the common Internet file system (CIFS; Microsoft Windows–type file sharing). The benefit of using CIFS is that it is a widely used standard in the industry and is platform independent. Linux and Unix servers work extremely well as CIFS file servers using the powerful and open-source Samba server. CIFS uses the universal naming convention (UNC) for addressing files, such as: \\servername\share\path\filename. The first portion of the UNC is the server name or network address at which the NAS server resides.

NAS is a relatively easy way to scale a storage system when the PACS application supports UNC. The database keeps a UNC pointer to the locations at which the files reside. One potential drawback of the NAS architecture is that data are sent on the same network twice, with one hop sending data from the NAS to the server and the second hop sending data from the server to the end-client workstation. This contrasts with a SAN, which typically uses a dedicated Fibre Channel network to afford fast, direct access to the storage devices. Some PACS applications avoid this potential problem by passing the UNC address to the end workstation to directly access the files from the NAS server. This architecture scales well, because it avoids having all system file access go through a central point that becomes a potential bottleneck. As an example, suppose there are 5 NAS devices, each containing 1 TB of storage on a 1 GB network, and the PACS software is architected to spread the storage load among the 5 NAS devices. The aggregate network bandwidth would be 5 × 1 GB or 5 GB, and the aggregate storage would be 5 TB. Both the amount of storage and the pipeline to the storage increase proportionally as (vendor-agnostic “inexpensive”) storage devices are added.



Distributed File System (DFS)

DFS redirects requests for files to assigned NAS devices. The primary advantage of redirecting client requests is to take advantage of replication. If two or more NAS devices are replicated and the primary NAS device fails, DFS will automatically direct all future client requests to one of the still functioning backup NAS devices.

Process Overview

In this example there are 4 NAS access points. Because the NAS devices present the underlining storage as a file system, the technology utilized by physical storage devices is irrelevant. To illustrate this point, this example utilizes 2 distinctly separate technologies (DAS and SAN). A simplistic way to look at this is that the operating system presents all disk storage as a file system. Storage vendors write plug-in software that interfaces the storage device to the operating system. This software is usually referred to as "device drivers." Therefore, if the operating system can mount (use) the storage device, it can then, in turn, offer the file system to other users on the network (in Windows-speak, file share).

FIGURE 16.5

Sample PACS (NAS) implementation.

Another exciting NAS-related technology is contained in new features offered by the distributed file system (DFS) that is standard with the Microsoft Windows 2003 operating system. A distributed file system can make several disparate NAS servers appear logically grouped and present them to the PACS application as a single source. A distributed file system provides replication between NAS servers that can act as a fault-tolerant backup to switch over to in the event 1 of the NAS servers fails. These features were once available only from high-end storage vendors.

Challenges involved in supporting a large NAS environment include the installation/support of replication/backup software, monitoring tools, antivirus measures (including daily signature updates), and, of course, those pesky service patches (because a NAS server is running an operating system with a file system, both will need to be maintained on all the NAS servers to avoid security vulnerabilities and file corruption).

GRID STORAGE

Grid storage creates fault-tolerant storage pools based on commodity "inexpensive" servers, often referred to as grid servers (Figure 16.6). Each grid server usually contains 4 internal serial ATA disk drives, 1 processor, and a modest amount of memory. Each of the servers participating in the grid runs vendor proprietary software that allows the devices to act as one. At least 1 server acts as the interface between the grid and external application (PACS) servers. This grid's interface server appears to the PACS servers as a NAS device presenting a standard file system. The interface server also hosts hierarchical storage manager (HSM)-like software that manages the location of files based on user configurable polices. These polices allow for defining $1 - n$ copies of a file, with each copy stored on a different grid server. This is how redundancy/fault tolerance is achieved. If 1 or more of the grid servers fails, the data can still be retrieved to the interface node (standard file system) from a copy located on a functioning grid server.

CONTENT ADDRESSABLE STORAGE

Content addressable storage (CAS) is functionally the same as grid storage. However, instead of presenting files to the application (PACS) servers via NAS, a proprietary interface is utilized (Figure 16.6). Moreover, the integrity of the files is guaranteed by elaborate algorithms that provide a unique value (token) based on the data in the file. The algorithm commonly employed is

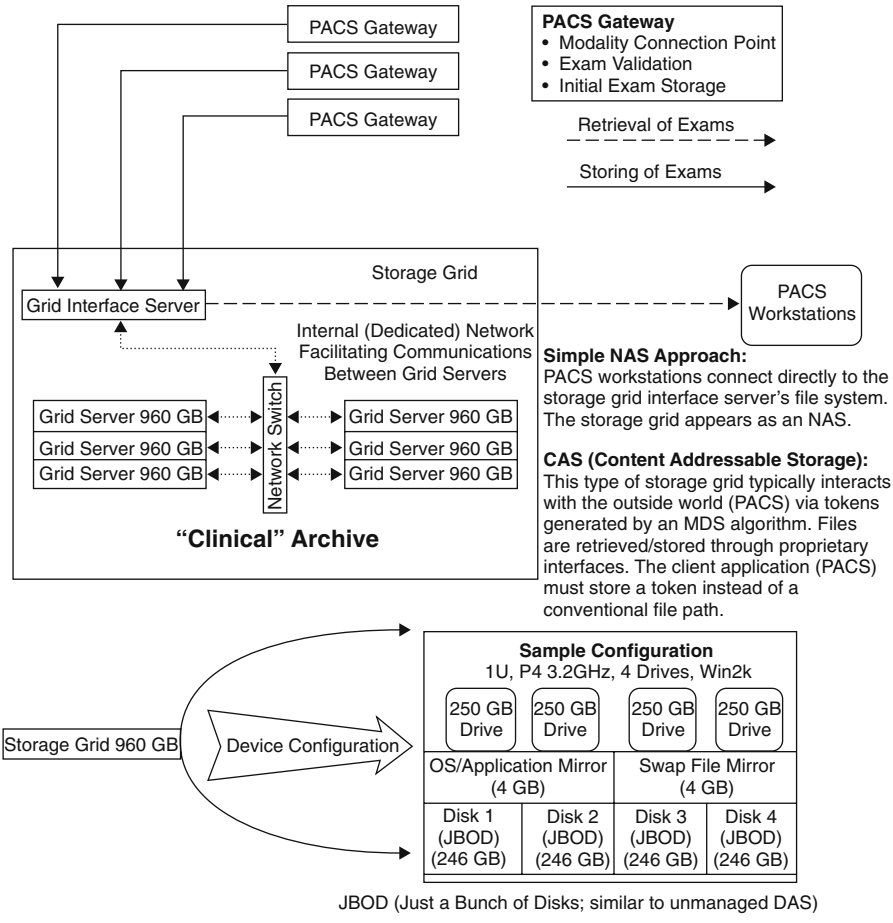


FIGURE 16.6

Grid storage.

MD5 (message-digest algorithm). The generated value acts as a unique key for retrieving the file and must be stored by the (PACS) application. If the data on the storage grid somehow changed, the file's unique value would no longer match the token stored in the (PACS) application database, resulting in an error being sent from the storage grid back to the requesting (PACS) application. This approach also supports the ability to logically group files together as related objects and store metadata (data about the data; for example, to store date, last access date, or information relating to the sources of the data.)

FOOD AND DRUG ADMINISTRATION AND STORAGE

A common myth is that storage is a U.S. Food and Drug Administration (FDA)-approved device and hence can be purchased only through a PACS vendor. The FDA has clarified this issue by stating that devices that need clearance are those that alter image data, such as image-processing algorithms. The bottom line is that storage does not have to be purchased from PACS vendors, especially when markup costs are large. This freedom to purchase is particularly useful in a NAS architecture where the PACS application is encapsulated from the storage by the CIFS standard. This should shorten the validation cycle time for testing new storage solutions and allow customers more flexibility in enjoying continued innovation from the storage industry.

NEAR-LINE TECHNOLOGY

OPTICAL STORAGE

Optical storage can also be used as an archive device for storing data. The compact disc (CD) is an aluminum disc encased in a transparent plastic layer, with binary information encoded as a series of bumps that start at the center of the disc and spiral out to the ends. A laser is focused onto the spiral path as the CD spins. In the absence of a bump, the laser is reflected back onto a photodetector that records a hit. Bumps on the disc cause the laser to deflect, and the absence of a hit on the photo detector is recorded as a miss (or 0). These series of hits and misses become 1s and 0s that form a massive data stream. The spiral path is a mere 50 μm wide and is separated from adjacent paths by just 1.6 μm. The linear length of a standard 700-MB, 12-cm CD path is more than 3.5 mi long. Commercial CD burning technology can make CDs by using an aluminum disc that is covered by a layer of photosensitive material. A write laser can stimulate the photosensitive material, which becomes opacified. Thus, when a read laser hits that area, there is no reflection from the underlying aluminum plate and a miss is recorded.

Digital versatile discs (DVDs) use the same underlying technology as CDs. The difference is that a DVD uses a higher-frequency laser that provides a smaller area on which to focus. This means that the bumps can be

smaller and packed more closely together. A DVD has a spiral path that is a mere 320 nm wide, separated from adjacent tracks by 740 nm. A DVD can store 4.7 GB of data, 7 times the capacity of CDs. DVDs can be written on both sides, providing twice the capacity (9.4 GB). A newer technology allows a DVD to store 2 layers on the same side by having a semitransparent layer made from gold overlaid on top of the base aluminum layer. A laser can focus either on the gold layer or through it on the aluminum layer. This technique requires slightly longer bump sizes to ensure accuracy, so the total capacity is slightly less than twice that of a regular DVD. A double-sided, double-layer DVD can hold 15.9 GB of data.

An interesting hybrid between magnetic and optical technology is magneto-optical (MO) drives. An MO drive writes to the media using a magnetic field that thermally changes the properties of the material. When polarized light is shone on the material, the material shifts the frequency of the reflected light. This gives MO faster write speeds than typical optical drives, up to 4 MB/s. MO disks are encased in a hard envelope and can be selected as rewriteable or "write once read many" (WORM). The WORM option provides an exceptionally long lifetime for the media, projected at more than 60 years. Magneto-optical media come in 5.2- and 9.1-GB capacities but cost several times as much per gigabyte as DVD media.

If optical media are selected for archive technology, it is likely that a jukebox will be part of the solution. A jukebox system uses a robotic arm to load and unload optical media from readers and stores the media in slots. Optical media are sensitive to extremes in temperature, humidity, and direct exposure to sunlight. A jukebox should be used, especially for exposed media such as CDs and DVDs that do not have enclosed hard cases and so are susceptible to scratches from manual handling. Today, large optical jukebox systems can hold up to 2000 slots and as much as 20 TB of storage using double-sided DVD or MO technology.

TAPE

Tape is based on the same magnetic technology used for hard drives. The difference is that tape encases the magnetic material in a flexible plastic strip that is 8 to 12 mm wide and holds multiple tracks of data. Tape is a sequential access device, and hard drives and optical media are random access devices. High-capacity cartridges hold more than 2000 linear feet of tape. A tape must wind to (seek) the location at which data are stored. This can take on the order of 30 to 60 seconds for larger-capacity tapes,

TABLE 16.3
Tape Capacity and Performance Matrix

Tape Type	Uncompressed Capacity (GB)	Transfer Rate (MB/s)	Transfer Rate (GB/h)	Average File Access Time (Seconds)
DLT	40	3	10.8	68
SDLT 320	160	16	57.6	70
SDLT 600	300	36	129.6	79
AIT-1	35	4	14.4	27
AIT-2	50	6	21.6	27
AIT-3	100	12	43.2	27
SAIT	500	30	108.0	70
LTO-1	100	15	54.0	52
LTO-2	200	35	126.0	52
SLR50	25	2	7.2	55
SLR100	50	5	18.0	58

whereas hard drives and optical drives can do this in well under a second (Table 16.3). The engineering trick is to wind the tape quickly with a minimal amount of tension that might cause wear and eventual failure. In a jukebox, an additional 10 to 30 seconds is usually needed to pull a tape from a slot and mount it into a reader. Tape jukebox systems have a MTBF (mean time between failures) of 200,000 hours but, in many cases, remain operational during component failures. Tape heads are the most common component to fail in a MTBF with approximately 40,000 hours of operation. In the event of a tape head failure, another drive unit can simply read the tape media.

Many different formats of tape are offered by vendors and/or consortiums of vendors. Formats are usually backwards-compatible over 2 or 3 previous generations. Be aware that several tape formats use the compressed capacity as the name of the format. For example, SDLT 320 holds only 160 GB of uncompressed data and 320 GB of data when compressed losslessly by a factor of 2.

ROLE OF THE HIERARCHICAL STORAGE MANAGER BETWEEN ONLINE AND NEAR-LINE STORAGE

The hierarchical storage manager software sits on servers, manages storage located in various storage devices, and provides a single view to the end user. The HSM has traditionally worked in PACS with a small portion of RAID and a large tape/optical jukebox. The HSM manages the process of synchronizing files on the RAID to 1 or more cartridges of removable media in the jukebox. As the RAID fills up past a point, known as a high-level watermark, the HSM will then flush files (a process where all but metadata about the file are deleted—enough information is retained on the RAID to know the location on a given cartridge where the file can be retrieved) from the RAID onto the near-line storage device.

Requests from the PACS for exams are intercepted by the HSM software and, if the exam has been flushed off the RAID, the HSM will restore the exam back to the RAID and then satisfy the request by allowing the requesting client to access the file. To help eliminate latency for prior exam requests, PACS pre-fetching algorithms use the scheduled procedure list for the next day to select and restore those patients' prior studies back to the RAID well before the newly scheduled procedure. For PACS applications, there is an 80% probability that a relevant prior is less than 6 months old. Physicians unlucky enough to request a patient's study that is not on the RAID typically have to wait between 40 seconds and several minutes to gain access to those patients' files.

There are 2 challenges to this multitier architecture. The first lies in understanding the trade-off that underlies the assumption behind HSMs: hard drives are fast but expensive, and the jukebox is slow but less expensive. This assumption is no longer compelling, as online storage devices continue to drop in price at a faster rate than enterprise jukebox technology. The second challenge is that predicting clinical need is not a perfect science. Waiting several minutes for a study to load from a jukebox when a clinician's expectation is that it should take only a few seconds might create a number problem reports to the PACS administrators.

As PACS vendors move toward storing all exams on spinning-disk solutions, the HSM is being reborn under a new name with a slightly different function. The new buzz names for HSM are life cycle storage management (LSM) or storage virtualization. This new approach to HSM helps manage the life cycle of the storage architecture and protect and maintain continuous availability of data. When new storage is added to the system, the storage

manager(SM) can have that additional storage appear seamlessly to the PACS application as 1 pool. An SM can also copy data from 1 disk array onto a new storage device in the background. This way, storage devices can be retired seamlessly, rebuilding and renaming new devices to take over their predecessors' roles. This can be very useful in ensuring that storage devices remain serviceable by the storage provider and that the department does not get caught with antiquated, unsupportable hardware. The HSM storage manager can be utilized to duplicate copies of exams stored in separate geographic locations to provide disaster recovery as well as better local performance. The logic of the SM can be placed at any level of storage architecture. It can reside as part of an SAN (integrated into the storage device level directly) or it could reside in the operating system at the file level. The SM could also reside as part of the PACS application. Since the functions of storage virtualizaton extend well beyond the PACS arena, this logic should ideally be processed at the file or block level and not by the application. This gives the PACS vendor one less piece of the puzzle to figure out and also allows customers to enjoy the economy-of-scale benefits of these applications that are used well beyond the medical community. This also gives more freedom to customers to select the best storage approach for their needs.

CONCLUSION

Storage is a fundamental enabling technology of the PACS industry. The only way to position ourselves to survive the onslaught of storage consumption is to be able to embrace the advances being made in the storage industry. Look for the continued trend toward commoditization through the use of well-accepted and proven industry standards such as copper Category 5 Ethernet network cable for storage networks in IP-based SANS. Expect PACS vendors to provide better encapsulation of their application from the storage architecture to allow customers to better enjoy the technological and economic benefits of this competitive market.

ABBREVIATIONS

AIT Advanced intelligent tape

ATA AT advanced

BIT Binary representation (0 or 1)

BYTE 8 bits

CAS Content addressable storage

CD Compact disk

CIFS Common Internet file system

DAS Direct attached storage

DAT Digital audio tape

DFS Distributed file system

DLT Digital linear tape

DVD Digital video disc

FC Fibre Channel

FCIP Fibre Channel over IP

Gigabyte 1000 megabytes

HBA Host bus adapter

HSM Hierarchical storage manager

IDE Integrated drive electronics

ISCSI Internet SCSI

JBOD Just a bunch of disks

LAN Local area network

LTO Linear tape open

MTTF Mean time to failure

NAS Network attached storage

RAID Redundant array of inexpensive disks

SAIT Super advanced intelligent tape

SAN Storage area network

SAS Serially attached SCSI

SATA Serial ATA

SCSI Small computer system interface

SMB Server message block

SMI-S System management interface

SNMP Simple network monitoring protocol

WAN Wide area network

REFERENCES

Christensen C. *The Innovator's Dilemma.* New York,: HarperCollins; 2003:3–68.

Coufal H, Grochowski E. The future of data storage, principles, potential, and problems. FAST Conference on File and Storage Technologies. January 2002. Available at: www.usenix.org/publications/library/proceedings/fast02/coufal.pdf.

Farley M. *Building Storage Networks.* 2nd ed. New York: McGraw-Hill; 2001: 160–207.

Gardner S, Hughes K. FDA ruling opens door to new PACS storage options. Auntminnie.com. July 2003. Available at: www.auntminnie.com/default.asp?Sec=sup&Sub=pac&Pag=dis&ItemId=58787&d=1. Accessed September 15, 2004.

Hennessy J, Patterson D. *Computer Architecture: A Quantitative Approach.* 3rd ed. New York: Morgan Kaufman Publishers; 2003:676–785.

Hitachi Global Storage 2004. Available at: www.hitachigst.com/hdd/support/7k400/7k400.htm. Accessed September 15, 2004.

Lyman P, Varian H. How Much Information? 2003. Available at: www.sims.berkeley.edu/research/projects/how-much-info-2003/index.htm. Accessed August 4, 2004.

Murray W, Maekawa K. Reliability evaluation of 3M magneto-optic media. Proceedings of Magneto-Optical Recording International Symposium. *J Magn Soc Jpn.* 1996;20(suppl S1):309–314.

Nagy P, Farmer J. Demystifying data storage: archiving options for PACS. *Appl Radiol.* May 2004:18–22.

Nice K. How DVDs work. Available at: http://electronics.howstuffworks.com/dvd.htm. Accessed September 15, 2004.

Webster J. Everything you need to know about SMI-S. *Infostor.* 2004;8:28–32.

SUGGESTED READING

Anderson M. The toll of downtime. *Healthcare Informatics.* April 2002. Available at: www.healthcare-informatics.com/issues/2002/04_02/leading.htm. Accessed September 15, 2004.

Andriole KP, Avrin DE, Yin L, Gould RG, Luth DM, Arenson RL. Relevant priors prefetching algorithm performance for a picture archiving and communication system. *J Digit Imaging.* 2000;13(2 suppl 1):73–75.

Avrin DE, Andriole KP, Yin L, Gould R, Arenson RL. Simulation of disaster recovery of a picture archiving and communications system using off-site hierarchal storage management. *J Digit Imaging.* 2000;13(2 suppl 1):168–170.

Avrin DE, Andriole KP, Yin L, Gould RG, Arenson RL. A hierarchical storage management (HSM) scheme for cost-effective on-line archival using lossy compression. *J Digit Imaging.* 2001;14:18–23.

Battaglia G, Maroldi R, Chiesa A. Evaluating the digital storage requirements for a partial picture archive and communication system. *J Digit Imaging.* 1990;3:54–59.

Bauman RA, Gell G, Dwyer SJ III. Large picture archiving and communication systems of the world. Part 2. *J Digit Imaging.* 1996;9:172–177.

Behlen F, Maughan J. Migrating data to the new PACS. *Health Imaging & IT.* May 2004. Available at: www.healthimaging.com/archives/HIIT/HIIT_2004/HIIT0205/HIIT020510.htm. Accessed September 15, 2004.

Benner A. *Fibre Channel for SANS.* New York: McGraw-Hill; 2001:5–12.

Blado ME. Management of the picture archiving and communications system archive at Texas Children's Hospital. *J Digit Imaging.* 2001;14(2 suppl 1):84–88.

Blado ME, Tomlinson A. Monitoring the accuracy of a PACS image database. *J Digit Imaging.* 2002;15(suppl 1):87–95.

Bui AA, McNitt-Gray MF, Goldin JG, Cardenas AF, Aberle DR. Problem-oriented prefetching for an integrated clinical imaging workstation. *J Am Med Inform Assoc.* 2001;8:242–253.

Chesson E. Long-term image archiving. *Health Imaging & IT.* June 2004. Available at: www.healthimaging.com/archives/HIIT/HIIT_2004/HIIT0206/HIIT020607.htm. Accessed September 15, 2004.

Channin D, Parisot C, Wanchoo V, Leontiev A, Siegel E. Integrating the Healthcare Enterprise: a primer: part 3. What does IHE do for ME? *Radiographics.* 2001;21:1351–1358.

D'Arcy S, D'Arcy J. Securing data from disaster. *Decis Imaging Economics.* June 2003. Available at: www.imagingeconomics.com/library/200306-02.asp. Accessed September 15, 2004.

Dumery B. Digital image archiving: challenges and choices. *Radiol Manage.* June 2002. Available at: www.ahra.com/AHRAArticles/AHRAArticles.dll/Show?ID=328. Accessed September 15, 2004.

Erickson BJ, Persons KR, Hangiandreou NJ, James EM, Hanna CJ, Gehring DG. Requirements for an enterprise digital image archive. *J Digit Imaging.* 2001;14:72–82.

Henderson M, Behlen F, Parisot C, Siegel E, Channin D. Integrating the Healthcare Enterprise: a primer: part 4. The role of existing standards in IHE. *Radiographics.* 2001;21:1597–1603.

Honeyman JC, Huda W, Frost MM, Palmer CK, Staab EV. Picture archiving and communication system bandwidth and storage requirements. *J Digit Imaging.* 1996;9:60–66.

McEnery KW, Suitor CT, Thompson SK, Shepard JS, Murphy WA. Enterprise utilization of "always on-line" diagnostic study archive. *J Digit Imaging.* 2002;15(suppl 1):81–86.

Mendenhall R, Dewey M, Soutar I. Designing fault-tolerant distributed archives for picture archiving and communication systems. *J Digit Imaging.* 2001;14(2 suppl 1):80–83.

Nagy P. The ABC's of storage technology. *SCAR University: Educating Healthcare Professionals for Tomorrow's Technology.* Great Falls, VA: Society for Computer Applications in Radiology; 2004:123–127.

Nagy P, Warnock M. Predicting PACS loading and performance metrics using Monte Carlo and queuing methods. In: *Proceedings of the International Society for Photo-Optical Engineering.* San Diego, CA: SPIE; 2003:46–55.

Nagy P, Daly M, Warnock M, Ehlers K. PACSPulse: A Web-based DICOM network traffic monitor and analysis tool. *Radiographics.* 2003;23:795–801.

National Electrical Manufacturers Association. DICOM resources 2004. Available at: http://medical.nema.org. Accessed September 15, 2004.

National Institute of Standards and Technology. Prefixes for binary multiples. Available at: http://physics.nist.gov/cuu/Units/binary.html. Accessed September 15, 2004.

Oosterwijk H. *PACS Fundamentals.* Aubrey, TX: Otech, Inc; 2004:53.

Pavlicek W, Zavalkovskiy B, Eversman WG. Performance and function of a high-speed multiple star topology image management system at Mayo Clinic Scottsdale. *J Digit Imaging.* 1999;12(2 suppl 1):168–174.

Piedad F, Hawkins M. *High Availability: Design, Techniques, and Processes.* Upper Saddle River, NJ: Prentice Hall; 2001:13–29.

Qi H, Snyder WE. Content-based image retrieval in picture archiving and communications systems. *J Digit Imaging.* 1999;12(2 suppl 1):81–83.

Shelton PD, Lyche DK, Norton GS, et al. Benchmark testing the Digital Imaging Network-Picture Archiving and Communications System proposal of the Department of Defense. *J Digit Imaging.* 1999;12:94–98.

Siegel E, Channin D. Integrating the Healthcare Enterprise: a primer: part 1. Introduction. *Radiographics.* 2001;21:1339–1341.

Smith EM, Wandtke J, Robinson A. The strategic and operational characteristics of a distributed phased archive for a multivendor incremental implementation of picture archiving and communications systems. *J Digit Imaging.* 1999;12(2 suppl 1):71–74.

Smith EM, Wright J, Fontaine MT, Robinson A. Archive selection for the MICAS, a multi-vendor incremental approach to PACS. *J Digit Imaging.* 1998;11(3 suppl 1):32–34.

Smith R. Rethinking disaster recovery. *Imaging Econ.* December 2001. Available at: www.imagingeconomics.com/library/200112-06.asp. Accessed September 15, 2004.

Tamm E, Thompson S, Venable S, McEnery K. Impact of multislice CT on PACS resources. *J Digit Imaging.* 2002;15(suppl 1):96–101.

Wong AW, Huang HK. Subsystem throughputs of a clinical picture archiving and communications system. *J Digit Imaging.* 1992;5:252–261.

Wong AW, Huang HK, Arenson RL, Lee JK. Digital archive system for radiologic images. *Radiographics.* 1994;14:1119–1126.