

# Storage Options for Document Management

Document management and imaging systems store large volumes of data, which must be maintained for long periods of time. Choosing storage is not simply a matter of reading the "System Requirements" on the side of the box. Capacity and other requirements must be considered specifically in terms of your application. Then, a combination of primary and backup storage components, properly supported by application or utility software, may be selected, which will meet your needs for many, many years.

## **Requirements**

Consider your requirements in terms of each of the following categories: capacity, performance, reliability, longevity, and compliance.

## **Capacity**

A typical document management system will require a storage capacity that is orders of magnitude greater than that of a database, accounting system, or other business application in the same organization.

The storage capacity requirement should be estimated before implementing the system. A skilled vendor or integrator can help, although the process is fairly straight forward.

Start by estimating the number of pages of scanned images and the number of electronic documents of various types that are received or generated on an ongoing basis. It is easy to forget some documents, so check file cabinets and file servers to help remember everything that will be stored. Then double your estimate-applications tend to grow.

Ok, you've estimated the number of pages, but how many Megabytes or Gigabytes will this translate too? Typical images of letter or legal sized paper are about 50 Kbytes in size, but this can vary. Forms with lines or shading can easily be two or three times this size. The sizes of your typical electronic documents (word processor files, spreadsheets, etc.) can be easily checked. Total the amount of data-scanned documents and electronic ones-which must be stored each month. Remember, there are approximately 1,000 kilobytes in a Megabyte and approximately 1,000 Megabytes in a Gigabyte.

Then consider the retention period. The period for which documents must be maintained may be determined by regulatory agencies, business partners, or internal policies.

Multiply the data generation rate by the period of time you wish to keep documents online to determine the online capacity needed.



Example: A third party administrator (TPA) of medical claims determines that they need to scan and store approximately 50,000 pages of claim information per month. They determine that their average scanned document size is 60 Kbytes. They want to keep at least three years of documents online. Their online capacity requirement is 50,000 pages/month x 60 Kbytes/page x 36 months = approximately 108 Gigabytes.

### **Offline Storage**

Consider keeping only your most commonly accessed documents online. Your less commonly accessed ones can be stored offline where they can still be accessed if needed. This is practical in many applications such as loan processing or orthopedic medicine in which documents will be accessed regularly for a period of time, but will then become relatively inactive-yet must be retained and available.

If this is the case in your application, compute your online storage needs using the maximum common "active period" for documents rather than the total retention period. Make sure your document management system tracks offline documents and lets you know precisely on which offline medium each is stored.

### **Performance**

The write speed and read speed of the storage device(s) each must meet the systems overall performance requirements. Often, there is a secondary (backup) device which records data more slowly than the primary storage device. Its ability to keep up with daily total capture rates (scanning plus importing) may be an issue.

Examples: Assume the same TPA mentioned above records up to 4,000 pages on a peak day. Furthermore, assume the system software allows the secondary device to be recorded at night to avoid impacting system performance during the day. It is reasonable to require that your backup storage device be capable of recording a whole day's worth of activity within a period of four hours. Thus, it must have a recording rate of 4,000 pages x 60,000 bytes/page divided by 4 hours. That's a very modest 60 MB per hour or 1 MB per minute. Even ten times this rate could be met by any modern backup storage device.

### **Reliability**

Reliability applies to both primary and backup storage devices.

Redundancy for your primary storage (see RAID, below) is so affordable today that it is a "no brainer". This combined with the use of quality components will make loss of data due to a storage device failure very unlikely. Beyond protecting against data loss, "hot spare" and "hot swap" options are available to eliminate or minimize down time in the event of failure.

Reliability of your backup device is also critical, because even a redundant primary device can be destroyed by flood, fire, a software virus, or an act of malice. Proper storage and handling of



backup media as well as proper technology selection can assure a high level of reliability. Periodic testing of backup media may be necessary for some technologies.

### **Longevity**

Not all storage media are equal in terms of longevity. Make sure the permanence of your backup medium safely exceeds your required retention period.

Alternately, you might schedule duplication of data to "refresh" your backups. However, doing so introduces additional work and additional media costs.

### **Compliance**

Aside from determining your retention period(s), regulations may dictate the type of media you use. In particular, broker/dealers and other entities regulated by Securities and Exchange Commission (SEC) Rule 17a-4 are required to use a non-rewritable medium if they use any digital medium to store documents. Note that using non-rewritable media to back up writable media is one way to comply with this rule.

### **Media**

The following list of storage devices is in no way exhaustive, but it does cover the devices most likely to be used in a document management system.

### **Magnetic Disks**

Once called a "Winchester disk", a fixed magnetic disk is usually referred to as a "hard disk" today. It is a common, affordable, high-performance, high-capacity storage device. They are fairly reliable, too, although they are electro-mechanical and inevitably will fail from time-to-time. Nevertheless, magnetic disks are used as the primary storage medium for essentially all document management systems, usually in one of the following configurations.

### **RAID**

A Redundant Array of Independent Disks uses two or more hard disks to store data with (usually) some level of redundancy, such that complete data restoration can take place automatically after the failure of a single disk. There are many different RAID configurations, a couple of which actually produce no data redundancy. These are the most common ones.

- RAID-0 or data striping simply spreads each file across multiple (two or more) disks to create a single virtual device with the combined capacity of all the disks. Performance is improved because reading and writing takes place in parallel on all disks. However, there is no redundancy or fault tolerance in a RAID-0 system.
- RAID-1 or mirroring writes data to exactly two disks in parallel, providing complete redundancy. Read performance may be better than a single disk because reading may take place in parallel from both, however there is no performance improvement on writes. The capacity of the RAID-1 system is the same as that of each disk.
- RAID-5 combines data striping with error correction (redundancy) using three or more disks. The level of redundancy is such that all data are preserved if a single disk fails.

Note that the chance of failure is greater than with a RAID-1, but the amount of total capacity lost to redundancy is lower. Also, reading and writing to the disks in parallel results in excellent performance. Five disks are common, providing a virtual device of four times the capacity of each disk; the remaining capacity is used for error correction info.

Many other schemes are possible, most of them combining aspects of these. See the references at the end of this document for more information.

### **NAS**

A Network Attached Storage device is generally a RAID system that is attached directly to a local area network. Client PC's access the NAS through the network without having to place any demands on a server, which might have other tasks to perform such as database operations. Server and storage performance are both improved. NAS systems have become very economical and easy to administer.

### **SAN**

A Storage Area Network (SAN) is a sub-network of shared storage devices. The sub-network communicates through a fiber-channel connection for very high speed. It includes backup devices and specialized servers dedicated to storage management.

A SAN is accessible directly through the local area network by client PC's with no involvement from database or application servers. Therefore, it shares the associated performance advantage of NAS devices.

However, a SAN stands apart from a simple NAS device in many respects. First, it can be expanded almost indefinitely by the addition of storage devices on the sub-network. Including backup devices within the high speed SAN sub-network avoids big blow to network performance that occurs when backups are performed through a LAN. Servers performing storage-intensive functions like an enterprise database can also be located on the SAN sub-network to boost their performance. These add up to big advantages for high-demand information systems, but the costs of implementing a SAN system can add up too

### **Tape**

Magnetic tapes used for computers are similar to tapes used to store music or video. Tapes are sequential-access media, which means that to get to a particular point on the tape, the tape must go through all the preceding points. In contrast, disks are random-access media because a disk drive can access any point at random without passing through intervening points.

Oxides and other contaminants in the tape drive can compromise reliability. Periodic cleaning of the drive is necessary, and periodic testing of backup tapes is recommended, to ensure that data are truly being written effectively.

Despite some drawbacks, tapes are an inexpensive, removable medium with high capacity,



making them ideally suited to backups.

Longevity of data on magnetic tape may be a concern. They are very susceptible to environmental factors (dust, humidity, and temperature swings). Reports of life expectancy-even in a controlled environment-vary from as little as 1 to 2 years to as long as 30 years. Most information technology professionals would be uncomfortable relying on accessibility of data on a tape that is more than a few years old. Obsolescence of a specific tape drive or format is generally not a concern because the media is not expected to last that long.

### **Optical**

There are many different kinds of optical disks offering different storage technologies, capacities, and sizes. All use a laser in some manner to read and write data. Speeds vary, but they generally fall between the speed of tapes and magnetic disks. Write speed is often slower than read speed.

Some optical disk technologies allow the disk to be erased and rewritten like a magnetic medium, but others-called WORM, or Write Once Read Many-inherently allow the medium to be recorded only once. WORM is favored or required by some regulatory agencies precisely because it prevents alteration.

A complete discussion of optical technologies is beyond the scope of this paper, but the references include sources of additional data.

In brief:

- Magneto-optical technology, which once dominated optical storage in document management applications, has generally given way to newer, more economical solutions. However, several vendors still offer magneto-optical drives and they do boast write speeds several times greater than other optical technologies.
- CD-R and DVD-R are each one-time recordable (WORM) technologies readily familiar to most people due to their prominence in entertainment (music and movie) applications. Third party recording software may be required, depending on your software application and operating system. DVD+R is a competing standard for one-time recordable DVDs.
- Many different rewritable DVD technologies exist, and their relative merits are hotly debated by their respective proponents. DVD-RAM was proposed and developed by the DVD Forum, and is the only DVD technology designed primarily for data processing applications, as opposed to entertainment (video, music). A DVD-RAM disk is normally used within a cartridge, which ensures that the medium remains clean and reliable. It also is the only DVD technology today that performs as a standard hard drive without special recording software.
- DVD-RW is a re-recordable medium that was also created by the DVD Forum. DVD+RW is a competing re-recordable medium created by the DVD+RW Alliance. Although they arrived on



the market later than DVD-RAM, each has gained significant share. DVD+RW, in particular, may be found in many popular brand PCs and servers. One advantage both of these technologies boast is the ability to be read in many home entertainment devices, although this is not important in your document management application.

Optical media are less particular about environmental conditions than magnetic tapes, but manufacturer's specifications should still be observed. Most drives and some application software automatically read back and verify data that have been written, eliminating the need to test backups, as is recommended for tapes. Periodic cleaning is required.

Magneto-optical, DVD-RAM, and other rewritable media generally can be expected to maintain data for at least 30 years while studies indicate that DVD-R and may last for 100 or more years. No data was found on the longevity of other DVD formats.

It should be noted that for long term storage, technology obsolescence becomes more important than degradation of the media. (It doesn't matter if the disk is "good" if you can no longer find a device that will read it.) Therefore, it is wise to evaluate the availability of the optical disk technology you are using every five years or so—maybe every leap day, for instance. If you determine that the availability of the technology you have been using is becoming questionable, you should copy data from their current storage to a newer technology. Don't wait until your only disk drive fails to determine that it can't be replaced.

## **Solutions**

In order to meet the capacity, performance, reliability, and compliance requirements for a given application, a solution must generally combine two different storage technologies.

### **Optical Library (Jukebox)**

Optical libraries, or Jukeboxes, mechanically move any one of a large number of optical disks into or out of a small number of optical disk drives. They thereby offer the capacity of dozens or even hundreds of disks. Optical library systems require management software and often include a magnetic disk which serves as a cache to more quickly access recent information.

The mechanical nature of optical jukeboxes, however, makes them prone to failure. Also, the time required to switch disks may be on the order of ten or twenty seconds, causing access to data in optical libraries to be painfully slow. Falling prices and rising capacity of magnetic disks have generally rendered the optical jukebox obsolete.

### **Magnetic with Tape Backup**

The speed and capacity of magnetic disks have made them the primary means of storage in practically all document management applications. However, the potential for loss due disaster, virus, etc. makes the recording of a backup essential.

As previously mentioned, tape is well suited to backup applications. Drives and media are available in a range of capacities, and software is available for scheduling backups



automatically.

Incremental backups, which record only new or changed information, are generally performed on a daily basis. However, restoration of data from incremental backup tapes requires restoring every disk, from the first. Periodically, a full backup is recommended to ensure a fresh set of data.

Tape drives capable of saving 160 GB of data cost under \$5,000 and media are under \$100.

### **Magnetic with Magnetic Backup**

The low cost of NAS devices or other magnetic storage devices has made possible the practice of just backing up your magnetic storage to another whole magnetic storage device. Backups and restoration are generally very fast and capacity is not an issue.

However, if the backup device is located on the LAN in the same premises as the primary device, it defeats some of the purposes of having a backup: protection from disaster, virus, or malice.

Protection from disaster may be achieved by backing up magnetic-to-magnetic through a wide area network (WAN) connection. The bandwidth needed must be computed as described under Requirements, Performance, above. The example given previously which required a bandwidth of 1 MB/minute could be supported by most WAN connection types, but larger applications might require a T1 or greater. Also, because it is "on line", such a backup may still be susceptible to a computer virus or digital tampering.

Once again, NAS devices can be very affordable. Although traditionally on the order of \$5,000, some solutions with several hundred Gigabytes of storage are now under \$2,000.

### **Magnetic with Optical Backup**

Optical disks are well suited to backing up magnetic disks in document management applications. Their longevity and reliability allow them to be moved offsite without ever having to be refreshed.

Convenient use of optical disks requires some level of support in the document management software. The software should be aware of the optical disk size so that data are written in folders that can readily be backed up.

Restoration from optical is easy and reasonably fast, because only the damaged or lost data needs to be recovered, unlike restoration from an incremental backup tape.

The various types of recordable DVD are all available for under \$500 per drive-some far under that price! Media prices vary from DVD-R, which are well under \$1.00 per disk, to DVD-RAM, which are under \$20 per disk-each with 4.7 GB capacities.



## References

<http://www.webopedia.com/>><http://www.webopedia.com>

<http://www.dvdforum.org>

<http://www.dvdrw.com/>

<http://www.storagesearch.com>

<http://www.opticaldisc-systems.com/> (See Technology Papers)

<http://www.dvddemystified.com/dvdfaq.html>

[http://www.iomega.com/dvd/dvd\\_faq.html](http://www.iomega.com/dvd/dvd_faq.html)

