

**CSCA0101  
COMPUTING BASICS**

**Chapter 5  
Storage Devices**

# CSCA0101 Computing Basics

## Storage Devices

1. Computer Data Storage
2. Types of Storage
3. Storage Device Features
4. Other Examples of Storage Device

### Storage Devices

- A **storage device** is used in the computers to store the data.
- Provides one of the core functions of the modern computer.

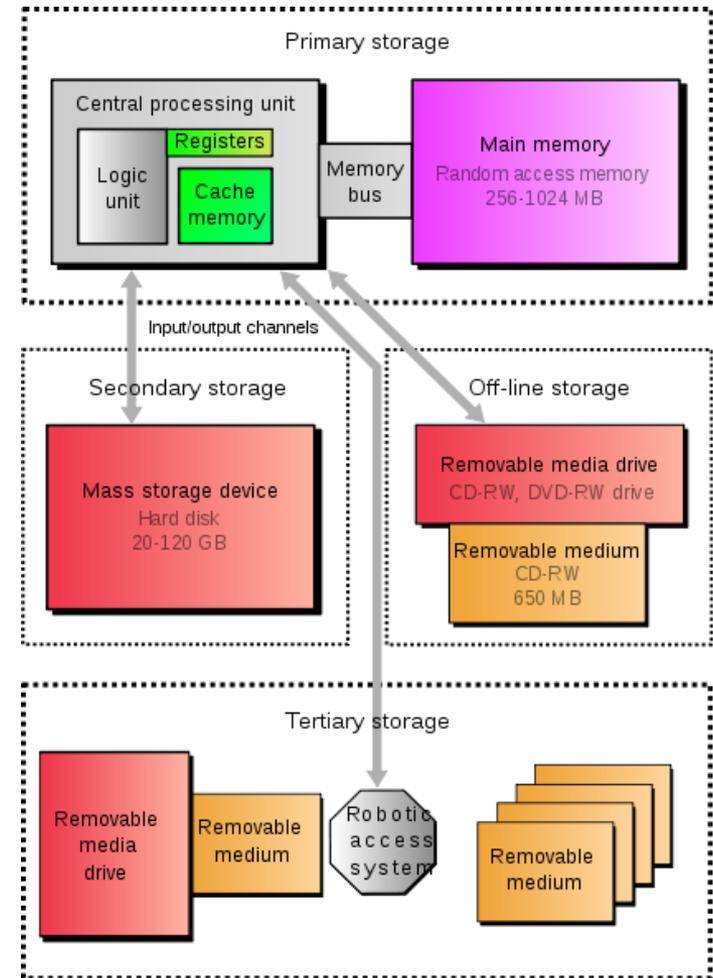
# CSCA0101 Computing Basics

## Storage Devices

### Types of Storage

There are four type of storage:

- Primary Storage
- Secondary Storage
- Tertiary Storage
- Off-line Storage



### Primary Storage

- Also known as **main memory**.
- Main memory is directly or indirectly connected to the central processing unit via a memory bus.
- The CPU continuously reads instructions stored there and executes them as required.
- Example:
  - RAM
  - ROM
  - Cache

### Primary Storage

#### RAM

- It is called Random Access Memory because any of the data in RAM can be accessed just as fast as any of the other data.
- There are two types of RAM:
  - DRAM (Dynamic Random Access Memory)
  - SRAM (Static Random Access Memory)

# CSCA0101 Computing Basics

## Storage Devices

### Primary Storage

#### RAM

Static RAM	Dynamic RAM
<ul style="list-style-type: none"><li>• Faster</li><li>• More expensive</li><li>• More power consumption</li><li>• does not need to be refreshed</li></ul>	<ul style="list-style-type: none"><li>• Slower</li><li>• Less expensive</li><li>• Less power consumption</li><li>• needs to be refreshed thousands of times per second</li></ul>
	

### Primary Storage

#### ROM

- This memory is used as the computer begins to boot up.
- Small programs called firmware are often stored in ROM chips on hardware devices (like a BIOS chip), and they contain instructions the computer can use in performing some of the most basic operations required to operate hardware devices.
- ROM memory cannot be easily or quickly overwritten or modified.



### Primary Storage

#### Cache

- **Cache** is a high-speed access area that can be either a reserved section of main memory or a storage device.
- Most computers today come with L3 cache or L2 cache, while older computers included only L1 cache.

### Secondary Storage

- It is not directly accessible by the CPU.
- Computer usually uses its input/output channels to access secondary storage and transfers the desired data using intermediate area in primary storage.
- Example:
  - Hard disk

### Secondary Storage

#### Hard Disk

- The hard disk drive is the main, and usually largest, data storage device in a computer.
- It can store anywhere from 160 gigabytes to 2 terabytes.
- Hard disk speed is the speed at which content can be read and written on a hard disk.
- A hard disk unit comes with a set rotation speed varying from 4500 to 7200 rpm.
- Disk access time is measured in milliseconds.

# CSCA0101 Computing Basics

## Storage Devices

### Secondary Storage

#### Hard Disk



Internal Hard disk



External Hard disk

# CSCA0101 Computing Basics

## Storage Devices

### Secondary Storage

#### Hard Disk

	<b>Internal Hard disk</b>	<b>External Hard disk</b>
Portability	No	Yes
Price	Less expensive	More expensive
Speed	Fast	Slow
Size	Big	Small

### Tertiary Storage

- Typically it involves a robotic mechanism which will mount (insert) and dismount removable mass storage media into a storage device.
- It is a comprehensive computer storage system that is usually very slow, so it is usually used to archive data that is not accessed frequently.
- This is primarily useful for extraordinarily large data stores, accessed without human operators.

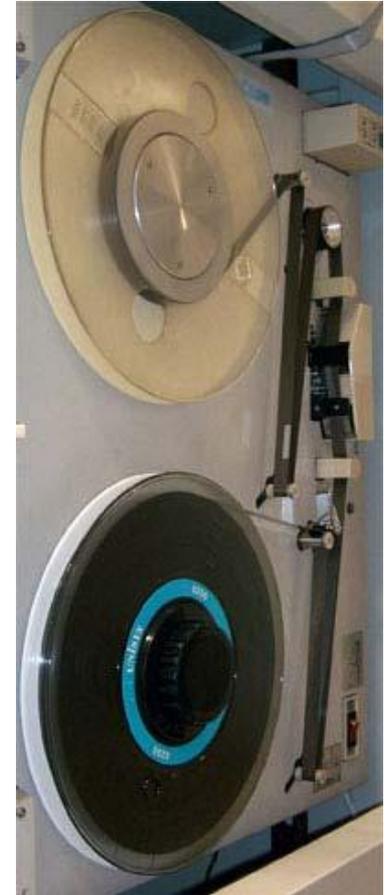
### Tertiary Storage

- Examples:
  - Magnetic Tape
  - Optical Disc

### Tertiary Storage

#### Magnetic Tape

- A magnetically coated strip of plastic on which data can be encoded.
- Tapes for computers are similar to tapes used to store music.
- Tape is much less expensive than other storage mediums but commonly a much slower solution that is commonly used for backup.



### Tertiary Storage

#### Optical Disc

- **Optical disc** is any storage media that holds content in digital format and is read using a laser assembly is considered optical media.
- The most common types of optical media are
  - Blu-ray (BD)
  - Compact Disc (CD)
  - Digital Versatile Disc (DVD)

# CSCA0101 Computing Basics

## Storage Devices

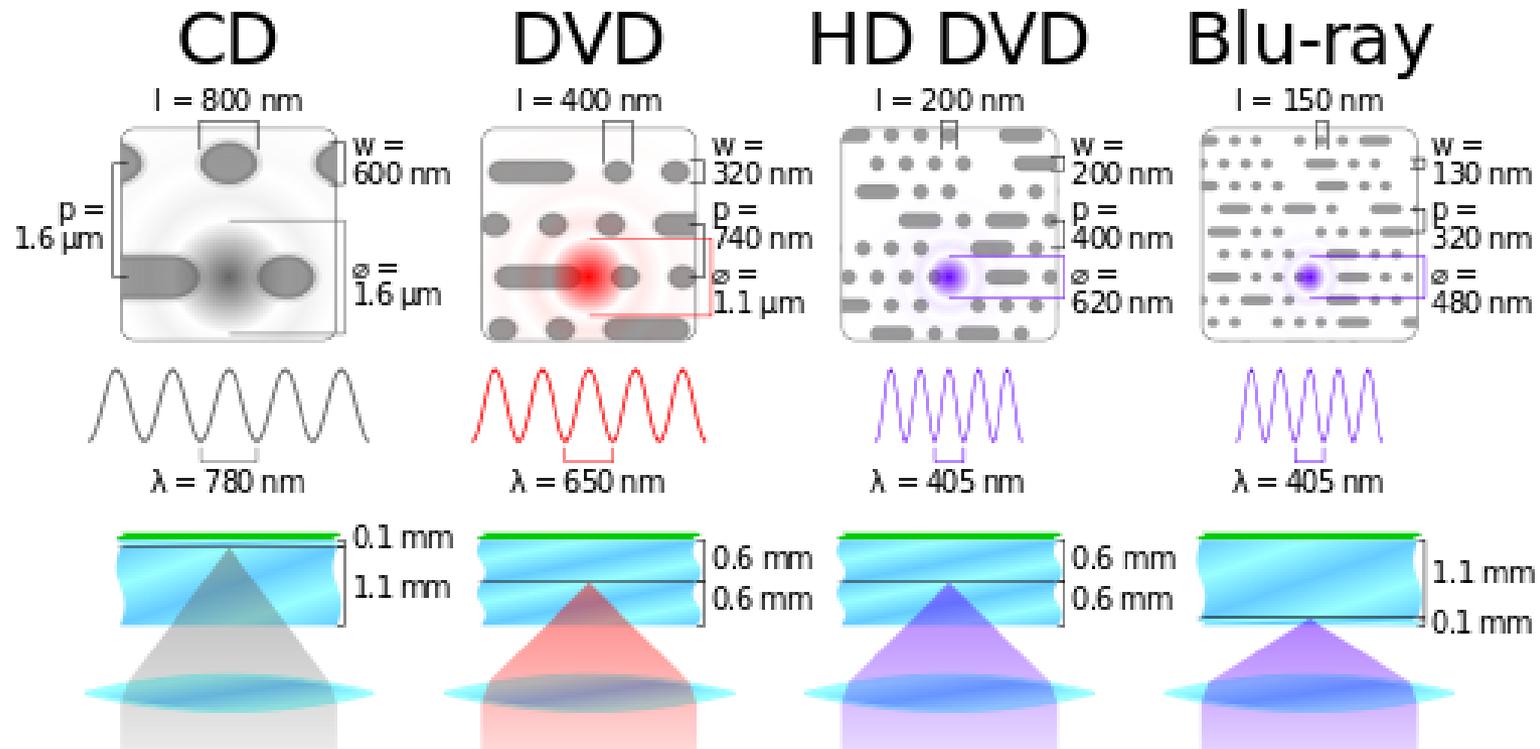
### Tertiary Storage

#### Optical Disc

	<b>CD</b>	<b>DVD</b>	<b>BD</b>
Capacity	700MB	4.7GB – 17GB	50GB
Wavelength	780nm	650nm	405nm
Read/Write Speed	1200KB/s	10.5MB/s	36MB/s
Example	<ul style="list-style-type: none"><li>• CD-ROM,</li><li>• CD-R</li><li>• CD-RW</li></ul>	<ul style="list-style-type: none"><li>• DVD-ROM</li><li>• DVD+R/RW</li><li>• DVD-R/RW</li><li>• DVD-RAM</li></ul>	<ul style="list-style-type: none"><li>• BD-R</li><li>• BD-RE</li></ul>

### Tertiary Storage

### Optical Disc



### Off-line Storage

- Also known as **disconnected storage**.
- Is a computer data storage on a medium or a device that is not under the control of a processing unit.
- It must be inserted or connected by a human operator before a computer can access it again.

### Off-line Storage

- Also known as **disconnected or removable storage**.
- Is a computer data storage on a medium or a device that is not under the control of a processing unit.
- It must be inserted or connected by a human operator before a computer can access it again.

### Off-line Storage

- Examples:
  - Floppy Disk
  - Zip diskette
  - USB Flash drive
  - Memory card

### Off-line Storage

#### Floppy Disk

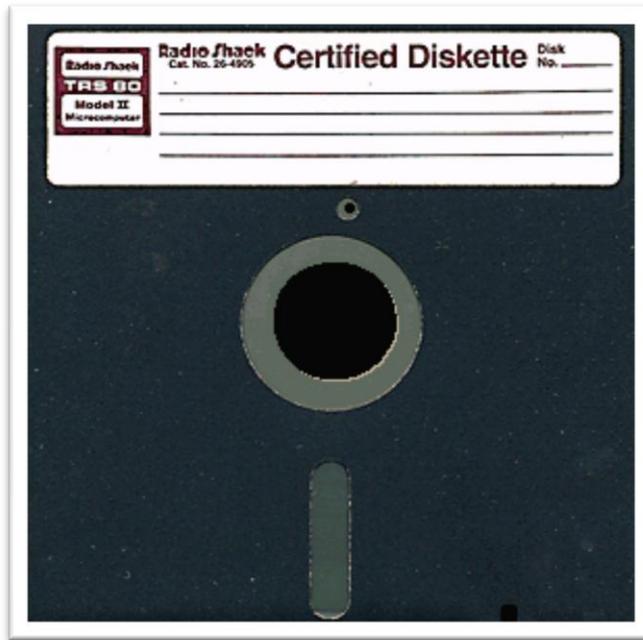
- A soft magnetic disk.
- Floppy disks are portable.
- Floppy disks are slower to access than hard disks and have less storage capacity, but they are much less expensive.
- Can store data up to 1.44MB.
- Two common sizes: 5 ¼” and 3 ½”.

# CSCA0101 Computing Basics

## Storage Devices

### Off-line Storage

#### Floppy Disk



5 1/4 inch Floppy Disk



3 1/2 inch Floppy Disk

### Off-line Storage

#### Zip Diskette

- Hardware data storage device developed by Iomega that functions like a Standard 1.44" floppy drive.
- Capable to hold up to 100 MB of data or 250 MB of data on new drives.
- Now it less popular as users needed larger storage capabilities.



# CSCA0101 Computing Basics

## Storage Devices

### Off-line Storage

#### USB Flash Drive

- A small, portable flash memory card that plugs into a computer's USB port and functions as a portable hard drive.
- Flash drives are available in sizes such as 256MB, 512MB, 1GB, 5GB, and 16GB and are an easy way to transfer and store information.



### Off-line Storage

#### Memory Card

- An electronic flash memory storage disk commonly used in consumer electronic devices such as digital cameras, MP3 players, mobile phones, and other small portable devices.
- Memory cards are usually read by connecting the device containing the card to your computer, or by using a USB card reader.

# CSCA0101 Computing Basics

## Storage Devices

### Off-line Storage

#### Memory Card



Secure Digital card (SD)



MiniSD



Compact Flash



Memory Stick



MultiMedia card



xD-Picture card



Memory card reader

### Storage Device Features

- Volatility
- Accessibility
- Mutability
- Addressability

### Volatility

- Two types of volatility:
  - Volatile Memory
  - Non-Volatile Memory

### Volatility

#### Volatile Memory

- Requires constant power to maintain the stored information.
- The fastest memory technologies.
- All contents are erased when the system's power is turned off or interrupted.
- It has been more popularly known as **temporary memory**.

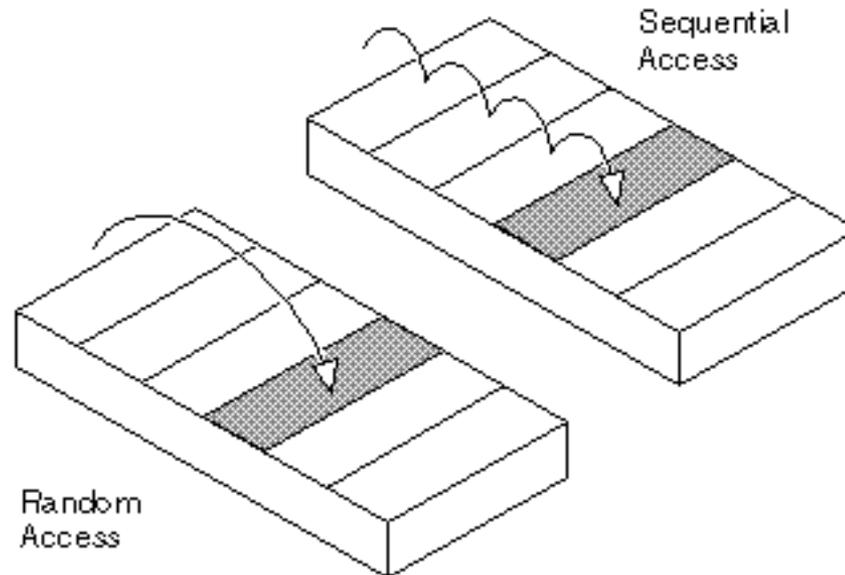
### Volatility

### Non-Volatile Memory

- Will retain the stored information even if it is not constantly supplied with electric power.
- Non volatile memory is the device which keeps the data even when the current is off.
- It is suitable for long-term storage of information.

### Accessibility

- Refers to reading or writing data records
- Two types of accessibility:
  - Random access
  - Sequential access



### Accessibility

### Random Access

- Any location in storage can be accessed at any moment in approximately the same amount of time.
- Such characteristic is well suited for primary and secondary storage.

### Accessibility

### Sequential Access

- The accessing of pieces of information will be in a serial order, one after the other; therefore the time to access a particular piece of information depends upon which piece of information was last accessed.
- Such characteristic is typical of off-line storage.

### Mutability

- Allows information to be overwritten at any time.
- A computer without some amount of read/write storage for primary storage purposes would be useless for many tasks.
- Three types of mutability:
  - Read/write storage or mutable storage
  - Read only storage
  - Slow write, fast read storage

### Mutability

#### Read/Write Storage or Mutable Storage

- Allows information to be overwritten at any time.
- A computer without some amount of read/write storage for primary storage purposes would be useless for many tasks.

### Mutability

### Read Only Storage

- Retains the information stored at the time of manufacture, and **write once storage** (WORM) allows the information to be written only once at some point after manufacture.
- These are called **immutable storage**.

### Mutability

### Slow Write, Fast Read Storage

- Read/write storage which allows information to be overwritten multiple times, but with the write operation being much slower than the read operation.

### Addressability

- Three types of addressability
  - Location-addressable
  - File addressable
  - Content-addressable

### Addressability

#### Location-addressable

- Each individually accessible unit of information in storage is selected with its numerical memory address.

### Addressability

#### File addressable

- Information is divided into files of variable length, and a particular file is selected with human-readable directory and file names.

### Addressability

#### Content-addressable

- Each individually accessible unit of information is selected based on the basis of (part of) the contents stored there.
- Content-addressable storage can be implemented using software (computer program) or hardware (computer device), with hardware being faster but more expensive option.
- Hardware content addressable memory is often used in a computer's CPU cache.

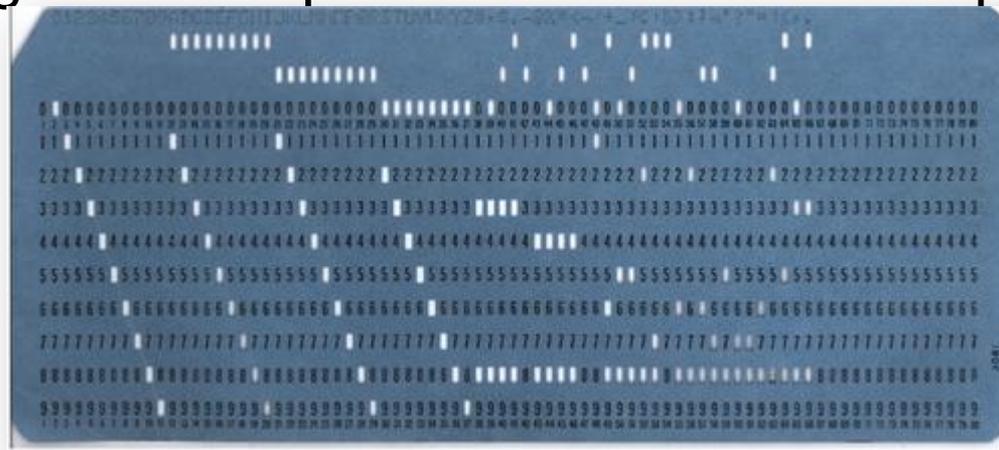
### Other Example of Storage Devices

- Punch card
- Cloud storage
- RAID

### Other Example of Storage Devices

#### Punched Card

- Early method of data storage used with early computers
- Punch cards also known as Hollerith cards
- Containing several punched holes that represents data



### Other Example of Storage Devices

#### Cloud Storage

- Cloud storage means "the storage of data online in the cloud," wherein a data is stored in and accessible from multiple distributed and connected resources that comprise a cloud.
- Cloud storage can provide the benefits of greater accessibility and reliability; rapid deployment; strong protection for data backup, archival and disaster recovery purposes.

### Other Example of Storage Devices

#### Cloud Storage

- Examples:
  - Google Drive
  - Flickr
  - Microsoft Sky Drive



Google Drive



### Other Example of Storage Devices

#### RAID

- RAID is short for **redundant array of independent** (or **inexpensive**) **disks**.
- It is a category of disk drives that employ two or more drives in combination for fault tolerance and performance.
- RAID disk drives are used frequently on servers but aren't generally necessary for personal computers.
- RAID allows you to store the same data redundantly (in multiple places) in a balanced way to improve overall storage performance.

# Types Of Storage Device

by

# Outline

- **Categorizing Storage Devices**
- **Magnetic Storage Devices**
- **Optical Storage Devices**

## **Categorizing Storage Devices**

- **Storage devices hold data, even when the computer is turned off.**
- **The physical material that actually holds data is called a storage medium. The surface of a floppy disk is a storage medium.**
- **The hardware that writes data to or reads data from a storage medium is called a storage device. A floppy disk drive is a storage device.**
- **The two primary storage technologies are magnetic and optical.**

## The primary types of magnetic storage are:

- **Diskettes (floppy disks)**
- **Hard disks**
- **High-capacity floppy disks**
- **Disk cartridges**
- **Magnetic tape**

## The primary types of optical storage are:

- **Compact Disk Read-Only Memory (CD-ROM)**
- **Digital Video Disk Read-Only Memory (DVD-ROM)**
- **CD-Recordable (CD-R)**
- **CD-Rewritable (CD-RW)**
- **PhotoCD**

# Magnetic Storage Devices

- **How Magnetic Storage Works**
- **Formatting**
- **Disk Areas**
- **Diskettes**
- **Hard Disks**
- **Disk Capacities**
- **Other Magnetic Storage Devices**

## How Magnetic Storage Works

- A magnetic disk's medium contains iron particles, which can be polarized—given a magnetic charge—in one of two directions.
- Each particle's direction represents a 1 (on) or 0 (off), representing each bit of data that the CPU can recognize.
- A disk drive uses read/write heads containing electromagnets to create magnetic charges on the medium.

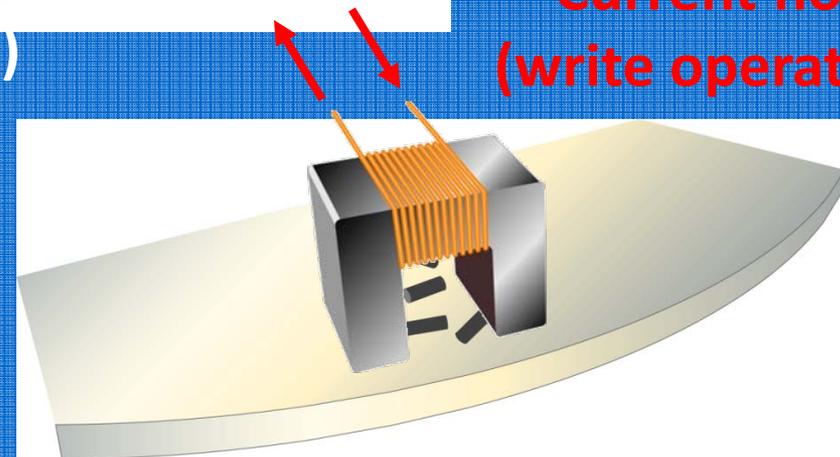
**Write head**

**Medium**

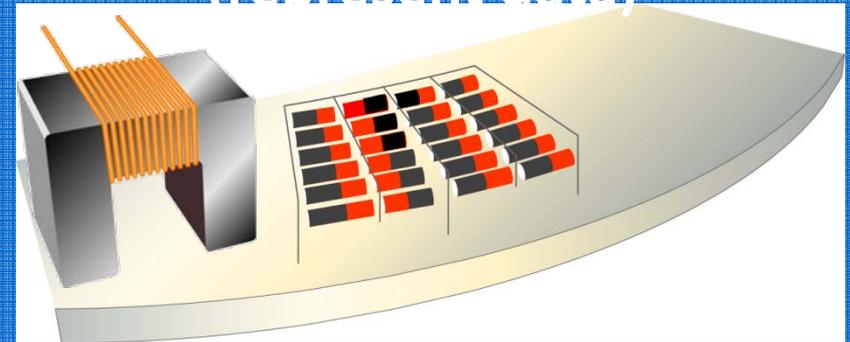
Random particles  
(no data stored)

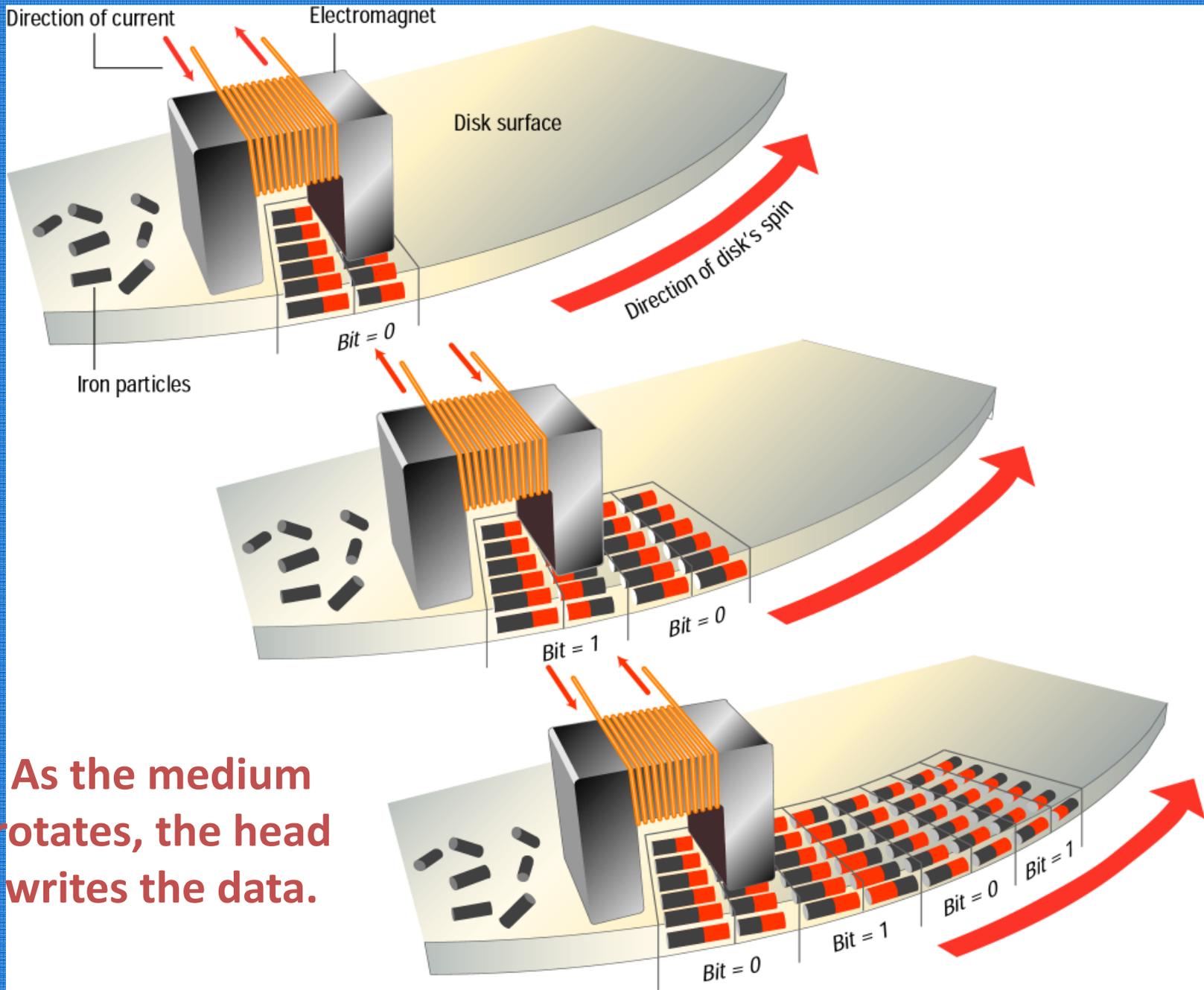


**Current flow  
(write operation)**



**Organized particles  
(represent data)**



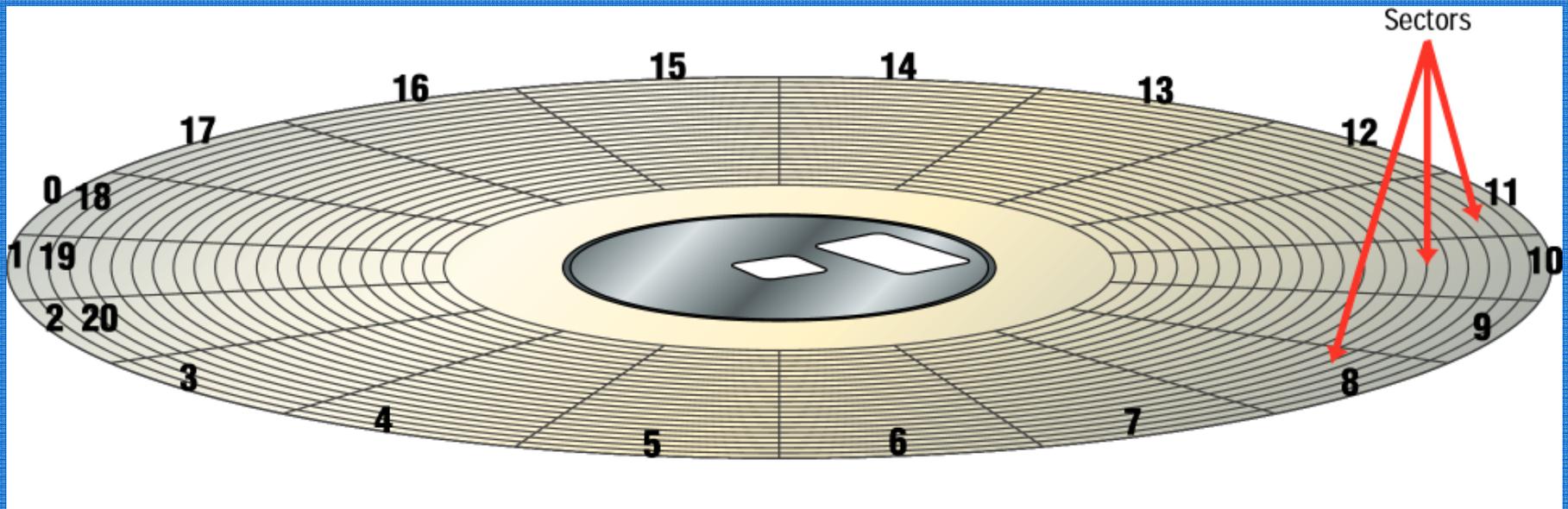


As the medium rotates, the head writes the data.

## Formatting

- **Before a magnetic disk can be used, it must be formatted—a process that maps the disk's surface and determines how data will be stored.**
- **During formatting, the drive creates circular tracks around the disk's surface, then divides each track into sectors.**
- **The OS organizes sectors into groups, called clusters, then tracks each file's location according to the clusters it occupies.**

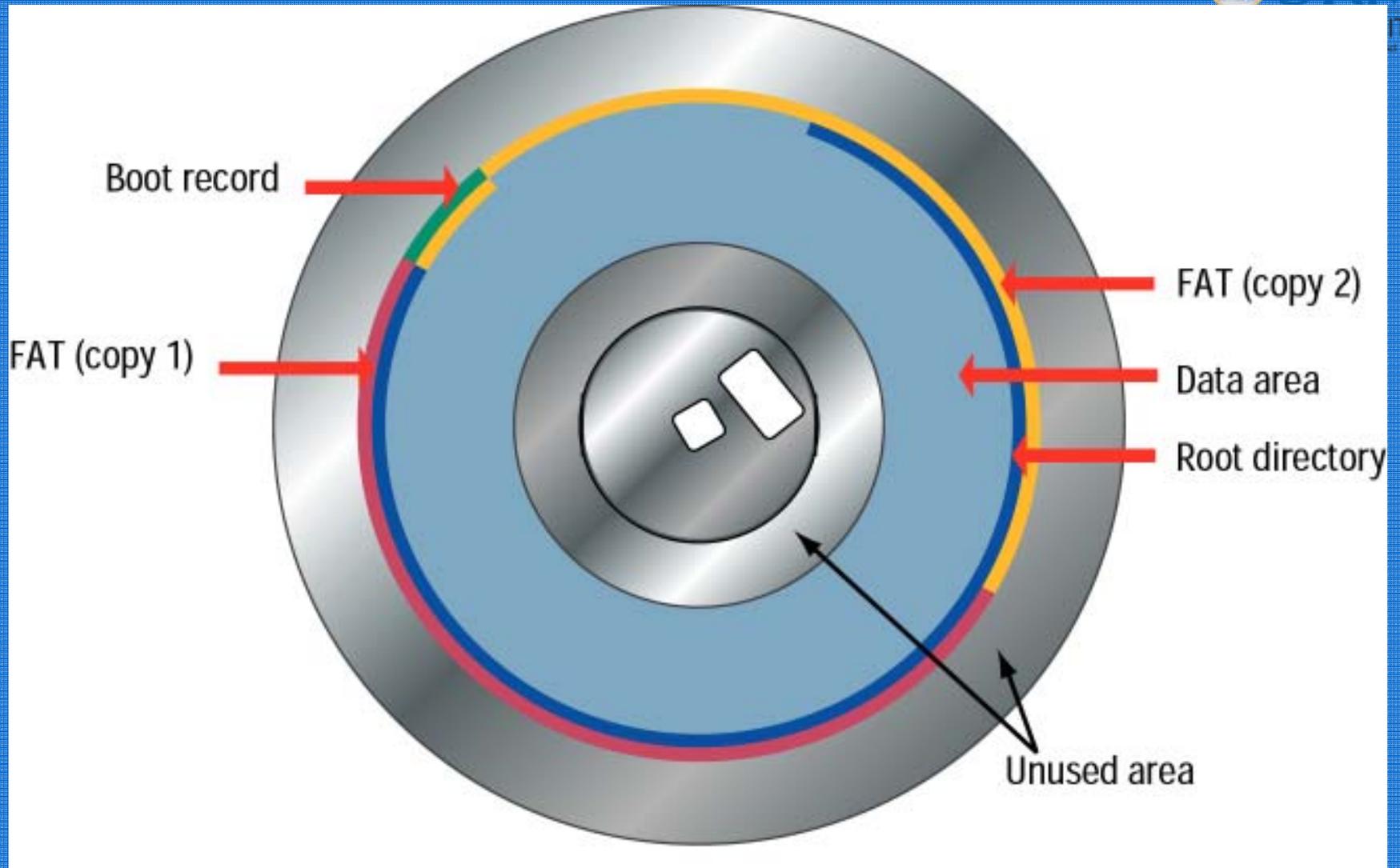
# Formatted Disk



# Disk Areas

**When a disk is formatted, the OS creates four areas on its surface:**

- **Boot sector** – stores the master boot record, a small program that runs when you first start (boot) the computer
- **File allocation table (FAT)** – a log that records each file's location and each sector's status
- **Root folder** – enables the user to store data on the disk in a logical way
- **Data area** – the portion of the disk that actually holds data

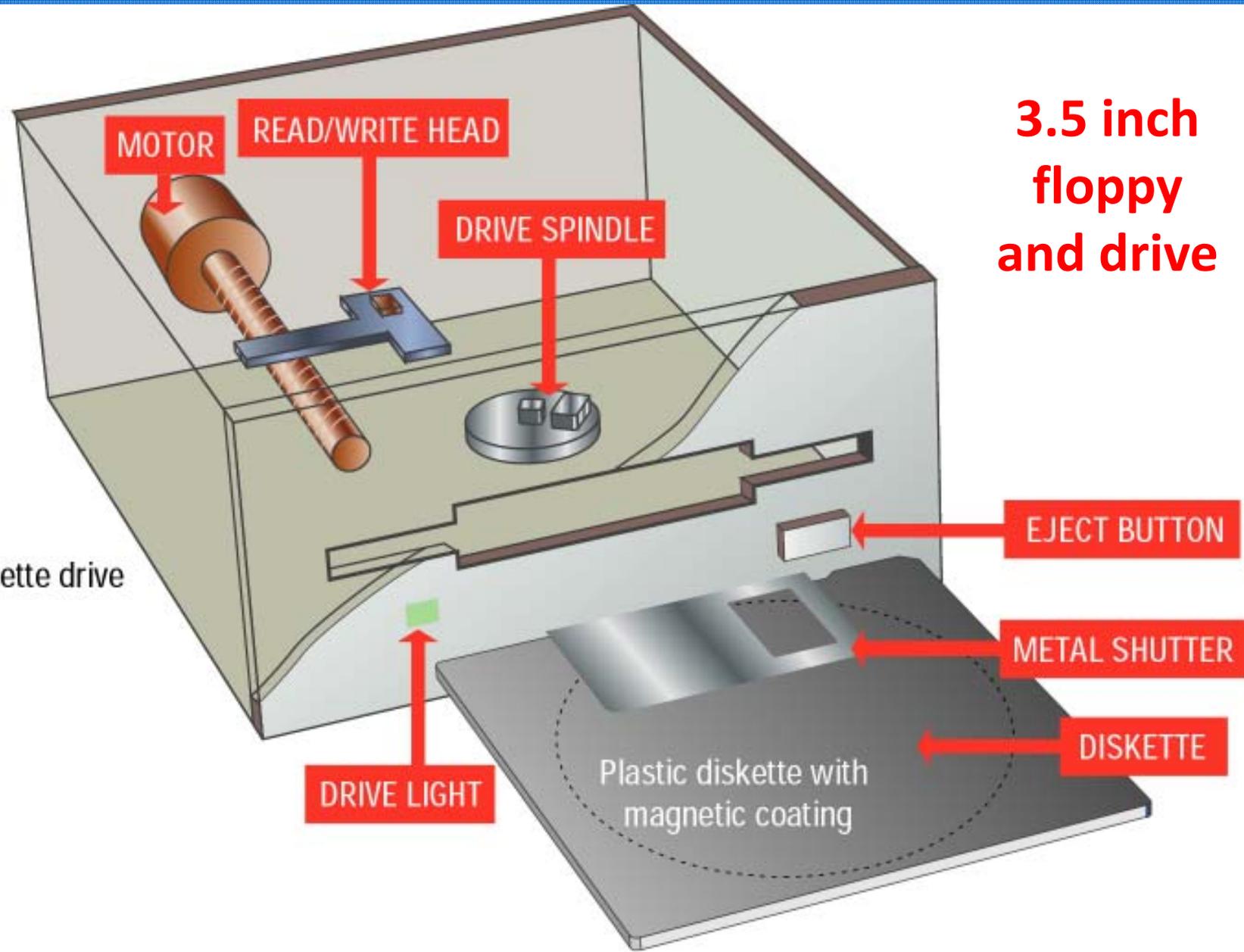


## **Magnetic Storage Devices - Diskettes**

- **Diskette drives, also known as floppy disk drives, read and write to diskettes (called floppy disks or floppies).**
- **Diskettes are used to transfer files between computers, as a means for distributing software, and as a backup medium.**
- **Diskettes come in two sizes: 5.25-inch and 3.5-inch.**

# 3.5 inch floppy and drive

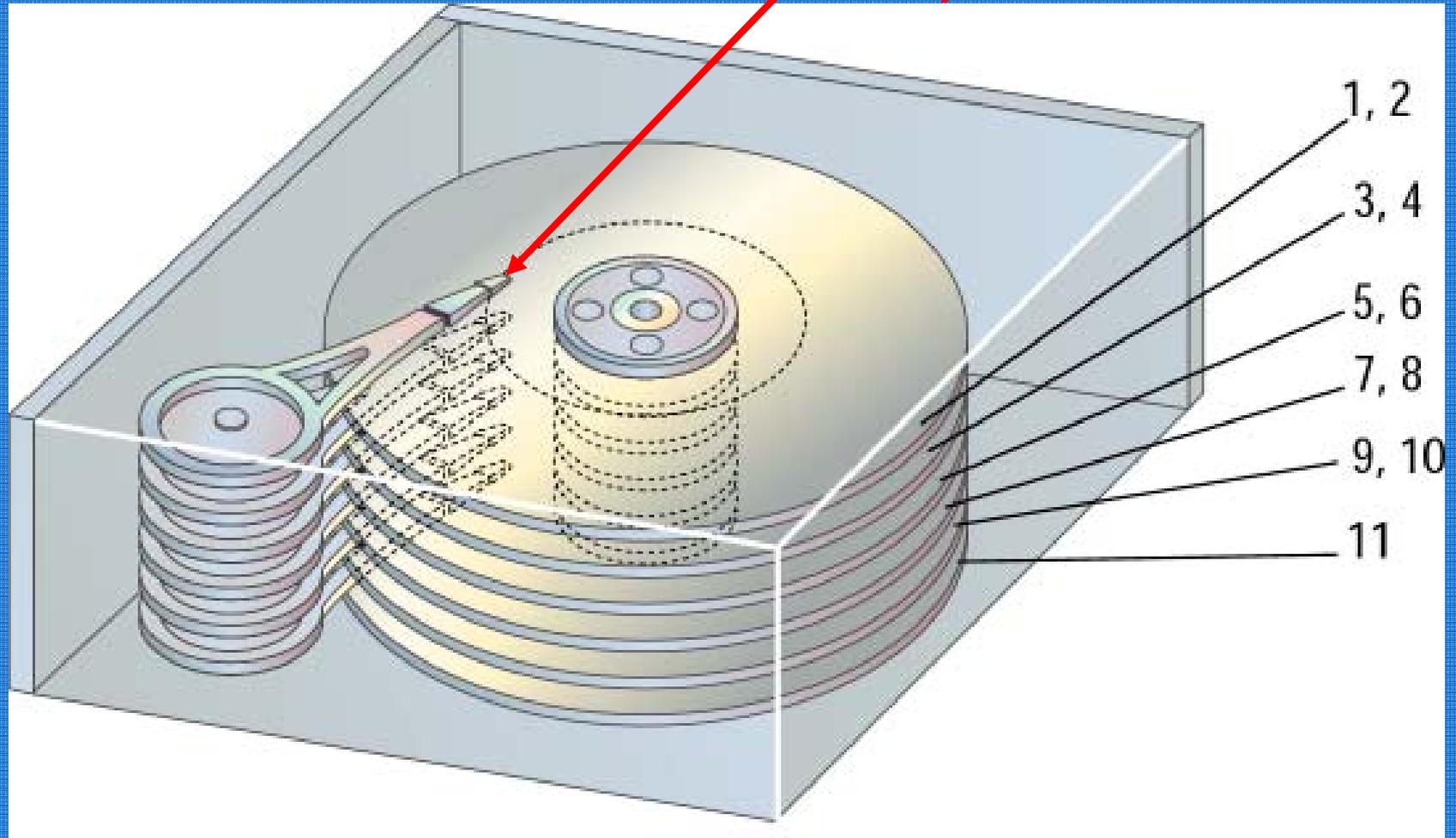
Diskette drive



## Hard Disks

- **Hard disks use multiple platters, stacked on a spindle. Each platter has two read/write heads, one for each side.**
- **Hard disks use higher-quality media and a faster rotational speed than diskettes.**
- **Removable hard disks combine high capacity with the convenience of diskettes.**

**Read/write heads**



## Disk Capacities

- **Diskettes are available in different capacities, but the most common store 1.44 MB.**
- **Hard disks store large amounts of data. New PCs feature hard disks with capacities of 10 GB and higher.**

## Other Magnetic Storage Devices

- **High-capacity floppy disks offer capacities up to 250 MB and the portability of standard floppy disks.**
- **Disk cartridges are like small removable hard disks, and can store up to 2 GB.**
- **Magnetic tape systems offer very slow data access, but provide large capacities and low cost.**

**Due to long access times, tape drives are used mainly for backups.**



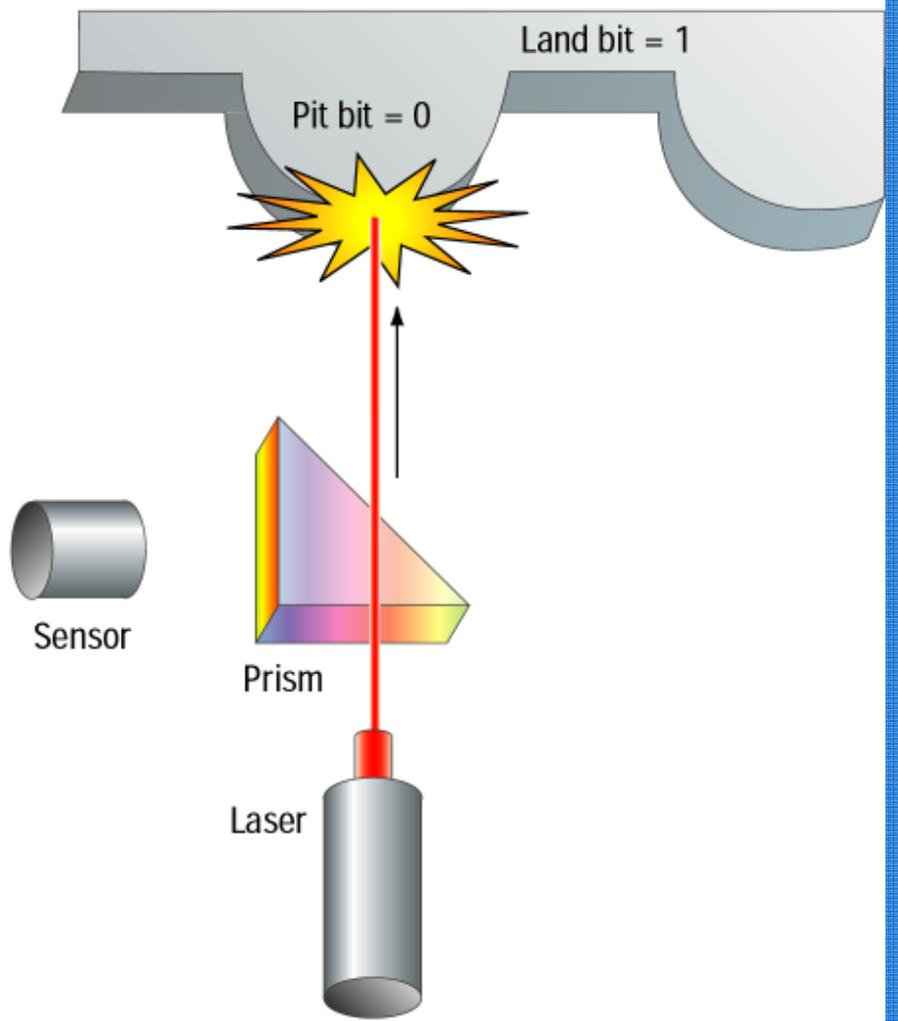
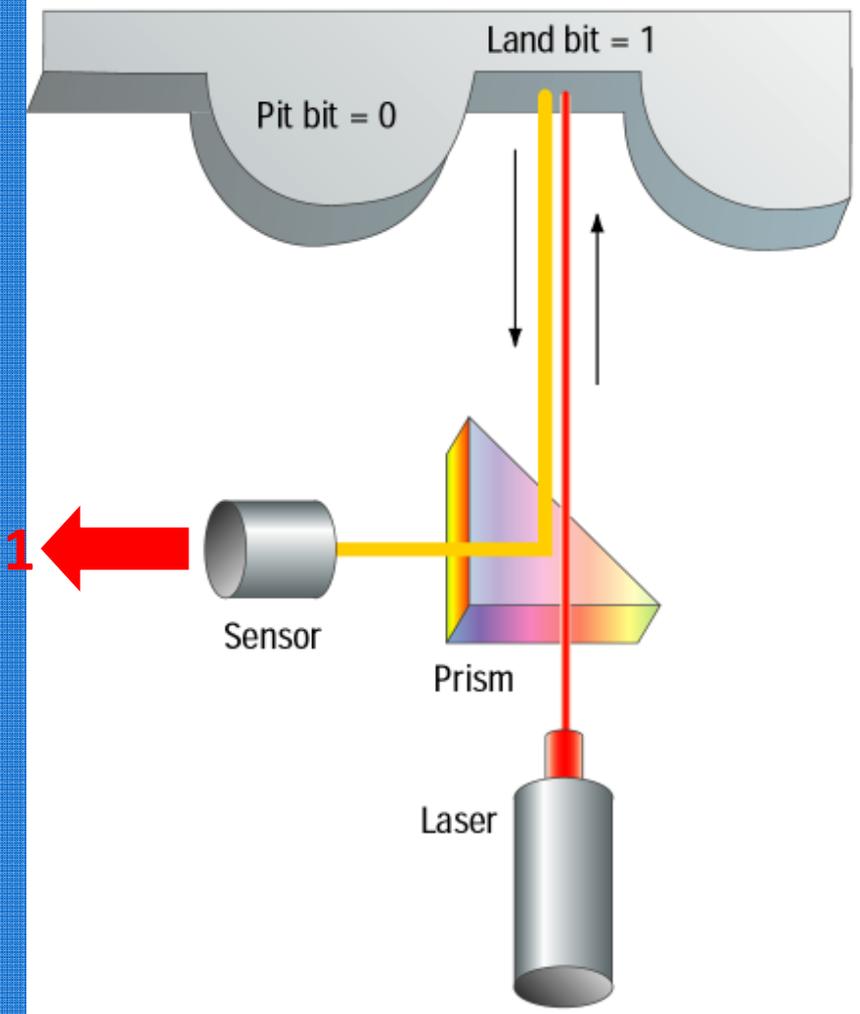
# Optical Storage Devices

- **How Optical Storage Works**
- **CD-ROM**
- **CD-ROM Speeds and Uses**
- **DVD-ROM**
- **Other Optical Storage Devices**

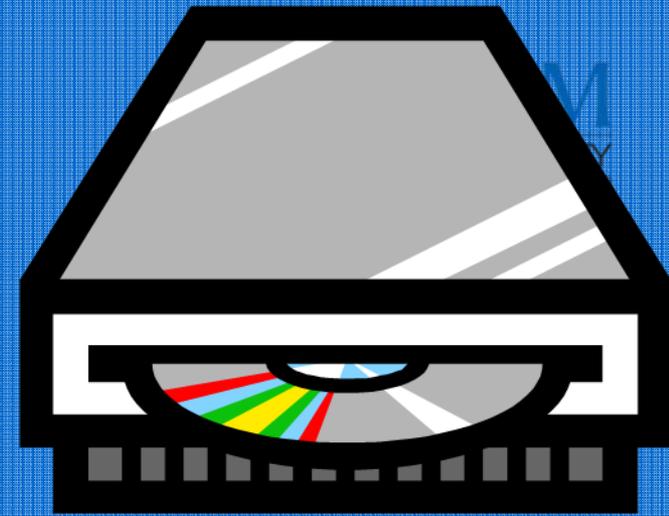
# How Optical Storage Works

- **An optical disk is a high-capacity storage medium. An optical drive uses reflected light to read data.**
- **To store data, the disk's metal surface is covered with tiny dents (pits) and flat spots (lands), which cause light to be reflected differently.**
- **When an optical drive shines light into a pit, the light cannot be reflected back. This represents a bit value of 0 (off). A land reflects light back to its source, representing a bit value of 1 (on).**

Rotation of disk 



## CD-ROM



- **In PCs, the most commonly used optical storage technology is called Compact Disk Read-Only Memory (CD-ROM).**
- **A standard CD-ROM disk can store up to 650 MB of data, or about 70 minutes of audio.**
- **Once data is written to a standard CD-ROM disk, the data cannot be altered or overwritten.**

## CD-ROM Speeds and Uses

- **Early CD-ROM drives were called single speed, and read data at a rate of 150 KBps. (Hard disks transfer data at rates of 5 – 15 MBps).**
- **CD-ROM drives now can transfer data at speeds of up to 7800 KBps. Data transfer speeds are getting faster.**
- **CD-ROM is typically used to store software programs. CDs can store audio and video data, as well as text and program instructions.**

## **DVD-ROM**

- **A variation of CD-ROM is called Digital Video Disk Read-Only Memory (DVD-ROM), and is being used in place of CD-ROM in many newer PCs.**
- **Standard DVD disks store up to 9.4 GB of data—enough to store an entire movie. Dual-layer DVD disks can store up to 17 GB.**
- **DVD disks can store so much data because both sides of the disk are used, along with sophisticated data compression technologies.**

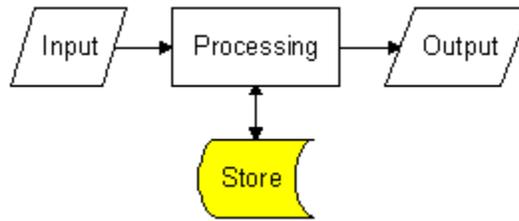
## Other Optical Storage Devices

- A **CD-Recordable (CD-R)** drive lets you record your own CDs, but data cannot be overwritten once it is recorded to the disk.
- A **CD-Rewritable (CD-RW)** drive lets you record a CD, then write new data over the already recorded data.
- **PhotoCD** technology is used to store digital photographs.

## SECTION 3: STORAGE DEVICES AND MEDIA

### STORAGE : Introduction

All information systems need to store data. This may be done temporarily whilst inputs are processed to produce outputs or for much longer periods of time.

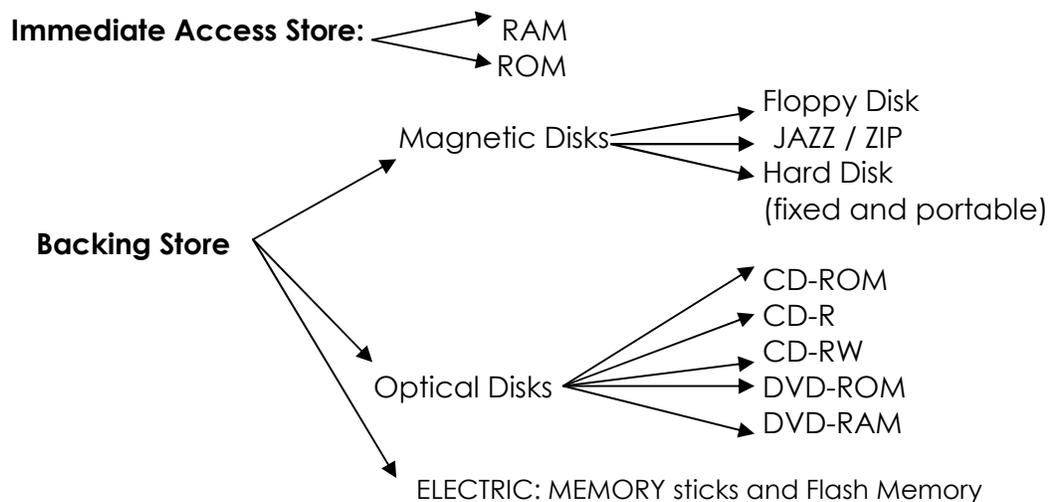


A storage device stores programs and data either temporarily or permanently. All information systems contain two different types of storage :

- **Immediate Access Store (IAS):** Immediate access store holds programs and data that the user is currently working with.
- **Backing Store:** Backing store keeps data and programs when the computer is turned off.

**Immediate access store is also known as main store or primary store( RAM). Backing store is also known as secondary store.**

The capacity (amount of data) that a storage device can hold varies significantly between different devices. Units such as bytes, kilobytes and megabytes are used to describe a storage device's capacity. Other factors such as speed of data access, cost and portability will also determine which storage device is the most appropriate one to use for a particular application.



Information stored on backing store is placed on a storage medium. The most common media which are used for backing store are: **Magnetic Disks, Magnetic Tapes, Optical Disks** such as CD-ROMs.

The data is read from or written to the storage medium by a piece of hardware known as a drive or a **storage device**.

Programs and data can not be used directly from backing store. They must be copied (loaded) into immediate access store (RAM) before they can be used. Any data which needs to be kept must be transferred back to the backing store from the immediate access store before the computer is switched off. This is called saving.

Sometimes data is compressed before it is stored. Compressing data reduces the storage space that a file uses without losing any of its contents.

## IMMEDIATE ACCESS STORE (IAS) OR MAIN INTERNAL MEMORY (RAM and ROM)

Immediate Access Store ( IAS ) holds programs and data that the user is currently working with. For example :

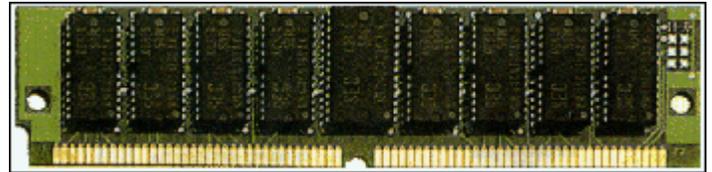
- A word processed document that is being edited will be loaded into IAS.
- An email program that is currently transmitting a message will be loaded into IAS.

There are two different types of IAS :

- Read Only Memory (ROM): The contents of ROM is permanent. It can not be altered by the user. The content is written onto the ROM when it is first made. ROM keeps its contents even when the computer is turned off and so is known as non-volatile memory. ROM is also often used in embedded systems where a small built-in computer is used to control a device such as a washing machine. The program that controls the machine is stored on ROM.



- Random Access Memory ( RAM ) : RAM is used to store programs and data that are being used by the computer. When the computer is turned on the RAM is empty. Data and programs can be put into RAM from either an input device or backing store. The data in RAM is lost when the computer is turned off so it is known as volatile memory. To keep data the user must save it to backing store before the computer is turned off.



The process of transferring data/programs from backing store into RAM so they can be used is known as **loading**.

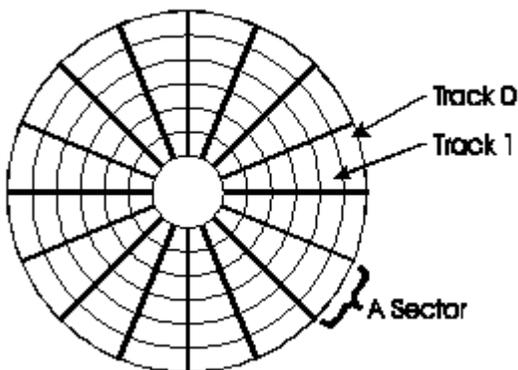
The process of transferring data/programs from RAM to backing store so that they will be retained when a computer is turned off is known as **saving**.

It is easy to add extra RAM to a microcomputer by inserting extra RAM cards called SIMM cards.

Main internal memory is located inside the computer. Data can be written to and read from RAM or ROM electronically at **very high speeds**, much faster than it can be written to or from backing store.

IAS is much more expensive to buy per Mb than backing store is.

## SECONDARY BACKING STORAGE **MAGNETIC DISKS: HARD DISK AND FLOPPY DISK**



Magnetic disks are the most common backing storage device. The two main types of magnetic disks are floppy disks and hard disks.

- 1.44Mb floppy disk.
- 120Gb hard disk drive.

Data stored on disks is arranged along a series of concentric rings called **tracks**. Each track is divided up into a number of **sectors**. Data is read to and written from a disk one sector at a time. A sector usually contains 512 or 1024 bytes of data.

The process of dividing a disk up into tracks and sectors so it can be used on a computer is known as **formatting**. You must format a new disk before you can use it.

Data is read from the disk using a disk head which moves mechanically about the disk (rather like a record player tone arm).

The disk head can move directly to any sector on the disk. Because of this a computer system can load a file or a record from a file very quickly. The system can move directly to the location of the record/file and

read it without having to read any other data from the disk. This is known as direct access. For most applications using a direct access medium is much faster than using a serial access medium.

## HARD DISKS



Hard disks are magnetic disks. They have much larger storage capacities than floppy disks. Data can be transferred to and from a hard disk much more quickly than from a floppy disk. Hard disks are usually fixed inside a computer and can not be moved between different machines. Some expensive hard disks can be moved between computers. These are called *exchangeable hard drives*.

A hard disk is made of a rigid disk which is coated with a magnetisable material. The magnetic material used is of a much higher quality than that found on floppy disks. Hard disks spin much more quickly than floppy disks and the disk head is positioned very close to the disk (thousandths of a millimetre away). Because the disk head is positioned so close to the disk hard drives can easily be damaged by dust or vibration. Therefore the disk, the drive head and all the electronics needed to operate the drive are built together into a sealed unit. This picture shows a hard disk drive with the case removed.

Typical hard disk capacities for a home PC now start at 80/120Gb and units storing up to 160Gb are available.

**Fixed Hard disk:** Used to store operating systems, software and working data. Any application which requires very fast access to data for both reading and writing to. Not for applications which need portability. Used for on-line and real time processes requiring direct access. Used in file servers for computer networks.

**Portable HARD DISK:** Used to store very large files which need transporting from one computer to another and price is not an issue. More expensive than other forms of removable media.

### Advantages:

Very fast access to data.

Data can be read directly from any part of the hard disk (random access).

The access speed is about 1000 KB per second.

## FLOPPY DISKS

Floppy disks are magnetic disks. They are portable (can be moved between computers) but have a small storage capacity. Reading and writing data from a floppy disk is very slow. The most common type of floppy disk is the 3.5" disk that can store 1.44Mb of data when it is used on a PC (enough to store about 350 pages of A4 text). Older disks were 5.25" or 8" in size but could store much less data.

A floppy disk is manufactured from a flexible plastic disk. This disk is coated with a magnetisable material. For protection the disk is encased in a plastic shell. All sizes of floppy disk have a write protect tab built into the shell. If this tab is set then data can be read from the disk but not written to it. The write protect tab can be used as a security measure to prevent important data being deleted or changed accidentally.

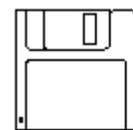
Most disks are now sold already formatted for PC's.

Floppy disks are useful for transferring data between computers and for keeping a back-up of small files.

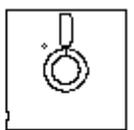
### Advantages/disadvantages:

- They are easily physically damaged if unprotected and magnetic fields can damage the data.
- They are slow to access.
- They have a small storage capacity.

Some hardware companies now produce storage devices which are very similar to floppy disks but can store 100/250Mb (ZIP) or even 1/2 Gb (JAZ) of data. These devices are also much faster than standard floppy disk drives. **ZIP AND JAZ** drives are similar to floppy drives because the individual disks are removable and portable but they hold much larger amounts of data.



3.5" Floppy



5.25" Floppy

## MAGNETIC TAPES

Magnetic tape comes in two different forms :

**Reels:** Large reels of tape (1/2 inch wide and 2400 feet long) which must be loaded into a reel-to-reel tape drive. This type of tape is usually used by mainframe computers.

**Cartridges:** The tape is supplied in a small cartridge rather like a music tape. The tape is typically 1/4 inch wide and 300 feet long. This type of tape is used on PCs (microcomputers) and the device used to read/write the tapes is called a tape streamer. The picture below is of a tape streamer for a PC. Capacities of cartridges vary from 10Gb to 200Gb.



Often files or records are stored on a tape in a particular order (e.g. sorted alphabetically by a key field). If this is the case then the tape is described as having sequential access. Because locating data on a tape takes a long time, magnetic tapes are not used as general purpose storage devices. They are only useful for a few applications. The two main applications tape is used for are :

- **Backup:** Often a tape streamer is used to make copies of data stored on a hard disk in case the data becomes corrupted. If this happens then the correct data can be restored from the tape. A backup copy of the contents of the hard disk could be made once every week. Tapes are more suitable for making backups than floppy disks or CD-ROMs. This is because the entire contents of a hard disk can be written onto one tape, producing the backup will be much quicker and tapes are cheaper to buy.
- **Batch Processing:** Tapes are preferred to disks in batch processing systems due to their relative low costs and fast read/write speeds.

### Advantages/disadvantages:

Accessing data is very slow and you cannot go directly to an item of data on the tape as you can with a disk. It is necessary to start at the beginning of the tape and search for the data as the tape goes past the heads (serial access).

Magnetic tape is relatively cheap and tape cassettes can store very large quantities of data.

## OPTICAL DISKS

Optical disks store data by changing the reflective properties of a plastic disk.

Like floppy disks, optical disks can be moved from one computer to another. They have much larger storage capacities than floppy disks but can not store as much data as a hard disk. Data can be read from an optical disk more quickly than from a floppy disk but hard disks are much quicker. As with a hard disk the drive head in an optical drive can move directly to any file on the disk so optical disks are *direct access*.

There are five types of optical disks that are currently in use. They are:

### **CD-ROM (Compact Disk - Read Only Memory)**

This is by far the most widely used type of optical disk. A CD-ROM disk can store up to 650/700Mb of data. The data is written onto the CD-ROM disk before it is sold and can not be changed by the user. Because of this CD-ROMs are often described as Write Once Read Many times (WORM) disks. CD-ROMs are used for applications such as distributing software, digital videos or multimedia products. They are also known as optical disks because the data is read by a laser beam reflecting or not reflecting from the disk surface.

CD's are available in 3 formats:

- CD-ROM's - ROM means Read Only Memory and this means you can only read from the disc, not write or store data onto it. This is the way most software programs are now sold.
- CD-R (Compact Disc - Recordable): A CD-R disk can store up to 650Mb of data. A CD-R disk is blank when it is supplied. The user can write data to it just once. After data has been written to the disk it can not be changed. A special CD-R drive which contains a higher powered laser than a CD-ROM drive is required to write to the disk. CD-Rs are often used for making permanent backups of data and distributing software when only a small number of copies are required. These CDs are

initially blank but you can use a special read/write CD drive unit to store programs and data onto the disc but they can only be written to once.

- CD-RW (Compact Disc - Rewriteable): A CD-RW disk can store up to 650/700Mb of data. CD-RW disks can be read from and written to just like a hard disk. CD-RWs can be used for any application that a hard disk can be used for but the time taken to access data is much longer than that for a hard disk. CD-RW - these are similar to the 'R' type above but you can read, write and delete files from the disc many times, just like a hard disk.

### **DVD-ROM (Digital Versatile Disk - Read Only Memory):**

DVD is the new standard for optical disks. By using a shorter wavelength laser, storing data on both sides of the disk and having more than one layer of data on each side of a disk. DVD disks are able to store much more data than CD disks. The DVD standard includes disk capacities up to 18Gb. Current DVD disks store far less than this. Because of their high capacity, DVD-ROM disks are used to store high quality video such as complete movies. Often extra data such as information about the making of the film or the actors and actresses who star in it are also stored on the disk. Unlike movies recorded on video tape, DVD-ROM movies can be interactive. The user can make selections on the screen and change what they see.

DVD drives are now replacing CD drives in computers due to the huge memory capacity of the disk and the high quality of stored images. A DVD single sided, single layer DVD can store up to 4.7 GB of data, the equivalent of 26 CD-ROMS. This means full-motion films with sound tracks and subtitles in multiple languages can easily be stored on one DVD disk.

A film stored on a DVD has significant advantages over magnetic VHS video tape because the digital images and sound tracks are of a higher quality and do not deteriorate with constant use. The user can also move to any part of the film immediately (random access).

Multi-layer and double sided DVD's can hold up to 17GB of data.

- DVD-RW drives (writable drives) are still quite expensive but may eventually replace home CD systems and VHS tapes as a way of recording films and music.
- DVD-RAM (Digital Versatile Disk - Random Access Memory): DVD-RAM disks have all of the benefits of DVD-ROM disks and can be written to as well. These very high capacity disks are ideal for producing backups. In the next five years they may replace video tapes for recording television programmes.

## **FLASH MEMORIES AND MEMORY STICKS**



**Flash memory:** is non-volatile computer memory that can be electrically erased and reprogrammed. It is a technology that is primarily used in memory cards and USB flash drives for general storage and transfer of data between computers and other digital products.

Flash memory is non-volatile, which means that no power is needed to maintain the information stored in the chip. In addition, flash memory offers fast read access times (although not as fast as volatile DRAM memory used for main memory in PCs) and better kinetic shock resistance than hard disks. These characteristics explain the popularity of flash memory in portable devices. Another feature of flash memory is that when packaged in a "memory card," it is enormously durable, being able to withstand intense

pressure, extremes of temperature, and even immersion in water.

A USB flash drive is a flash memory data storage device integrated with a USB (universal serial bus) connector. USB flash drives are typically removable and rewritable, much shorter than a floppy disk (1 to 4 inches or 2.5 to 10 cm), and weigh less than 2 ounces (60 g). Storage capacities typically range from 64 MB to 32 GB or more.

USB flash drives offer potential advantages over other portable storage devices, particularly the floppy disk. They are more compact, faster, hold much more data, have a more durable design, and are more reliable for lack of moving parts. Additionally, it has become increasingly common for computers to ship without floppy disk drives.

A flash drive consists of a small printed circuit board typically in a plastic or metal casing and more recently in rubber casings to increase their robustness.

To access the data stored in a flash drive, the drive must be connected to a USB port through either a host controller built into a computer, a USB hub, or some other device designed to access the data, such as an mp3 player with a USB-in port.

**MEMORY STICK:**

Memory Stick is a removable flash memory card format, launched by Sony, and is also used in general to describe the whole family of Memory Sticks

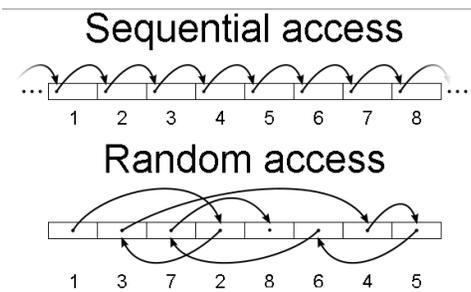
Typically, Memory Sticks are used as storage media for a portable device, in a form that can easily be removed for access by a personal computer. For example, Sony digital compact cameras use Memory Sticks for storing image files. With a Memory Stick-capable reader (typically a small box that connects via USB or some other serial connection), a user can copy the pictures taken with the Sony digital camera onto his or her computer. Sony uses and has used Memory Sticks in digital cameras, digital music players, PDAs, cellular phones, the PlayStation Portable (PSP), and in other devices, and the Sony VAIO line of personal computers has long included Memory Stick slots.



**COMPUTER DATA STORAGE:**

Often called storage or memory, refers to computer components, devices, and recording media that retain digital data used for computing for some interval of time. Computer data storage provides one of the core functions of the modern computer, that of information retention. It is one of the fundamental components of all modern computers

**TYPE OF ACCESS:**



**SERIAL or SECUENTIAL ACCESS:** sequential access means that a group of elements (e.g. data in a memory array or a disk file or on a tape) is accessed in a predetermined, ordered sequence. Sequential access is sometimes the only way of accessing the data, for example if it is on a tape. It may also be the access method of choice, for example if we simply want to process a sequence of data elements in order.

**RANDOM or DIRECT ACCESS:** is the ability to access an arbitrary element of a sequence in equal time. The opposite is sequential access, where a remote element takes longer time to access. A typical illustration of this distinction is to compare an ancient scroll (sequential; all material prior to the data needed must be unrolled) and the book (random: can be immediately flipped open to any random page. A more modern example is a cassette tape (sequential—you have to fast-forward through earlier songs to get to later ones) and a compact disc (random access—you can jump right to the track you want).

**ACCESS SPEEDS:** Main internal memory (RAM or ROM) is located inside the computer. Data can be written to and read from RAM or ROM electronically at **very high speeds**, much faster that it can be written to or from backing store.

**BACKUPS : INTRODUCTION**

No-one likes losing data that they have entered into a computer. Lost data represents a waste of time and effort. For companies the loss of data can be especially serious. Many companies store essential information such as customer accounts or stock databases on computer. Loss or corruption of this information could cost a company a lot of money.

If data is lost then it will have to be restored from the **backup**

If data is very important then appropriate security measures should be put in place to try and avoid any data loss or corruption. However problems can still occur :

- An employee could accidentally delete a file or a storage device could fail resulting in data being completely lost.

- Data could be changed accidentally. For example an employee may incorrectly update a set of records in a database. A virus could deliberately change the contents of a file.

Therefore extra copies of important data should be made on a regular basis. **These copies are known as backups.** If the original files become corrupt then the data can be restored from a backup copy. A company should have a backup strategy which sets out how backups will be made.

#### **How often should data be backup up ?**

Backups should therefore be made as regularly as possible.

Unfortunately backups cost money to produce. Storage media must be purchased and staff time is taken up. Therefore a balance has to be struck between making backups very frequently and keeping backup costs down. As a general rule files which change frequently should be backed up often whereas files which do not change very much can be backed up less frequently.

#### **When should data be backed up ?**

Backing up data can tie up a computer so that it can not be used whilst the backup is in progress. Therefore backups are usually made overnight or at weekends when a computer system is less likely to be in use.

#### **What storage medium should the backups be stored on ?**

There are a variety of different storage media that backups can be stored on. These include magnetic and optical disks and magnetic tapes. The best medium to use in a particular situation will probably depend upon the volume of data to be backed up and the speed at which the backup must be performed.

- Floppy Disks: Only suitable for backing up small amounts of data (1.44Mb) and very slow. Some high capacity floppy disks which can store up to 120Mb are now available. These can store more data but are still relatively slow.
- Optical Disks: Optical disks such as CD-RWs can back up much more data than floppy disks but not as much data as a tape. Typical capacities are around 650Mb. They are much faster than floppy disks but not as fast as magnetic tapes in most circumstances.
- Magnetic Tapes: These are suitable for backing up large volumes of data (tens of gigabytes). The backups can be performed very quickly. Tapes are very cheap to buy but the initial cost of the hardware required can be high.

#### **Where should backups be stored ?**

Store at least one backup at a site away from the main computer system. If the main system is damaged by a natural disaster such as a fire then the off site backup will be safe. Any backups kept on site may be destroyed.

Control access to backups so that unauthorised personnel can not remove them or steal data from them. Keep backups in a clean, dust-free environment so that the medium that the backups are stored on will not be damaged.

It is a good idea to make more than one backup copy.



# Electronic Storage Media

See also

- [Video Preservation](#)
- [AMIA-L](#)
- [ARSClist](#)
- [AV-Media-Matters](#)
- [Conservation DistList](#)

---

[Optical disks, CD-Roms, etc.](#)

[Various Media](#)

[NML](#)

[Holographic storage](#)

[Responses to "Ensuring the Longevity of Digital Documents"](#)

[Organizations](#)

[Miscellany](#)

[Frequently Asked Questions](#)

---

## Overview

[SearchStorage.com](#)

**Public Records Office (PRO)**

[Selecting Storage Media for Long-Term Preservation](#) (Word document)

## Optical disks, CD-Roms, etc.

See also [Glossaries](#), and [Holographic storage](#)

### The CD-Info Company

[CD-R Media Longevity](#)

### Das PhotoCD-Forum

In German. Das Wichtigste zur PhotoCD.

### Luis Facelli

[Disco Video Digital](#) (Digital Video Discs), 2004. PowerPoint presentation, in Spanish

### Anne R. Kenney and Oya Y. Rieger

[Using Kodak Photo CD Technology for Preservation and Access: A Guide for Librarians, Archivists, and Curators](#)

### Kodak

[Permanence, Care, and Handling of CDs](#)

[Permanence of Kodak Photo CD and Writable CD Media with Infoguard Protection System](#),

June 1993

[KODAK: Photo CD Technical Papers](#)

[General Information on Kodak CD-R Media](#)

### Media Sciences

[Frequently Asked Questions About Compact Discs](#)

[ISO Standards](#)

[Links for Professionals](#) (News and Technology, Technical Support, Test Equipment, CD-R Vendors)

### Jerome L. Hartke

[Measures of CD-R Longevity](#)

[CD-R Media Survey](#)

### National Archives and Records Administration

[Long-Term Usability of Optical Media The National Archives and Records Administration and the Long-Term Usability of Optical Media for Federal Records: Three Critical Problem Areas](#)

[Development of a Testing Methodology to Predict Optical Disk Life Expectancy Values \(Summary\)](#)

### National Institute of Standards and Technology

[Government Information Preservation Working Group \(NIPWG\)](#)

[Stability Study of Optical Discs](#) (PDF)

## Sanyo

[DVD FAQ](#)

[CD-Rom/CD-Audio/Game Console Discs](#)

## Linda Schamber

[Optical Disk Formats: A Briefing. ERIC Digest](#)

## Various Media

### [Ampex Virtual Museum and Mailing List](#)

"Dedicated to preserving the history of the most important manufacturer of magnetic recorders, the Ampex Virtual Museum provides online access to:

[Manuals, Schematics, and Service Bulletins](#)

[Repair, Maintenance, and Modification Tips](#)

[Parts Sources](#)

[Catalogs, Sales Brochures, and Similar Literature](#)

[Pictures](#)

[Ampex history](#) & [Other Historical Information](#)

[Alignment Instructions](#)

[FTP Server for Uploading and Downloading Ampex-related Files"](#)

### [Audio Engineering Society Historical Committee](#)

[An Audio Timeline](#): A selection of significant events, inventions, products and their purveyors, from cylinder to DVD

[Historical Interviews and Talks](#)

[3M Analog Magnetic Tape Technology](#)

Includes:

**Delos A. Eilers**

[Introduction to 3M Audio Open Reel Tape List](#)

[3M Audio Open Reel Tapes. 2000](#)

## **Jay McKnight**

[A Selected Bibliography of Histories of Magnetic Tape Sound Recording](#)

[Oral History Project.](#)

[Analog Audio Mastering Tape Print-Through.](#) Technical Bulletin A011194 (PDF)

[Stanford Acquires Ampex History Collections.](#) Nov. 2001

## **Ellen McCrady**

[NARA Conference on Preserving Tapes & Disks, March 1996: Facts and Advice from the Speakers](#) (*Abbey Newsletter* Volume 20, Number 6 Nov 1996)

## **[Quick-Data-Recovery.com](#)**

---

# **NML (Now Imation Government Services Program)**

The National Media Lab (NML), is "an industry resource supporting the U.S. Government in the evaluation, development, and deployment of advanced storage media and systems. NML endeavors to provide a broad perspective of current progress in information technology issues, both from a commercial and a government perspective."

"Imation is continuing the work of the National Media Lab (NML) through the Imation Government Services Program."

## **NML Archived Documents**

### **Imation**

[Imation Supports NASA on Space Shuttle Columbia Data Recovery](#) April 2003

### **Koichi Sadashige**

[Data Storage Technology Assessment - 2002 Projections through 2010](#) March 2003

### **Roger J. Anderson**

[High Speed Tape Mechanics](#) December 1996

### **Dr. John W. C. Van Bogart; John Merz, eds.**

[St. Thomas Electronic Records Disaster Recovery Effort](#) November 1995

### **Dr. John W. C. Van Bogart**

[Magnetic Tape Storage and Handling — A Guide for Libraries & Archives](#) The Commission on Preservation and Access June 1995

### **Martin Vos; Gary Ashton; John Van Bogart; Ron Ensminger, eds.**

[Heat & Moisture Diffusion in Magnetic Tape Packs](#) March 1994. Originally published in *IEEE Transactions on Magnetics*, Vol. 30, No. 2, March 1994. IEEE Log Number 9214603

### **Devora Molitor**

[Cleaning Methods for Helical Scan Recorders](#)

**John W.C. Van Bogart & Leon D. Wald**

[NML Storage Technology Assessment Final Report](#). See especially: [Archival Stability of Digital Storage Media](#)

**John W.C. Van Bogart**

[Recovery of Damaged Magnetic Tape and Optical Disc Media](#). Presented at "Emergency Preparedness and Disaster Recovery of Audio, Film, and Video Materials" A Library of Congress Symposium, September 21, 1995

**Koichi Sadashige**

Data Storage Technology Assessment 2000

[Part 1. Current State and Near-Term Projections for Hardware Technology. Part 1](#)

[Part 2. Storage Media Environmental Durability and Stability](#)

[Recovery of Damaged Magnetic Tape and Optical Disk Media](#)

Presented at "Emergency Preparedness and Disaster Recovery of Audio, Film, and Video Materials", Library of Congress, September 21, 1995

**Robert D. Lorentz**

[NML Thin Film Media Final Report](#)

**SIGCAT Foundation and Doculabs, Inc.**

[CD Recording: A Troubleshooting Handbook](#)

**Doculabs, Inc.**

[Compatibility of CD-R Media, Reader and Writers](#)

[Recommended Storage Conditions for Magnetic Tape](#). 1997.

[Information on Storage Media Longevities](#): Disposition Charts (life Expectancy of Various Information Storage Media (magnetic tape, optical disk, paper, microfilm) for storage at various Temperature and RH levels)

[Questions & Answers](#)

- How long do digital data storage media last?
- I have several tapes ranging in age from 2 - 8 years which are unreadable halfway through the tape due to sticky shed. I would like to get more information regarding the tape baking process for the recovery of degraded tapes. Do you have any suggested sources?
- I have a collection of audio masters recorded in the 1970's. They have developed the "sticky shed" syndrome. If I store these tapes in a sealed container over a desiccant, will this cure the problem?
- What are the life expectancies of 3.5" and 5-1/4" floppy disks? What are the optimum storage conditions for floppy disks? (We know that floppy disks are not recommended for archival storage, but they are commonly used to store data from PC's.)
- What is the sensitivity of magnetic media to temperature extremes? For

example, at what temperature is magnetic tape likely to be damaged by fire?

[Resources for Transfer and Restoration of Video and Audio Tape](#)

[Metal Particle Tape and AMPEX DST \(tm\) Media Guide](#)

[Optical Tape Technology Final Report](#)

## Other NML documents

Some NML documents no longer be available from NML, but are available at other sites:

### Anonymous

[Overview of Archival Stability of Recording Media](#)

### Council on Library and Information Resources (CLIR)

[Magnetic Tape Storage and Handling: A Guide for Libraries and Archives](#), June 1995.

Includes

[What Can Go Wrong with Magnetic Media?](#)

[Preventing Information Loss: Multiple Tape Copies](#)

[Life Expectancy: How Long Will Magnetic Media Last?](#)

[How Can You Prevent Magnetic Tape from Degrading Prematurely?](#)

[Ampex Guide to the Care and Handling of Magnetic Tape](#)

[Estimation of Magnetic Tape Life Expectancies \(LEs\)](#)

[Further Reading](#)

[Glossary](#)

---

## Holographic storage

### Colossal Storage Corp.

[3D Volume Holographic Optical Data Storage NanoTechnology](#)

### John H. Hong & Demetri Psaltis

[Dense holographic storage promises fast access](#), Adapted from *Laser Focus World*, April 1996  
p. 119.

### IBM Almaden Research Center

[Holographic Storage DEMONstrator](#)

### Lucent Technologies

[Lucent, Imation Developing Bell Labs Holographic Storage Technology](#)

### Rene Millman

## [Storage enters the third dimension](#)

### **Margaret Quan**

[Holographic storage nears debut](#), *EE Times*, April 26, 2001

### **Tom Thompson**

[Creating Holographic Storage](#). Sidebar to [When Silicon Hits Its Limits, What's Next?](#), Byte, 1996.

**Caltech Optical Information Processing Group** [Holographic Data Storage](#)

[Hesselink's Research Group \(Stanford\)](#)

## **Responses to "Ensuring the Longevity of Digital Documents"**

In the January 1995 issue of *Scientific American*, an article by Jeff Rothenberg appeared, with the headline "The digital medium is replacing paper in a dramatic record-keeping revolution. But such documents may be lost unless we act now". The following responses were sent as letters to the editor of *Scientific American* and are reproduced here with the permission of the authors.

- Jim Wheeler's [Unscientific American](#)
- John W.C. Van Bogart's [Mag Tape Life Expectancy 10-30 years](#)
- Ross Harvey's comments in [From Digital Artefact to Digital Object](#)

## **Organizations**

### [Advanced Television Systems Committee \(ATSC\)](#)

The Advanced Television Systems Committee (ATSC) was formed by the Joint Committee on Inter-Society Coordination (JCIC) to establish voluntary technical standards for advanced television systems, including digital high definition television (HDTV). ATSC suggests positions to the Department of State for their use in international standards organizations. ATSC proposes standards to the Federal Communications Commission.

### [International Federation of Television Archives \(FIAT/IFTA\)](#)

The International Federation of Television Archives (FIAT/IFTA) is a non-profit association of Television Archives, set up on June 13th, 1977 in Rome by the BBC, RAI, ARD, and INA.

## **Miscellany**

- [Quantegy Inc. \(Ampex\)](#), Source for audio, video, and instrumentation data sheets
- [Dead Media](#)
- [Open Media Framework \(OMF®\) Interchange](#)

## Frequently Asked Questions

- Andy McFadden's [CD-Recordable FAQ](#)
- [CDR Forum](#)
- [comp.arch.storage](#)
- [Laserdisc](#)
- [JPEG](#)
- [MPEG](#)
- [Sanyo DVD FAQ](#)

## Glossaries

See also [Video Preservation](#)

### **John W.C. Van Bogart**

[Magnetic Tape Storage Glossary](#)

### **The CD-Info Company**

[Compact Disc Terminology](#)

### **Steve Cunningham and Judson Rosebush**

[CD-ROM Glossary](#)

### **Octave Systems, Inc.**

[The CD Recordable Glossary](#)

### **Sanyo**

[CD-Rom Glossary](#)

[DVD Glossary](#)

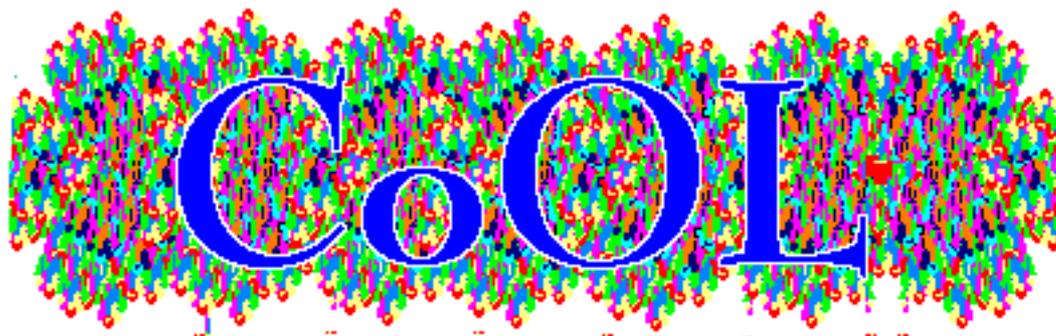


---

[\[Search all CoOL documents\]](#)

[\[Feedback\]](#)

This page last changed: June 22, 2005



## Conservation OnLine

### Resources for Conservation Professionals

#### Welcome to CoOL

CoOL, a project of the Preservation Department of Stanford University Libraries, is a full text library of conservation information, covering a wide spectrum of topics of interest to those involved with the conservation of library, archives and museum materials.

The content of CoOL comes from a variety of sources and we hope that all users will consider contributing some material to the project. As you use the server please pay attention to lacunae that you might be able to help fill. If this is your first time here, please read an important message about [copyright](#). If you would like to contribute material to CoOL, please send a note to [wh](#). To report problems or offer suggestions, select the "Feedback" buttons at the bottom of each page

#### News

This area has News, New items, and Time-sensitive information

Updated: Saturday, 17-Sep-05 12:02:28

See [Silver mirroring on silver gelatin glass negatives](#) by Giovanna Di Pietro

---

See Chris Stavroudis's software package: [The Modular Cleaning Program](#). For Windows and Macintosh.

See Mark Ormsby's software package: [Solvent Solver: A Calculator for Working with Teas Fractional Solubility Parameters](#). For Windows 95 and later.

# Hurricane Katrina

## Emergency information and Resources

<p><b>American Institute for Conservation (AIC)</b></p>	<p><a href="#">National Collections Emergency News (NCEN)</a></p> <p>"The American Institute for Conservation (AIC) established this website to provide a centralized repository of news and other information useful or those involved in efforts to preserve cultural material impacted by the hurricane as well as related health and safety issues."</p> <hr style="width: 10%; margin: 20px auto;"/> <p><a href="#">Emergency Preparedness, Response, And Recovery Committee</a></p>
<p><a href="#">The American Association of museum (AIM)</a></p>	<p>Info on Federal resources, Technical Assistance, etc</p> <p><a href="#">Reports on affected museums</a></p>
<p><b>The International Council of Museums (ICOM)</b></p>	<p><a href="#">Disaster Relief for Museums (US Mirror site in <i>Conservation OnLine</i>)</a></p> <p>ICOM home site at <a href="http://icom.museum">icom.museum</a></p> <p><a href="#">Disaster Relief for Museums</a></p> <p>Includes sections on:</p> <p>General Information:</p> <ul style="list-style-type: none"> <li>● Disaster Reporting Form</li> <li>● Bibliography</li> <li>● ICOM Publications</li> <li>● Museums Emergency Programme</li> <li>● MEP Bibliography</li> <li>● Useful Tools</li> <li>● Training</li> <li>● Conferences</li> <li>● Useful Links</li> </ul> <p>News:</p> <ul style="list-style-type: none"> <li>● Latest News</li> <li>● Disaster Relief Task Force</li> <li>● ICBS on the impact of Hurricane Katrina</li> </ul> <p>Help:</p>

	<ul style="list-style-type: none"> <li>• How can you help</li> <li>• Disaster Relief Fund</li> </ul> <p>Latest catastrophes :</p> <ul style="list-style-type: none"> <li>• Hurricane Katrina August 2005</li> <li>• Tsunami December 2004</li> </ul>
<p><b>American Library Association</b></p>	<p><a href="#">ALA Hurricane Katrina News</a></p>
<p><a href="#">Heritage Emergency National Task Force</a></p>	<p>See especially <a href="#">updates</a></p>
<p><b>The Society of Southwest Archivists (SSA)</b></p>	<p><a href="#">Archivists and Archives affected by Hurricane</a></p> <p>"The Society of Southwest Archivists (SSA) has established [a] weblog to capture and share information about our colleagues and friends from Louisiana and Mississippi, who have been affected by Hurricane Katrina. For those individuals from these states, let us hear from you directly. Secondly, if anyone knows anything about the individuals from these states, please share that with us."</p> <ul style="list-style-type: none"> <li>• <a href="#">Colleague check-in</a> If you're from the affected areas let us hear from you.</li> <li>• <a href="#">Share Zone</a> What do you know about Louisiana and Mississippi archivists?</li> <li>• <a href="#">Expressions</a> Have anything to say to the Louisiana and Mississippi archivists? Say it here.</li> <li>• <a href="#">Repository Information</a> Information about repositories in the areas affected by the hurricane.</li> <li>• <a href="#">Needs</a> For affected</li> </ul>

	<p>repositories, tell us what you need to rebuild and restore.</p> <ul style="list-style-type: none"> <li>• <a href="#">Photos</a> See and share photos of affected institutions and collections.</li> <li>• <a href="#">Jobs</a> Temporary, short-term, and contract work.</li> <li>• <a href="#">Recovery Vendors</a> Vendors can register their services for repositories in need.</li> <li>• <a href="#">Supply and Space Donations</a> Donations of supplies and space for repositories in need.</li> </ul>
<p><b>Society of American Archivists (SAA)</b></p>	<p><a href="#">Information About Archivists and Archives Affected by Hurricane Katrina</a></p> <ul style="list-style-type: none"> <li>• <a href="#">Joint Statement on Hurricane Relief</a></li> <li>• <a href="#">An Open Letter to Louisiana Gov. Blanco Regarding Hurricane Katrina and Archival Records</a></li> <li>• <a href="#">Updates from SAA Preservation Section Chair Gregor Trinkaus-Randall on Recovery Efforts</a></li> <li>• <a href="#">Heritage Emergency National Task Force "Hurricane Resource Page"</a></li> <li>• <a href="#">Northeast Document Conservation Center "Hurricane Recovery" Info</a></li> <li>• <a href="#">Hurricane Recovery</a></li> </ul>

[Resources](#)

[Available from  
State Archives](#)

- [Hurricane-  
response  
volunteer list](#)

For additional  
opportunities to serve,  
see also:

- [FEMA Issues  
Call for  
Historic  
Preservation  
Specialists](#)
- [Information on  
State  
Emergency  
Management  
Agencies](#)

**Western Association for Art  
Conservation (WAAC)**

[Disaster Salvage and Response: A  
Special Issue of \*The WAAC Newsletter\*](#)

In response to recent events, the Western Association for Art Conservation (WAAC), has devoted the September 2005 (vol 27, no 3) issue of the *WAAC Newsletter* to articles on salvage and response. In order to facilitate recovery efforts WAAC is making this issue available online for a limited time (until January 15, 2006). See <http://palimpsest.stanford.edu/waac/ttl/>

This issue includes information on health and safety for salvage operations, a reprint with a new introduction of Betty Walsh's "Salvage Operations for Water Damaged Archival Collections" and the "Salvage at a Glance" chart (in the print version this is printed on waterproof synthetic paper), a basic primer on mold, as well as a collection of new information and reprinted materials from a number of sources.

This online version will be available until January 15, 2006, but print copies are available at any time. See [back issues](#)

---

[Betty Walsh. \*Salvage Operations for Water\*](#)

	<p><a href="#"><u><i>Damaged Archival Collections: A Second Glance</i></u></a></p> <p>Absolutely <i>must have</i> information on handling a wide range of cultural property.</p>
<a href="#"><u>California Preservation Program</u></a>	Consulting services, technical information
<a href="#"><u>Conservation OnLine Disaster Page</u></a>	Technical info an links to services, resources, etc
<b>Centers for Disease Control and Prevention (CDC)</b>	<p><a href="#"><u>Hurricane Information for Response and Cleanup Workers</u></a></p> <p>Includes</p> <ul style="list-style-type: none"> <li>• Guidance for Hurricane Katrina</li> <li>• Safety</li> <li>• Cleanup</li> <li>• Handling Human Remains</li> <li>• Hurricane-Related Information for Health Care Professionals</li> </ul>
<b>National Institute for Occupational Safety and Health (NIOSH)</b>	<a href="#"><u>Emergency Response Resources: Natural Disasters</u></a>
<a href="#"><u>Heritage Preservation <i>Emergency Response and Disaster Wheel</i></u></a>	Easy to use guidance for handling a range of cultural property damaged by water, etc. See also <a href="#"><u>Disaster Resources</u></a> for links to a number of organizations that can provide information and assistance.
<a href="#"><u>Center for Great Lakes Culture Disaster Mitigation Planning Assistance</u></a>	Sample disaster plans, an excellent searchable database of disaster supplies, experts, services, and other resources.

[Regional Alliance for Preservation \(RAP\)](#)

"The members of the Regional Alliance for Preservation (RAP) are poised to offer technical assistance by phone to help in the recovery of important collections in museums, libraries, archives, historical societies and other cultural institutions. Once personal safety has been established and institutions are able to access their collections, assessment of damage can begin">

**From the news:**

- *New York Times* [Culture: Toll Is Also Exacted on Gulf Region's Historical and Cultural Treasures](#),  
by Daniel J. Wakin
- [Hurricane Center \(nola.com\) Hurricane Katrina a capricious visitor to cultural sites](#) (AP)
- See [NCEN](#) above for more news items.

**Finding People**

This area has the **ConsDir** and other tools for finding people involved with conservation as well as allied professionals.

Updated: Tuesday, 01-Nov-05 15:55:27

**Author Index**

Updated: Friday, 21-Oct-05  
17:21:23

**Search**

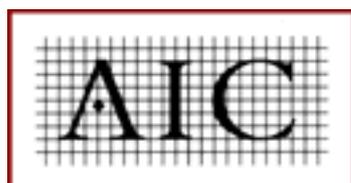
**Conservation Topics**

- [Audio materials](#)
- [Copyright and Intellectual Property](#)
- [Digital Imaging](#)
- [Disaster planning and response](#)
- [Documentation](#)
- [Education and Training](#)
- Electronic Materials
  - [Electronic media](#)
  - [Electronic records, digital archives](#)
  - [Video preservation](#)
- [Environment](#)
- [Ethics](#)
- [Conservation/Preservation Information for the General Public](#)
- [Health & Safety](#)
- [Library Binding](#)
- [Mass Deacidification](#)
- [Motion Picture Film](#)
- [Mold](#)
- [Preservation-related organizations](#) (see also the [Organizations pages](#) below.)
- [Pest Management](#)
- [Presentations and Training Tools](#)
- [Reprographics](#)
- *Sound Recordings* See [Audio materials](#)
- [Suppliers](#)
- [Survey Results](#)
- Bibliographies, etc.
  - [Bibliographies & Resource Guides](#)
  - [Conservation Fiction](#)
  - [Dictionaries, thesauri, glossaries, abbreviation lists, etc.](#)
- [Conservation Resources at Other Sites](#) ( Updated: Saturday, 15-Oct-05 11:54:41 )

## Organizations

Please see [disclaimer](#)

### Organizations with their home pages (or mirrors) in CoOL



## Albumen



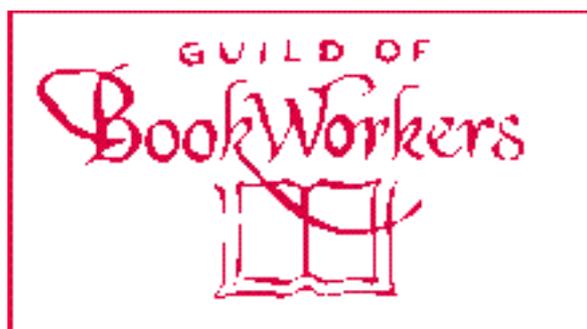
history, science & preservation



[CalPreservation.org](http://CalPreservation.org)  
Helping preserve libraries and archives



Chicago Area Conservation Group





Institute of Paper Conservation (Mirror)



(Mirror)



(Mirror)



## Organizations with Documents in CoOL

- [Chicora Foundation](#)

- [Georgia Department of Archives and History](#)

## Other organizations

- [Organizations involved with/allied to conservation](#)
- [Museum, Library and Archives Conservation/Preservation Departments](#)
- Information [about](#) some other preservation-related organizations is also available.
- List of network resources offered by [commercial firms, vendors, etc.](#)

## Mailing list archives

- Conservation DistList Archives



- [AMIA-L](#) (Association of Moving Image Archivists)
- [ARSCLIST](#) (Association for Recorded Sound Collections (ARSC) Discussion List)
- [AV Media Matters](#)
- [Book Arts-L](#)
- [ExLibris](#)
- [Conservation Framer's Mailing List](#)
- [IRRdistlist: The Infrared Reflectography DistList](#)
- [MICAT-L Archives](#). Musical Instrument Conservation and Technology. (This list is no longer in operation. Archives only)
- OSG-L (AIC Objects Specialty Group) (No longer available; access restricted to OSG members)
- [Australian Conservation Research Mailing List \(OzCons\)](#)
- [PhotoHst: History and Criticism of Photography](#) (New archives are at ASU)
- [Preservation Administration Discussion Group \(PADG\)](#)
- [TexCons \(Textiles Conservation Discussion List\)](#)
- [Bay Area Preservation Network \(BAPNet\) Archives](#) (This list is no longer in operation. Archives only)

## Misc.

[Online conservation/preservation serials](#)

[Conservation Resources at Other Sites](#) ( Updated: Saturday, 15-Oct-05 11:54:41 )

## About this server

If you're interested in some [background information](#) on CoOL, it is being prepared, slowly.

[Walter Henry](#)

[\[Search all CoOL documents\]](#) [\[Feedback\]](#)

**This page last changed: Saturday, 29-Oct-05 19:08:47**



# Video Preservation

Contributing Editor: [Hannah Frost](#)

See also

- [Electronic Storage Media](#)
- [Audio Preservation](#)
- [Motion Picture Film Preservation](#)

---

[Overview](#)

[Standards, Guidelines and Best Practices](#)

[Bibliographic Resources](#)

[Glossaries and Format Identification](#)

[Tools](#)

[Digital Video](#): Guides, Blogs, Books & Articles, Projects & Initiatives, Formats & Standards

[Organizations Concerned with Video](#)

[Preservation](#)

[Mailing Lists](#)

---

## Overview

### [Bay Area Video Coalition](#)

Abstracts, slide presentations, and transcripts from [playback 1996](#), a conference on video preservation, held in San Francisco March 29-30, 1996 and hosted by BAVC, the nation's first independent, non-commercial video preservation facility.

See also: [PLAYBACK: Preserving Analog Video](#), "an interactive DVD [produced and for sale

by BAVC] that invites users to view the technical practices of video preservation and to experience the complex decision-making process artists, conservators and video engineers engage in when the reconstruction of video artwork occurs."

### **Experimental Television Center, Ltd.**

#### [Video History Project](#)

"An on-going research initiative which documents video art and community television, as it evolved in rural and urban New York State, and across the US." Includes a useful section on "[Video Preservation - The Basics](#)."

### **Independent Media Arts Preservation (IMAP)**

#### [Preservation 101](#)

### **Jim Lindner**

[Confessions of a Videotape Restorer, or how come these tapes all need to be cleaned differently?](#)

[Digitization Reconsidered](#)

[The Proper Care and Feeding of Videotape](#)

[Magnetic Tape Deterioration: Tidal Wave at Our Shores](#)

[Videotape Restoration--Where Do I Start?](#)

### **Library of Congress**

#### [Television/Videotape Study](#)

### **[Videotape Preservation Fact Sheets](#)**

This resource, prepared by the Preservation Committee of [The Association of Moving Image Archivists](#), offers guidance to custodians of archival video collections of any size. The coverage of topics aims to be comprehensive and the discussion uses non-technical language to focus on the fundamental issues concerning the long-term care and handling of videotape.

### **[VidiPax](#)**

Informative site from a vendor that provides magnetic media duplication, restoration, digitization, and consulting services. Includes images of historic a-v equipment and a list of resources.

### **Jim Wheeler**

[The Do's and Don'ts of Videotape Care](#)

[Videotape Preservation](#)

## **Standards, Guidelines, and Best Practices**

### **Canadian Conservation Institute**

[How to Care for Video Tapes](#)

## **International Organization for Standardization (ISO)**

Standards, such as ISO 18923:2000 titled "Imaging materials -- Polyester-base magnetic tape -- Storage practices", are [available for purchase](#).

See also: [Standards relating to telecommunications, audio and video engineering](#)

## **Library of Congress**

[Specifications for Plastic Containers for Long-term Storage of Motion Picture Film and Magnetic Tape Media](#)

## **National Screen and Sound Archive (formerly ScreenSound Australia)**

[How to Care for Your Video](#)

## **Society of Motion Picture and Television Engineers (SMPTE)**

[Standards and Recommended Practices in Print](#)

Including SMPTE RP 103-1982 (Reaffirmed 1987), Care and Handling of Video Magnetic Recording Tape.

# **Bibliographic Resources**

## **Deirdre Boyle**

*Video Preservation: Securing the Future of the Past*. New York: Media Alliance. 1993.

## **Ray Edmondson**

[Audiovisual Archiving: Philosophy and Principles](#) [PDF]. Written in commemoration of the 25th anniversary of the UNESCO Recommendation for the Safeguarding and Preservation of Moving Images. 2004.

## **Jim Feeley**

"Tape Formats Compared: How do DV formats measure up with Betacam SP and 601?" In [Digital Video Magazine](#) May 1999.

## **Steven Davidson and Gregory Lukow**

*The Administration of Television Newsfilm and Videotape Collections: A Curatorial Manual*. Los Angeles and Miami: American Film Institute and the Louis Wolfson II Media History Center. 1997. Covers the range of archival issues, including several chapters on preservation.

## **Barbara L. Grabowski**

[Interactive Videodisc: An Emerging Technology for Educators](#). ERIC Digest. 1989.

## **Erich Kesse**

[Archival Copies of Video Tapes](#). University of Florida, George A. Smathers Libraries policies and procedures.

## **Helen P. Harrison, ed.**

[Audiovisual Archives: A Practical Reader](#). Compiled for the General Information Programme and UNISIST under the auspices of United Nations Educational, Scientific and Cultural

Organization (UNESCO). Section VII includes: "Preservation of Audio and Video Materials in Tropical Countries" and "Strategies for the Safeguarding of Audio and Video Materials in the Long Term" by Dietrich Schüller, Phonogrammarchiv, Vienna. Section XII includes "Emergency Preparedness and Disaster Recovery in Audiovisual Collections" by Gerald D. Gibson, Library of Congress, Washington DC. 1987.

### **Peter Hodges**

*An Introduction to Video and Audio Measurement*. Boston: Focal Press. Third edition, 2004.

### **Mika Iisakkila**

[Video Recording Formats](#)

### **Pip Laurenson**

"[The Conservation and Documentation of Video Art](#)" [pdf] In *Modern Art: Who Cares?* (Amsterdam: International Network for the Conservation of Contemporary Art): 263-271. 1999. From a symposium organized by INCCA on the unique preservation and conservation challenges posed by installation art and other contemporary art forms. Other papers from the symposium are available from [INCCA's Web site](#).

### **Library of Congress Preservation Directorate**

[Research and Testing Publications](#).

Provides a listing of printed reports which are available at no charge to the public upon request. Several address audiovisual materials specifically, including the most recent report issued in February 2002, "Bibliography on the Preservation of Magnetic Media" by Gerald D. Gibson.

### **Paul Messier**

"Dara Birnbaum's *Tiananmen Square: Break-In Transmission*: A Case Study in the Examination, Documentation, and Preservation of a Video-Based Installation". In *Journal of the American Institute for Conservation* Vol. 40, No. 3: 193-209. From the TechArcheology Symposium organized with the Bay Area Video Coalition at the San Francisco Museum of Modern Art in January 2000. This issue features other articles on electronic media preservation which emerged from discussions which took place at this pioneering event. 2001.

### **Jerry Rodgers**

"Preservation and Conservation of Video Tape." In *Care of Photographic, Moving Image and Sound Collections: Conference Papers, York, England, July 20-24, 1998*, edited by Susie Clark (Worcestershire: Insitute of Paper Conservation): 6-10. 1999.

### **Steve Schoenherr**

[Recording Technology History](#)

### **Steve Seid**

"[The Terrible Tenets of Video Preservation](#)"

### **VidiPax**

["Recovery of Mold Damaged Magnetic Tape at Vidipax: A Procedure Sheet."](#) In *Mold Reporter* vol. 1, no.6. 2001.

## **Richard Wright**

["Broadcast Archives: Preserving the Future."](#) [pdf] Describes the results of a survey of holdings within ten major European public service broadcasting archives and their preservation needs.

# **Glossaries and Format Identification Tools**

## **Association of Cinema and Video Laboratories**

[Glossary of Video Terms and Definitions](#)

## **Kodak**

[Glossary of Film / Video Terms](#)

## **National Screen and Sound Archive (formerly ScreenSound Australia)**

[Technical Glossary of Common Audiovisual Terms](#)

## **Rebecca Bachman**

[Video Preservation: Glossary of Terms](#), handout from [playback 1996](#) a conference on preservation of video, held in San Francisco March 29-30, 1996.

## **Sarah Stauderman**

[Video Format Identification Guide](#)

An invaluable tool, complete with color photographs, format obsolescence ratings, and a glossary of terms. Created with the assistance of the American Institute for Conservation's Electronic Media Group and Jim Lindner/VidiPax.

## **Texas Commission on the Arts**

[Videotape Identification and Assessment Guide](#)

# **Digital Video**

## **Guides**

### **[The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials](#)**

While chapter seven specifically addresses [Audio/Video Capture and Management](#), other sections of this guide provide information relevant to digital video and preservation, such as the appendices on equipment, metadata, digital data capture by sampling and a very extensive bibliography. Produced by the Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, and the National Initiative for a Networked Cultural Heritage (NINCH). 2002.

## Blogs

### [Digital Audiovisual Archiving \(DAVA\)](#)

"Focused on the digital transformation and preservation of audiovisual material"

## Books and Articles

### **Jerome McDonough**

[Preservation-Worthy Digital Video; or, How to Drive your Library into Chapter 11](#) [PDF]

Paper presented at the AIC [Electronic Media Group 2004 meeting](#) in Portland, Oregon.

### **Judith Thomas**

"[Digital Video, the Final Frontier](#)." In *Library Journal netConnect*. January 2004.

### **Howard D. Wactlar and Michael G. Christel**

"[Digital Video Archives: Managing Through Metadata](#)." In [Building a National Strategy for Preservation: Issues in Digital Media Archiving](#). Washington, DC: Council on Library and Information Resources. 2002.

## Projects, Workshops, and Initiatives

### [Dance Heritage Coalition's Digital Video Preservation Reformatting Project](#)

### [Getting to Disk-based Lossless Digital Video Preservation](#)

A one-day meeting held at the National Library of Medicine in Bethesda, MD in August 2005. Includes link to presentations and discussions.

### [Managing Digital Video Content](#)

Documents from a two-day workshop on current and emerging standards for managing digital video content. 2001.

### [The Open Video Project](#)

### [SMPTE Metadata Dictionary and Related Items](#)

### [Universal Preservation Format](#)

An effort initiated by WGBH-TV staff

## [ViDe](#)

"Five institutions . . . established the Video Development Initiative (ViDe) in 1998 to identify and prioritize digital video development issues." Offers access to [resources on video, especially metadata](#), and includes a publication titled, "[Digital Video for the Next Millenium](#)",

which "provides an overview of digital video on demand -- the underlying technology, the client/server capabilities currently available and development areas for the near future."

## Formats and Standards

### AAF Association

[Advanced Authoring Format](#)

### Bruce Devlin

[MXF -- the Material eXchange Format](#) [pdf]

### Diffuse

[Video Interchange Standards](#) (NOTE: This URL points to the Diffuse site preserved by the [Internet Archive](#) in June 2003.)

## MPEG

### Adam Wilt

[DV, DVCAM, DVCPRO Formats](#)

## Organizations

### Electronic Media Group (EMG)

A specialty group of the American Institute for Conservation of Historic and Artistic Works (AIC).

### Electronic Arts Intermix

EAI, which houses and distributes one of the country's largest historical collections of independent video, has a [preservation program](#) and sponsors [IMAP](#).

### Advanced Television Systems Committee (ATSC)

The Advanced Television Systems Committee (ATSC) was formed by the Joint Committee on Inter-Society Coordination (JCIC) to establish voluntary technical standards for advanced television systems, including digital high definition television (HDTV). ATSC suggests positions to the Department of State for their use in international standards organizations. ATSC proposes standards to the Federal Communications Commission.

### IMAP (Independent Media Arts Preservation)

"A service, education, and advocacy consortium, IMAP was organized in 1999 to ensure the preservation of independent electronic media for cultural and educational use by future generations."

### International Federation of Television Archives (FIAT/IFTA)

The International Federation of Television Archives (FIAT/IFTA) is a non-profit association of television archives, set up on June 13th, 1977 in Rome by the BBC, RAI, ARD, and INA.

## [National Television & Video Preservation Foundation](#)

" . . . an independent, non-profit organization created to fulfill a long-standing need by raising private funds and providing grants to support preservation and access projects at institutions with television and video collections throughout the United States."

## [PRESTO](#)

Preservation Technology, a collaborative effort on the part of several European Broadcast Archives, aims to develop state of the art technology in the preservation of film, video and audio media.

## [Southeast Asia-Pacific Audiovisual Archive Association](#)

"SEAPAVAA . . . is an association of organizations and individuals involved in, or interested in the development of audiovisual archiving in a particular geographic region - the countries of Southeast Asia . . . , Australasia . . . , and the Pacific Islands."

## Mailing Lists

### [amia-l@lsv.uky.edu](mailto:amia-l@lsv.uky.edu)

#### [Archives in Conservation OnLine](#)

AMIA-L is an e-mail discussion list sponsored by the Association of Moving Image Archivists (AMIA). It is intended to facilitate communication among AMIA members and professionals in related disciplines interested in issues relevant to the association, to archival issues involving all aspects of moving image materials and moving image archives, and to any related technologies or special interests of the profession. Appropriate postings from AMIA non-members are welcome, and subscription to AMIA-L is open to the public.

To subscribe to AMIA-L, send the following message to Listserv@LSV.UKY.EDU:  
Subscribe AMIA-L Your Name

### [AV-Media-Matters@topica.com](mailto:AV-Media-Matters@topica.com)

#### [Archives in Conservation OnLine](#)

---

[\[Search all CoOL documents\]](#)

[\[Feedback\]](#)

This page last changed: October 16, 2005



## Individual messages

### 2005

[November](#)

[October](#)

[September](#)

[August](#)

[July](#)

[June](#)

[May](#)

[April](#)

[March](#)

[February](#)

[January](#)

# AMIA-L

## Association of Moving Image Archivists

### 2004

[December](#)

[November](#)

[October](#)

[September](#)

[August](#)

[July](#)

[June](#)

[May](#)

[April](#)

[March](#)

[February](#)

[January](#)

### 2003

[December](#)

[November](#)

[October](#)

[September](#)

[August](#)

[July](#)

[June](#)

[May](#)

[April](#)

[March](#)

[February](#)

[January](#)

## 2002

[December](#)  
[November](#)  
[October](#)  
[September](#)  
[August](#)  
[July](#)  
[June](#)  
[May](#)  
[April](#)  
[March](#)  
[February](#)  
[January](#)

## 2001

[December](#)  
[November](#)  
[October](#)  
[September](#)  
[August](#)  
[July](#)  
[June](#)  
[May](#)  
[April](#)  
[March](#)  
[February](#)  
[January](#)

## 2000

[December](#)  
[November](#)  
[October](#)  
[September](#)  
[August](#)  
[July](#)  
[June](#)  
[May](#)  
[April](#)  
[March](#)  
[February](#)

[January](#)

**1999**

[December](#)

[November](#)

[Search Association of Moving Image Archivists Mailing List Archive archives](#)

[AMIA web site](#)

---

## AMIA-L, The Association's E-mail Discussion List

This is a secondary archive site for AMIA-L. The primary archive is at <http://lsv.uky.edu/archives/amia-l.html>

AMIA-L is an online discussion list focusing on topics relating to the preservation of moving images. Utilizing this electronic forum, [AMIA](#) members, and others interested in the preservation of motion picture film, video, and other moving image formats, are able to communicate on a daily basis. Through messages posted to AMIA-L, list subscribers are able to exchange information relating to all aspects of moving image preservation and the operation of moving image archives. Appropriate postings can include, but are not limited to: queries concerning archival holdings, preservation activities, availability of equipment and services; announcements of job openings, conferences and meetings, new acquisitions, new publications, etc.

To subscribe to AMIA-L, send the following message to **Listserv@LSV.UKY.EDU**

### **Subscribe AMIA-L Your Name**

Your name and the address from which the subscription message is sent will be added to the AMIA-L subscribers list. Please be sure to send the subscription message while connected to the account to which you want AMIA-L postings delivered.

You will receive a "Welcome" message and further instructions. Messages posted on AMIA-L are saved in archive files. Once you have subscribed you can retrieve any of the previous postings.

You may also start the AMIA-L subscription process by visiting the AMIA-L Web Site. Click on the link [Join or leave the AMIA-L list](#) and follow the instructions from there.

For more information, contact the list administrators [AMIA-L-request@LSV.UKY.EDU](mailto:AMIA-L-request@LSV.UKY.EDU)

# ARSC Recorded Sound Discussion List



## Messages

### 2005

- [November](#)
- [October](#)
- [September](#)
- [August](#)
- [July](#)
- [June](#)
- [May](#)
- [April](#)
- [March](#)
- [February](#)
- [January](#)

### 2004

- [December](#)
- [November](#)
- [October](#)
- [September](#)
- [August](#)
- [July](#)
- [June](#)
- [May](#)
- [April](#)
- [March](#)
- [February](#)
- [January](#)

### 2003

- [December](#)
- [November](#)
- [October](#)
- [September](#)
- [August](#)
- [July](#)
- [June](#)

[May](#)  
[April](#)  
[March](#)  
[February](#)  
[January](#)

**2002**

[December](#)  
[November](#)  
[October](#)  
[September](#)  
[August](#)  
[July](#)  
[June](#)  
[May](#)  
[April](#)  
[March](#)  
[February](#)  
[January](#)

**2001**

[December](#)  
[November](#)  
[October](#)  
[September](#)  
[August](#)  
[July](#)  
[June](#)  
[May](#)  
[April](#)  
[March](#)  
[February](#)  
[January](#)

**2000**

[December](#)  
[November](#)  
[October](#)  
[September](#)  
[August](#)  
[July](#)

[June](#)

[May](#)

[April](#)

[March](#)

[February](#)

[January](#)

**1999**

[December](#)

[November](#)

[October](#)

[Search ARSC Recorded Sound Discussion List archives](#)

---

## ARSC Recorded Sound Discussion List

ARSCLIST is an unmoderated mail reflector to facilitate the exchange of information on sound archives and promote communication among those interested in preserving, documenting, and making accessible the history of recorded sound. The list is sponsored by the Association for Recorded Sound Collections (ARSC) as a service to ARSC members and the archival community at large.

## How to Subscribe to the List

To subscribe, send an e-mail to:

[listserv@listserv.loc.gov](mailto:listserv@listserv.loc.gov)

Leave the "Subject:" blank. In the first line of the body of the message, type:

SUBSCRIBE ARSCLIST [your name]

Please type your name after the name of the list, as in: "SUBSCRIBE ARSCLIST Joe H. Smith". Alternatively, if you want to subscribe anonymously, send the command: "SUBSCRIBE ARSCLIST Anonymous". Your subscription will then be hidden automatically.

Then, send the message normally. You will then be subscribed to the list. Only subscribers may post to the list.

## The Purpose of the List

Topics appropriate for discussion may include discussion about recorded sound research, history, innovations, preservation, archiving, copyrights and access and announcements about ARSC activities and publications. Other lists may be more appropriate forums for subjective discussions of particular recordings or artists, restoration of antique equipment, buying and selling recordings and the collecting of ephemera. All messages posted to the list will be archived permanently. By posting to this list the subscriber agrees to have his or her message become part of the permanent public archive.

## ARSClist Archive

Contributions sent to this list are automatically archived in two locations. Archives from January 2003 are available at the following site maintained by the Library of Congress at <http://listserv.loc.gov/listarch/arsclist.html>. You can request a list of the available archive files using email by sending an "INDEX ARSCLIST" command to [LISTSERV@LISTSERV.LOC.GOV](mailto:LISTSERV@LISTSERV.LOC.GOV). You can then order these files with a "GET ARSCLIST LOGxxxx" command (where xxxx stands for a file number you select,) or by using listserv's database search facilities.

Send an "INFO DATABASE" command for more information on the latter. The list is also available in digest form. If you wish to receive the digested version of the postings, just issue a SET ARSCLIST DIGEST command.

The complete ARSCLIST archives are kept at the Conservation OnLine site maintained by Stanford University at <http://palimpsest.stanford.edu/byform/mailling-lists/arsclist/>. You must have a web client to use the archives at Stanford. These archives are courtesy of Stanford University Libraries, who set them up and maintain them. As with any web based resource, the Stanford archives are often available on other web sites because of the activities of web robots or other search engines. By joining ARSCLIST, you acknowledge that you understand any message you post or anything said about you is public information, may be spread over the entire world by way of the world wide web, and that anyone with a web browser may access the Stanford archives. Once archived, messages become part of the historical record of discourse in this field and will not be removed from the archives.

If you have questions or problems accessing the archives, contact [Mary Bucknum](#).

## Host Institution

The Library of Congress hosts the list. They provide the server, disk space, network connections, software and technical support to set up and run the list. [Mary Bucknum](#), Sound Recording Curator at the Library of Congress, coordinates the day to day management of the list, assisted by [Larry Appelbaum](#) of the Library of Congress and [David Seubert](#) of the University of California, Santa Barbara.

## The Fine Print

Messages posted to this list do not necessarily represent the opinions of ARSC, the ARSC Board of Directors, the list owners or the host institution.

Flaming and personal attacks are not appropriate for this list. Such behavior will result in removal from the list at the sole discretion of the list owner. Spamming and commercial advertising are also inappropriate and will not be tolerated and can result in removal from the list. Personal messages, subscription requests and other business should be sent to the appropriate address and not posted to the list.

For the consideration of users with systems of varying capability list members are strongly encouraged to include the text of any MIME attachments within the body of an email message or to post files to a web site and post the URL.

## To unsubscribe from the list

Send an e-mail to:

[listserv@listserv.loc.gov](mailto:listserv@listserv.loc.gov)

Leave the "Subject:" blank. In the first line of the body of the message, type in:

signoff arsclist

You will then get an automated message returned to you saying you have been unsubscribed.

# AV Media Matters

## List Archives

### Individual messages

#### 2005

[November](#)

[October](#)

[September](#)

#### 2004

[April](#)

#### 2003

[November](#)

[August](#)

[June](#)

[May](#)

[April](#)

[March](#)

[February](#)

[January](#)

#### 2002

[December](#)

[November](#)

[October](#)

[September](#)

[August](#)

[July](#)

[June](#)

[May](#)

[April](#)

#### 2001

[July](#)

[June](#)

[May](#)

[April](#)

[March](#)

[February](#)

[January](#)

**2000**

[December](#)

[November](#)

[October](#)

[September](#)

[August](#)

[July](#)

[June](#)

[May](#)

[April](#)

[March](#)

[February](#)

[January](#)

**1999**

[December](#)

[November](#)

[October](#)

[September](#)

[August](#)

[July](#)

[June](#)

[May](#)

[April](#)

**[Search AV-Media-Matters archives](#)**

This is a moderated list which means that submissions will be reviewed before posting - posting will therefore NOT be automatic and there very well may be a delay between submission and posting to the group. This is being done in order to minimize spam, off-topic posts, as well as flames.

This list is for those who are interested in issues related to magnetic, optical, and other media - past, present, and future - both professionals and amateurs or "newbies" are welcome. We hope that professionals representing manufacturers as well as users will contribute their valuable expertise and share it with the list members. This list is NOT concerned with content (there are many other lists for that purpose). Off topic submissions will not be posted. Commercial information or press releases are allowed provided that they specifically relate to AV Media Matters. Warning - the decision of who is on the list or not on the list as well as which submissions are posted is decided by the moderator - this is done to eliminate off topic and other wastes of time - membership and participation in the list is therefore up to the SOLE discretion of the moderator.

## Subscribing

To subscribe/unsubscribe see <http://www.topica.com/lists/AV-Media-Matters/> or send mail to AV-Media-Matters-subscribe@topica.com

## Archives

Archives are kept at Stanford University in the web site Conservation OnLine (CoOL) at . These archives are courtesy of Stanford University Libraries, who set them up and maintain them. As with any web based resource, the Stanford archives are often available on other web sites because of the activities of web robots or other search engines. Topica also maintains publically accessible archives of the list traffic. By joining AV-Media-Matters, you acknowledge that you understand any message you post or anything said about you is public information, may be spread over the entire world by way of the world wide web, and that anyone with a web browser may access the Stanford and/or Topica archives.



## Conservation DistList Archives

### Individual messages (analytics) of the *Conservation DistList*

#### Browse by year

[2005](#) [2004](#) [2003](#) [2002](#) [2001](#) [2000](#) [1999](#) [1998](#) [1996](#) [1995](#) [1994](#) [1993](#)  
[1992](#) [1991](#) [1990](#) [1989](#) [1987-1988](#)

Browse [Author Index](#)

[Search](#)

### *Conservation DistList* back instances

All [back mailings of the \*Conservation DistList\*](#) are available for browsing. These are the complete versions, as received by DistList participants.

### Instructions for participants

Please [read the instructions](#) before posting. This document covers list rules and policies, as well as giving instructions for changing your list info (ConsDir entry), setting nomain, etc.

---

## Participating in the Conservation DistList

If you are professionally involved with the conservation of museum, archive, or library materials, please consider [participating in the \*Conservation DistList\*](#). An interdisciplinary forum, the DistList is open to conservators, conservation scientists, curators, librarians, archivist, administrators, and others whose work life touches on the preservation of cultural property. In addition, if you are a student in museum, library, or archive program, or considering conservation as a career, you may find the DistList an interesting introduction to the work of our field(s).

Participants are asked to fill out a brief questionnaire, in order to be included in the [ConsDir](#). To receive the questionnaire and sign on to the DistList, send a one line note

```
subscribe consdist YourFirstName YourLastName
```

to consdist-request@lindy.stanford.edu



This page last changed: May 24, 2005

ADVERTISEMENT

[NEWS](#) • [TOPICS](#) • [ITKNOWLEDGE EXCHANGE](#) • [TIPS](#) • [ASK THE EXPERTS](#) • [WEBCASTS](#) • [WHITE PAPERS](#) • [PRODUCTS](#) • [CAREERS](#)

SEARCH this site and the web

[ADVANCED SEARCH](#) | [SITE MAP](#)

Search Powered by 

**TODAY ON SEARCHSTORAGE.COM**

ADVERTISEMENT

**[Users say virtualization switch worth the leap](#)**

ARTICLE - According to users, installing Acopia's virtualization switch requires new thinking around storage, but they say they expect it to save them big money in the long run. (SearchStorage.com)

[→ MORE ON VIRTUALIZATION](#)

**[Guide to buying and implementing backup software](#)**

ARTICLE - With so much riding on enterprise backups, it's important to choose the best product and ensure that any new product will easily integrate into your environment. (SearchStorage.com)

[→ MORE ON BACKUP SOFTWARE](#)

**[Pop Quiz: An IT Thanksgiving](#)**

LEARNING GUIDE - Please join us for our traditional Thanksgiving quiz-feast of tasty IT terms. (SearchStorage.com)

**WHAT'S NEW**

▶ **[Storage Decisions presentations online](#)**

Take advantage of exclusive access to the presentations from Storage Decision Las Vegas "Assembling the Tiered Storage Enterprise".



▶ **[Check out upcoming storage events](#)**

**GET STORAGE MAGAZINE**

**Free Subscription Offer - [Apply today for a FREE subscription](#) to the leading monthly magazine for IT pros charged with building and managing the growth of storage networks.**



## STORAGE TOPICS

### [SAN \(storage area network\)](#)

SAN (storage area network) (General), SAN planning and design, SAN components, Switches, SAN management and administration, SAN routing and partitioning, SAN Protocols, Connectivity

### [NAS \(network-attached storage\)](#)

NAS (network-attached storage) (General), NAS backup, NAS clusters, NAS hardware, NAS management, NAS protocols

### [Primary storage/Storage hardware](#)

Primary storage/Storage hardware (General), Disk arrays, Drives, Connectivity, RAID technology

### [Backup/Data protection](#)

Backup/Data protection (General), Backup process, Backup software, Business continuance, By vendor, Disk-based backup, Tape backup

## STORAGE INFO CENTER

SearchStorage.com Info Centers offer IT professionals in-depth news and technical advice on the hottest topics in the Storage industry.

SPONSORED BY: EMC

## SITE HIGHLIGHTS

### EDITOR

[Cathleen Gagne](#)

New to SearchStorage? Meet our editorial team



### [Data management/Storage management](#)

Data management/Storage management (General), Compliance, Data management process, Data management tools, Virtualization

### [Storage strategy \(Planning, buying, vendors\)](#)

Storage strategy (Planning, buying, vendors) (General), Buying advice, Enterprise level, Professional issues, Strategic vendors



### Select newsletters you would like to receive via e-mail!

Today's News

Storage Management

TechTarget Conference Updates

Not a member? We'll activate your membership with your subscription.

### MEMBER BENEFITS

- [Apply for a FREE subscription to Storage magazine!](#)
- 

**TechTarget**  
Storage Media

**STORAGE**

[View this month's issue and subscribe today.](#)

**Storage Decisions**

[Apply online for free conference admission.](#)

 **SearchStorage.com**

[NEWS](#) • [TOPICS](#) • [ITKNOWLEDGE EXCHANGE](#) • [TIPS](#) • [ASK THE EXPERTS](#) • [WEBCASTS](#) • [WHITE PAPERS](#) • [PRODUCTS](#) • [CAREERS](#)

[About Us](#) | [Contact Us](#) | [For Advertisers](#) | [For Business Partners](#) | [Reprints](#) | [RSS](#)

**SEARCH**

SearchStorage.com is part of the TechTarget network of industry-specific IT Web sites

#### WINDOWS

[SearchExchange.com](#)

[SearchSQLServer.com](#)

[SearchVB.com](#)

[SearchWin2000.com](#)

[SearchWindowsSecurity.com](#)

[SearchWinSystems.com](#)

[Labmice.net](#)

#### APPLICATIONS

[SearchCRM.com](#)

[SearchSAP.com](#)

#### ENTERPRISE IT MANAGEMENT

[SearchCIO.com](#)

[SearchDataCenter.com](#)

[SearchDataManagement.com](#)

[SearchSMB.com](#)

#### CORE TECHNOLOGIES

[SearchEnterpriseVoice.com](#)

[SearchMobileComputing.com](#)

[SearchNetworking.com](#)

[SearchOracle.com](#)

[SearchSecurity.com](#)

[SearchStorage.com](#)

[SearchWebServices.com](#)

#### PLATFORMS

[Search390.com](#)

[Search400.com](#)

[SearchDomino.com](#)

[SearchOpenSource.com](#)



[WhatIs.com](#)

[TechTarget Expert Answer Center](#) | [TechTarget Enterprise IT Conferences](#) | [TechTarget Corporate Web Site](#) | [Media Kit](#) | [Site Map](#)

Explore **SearchTechTarget.com**, the guide to the TechTarget network of industry-specific IT Web sites.

All Rights Reserved, [Copyright 2000 - 2005, TechTarget](#)

[Read our Privacy Statement](#)

- Home
- About us
- Visit us
- Research, education & online exhibitions
- Search the archives
- Services for professionals
- News
- Shop online

You are here: [Home](#) > [Services for professionals](#) > Preservation

## Preservation

- Introduction
- Digital Preservation
- Collection Care
- Contact us



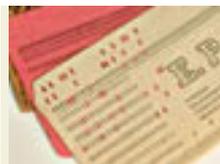
### Preservation



The National Archives aims to ensure that public records are preserved for present and future access. We work to raise awareness amongst readers and staff of the joint responsibility for material in our care.

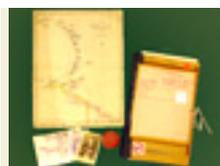
Records for preservation dating from 1817 to present day

The National Archives' preservation service aims to preserve the records through:



### Digital Preservation

The National Archives are playing an active role in storing and preserving digital material. [Go >](#)



### Collection Care

Including preservation responsibilities, conservation techniques and conservation research projects. [Go >](#)

# CD-R Media Longevity

## New Update

A major news magazine misquotes the National Media Lab's data. Read our [Special Report](#) on *CD Media Longevity Misrepresented in US News & World Report*.

See [the report on the National Media Lab-funded media compatibility tests](#) for more insight into testing methods and some recent results.

## Background

[CD-Recordable](#) is a young and rapidly growing field within the larger Compact Disc industry. The technology is complex and exacting, and is still evolving. There are many questions in people's minds about the technology and its uses, but this section focuses on the expected lifespan of the recordable discs themselves. Please refer to [the Technology Overview](#) section in this site for articles on other topics relating to CD-R. The [glossary of CD and CD-Recordable terminology](#) provides some basic information about the technology.

## The Issues

One of the primary applications for CD-Recordable is archiving. To consider CD-R a viable medium for this purpose, obviously some data on the reliability and expected lifespan of discs is required by those responsible for preserving important information. CD-R is a new technology, however, and estimating lifespans expected to be many decades for a medium that is less than fifteen years old requires sophisticated accelerated aging techniques and rigorous testing methodologies whose results may not be easy to interpret. Also, because the [media](#) itself has been relatively expensive, testing by independent agencies has been rare. Since prices have dropped significantly, that situation may change soon.

Recently two organizations, the [Optical Storage Technology Association \(OSTA\)](#) and [the Special Interest Group for CD Applications and Technology \(SIGCAT\)](#) have conducted tests for compatibility with recorders and players, but longevity testing remains relatively uncommon.

## The Manufacturers

Several of the leading manufacturers of CD-R media have provided information about their products for this section. These links will be made active as the material is released. Thank you for your patience!

 [Kodak](#)



 Taiyo Yuden

 [TDK](#)

 Verbatim

## Commentary

The controversy surrounding CD & CD-R media longevity and readability continues. See our [Special Report](#) about a recent article in US News & World Report that misstates the facts.

Another [comment on an earlier print media brouhaha](#) about CD-R media life expectancy is provided here by .

[Home](#) | [Applications](#) | [Bibliography](#) | [CD-Lite](#)  
[History](#) | [Industry](#) | [Sponsors](#) | [Technology](#)

© 1995, 1996, 1998

Send comments to [Email](#)

030402



[Archiv-Portal](#) bei  
ART & SCIENCE

## Das ImagePac Forum

Grundlagen und Anwendungen der Kodak Photo CD



Das Wichtigste zur Photo CD



Arbeiten mit der Photo CD



Die Zukunft des Bildes



Support und Software



Technologie der Photo CD



Wer liefert die Photo CD?



Kontakt



Systemintegratoren+ Berater

Ein Klick auf die Titelgrafik führt Sie zur detaillierten [Inhaltsübersicht](#)

Letzte Änderung am 29.12.2004

Copyright für die deutschen Texte: Roland Dreyer, [Art & Science](#) Stuttgart

# Using Kodak Photo CD Technology for Preservation and Access

## A Guide for Librarians, Archivists, and Curators

Anne R. Kenney and Oya Y. Rieger

*Department of Preservation and Conservation, Cornell University Library  
for  
New York State Education Department, Program for the Conservation and  
Preservation of Library Materials*

*May 1998*

---

- [Adobe Acrobat \(PDF\) Version of the Brochure](#) (requires [Adobe Acrobat](#) reader)
- [HTML Version of the Brochure](#)
- **Figures**

[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#)

- **Forms**

[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#)

---

To obtain a hard copy of this report, please send a **self-addressed** and **stamped** 9x12 envelope (*first class* postage for a 4.5 ounces brochure -- \$1.24 for US and \$3.80 for overseas) to:

[Mary Arsenault](#)

Department of Preservation and Conservation  
215 Olin Library  
Cornell University  
Ithaca, NY 14853

---

[Return to Publications Page](#)

---

Last updated October 1, 1998, oyr\_  
[Send comments to cd58@cornell.edu](mailto:cd58@cornell.edu)

# Permanence of Kodak Photo CD and Writable CD Media with InfoGuard Protection System

Eastman Kodak Company  
343 State Street  
Rochester, NY 14650  
June 1993

As new applications for writable CD products are identified and implemented, the issue of permanence of information recorded on CDs must be addressed. As with other information-recording media, including photographic film, magnetic tape, and optical discs, different manufacturers produce products having different quality, reliability, and lifetime properties. It is important that customers carefully select their suppliers.

Kodak manufactures Kodak Writable CD Media with InfoGuard Protection System, and Kodak Photo CD media\* , using materials and manufacturing processes unlike those that are used for mass-produced audio or data discs. The reflector Kodak uses is gold, a material that is not adversely affected by moisture, oxygen, or solvents. The data-recording layer Kodak uses is a carefully selected laser-sensitive dye that does not change significantly over time, even when exposed to extreme light, heat, and humidity conditions.

Kodak knows there is a large diversity in composition of the dye layers used by various media manufacturers. Data stability characteristics in other manufacturers' media vary widely. Stability is determined primarily by the types of dyes used to make the discs. We have tested Kodak Photo CD and Kodak Writable CD Media with InfoGuard Protection System, and other manufacturers' media. In our accelerated lifetime testing, over a very broad range of conditions, Kodak media were equivalent to, or superior to, all other writable CD products tested. A variety of methods are used to project product lifetimes. Typically, products are stored under a range of temperatures and humidities, and subjected to cold/hot and dry/humid cycling.

Kodak CD media, under the conditions tested to date, do not change significantly, even after testing over extended periods of time. We predict the lifetime of Kodak Photo CD, and Kodak Writable CD Media with InfoGuard Protection System, under normal storage conditions in an office or home environment, should be 100 years or more. When comparing lifetime claims from different CD-ROM media manufacturers, it is important to compare error rate data and testing conditions. We continue testing, covering many production lots of material, under a variety of testing conditions. Additional results will be reported when testing is completed. It is important to note that results from any accelerated aging tests are predictions that must be confirmed in real-life keeping tests.

Kodak Photo CD, and Kodak Writable CD Media with InfoGuard Protection System, also have been specially treated to dramatically reduce damage caused by handling. Our scratch-tolerant surface resists damage even when subjected to forces twice as great as those that would cause data loss on conventional mass-produced CDs, and five times greater than those that would cause data loss on

some other recordable CD media. The resultant Kodak product has excellent physical durability characteristics.

For more information on Kodak Photo CD and Kodak Writable CD Media with InfoGuard Protection System, consumers can call the Kodak Customer Assistance Center at 800-242-2424, ext. 53. Consumers also can reach Kodak technical support staff in the Kodak CompuServe Forum by typing GO: KODAK.\*

Kodak Photo CD Master, Pro Photo CD Master, Photo CD Portfolio, and Photo CD Catalog disc formats####(Note: Kodak and InfoGuard are trademarks of Eastman Kodak Company.)



---

[\[Search all CoOL documents\]](#)

[\[Feedback\]](#)

**This page last changed: August 24, 2005**

## Frequently Asked Questions

### General Information on KODAK CD-R Media

---

#### Contents

- [Capacity](#)
- [General Product Information](#)
- [Packaging](#)
- [Recording Speed](#)

These Frequently Asked Questions (FAQs) were last updated 28 March, 2002

---

#### Capacity

**1. How much data can I store on a 74 minute CD-R?**

Although the stated capacity of a 74-minute CD-R is 650 MB, keep in mind that the exact amount of information that can be stored on any disc depends on the mode of recording, the length of individual files, and the number of tracks recorded, etc. The overhead associated with these variables uses up some of the available capacity. The official capacity of a CD-R disc is determined by its Maximum Start of Lead Out (MSLO) time. The MSLO of a Kodak 74 minute disc is 74 minutes, 5 seconds and 1 frame.

**2. How much data can I store on an 80-minute CD-R?**

You can store 700 MB of data on an 80-minute CD-R. However, keep in mind that the exact amount of information that can be stored on any disc depends on the mode of recording, the length of individual files, and the number of tracks recorded, etc. The overhead associated with these variables uses up some of the available capacity. The following calculation may be useful:

- o In a standard (Mode 1) CD-ROM, there are 2048 bytes of user data per sector.
- o There are 75 sectors per second.
- o There are 60 seconds per minute.
- o There are 80 minutes per disc.

This equates to 737,280,000 bytes of user data per disc. If you define a megabyte as one million bytes, then there are 737.28 MB of storage per disc. If you define a megabyte as 1,048,576 bytes (1024\*1024), then there are 703.125 MB of storage per disc.

**3. How is the additional capacity achieved between KODAK 74-minute (650 MB) media and 80-minute (700 MB) media?**

The 6% increase in capacity is obtained by decreasing the track pitch by 6%, from 1.6 um to 1.5 um. 80-minute media was always allowed in the Orange Book specification, as was the reduced track pitch. CD-R media with record times longer than 80 minutes are not compliant with industry standards for CD-R media.

**4. Are there any compatibility issues when using 80-minute media instead of 74-minute media?**

As long as you are running the most up-to-date firmware in your writer, there should be no issues at all. If you have an older drive, and you have verified that you are running the most up-to-date firmware, contact your drive manufacturer or visit their Web site to determine whether they are working on firmware that enables 80-minute recording.

**5. Does Kodak 80-minute media cost more than the 74-minute media?**

80-minute media is priced at industry-competitive prices, similar to the pricing of 74-minute media.

**6. Can you get CD-R's with greater than 80 minutes capacity? Does Kodak make**

**them?**

The maximum capacity CD-R media Kodak produced conforms to the maximum capacity allowed by the official specification: 80 minutes. The Maximum Start of Lead Out (MSLO) of a KODAK CD-R Ultima 80-minute disc is 79 minutes, 59 seconds, 74 frames (364,499 sectors). The grooves on KODAK CD-R Ultima 80 media end at 83 minutes, 13 seconds. How much of the disc you can use beyond the nameplate 80-minute capacity depends on your system and your application. Kodak doesn't warrant anything beyond 80 minutes.

You will find good responses to your first questions here:

- o <http://www.cdrfaq.org/faq03.html#S3-8-2>
- o <http://www.cdrfaq.org/faq03.html#S3-8-3>

## General Product Information

### 1. How do I know what my writer's most recent firmware version is?

Visit your writer manufacturer's website or you can [check](#) at <http://www.ahead.de/en/firmware.htm>.

### 2. How do I know what version of firmware is running on a writer?

You can find the firmware version for your writer by accessing the writer's properties through the Control Panel.

Your computer may have a slightly different procedure, depending on the MACINTOSH or Personal Computer you have. This is a generic procedure, to be used as a guide to finding firmware versions.

1. Click to open My Computer > Control Panel > System.
2. Select the Device Manager tab.
3. Double-click CDROM.
4. Highlight the CD writer and right-click.
5. Select Properties.
6. Select the Settings tab.

*The firmware version is listed here.*

### 3. Does Kodak make a CD-R disc with a three-inch form factor? What about the "business card" style CD-R?

No, Kodak offers only 120 mm. (diameter) CD-R media.

Media is available from other sources in an 80 mm. diameter circular format and a variety of custom shapes. An Internet search will provide numerous sources.

NOTE: Some high speed recorders will not accept these smaller CD-R formats, or will restrict recording to very low speeds because of the lower mass of the disc. They do so because the drive spin motor control circuits are designed for operation with a 120 mm. disc; the lower mass discs cause spin speed control problems at higher data rates.

### 4. Do X-rays in airport scanning units affect Kodak CD-R's?

No, CD-R media is insensitive to X-ray radiation.

CD-R recording is heat driven. The use of a laser for recording allows high power to be delivered to a very small spot. The dye in the recording layer converts the laser light to heat which actually causes the recording to occur. In contrast, the energy in X-ray systems is quite diffuse and causes no significant heating to occur.

### 5. What is the difference between the human readable code and the bar code on an Ultima CD-R?

A unique 12-digit human-readable code is printed on each KODAK CD-R Ultima disc. On discs with bar code, this number is laser-scribed onto the disc in a bar code format so that the number is machine-readable as well. It uses an interleaved 2 of 5 type form barcode. The PCD series CD-R recorders sold by Kodak are capable of reading this

barcode. This product line has been discontinued.

6. **What kind of data can I store on CD-R media?**

You can record audio, text, images, graphics and drawings on CD-R media.

7. **Can I erase data written on CD-R media?**

Once recorded on a CD-R disc, data cannot be erased.

8. **What is the difference between the No Brand CD-R media and KODAK's Ultima CD-R media?**

KODAK Ultima media features the INFOGUARD Protection System, which includes one of the most stable dye layers available and a super-tough overcoat to help protect the surface of the KODAK CD-R discs from scratches and fingerprints. The reflective layer of every disc contains real gold to provide superior stability. Kodak has run many accelerated incubation tests on CD-R media from a variety of suppliers (including a number of no brands) that confirm that no other media surpasses KODAK CD-R Ultima media for stability.

KODAK Ultima media's design meets all Compact Disc specifications. Interchange testing involving dozens of varieties of CD recorders and players assured that KODAK CD-R media has the highest level of interchange possible.

These test results show that you can't get the same reliability and longevity from no-brand media.

9. **Where can I get catalog numbers for KODAK Ultima CD-R, KODAK Digital Audio Gold, and KODAK CD-RW, DVD-R, DVD-RAM products?**

Please call the Kodak Digital Imaging Support Center at 1-800-235-6325 and a customer support representative will assist you.

10. **What is the "Orange Book"?**

The 'Orange Book' is the common name for the industry specification document for CD-R media issued by Sony and Philips. Its complete title is "Compact Disc Recordable System Description". In its 100 or so pages all the requirements for CD-R media are spelled out. Only media that meet these requirements should bear the 'CD Recordable' logo. All KODAK CD-R Ultima media is compatible with this specification.

11. **What does INFOGUARD mean?**

INFOGUARD represents a set of features unique to KODAK CD-R media including one of the most stable dye layers available today, and a super-tough durability overcoat to help protect the surface of the media from scratches and fingerprints. In addition, the silver-gold alloy in the reflective layer of KODAK'S CD-R Ultima media provides enhanced stability compared to media that has a reflective layer of pure silver. In a Kodak non-printable product, the InfoGuard Protection System is comprised of a single durability enhancement layer. In the printable products, the InfoGuard Protection System is comprised of two layers: a durability enhancement layer and, coated on top, the printable layer.

12. **What does Ultima mean?**

KODAK Ultima products are those that have a silver-gold alloy as the reflective layer. This silver-gold alloy provides enhanced stability compared to media that has a reflective layer of pure silver.

13. **What does 'KSP' mean?**

KODAK KSP (KODAK Screen Print) products are screen printed with the Kodak logo, the name of the media and lines for the user to write information on.

14. **What does 'NSP' mean?**

NSP (No Screen Print) describes KODAK products that are not screen printed with the Kodak logo, etc.

15. **Is the gold color of your media due to the presence of actual, metallic gold (Au)?**

The reflective layer of every KODAK CD-R disc contains real, elemental gold. In our KODAK CD-R Gold Ultima products and our KODAK CD-R Digital Audio Gold product, the reflective layer is pure gold, 24 karat. These products look gold because the reflective layer is gold.

In KODAK CD-R Ultima products, the reflective layer is a gold/silver alloy. This alloy, by itself, is silver-colored.

Retail KODAK CD-R Ultima products look light gold because the durability overcoat used on them contains a gold tint. This durability overcoat does not contain real gold. However, the discs appear gold because of it.

Commercial KODAK CD-R Ultima products look silver because the durability overcoat used on them is clear (transparent). The discs look silver because the gold/silver alloy used as the reflective layer is silver-colored.

**16. How thick is each layer on a CD-R?**

The reflective layer is typically between 50 and 100 nanometers thick. The same is true for the dye-recording layer. The lacquer protective layer applied to all CD-R media is significantly thicker, usually in the range of 3 to 10 micrometers. Many CD-R manufacturers, including Kodak, apply an additional protective layer to improve resistance to handling damage or to allow thermal or inkjet printing. This layer can be 5 - 20 micrometers thick. For more information about CD-R construction, see:

- o <http://www.cdrfaq.org/>
- o <http://kodak.com/US/en/digital/dlc/book3/chapter6/index.shtml>

**17. If metallic gold is used, approximately how much is coated on each disc?**

The reflective layer on a CD-R disc is pretty thin, on the order of 100 nanometers. That's thinner than a wavelength of light. If you laid 10,000 of these layers on top of each other, you'd have less than a millimeter thick layer of gold (for our KODAK CD-R Gold Ultima products). The amount of gold in a Gold Ultima disc reflective layer is approximately  $75 \text{ nm} \times 100 \text{ cm}^2 \times 19.3 \text{ g/cc} = 14 \text{ mg}$ . In KODAK CD-R Ultima products, the reflective layer is a gold/silver alloy with a gold content of approximately 5%. Therefore, the amount of gold in an Ultima disc is <1 mg.

**18. How is the metal, gold or not, coated onto the media?**

Both the gold and the gold/silver alloy are applied using a vacuum vapor deposition process referred to as sputtering.

**19. What dye is used on KODAK CD-R media?**

Kodak uses their version of phthalocyanine dye in all their general purpose and hybrid media. Kodak uses a metal-stabilized cyanine dye in their audio media to enhance low-speed recording performance.

**20. Do KODAK CD-R Gold Ultima and KODAK CD-R Ultima media use the same dye?**

A phthalocyanine dye is used in all KODAK general purpose and hybrid media.

**21. Are the materials used to manufacture CD-Rs hazardous? Are they recycled?**

Kodak is very concerned about the environment and about the impact our products have on it. Our CD-R discs are manufactured under strict guidelines to minimize environmental impact. There is very little waste generated in CD-R manufacture as most materials are recycled and reclaimed. A CD-R disc is not considered a hazardous waste; by far the largest component is the substrate, which is made from polycarbonate, a rather universal material. Any Gold Ultima product that is returned from distribution channels to Kodak in Rochester is sent out for gold reclamation. Other products that are returned through distribution channels are sold at low cost for use as fuel.

**22. Does Kodak sell its rejected CDs?**

Kodak does not sell its 'rejected' CDs. Kodak prefers to keep rejected CD media out of any type of circulation.

**23. What is the warranty on your CD-R products?**

Kodak offers a lifetime warranty on its INFOGUARD products. If any disc is found to be defective in manufacture or packaging, it will be replaced.

**24. What is the declared Block Error Rate (BLER) for KODAK CD-R media?**

BLER is not just a media property. BLER depends on the speed and condition of the recorder and on the test system used to measure it. For this reason, Kodak does not quote a BLER specification for its media. Kodak bench marked its CD-R media against the major industry brands on multiple occasions. These studies were conducted using the same recording conditions and test equipment for each brand. In these tests, the

BLER of Kodak media ranked among the leaders.

25. **Does each CD-R disc have a unique serial number that could be used with programming to prevent duplication?**

Conventional CD-R discs do not have unique serial numbers recorded on them. Some discs have unique identification numbers printed on them and/or barcoded on them. However, a standard reader or recorder cannot read these numbers and so they are not useful for copy protection.

---

## Packaging

1. **What packaging options are available for bulk media?**

Kodak commercial bulk packaging configuration consists of 100 discs stacked with a blank substrate at each end, shrink-wrapped. This configuration allows easy loading of writer system spindle feeds. Case quantity is 4x100.

Kodak retail bulk packaging configurations consists of 25, 50, or 100 discs stacked on a spindle with endcaps, shrink-wrapped. Case quantities are 12x25, 6x50, 4x100.

2. **Are individual bulk units (i.e., 1x25 or 1x100 pack) shippable?**

Although cases are shippable, individual bulk units are not shippable. If cases are broken up into individual units, they must be carefully and thoroughly overpacked to prevent damage during shipment. Kodak only warrants media shipped in its original cases, packaging and case quantities.

3. **Does Kodak recycle or reuse beehives (50 count bulk containers)?**

No, Kodak has no program in place to accept back and recycle or reuse CD-R beehives.

The beehive bases, covers and nuts all have the universal recycling code/icon of HDPE (high-density polyethylene) molded onto them. This material is widely accepted in city/town recycling programs.

---

## Recording Speed

1. **What does 'recording speed' mean?**

Recording speed is used to calculate how long it will take to write your data. For example, it takes 74 minutes to write 650 MB of data at a 1x recording speed. But, if the recording speed of your drive is 12x, you can write 650 MB of data in about 6 minutes. If you have half the data, it will finish in (about) half the time. You have to add a minute or two to 'finalize' your disc as well.

2. **What recording speeds is your media compatible with?**

You can currently record at speeds of 1X to 12X to create discs that comply with the following standards: CD-DA, CD-ROM, CD-ROM XA and CD-I. Kodak most recently released CD-R Ultima 80 product that is 24X compatible.

3. **Was KODAK CD-R Ultima media certified for 12X recording?**

Yes, when writer manufacturers develop higher speed drives, Kodak sends media to them for certification. The following web sites illustrate Kodak's 12X certification by a few major writer manufacturers for your reference.

- o [http://www.plextor.com/english/support/support\\_compat\\_media.html](http://www.plextor.com/english/support/support_compat_media.html)  
*Media compatibility chart with recording speed certifications by Plextor*
- o <http://www.sannet.ne.jp/burn-proof/media/supportmedia-bp2.html>  
*(Sanyo certifications with Kodak CD-R Ultima specifically cited)*

4. **Why does some of the older KODAK CD-R Ultima media packaging state the media only compatible with 1X to 8X recording?**

Kodak made no changes to its media to make it 12X compatible.

Before 12X writers were developed, the packaging reflected our media's compatibility with all recording speeds available at the time.

Kodak chose to deplete its supply of 1X-8X packaging before printing new packaging even after being certified as 12X compatible.

**5. Can I record KODAK CD-R Ultima 80 at 24X?**

The manufacturing plants began producing 24X-compatible KODAK Ultima 80 media during the latter part of 2001. Because there was a significant quantity of 1X-12X packaging materials in inventory, the new 24X media was shipped out labeled as 12X. Eventually, the old packaging materials were used up and new, 24X-marked packaging was used.

If you own some KODAK Ultima 80 media in 12X packaging, you can run a program like CDR Identifier to read the ATIP code molded into each blank CD-R. The distinguishing feature between Kodak's 12X and 24X-compatible media is the ATIP start-of-lead-in. Kodak's 1X-12X media had a start-of-lead-in code of 97m27s45f. Kodak's 1X-24X media has a start-of-lead-in code of 97m27s46f. The final and most recent version of CDR Identifier can be downloaded as freeware from: <http://www.gum.de/it/download>

\* Kodak makes no representation with respect to the freedom to use this product with inkjet printers in the U.S.A.

**Kodak and InfoGuard are trademarks of Eastman Kodak Company.**

**Frequently Asked Questions provide information of limited or specific application. Responsibility for judging the applicability of the information for a specific use rests with the end user.**

**FAQ1630**

---



[Home](#) | [Privacy \(Sept 2005\)](#) | [Website Terms of Use](#)

[Home](#) | [Testing](#) | [Seminars](#) | [FAQs](#) | [Contact](#) | [About Us](#) | [Site Map](#)

[Services](#) | [Certification](#) | [Publications](#) | [Standards](#) | [Links](#) | [Freeware](#)



# Frequently Asked Questions About Compact Discs



## Interchangeable Media for Computer Mass Storage

- DVD and CD Optical Discs • Diskettes •
- Quality Testing • Training • Research • Product Certification •

**Please contact Media Sciences if your questions are not answered on these pages.**

[How to Contact Media Sciences](#)

[Return to the FAQ Index](#)

## What do the tests referenced in CD standards really mean?

[CD standards](#) exist for the purpose of assuring interchange. This requires that every disc is readable in all drives, except those that are defective. Many specifications are physical, such as outer diameter, inner diameter, thickness, weight, [unbalance](#), eccentricity, deviation, and deflection. Other requirements are optical such as index of refraction and birefringence. Measurements of these are either manual, or require expensive dedicated equipment, and are not discussed further in this FAQ.

Affordable, computerized [equipment](#) is available that evaluates many electrical requirements of CD standards by evaluating intensity variations of the laser beam that is reflected from the information layer of all CD discs. One beam that carries analog information is split both optically and electrically into various paths in the read drive. Standards assure interchange by specifying performance in each path.

The first path is total reflected beam intensity. Standards require minimum reflectance from a mirror region of the disc to assure sufficient signal strength for all paths. CD-R and CD-RW discs have an additional requirement,  $R_{top}$ , that measures on-track total beam intensity.  $R_{top}$  is lower than reflectivity because of diffraction loss from the pre-groove. Cross talk measures the ratio of off-track to on-track beam intensity, ensuring that signals from adjacent tracks do not interfere with data from the desired track.

The second drive path is radial tracking. This servo loop uses a radial error signal to center the focussed laser beam onto the track. Radial tracking, or push-pull, evaluates the sensitivity of the error signal to radial position, and must fall between upper and lower limits to assure proper servo loop

operation. Radial eccentricity evaluates the radial runout of the tracks, assuring that variations remain within the range of the servo loop. Radial noise measures variations in radial track location that are at frequencies higher than the upper cutoff of the radial servo. Radial acceleration limits sudden radial track jumps to values within the capabilities of the electro-mechanical servo.

The third drive path is read data. Peak-to-peak signal strengths must be within limits at both the lowest, I11, and highest, I3, data frequencies. This analog information is then converted into binary data utilizing intervals between times when the analog signal crosses a d.c. decision level. Asymmetry assures that this decision level is within acceptable bounds. Effect length, or length deviation, verifies that averages of eighteen different time intervals are within tolerance. [Jitter](#) confirms that random variations of each time interval are not excessive. Out-of-tolerance read data usually results in read errors.

Very high data densities, noise, and physical defects can generate hundreds-of-thousands of read errors from even the best discs. Such errors are routinely detected and corrected using a Cross-Interleave Reed Solomon code (CIRC.) Before recording, data is organized into frames, each containing sync, subcode, 24 data bytes, and eight parity bytes. Four Q parity bytes are appended at the C2 level, frames are interleaved, then four P parity bytes are appended at the C1 level, and frames are again interleaved, but with a different pattern. C1 and C2 can each detect and correct two erroneous bytes in one frame. Various tests confirm that all errors are readily correctable.

De-interleaved read data is first sent to the C1 decoder for error detection and correction. [BLER](#) measures the rate (frames per second) of frames arriving with one or more errors. E11 represents the rate of correctable C1 frames having exactly one erroneous byte, E21 is the rate for frames with two bad bytes, and E31 is the rate for uncorrectable C1 frames having three or more errors. Good discs would have moderate E11 rates, low E21 rates, and very low E31 error rates.

Frames leaving the C1 decoder are de-interleaved again, distributing concentrated errors over many other error-free frames, and then go to the C2 decoder. E21 is the rate for correctable C2 frames having exactly one erroneous byte, while E22 and E32 are rates of uncorrectable frames having two and three-or-more errors respectively. E22 and E32 rates must be zero.

Burst is a different error test for scratches or other tangential defects. Standards require that no more than six successive C1 frames can each contain two or more erroneous bytes.

Forgiving read drives may tolerate an out-of-spec disc, but other drives will not. Discs that fail radial parameter tests can have acceptable error rates in one drive, but unacceptable rates in another because of servo loop differences. Drives may or may not be able to detect and correct E22 and E32 errors, but standards forbid them. Some players will ignore high radial acceleration while others will not. Every one of the above requirements must be confirmed by comprehensive [testing](#) in order to achieve confidence in predictable interchange.

If it would help, Media Sciences will test one recorded sample at no charge. Please follow the [free test](#) instructions on our web site.

[Return to Top of Page](#)

---

## Media Sciences, Inc. — Dedicated to Quality



[Home](#) | [Testing](#) | [Seminars](#) | [FAQs](#) | [Contact](#) | [About Us](#) | [Site Map](#)  
[Services](#) | [Certification](#) | [Publications](#) | [Standards](#) | [Links](#) | [Freeware](#)



# ISO Standards



## Interchangeable Media for Computer Mass Storage

- DVD and CD Optical Discs • Diskettes •
- Quality Testing • Training • Research • Product Certification •

[How to Contact Media Sciences](#)

## Information About ISO Standards

[ISO Standards:](#) Standards for interchangeable computer media.

[How to Obtain ISO Standards.](#)

[Tips](#)

## Helpful Topics

[Home Page:](#) What's new, quality alerts.

[Quality Testing:](#) Obtain test results for indicators that predict successful interchange.

[Training Seminars:](#) Learn fundamentals and the newest techniques from experienced instructors.

[FAQs:](#) Answers to frequently asked questions about interchangeable computer media.

[How to Contact Media Sciences:](#) by phone, fax, e-mail, Internet, or mail.

[About Media Sciences:](#) History and personnel.

[Site Map:](#) Guide to all resources on this web site.

[Other Services:](#) Personalized support, specs, reference and calibration discs, and a torque kit.

[Product Certification:](#) Vendor certification based on ISO Standards.

[Publications:](#) A Quality Tips book, papers, and a newsletter offer timely information.

[ISO Standards:](#) Listings and sources of international Standards for interchange.

[Links for Professionals:](#) Other web sites containing useful quality and technology information.

[Freeware and Shareware for Computer Media:](#) Conduct quality tests using these programs.

---

## ISO STANDARDS

Media interchange and quality management are supported by the work of the International Organization for Standardization (ISO) and its member organizations such as the American National

Standards Institute (ANSI), the Association française de normalisation (AFNOR), the British Standards Institution (BSI), the Deutsches Institut für Normung (DIN), the European Computer Manufacturers Organization (ECMA), the Japanese Industrial Standards Committee (JISC), and the Standards Council of Canada (SCC). The following published Standards detail the requirements necessary for interchange and minimum acceptable quality. Equivalent ECMA Standards are listed in parenthesis, and can be downloaded from the ECMA website.

## ICS Code 35.220: Information Technology; Office Machines: Data Storage Devices

### DISKETTES ICS Code 35.220.21

- ISO 5654/1:1984, 200 mm (8"), 1,9 tpmm (48 tpi), one-sided, 13 262 ftpr (SS/SD); ISO 5652/2:1984 Formatted (ECMA-54)
- ISO 7065/1:1985, 200 mm (8"), 1,9 tpmm (48 tpi), two-sided, 13 262 ftpr, (DS/DD); ISO 7065/2:1985 Formatted (ECMA-69)
- ISO 6596/1:1985, 130 mm (5¼"), 1,9 tpmm (48 tpi), one-sided, 7 958 ftpr, (SS/SD); ISO 6596/2:1985 Formatted (ECMA-66)
- ISO 7487/1:1986, 130 mm (5¼"), 1,9 tpmm (48 tpi), two-sided, 7 958 ftpr, (DS/DD) (MD-2DD); ISO 7487/2:1986 Track Format A; ISO 7487/3:1986 Track Format B (IBM) (ECMA-70)
- ISO 8378/1:1986, 130 mm (5¼"), 3,8 tpmm (96 tpi), two-sided, 7 958 ftpr, (DS/QD); ISO 8378/2:1986 Track Format A; ISO 8378/3:1986 Track Format B (IBM) (ECMA-78)
- ISO 8630/1:1987, 130 mm (5¼"), 3,8 tpmm (96 tpi), two-sided, 13 262 ftpr, (DS/HD) (MD-2HD); ISO 8630/2:1987 Track Format A (77 Tracks); ISO 8630/3:1987 Track Format B (IBM) (ECMA-99)
- ISO 8860/1:1987, 90 mm (3½"), 5,3 tpmm (135 tpi), two-sided, 7 958 ftpr, (DS/DD) (MF-2DD); ISO 8860/2:1987 (ECMA-100)
- ISO 9529/1:1989, 90 mm (3½"), 5,3 tpmm (135 tpi), two-sided, 15 916 ftpr, (DS/HD) (MF-2HD); ISO 9529/2:1989 (ECMA-125)
- ISO 10994/1:1992, 90 mm (3½"), 5,3 tpmm (135 tpi), two-sided, 31 831 ftpr, (DS/ED) (MF-2ED); ISO 10994/2:1992 (ECMA-147)

### OPTICAL DISKS ICS Code 35.220.30

- IEC 908:1987, Compact Disc Digital Audio System (CD-DA)
- ISO 9660:1988, Volume and File Structure (CD-ROM) (ECMA-119)
- ISO/IEC 10149:1995, Read-Only 120 mm Optical Data Disks (CD-ROM) (ECMA-130)
- ISO/IEC 13346, Recordable/Rewritable Volume and File Structure (ECMA-167)
- ISO/IEC 16448, 120 mm DVD - Read-Only Optical Disks (ECMA-267)
- ISO/IEC 16449, 80 mm DVD - Read-Only Optical Disks (ECMA-268)
- ISO/IEC 16824, DVD-RAM - Rewritable Optical Disks, 2.6 GB (ECMA-272)
- ISO/IEC 16825, DVD-RAM Case - DVD-RAM Cartridge (ECMA-273)
- ISO/IEC 16969, DVD+RW - Rewritable Optical Disks, 3 GB (ECMA-274)
- ISO/IEC 17341, DVD+RW - Rewritable Optical Disks, 4.7 GB (ECMA-337)

- ISO/IEC 17342, DVD-RW - Rewritable Optical Disks, 4.7 GB (ECMA-338)
- ISO/IEC 17344, DVD+R - Recordable Optical Disks, 4.7 GB, 8X (ECMA-349)
- ISO/IEC 17592, DVD-RAM - Rewritable Optical Disks, 4.7 GB (ECMA-330)
- ISO/IEC 17594, DVD-RAM Case - DVD-RAM Cartridge (ECMA-331)
- ISO/IEC 20563, DVD-R - Recordable Optical Disks, 3.95 GB (ECMA-279)
- ISO/IEC 23912, DVD-R - Recordable Optical Disks, 4.7 GB (ECMA-359)
- DVD+R DL - Recordable Optical Disks, 4.7 GB (ECMA-364)
- ISO/IEC DTR 18002 - DVD File System Specifications
- ECMA Technical Report TR/71 - DVD Read-Only Disk File System Specifications
- OSTA Universal Disk Format Specifications (UDF)
- ANSI/NAPM IT9.21-1996 - Life Expectancy of Compact Discs (CD-ROM)-Method for Estimating Based on Effects of Temperature and Relative Humidity

## **HELICAL SCAN MAGNETIC TAPE CARTRIDGE** ICS Code 35.220.22

- ISO/IEC 10777:1991, 3,81 mm Wide, DDS Format (ECMA-139)
- ISO/IEC 11319:1991, 8 mm Wide, Single Azimuth Format (ECMA-145)
- ISO/IEC 11321:1992, 3,81 mm Wide, DATA/DAT Format (ECMA-146)
- ISO/IEC 11557:1992, 3,81 mm Wide, DDS-DC Format (ECMA-150)
- ISO/IEC 11558:1992, 3,81 mm Wide, DCLZ Format
- ISO/IEC 12246:1993, 8 mm Wide, Dual Azimuth Format (ECMA-169)
- ISO/IEC 12447:1993, 3,81 mm Wide, 60 m and 90 m DDS Format (ECMA-170)
- ISO/IEC 12448:1993, 3,81 mm Wide, 60 m and 90 m DATA/DAT Format (ECMA-171)
- ISO/IEC 13923, 3,81 mm Wide, 120 m DDS-2 Format (ECMA-198)
- ISO/IEC 15521, 3,81 mm Wide, 125 m DDS-3 Format (ECMA-236)

*Link to ISO Site Listing International Standards* <http://www.iso.org/>

*Link to ISO Standards Catalog* <http://www.iso.org/iso/en/CatalogueListPage.CatalogueList>

*Link to ISO Site Listing Magnetic Disk Storage Standards* <http://www.iso.org/iso/en/CatalogueListPage.CatalogueList?ICS1=35&ICS2=220&ICS3=21>

*Link to ISO Site Listing Optical Storage Standards* <http://www.iso.org/iso/en/CatalogueListPage.CatalogueList?ICS1=35&ICS2=220&ICS3=30>

*Link to Current Standards Activity* <http://www.y-adagio.com/public/>

*Link to IEC Site Listing International Standards* <http://www.iec.ch/>

*Link to ANSI/NSSN Gateway* <http://www.nssn.org/>

*Link to ECMA Site Listing National Standards* <http://www.ecma-international.org/>

*Link to T10 Site Listing SCSI Multimedia Commands (CD and DVD) (Mt. Fuji)* <http://www.t10.org/>

*Link to OSTA Specifications* <http://www.osta.org/specs/>

*Link to Joliet Site* <http://www-plateau.cs.berkeley.edu/people/chaffee/jolspec.html>

*Link to Unicode Consortium Site* <http://www.unicode.org/>

*Link to Rock Ridge Site* <ftp://ftp.yimi.com/pub/rockridge/>

*Link to Apple HFS Sites* [http://developer.apple.com/technotes/fl/fl\\_36.html](http://developer.apple.com/technotes/fl/fl_36.html) and <http://developer.apple.com/technotes/tn/tn1150.html>

also [http://developer.apple.com/techpubs/mac/Devices/Devices\\_119.html](http://developer.apple.com/techpubs/mac/Devices/Devices_119.html) and [http://developer.apple.com/techpubs/mac/Devices/Devices\\_119.html](http://developer.apple.com/techpubs/mac/Devices/Devices_119.html)

[apple.com/techpubs/mac/Files/Files-99.html](http://apple.com/techpubs/mac/Files/Files-99.html)

Link to El Torito Bootable CD Site <http://www.phoenix.com/PlatSS/PDFs/specs-cdrom.pdf>

## PRIVATE STANDARDS FOR OPTICAL DISKS

Link to Philips Color Books <http://www.licensing.philips.com/>

Link to Mt. Rainier (CD-RW) <http://www.licensing.philips.com/information/mtr/>

Link to DVD Books <http://www.dvdforum.org/>

[Return to Top of Page](#)

---

## HOW TO OBTAIN ISO STANDARDS

ISO publications are available from many ISO member organizations. The organization for your area may be obtained from:

ISO Central Secretariat

1, rue de Varembé, Case postale 56

CH-1211 Genève 20, Switzerland

Telephone: +41 22 749 01 11    Telefax: +41 22 733 34 30

Telex: +41 22 05 iso ch    Telegrams: isorganiz

E-mail: [central@iso.org](mailto:central@iso.org)    URL: <http://www.iso.org/>

European Computer Manufacturers Association publications or information may be obtained from:

ECMA

114, Rue du Rhone

CH-1204 Geneva, Switzerland

Telephone: +41 22 849 60 00    Telefax: +41 22 840 60 01

E-mail: [documents@ecma.ch](mailto:documents@ecma.ch)    URL: <http://www.ecma-international.org/>

A comprehensive list of ISO member bodies may be found on the Internet:

URL: <http://www.iso.org/iso/en/aboutiso/isomembers/MemberList.MemberSummary?MEMBERCODE=10>

Providers are also listed on the ANSI/National Standards System Network:

URL: <http://www.nssn.org/>

A partial list follows:

### Canada (SCC)

Standards Council of Canada

45 O'Connor Street, Suite 1200

Ottawa, Ontario K1P 6N7

Telephone: +1 613 238 32 22    Telefax: +1 613 995 45 64

Telex: 053 44 03 stancan ott    Telegrams: stancan ottawa

URL: <http://www.scc.ca/>

### **France (AFNOR)**

Association française de normalisation  
Tour Europe  
F-92049 Paris La Défense Cedex  
Telephone: +33 1 42 91 55 55    Telefax: +33 1 42 91 56 56  
Telex: 61 19 74 afnor f    Telegrams: afnor courbevoie

### **Germany (DIN)**

DIN Deutsches Institut für Normung  
Burggrafenstrasse 6  
D-10787 Berlin  
Telephone: +49 30 26 01-0    Telefax: +49 30 26 01 12 31  
Telex: 18 42 73 din d    Telegrams: deutschnormen berlin  
Internet: postmaster@din.de X.400: c=de; a=d400; p=din; s=postmaster  
URL: <http://www.din.de/>

### **Japan (JISC)**

Japanese Industrial Standards Committee  
c/o Standards Department  
Ministry of International Trade and Industry  
1-3-1, Kasumigaseki, Chiyoda-ku, Tokyo 100  
Telephone: +81 3 35 01 92 95    Telefax: +81 3 35 80 14 18  
Telex: 02 42 42 45 jsatyo j    Telegrams: mitijisc tokyo  
URL: <http://www.hike.te.chiba-u.ac.jp/ikeda/JIS/index.html>

### **United Kingdom (BSI)**

British Standards Institution  
389 Chiswick High Road  
GB-London W4 4AL  
Telephone: +44 181 996 90 00    Telefax: +44 181 996 74 00

### **USA (ANSI)**

American National Standards Institute  
11 West 42nd Street, 13th Floor  
New York, NY 10036 U.S.A.  
Telephone: +1 212 642 49 00    Telefax: +1 212 398 00 23  
Telex: 42 42 96 ansi ui    Telegrams: standards, new york  
URL: <http://www.ansi.org/>

### **ISO documents can be ordered from:**

Global Engineering Documents  
15 Inverness Way East

Englewood, CO 80112 U.S.A.

Telephone: +1 800 854 7179    Canada: +1 800 387 4408

Telephone: +1 303 790 0600    Telefax: +1 303 397 2740

International: Telephone: +1 303 397 7956

Telefax: +1 303 397 7935 or +1 303 397 2740

URL: <http://www.ihs.com/>

Document Center Inc.

111 Industrial Road Suite 9

Belmont, CA 94002 U.S.A.

Telephone: +1 650 591 7600    Telefax: +1 650 591 7617

URL: <http://www.document-center.com/>

[Return to Top of Page](#)

---

**Media Sciences, Inc. — Dedicated to Quality**



[Home](#) | [Testing](#) | [Seminars](#) | [FAQs](#) | [Contact](#) | [About Us](#) | [Site Map](#)

[Services](#) | [Certification](#) | [Publications](#) | [Standards](#) | **Links** | [Freeware](#)



## Links for Professionals



### Interchangeable Media for Computer Mass Storage

- DVD and CD Optical Discs • Diskettes •
- Quality Testing • Training • Research • Product Certification •

[How to Contact Media Sciences](#)

## Links for Professionals

[Quality](#): Links to sites containing quality information.

[News and Technology](#): Links to sites containing industry news and technology updates.

[Technical Support](#): Links to sites providing helpful information.

[Test Equipment Suppliers](#): Links to sites of test equipment vendors.

[CD-R Vendors](#): Links to sites of disc, software, and recorder manufacturers.

[Please notify us](#) of new, helpful sites that we can post.

### [Tips](#)

## Helpful Topics

[Home Page](#): What's new, quality alerts.

[Quality Testing](#): Obtain test results for indicators that predict successful interchange.

[Training Seminars](#): Learn fundamentals and the newest techniques from experienced instructors.

[FAQs](#): Answers to frequently asked questions about interchangeable computer media.

[How to Contact Media Sciences](#): by phone, fax, e-mail, Internet, or mail.

[About Media Sciences](#): History and personnel.

[Site Map](#): Guide to all resources on this web site.

[Other Services](#): Personalized support, specs, reference and calibration discs, and a torque kit.

[Product Certification](#): Vendor certification based on ISO Standards.

[Publications](#): A Quality Tips book, papers, and a newsletter offer timely information.

[ISO Standards](#): Listings and sources of international Standards for interchange.

**Links for Professionals**: Other web sites containing useful quality and technology information.

[Freeware and Shareware for Computer Media](#): Conduct quality tests using these programs.

## Links to Quality Sites

*American Society for Quality* <http://www.asq.org/>

*ANSI/National Standards System Network* <http://www.nssn.org/>

*Document Center Inc.* <http://www.document-center.com/>

*DVD Forum* <http://www.dvdforum.org/>

*European Computer Manufacturers Association, ECMA* <http://www.ecma-international.org/>

*IHS Engineering (Documents)* <http://www.ihs.com/>

*International Electrotechnical Commission, Geneva, IEC* <http://www.iec.ch/>

*International Organization for Standardization, Geneva, ISO* <http://www.iso.org/>

*ISO 9000 Guide* <http://www.isoeasy.org/>

*ISO 17025 (Was Guide 25)* <http://www.fasor.com/iso25/>

*Philips Licensing & Standards* <http://www.licensing.philips.com/>

*Quality Resource Guide* <http://deming.eng.clemson.edu/onlineq.html>

*Registrar Accreditation Board, RAB* <http://www.rabnet.com/>

[Return to Top of Page](#)

---

## Links to News and Technology Sites

*Blu-ray Disc Association* <http://www.bluraydisc.com/>

*Blu-ray Disc License* <http://www.blu-raydisc.info/>

*CD Freaks* <http://www.cdfreaks.com/>

*CD Information Center* <http://www.cd-info.com/>

*CD Media World* <http://www.cdmediaworld.com/hardware/cdrom/cd.shtml>

*CDR Info* <http://www.cdrinfo.com/>

*Compact Disc Consulting (Dana Parker, Robert Starrett)* <http://www.cdpage.com/>

*DualDisc* <http://www.dualdisc.com/>

*DVD Association* <http://www.dvda.org/>

*DVD Insider* <http://www.dvdinsider.com/>

*DVDplusRW.org* <http://www.dvdplusrw.org/>

*DVD-RAM Central* <http://www.ramprg.com>

*DVD+RW Alliance* <http://www.dvdrw.com/>

*Emedia Professional Magazine* <http://www.emedialive.com/>

*Freeman Data Storage Markets* <http://www.freemaninc.com/>

*HD DVD* <http://www.hddvd.org/>

*HD DVD Promotion Group* <http://www.hddvdprg.com/>

*IFPI, International Federation of the Phonographic Industry (anti-piracy, SID)* <http://www.ifpi.org/>

*International Disc Duplication Association* <http://www.discdupe.org/>

*medialine* <http://www.medialinenews.com/>

*National Institute of Standards and Technolgy, NIST* <http://www.nist.gov/>

*One to One Magazine* <http://www.oto-online.com/>

*Optical Disc Systems Magazine* <http://www.opticaldisc-systems.com/>

*Optical Storage Technology Association* <http://www.osta.org/>

*ProNET Webguide* <http://www.pronetguide.com/>

*Recordable DVD Council* <http://www.rdvdc.org/english/>

*Recording Industry Association of America* <http://www.riaa.com/>

*RLG Digital Image Preservation* <http://www.rlg.org/>

*Webcom Communications (Software Business Magazine)* <http://www.infowebcom.com/>

[Return to Top of Page](#)

## Links to Technical Support Sites

*advanced\_cdr Discussion Group* [http://groups.yahoo.com/group/advanced\\_cdr/](http://groups.yahoo.com/group/advanced_cdr/)

*Andy McFadden's CD-Recordable FAQ* <http://www.cdrfaq.org/>

*Association for Recorded Sound Collections Discussion List* <http://www.arsc-audio.org/arsclist.html>

*New host: e-mail (no subject) INFO REFCARD command to LISTSERV@LISTSERV.LOC.GOV.*

*Audio Visual Media Discussion Group* <http://lists.topica.com/lists/AV-Media-Matters/>

*Bootable CD for WinNT4/2000/XP* <http://bink.nu/Bootcd/default.htm>

*CD and DVD Technology (Deluxe Global Media)* <http://www.disctronics.co.uk/technology/index.htm>

*CD Freaks Recording Forum* <http://club.cdfreaks.com/>

*CD Related Links* <http://www.fokus.gmd.de/research/cc/glone/employees/joerg.schilling/private/cdr.html>

*CD-R/CD-RW Technical Information* <http://www.5starsupport.com/info/cdrinfo.htm>

*Disctronics CD & DVD Technology* <http://www.disctronics.co.uk/technology/>

*DMI Introduction to ISO 9660 for DOS, MacIntosh, UNIX* <http://www.mp3ar.com/Literature/iso9660.pdf>

*DVD List Discussion Group and Archive* <http://www.tully.com/DVDList/>

*DVD FAQ* <http://www.dvddemystified.com/dvdfaq.html>

*DVD Technical Guide (Pioneer)* <http://www.pioneer.co.jp/crdl/tech/index-e.html>

*DVD Workshop* <http://stream.uen.org/medsol/dvd/home.html#2>

*Internet FAQ Archives-cdrom* <http://www.landfield.com/faqs/cdrom/>

*ISO9660 Simplified for DOS/Windows* <http://alumnus.caltech.edu/~pje/iso9660.html>

*Mac-PC Data Transfer Solutions* <http://www.macdisk.com/>

Microsoft Technet and Knowledgebase <http://www.microsoft.com/technet/>

Microsoft Bootable CD-ROM <http://support.microsoft.com/default.aspx?scid=kb;EN-US;q167685>

Mike Richter's Resources for CD-R <http://www.mrichter.com/>

NIST's "Care and Handling Guide for the Preservation of CDs and DVDs" <http://www.itl.nist.gov/div895/carefordisc/index.html>

Official DVD FAQ by Jim Taylor <http://www.thedigitalbits.com/officialfaq.html>

OSTA's Understanding CD-R & CD-RW <http://www.osta.org/technology/>

OSTA's OpticalU <http://www.opticalu.org/>

PCGuide-CD <http://www.pcguides.com/ref/cd/>

Plextor University [http://www.plextor.com/english/news/news\\_university.html](http://www.plextor.com/english/news/news_university.html)

Roxio (formerly Adaptec) <http://www.roxio.com>

Roxio Discussion List and Archives <http://www.roxio.com/en/interest/community/>

Professional Multimedia Test Centre (Compatibility & Functional Testing) <http://www.pmtctest.com/>

Storage Resource Cornucopia (SCSI, IDE) <http://www.bswd.com/cornucop.htm>

Training Resources <http://www.customguide.com/search/index.htm>

UDF White Paper [http://www.softarch.com/us/dvd/UDF\\_whitepaper.pdf](http://www.softarch.com/us/dvd/UDF_whitepaper.pdf)

Univ. of Washington CD Lecture <http://www.ee.washington.edu/conselec/CE/kuhn/doi96/dhome.htm>

VideoHelp <http://www.videohelp.com/>

[Return to Top of Page](#)

---

## **Links to Test Equipment Sites**

Adivan High Tech <http://www.adivan.com/>

AudioDev <http://www.audiodev.com/>

Basler <http://www.baslerweb.com/>

Clover Systems <http://www.cloversystems.com/>

DaTARIUS (formerly kdg/Koch) <http://www.datarius.com>

Doug Carson Associates <http://www.dcainc.com/>

Dr. Schenk <http://www.drshenk.com/>

dr. schwab <http://www.schwabinspection.com/>

Eclipse <http://www.eclipsedata.com/>

Expert Magnetics <http://www.expertmg.co.jp/>

Infinadyne (formerly Arrowkey) <http://www.infinadyne.com/>

Pulstec <http://www.pulstec.co.jp/Epulstec/>

Quantized Systems <http://www.quantized.com/>

Sony Precision Technology <http://www.sonypt.com/>

Stagetech <http://www.stagetech.se/>

[Return to Top of Page](#)

---

## **Links to CD and DVD Vendors**

ahead software (Nero) <http://www.ahead.de/>

Adaptec <http://www.adaptec.com/>

BenQ <http://www.benq.com/>

CD Recordable Database [http://www.instantinfo.de/index\\_cdrohlinge\\_e.php/](http://www.instantinfo.de/index_cdrohlinge_e.php/)

CD Stomper (Labels) <http://www.cdstomper.com/>

Discmatic <http://www.discmatic.com/>

Gear Software <http://www.gearsoftware.com/>

Golden Hawk Technology (CDRWIN) <http://www.goldenhawk.com/>

*Hewlett Packard* <http://products.hp-at-home.com/home/home.php>

*HHB* <http://www.hhb.co.uk/>

*HyCD* <http://www.hycd.com/>

*Imation Enterprises* <http://www.imation.com/>

*Infinadyne (formerly Arrowkey)* <http://www.infinadyne.com/>

*LG Electronics* <http://www.lge.com/>

<http://us.lge.com/>

<http://us.lgservice.com/>

*Lite-On* <http://www.liteon.com/>

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

*Maxell* <http://www.maxell.com/>

*Memorex* <http://www.ememorex.com/>

*Mitsui Toatsu Chemicals* <http://www.mitsuigold.com/>

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

*Mitsumi* <http://www.mitsumi.com/>

*Neato (Labels)* <http://www.neato.com/>

*NewTech Infosystems* <http://www.ntius.com/>

*NuTech* <http://www.nu-global.com/>

*Optical Storage Directory* [http://www.instantinfo.de/links\\_e.php](http://www.instantinfo.de/links_e.php)

*Orange (Book) Forum* <http://www.orangeforum.or.jp/>

*Panasonic* <http://www.panasonic.com/>

*Philips* <http://www.philips.com/>

*Pinnacle Systems* <http://www.pinnaclesys.com/>

*Pioneer Electronic* <http://www.pioneer.co.jp/>

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

*Plasmon* <http://www.plasmon.com/>

<http://www.plasmon.co.uk/pdsl/>

*Plextor* <http://www.plextor.com/>

*Princo* <http://www.princo.com.tw/>

*Ricoh* <http://www.ricoh.com/>  
<http://www.ricoh.com/drive/index.html>

*Roxio* <http://www.roxio.com>

*Samsung* <http://www.samsungusa.com/>

*Sanyo (BURN-Proof)* <http://www.digital-sanyo.com/BURN-Proof/>  
<http://www.burn-proof.com/>  
[http://www.digital-sanyo.com/main\\_e.html](http://www.digital-sanyo.com/main_e.html)  
<http://www.sannet.ne.jp/>

*Sonic Solutions* <http://www.sonic.com/>

*Sony* <http://www.sonychemicals.com/>  
<http://www.sony.com/>

*Taiyo Yuden* <http://www.yuden.co.jp/>  
<http://www.t-yuden.com/>

*TDK* <http://www.tdk.com/recmedia/>

*TEAC* <http://www.teac.com/>

*Toshiba* <http://www.toshiba.com/>

*ULead* <http://www.ulead.com/>

*Verbatim* <http://www.verbatimcorp.com/>

[Return to Top of Page](#)

---

**Media Sciences, Inc. — Dedicated to Quality**



# Measures of CD-R Longevity

**Jerome L. Hartke, Media Sciences, Inc.**

**July 17, 2001**

## **Contents of this Document**

[Introduction:](#) Background of this study.

[Objective:](#) Outline of purpose and approach.

[Degradation Modes:](#) Longevity limitations.

[Test Methods:](#) Approach to destructive aging.

[Test Results:](#) Observed defects for parameters and errors.

[End-of-Life Indicators:](#) Measures of longevity.

[Conclusions:](#) Findings of this CD-R longevity study.

[Return to Home Page](#)

## **INTRODUCTION**

Each media type has distinctive capabilities and limitations, including environmental instability. CD-R has become popular for mass storage because of its high capacity, low cost, reliable interchange, and a large installed base of CD-ROM and CD-R drives. Interchange is not assured by readability in one or a few drives. Interchange requires that every disc must be readable in every drive, and is achieved by conformance to ISO and Philips Standards. Comparable standards do not exist for longevity that is defined as satisfactory interchange over some unspecified period of time.

In the absence of an industry standard, CD-R longevity is defined by user requirements. Application dependent expectations and murky vendor claims lead to uncertainty, especially when media is used for long term archiving of valuable information. Longevity can be limited both by media deterioration and by technological obsolescence. Product support cycles are typically 5-10 years, while computer system use rarely exceeds 20 years. Information transfer to upgraded platforms may therefore occur every 10-20 years, requiring significantly longer media longevity.

CD-R discs and drives are technically sophisticated. Users may not fully understand complex requirements that must be satisfied in order to attain interchange and longevity. Because these products are sold as commodities, users may find that performance claims originated by marketing departments may not be supported by accurate test results.

[Return to Top of Page](#)

## **OBJECTIVE**

Standards specify methods and limits for a broad scope of tests that confirm interchange. The purpose of this independent study by Media Sciences was to distinguish those tests that can accurately predict media end-of-life. Electrical parameter tests are important because they directly evaluate media compatibility with read drive servos. Soft errors and unpredictable read failure can result from parameters that are outside of tolerance limits. Electrical error tests measure correctable error rates and detect uncorrectable errors caused by local defects. Visual and mechanical inspections discover flaws that may not be detected by high quality test drives.

Evaluations were conducted using destructive test methods that intentionally induced failures. This approach identified end-of-life tests that are important to longevity, and isolated other tests that might produce misleading results. This study did not attempt to evaluate the longevity of various dye types or brands, since these are changing issues where manufacturers must be responsible for claims based upon accurate, objective test results on their products.

[Return to Top of Page](#)

## **DEGRADATION MODES**

Exposure to radiation, inks, other chemicals, water, or pollutants can adversely affect CD-R quality. Although environmental degradation should be avoided, a major cause of deterioration is improper user handling based upon overconfidence in the robust construction and error correction of CD-R media. Handling or storage conditions may degrade good discs, especially if the thin, vulnerable protective coating on the label surface is flawed or damaged. In addition, high quality media that is recorded in poor writers may fail interchange standards. Such issues are outside the scope of this study that is focussed on the longevity of media that has been properly recorded and handled.

CD-R discs normally contain four layers consisting of a pre-grooved substrate, a dye layer, a metal layer, and a protective coating. All but the metal layer contain organic compounds that can degrade as a result of changes in their chemical structures or because of unstable defects that grow in size. Mechanical stress induced by rapid environmental changes may result in excessive differential expansion of the various layers and delamination or excessive birefringence. These and other ageing modes limit media longevity.

CD-R ageing processes can be accelerated by high temperature and humidity environments. Media manufacturers may estimate longevity by evaluating discs aged at elevated temperatures and humidities. Extrapolation to ambient conditions, typically 21 C-23 C and 15%-60% RH, then provide a lifetime for the product. Such methods are valid only when a proper end-of-life criteria is applied. BLER is commonly used for this purpose, although supporting evidence has not been published. Some vendors have used a maximum BLER of 50 per sec. to determine end-of-life. Standards ISO/IEC 10149, ANSI/NAPM IT9.21-1996, and others use a BLER limit of 220 per sec. Draft Standard ISO/DIS 12024 required zero E22 and E32 errors in addition to BLER limits.

[Return to Top of Page](#)

## TEST METHODS

All test methods of Media Sciences were based upon ISO/IEC 10149, and employed test equipments that had been directly correlated to multiple Philips test discs. Discs of various dye types from different manufacturers were tested to ensure that conclusions were representative of current CD-R technology. Ageing was accelerated by storage at elevated temperature and humidity. One set of twenty samples was recorded in a high quality drive prior to storage and testing. Another set of four samples was stored in their unrecorded condition, after which the unrecorded discs were tested, recorded in a high quality drive, and then tested for their recorded properties.

Both pre-recorded and unrecorded sets were tested initially, were loaded into special fixtures, and were then destructively aged by storage for 100 hours at 85 C and 85% relative humidity. Non-condensing ramp-up and ramp-down conditions were maintained in the environmental chamber. After interim testing, the samples were subjected to an additional 100 hours at 85 C and 85% RH, after which final testing was conducted.

[Return to Top of Page](#)

## TEST RESULTS

Longevity results were not the same for all discs. Significant differences were observed between manufacturers and also between samples from the same manufacturer. No clear differences were observed between dye types within the limited sample. Degradation was more severe for discs of very poor initial quality than for high quality samples, indicating that initial recorded quality was important to longevity for multiple reasons.

Identification of specific degradation mechanisms were not studied, and is the responsibility of each manufacturer. Mechanical tests did not disclose any warping or other physical deformation. Visual examination disclosed penetration of label ink into the dye layer of two samples. This caused severe defects and unreadable discs. Further testing was discontinued on these two samples. The protective coating delaminated on two different samples, resulting in uncorrectable errors where the metal layer was exposed. Comprehensive testing was conducted on these samples. The following tables summarize test results of the two sets of samples.

**Pre-Recorded CD-R Disc Test Results After Storage**

Quality Indicator (p-parameter, e- error)	Pass (percent)	Percent Defective		
		Minor Defect	Major Defect	Critical Defect
Reflectance (p)	100	0	0	0
I3/Itop (p)	88	6	0	6
I11/Itop (p)	88	0	6	6
Asymmetry (p)	88	6	0	6

Radial Tracking (p)	94	6	0	0
Radial Noise (p)	100	0	0	0
Radial Acceleration (p)	100	0	0	0
Cross Talk (p)	100	0	0	0
Radial Contrast After (p)	100	0	0	0
Jitter (p)	59	0	0	41
BLER (1 sec. avg.) (e)	53	0	12	35
E22 (e)	18	23	12	47
E32 (e)	53	0	0	47
BURST (e)	35	6	0	59
ALL	6	18	6	70

*Minor defects indicate significant, but marginal flaws. Major defects reflect significant interchange risks. Critical defects are expected to cause interchange failures. Parameters are time averages. Errors evaluate local defect density and size.*

The following changes, **bold** designating serious changes, were observed along with attribute results tabulated above:

Reflectance: 12% of the samples changed less than 1%, **12% decreased by 3-6%, and 76% increased by 3% to 5%.**

I3/Itop: 76% changed less than 0.01, 18% decreased by 0.03 to 0.09, **6% decreased by 0.22.**

I11/Itop: 76% changed less than 0.01, 12% decreased by 0.03 to 0.09, **12% decreased by 0.17 to 0.27.**

Asymmetry: 82% changed less than 1%, 6% decreased by 4%, 6% increased by 4%, **6% increased by 19%.**

Radial Tracking: 59% changed less than 0.02, 12% increased by 0.03, **29% decreased by 0.06.**

Radial Noise: 35% changed less than 1 nm rms, 65% increased by 1 nm rms to 5 nm rms.

Cross Talk: 88% changed less than 0.02, 6% decreased by 0.07, 6% decreased by 0.21.

Radial Contrast After: 82% changed less than 0.02, 18% decreased by 0.05 to 0.12.

Jitter: 53% changed less than 2 ns. **35% increased by 4 ns to 7 ns, 6% increased by 18 ns, 6% increased by 48 ns.**

### Unrecorded CD-R Disc Test Results After Storage

Quality Indicator (p-parameter, e-error)	Pass (percent)	Percent Defective		
		Minor Defect	Major Defect	Critical Defect
<b>Unrecorded Samples After Storage</b>				
Normalized Wobble Amplitude (p)	75	25	0	0

Wobble Carrier-to-Noise (p)	75	25	0	0
ILAND (p)	100	0	0	0
IGROOVE (p)	50	25	25	0
Radial Contrast Before (p)	100	0	0	0
ATIP Error Rate (e)	100	0	0	0
<b>Samples Recorded After Storage</b>				
Reflectance (p)	100	0	0	0
I3/Itop (p)	50	25	25	0
I11/Itop (p)	0	50	50	0
Asymmetry (p)	50	25	25	0
Radial Tracking (p)	100	0	0	0
Radial Noise (p)	100	0	0	0
Radial Acceleration (p)	100	0	0	0
Cross Talk (p)	100	0	0	0
Radial Contrast After (p)	100	0	0	0
Jitter (p)	50	25	0	25
BLER (1 sec. avg.) (e)	75	0	0	25
E22 (e)	0	0	0	100
E32 (e)	25	0	0	75
BURST (e)	0	0	0	100
ALL	0	0	0	100

*Minor defects indicate significant, but marginal flaws. Major defects reflect significant interchange risks. Critical defects are expected to cause interchange failures. Parameters are time averages. Errors evaluate local defect density and size.*

The following changes, **bold** designating serious changes, were observed along with attribute results tabulated above:

Normalized Wobble Amplitude: 75% changed less than 0.002, 25% decreased by 0.003 to 0.005.

Wobble Carrier-to-Noise Ratio: 50% changed less than 1 dB, 50% decreased by 1 dB to 2 dB.

ILAND: 50% changed less than 1%, 50% increased by 1% to 2%.

IGROOVE: 25% changed less than 1%, 25% increased by 1-2%, 25% decreased by 1-2%, **25% decreased by 2% to 4%.**

Radial Contrast Before: 75% changed less than 0.01, **25% increased by 0.06 to 0.08.**

ATIP Error Rate: 100% changed less than 1%.

Even though most unrecorded discs passed appropriate quality tests after storage, E22, E32, and

BURST errors were more severe in samples that were recorded after storage than in discs recorded prior to storage. It was clear that severe changes in properties of unrecorded discs had degraded media quality so as to seriously affect recording quality.

Further interpretations of these test results are inappropriate because of the variety of manufacturers and dye types that were present in this study. Instead, it should be realized that both attribute and variables data are important in evaluating both the interchange and longevity capabilities of CD-R media, and that one, common degradation mode and end-of-life indicator could not be identified.

[Return to Top of Page](#)

## **END-OF-LIFE INDICATORS**

Test data did not identify one universal end-of-life indicator, probably because of the diverse media sources included in this study. BLER was a poor indicator, showing acceptable results for samples that failed other important quality requirements. This is not surprising, since BLER does not distinguish between easily correctable errors and severe uncorrectable errors, and fails to evaluate important parameters that can affect read drive servos. Only 47% of the failed discs had unacceptable BLER values in excess of 220 per sec., 14% of the failed discs had BLER values between 100 and 220, and 14% had BLER values between 50 and 100. BLER values below 50 per sec., usually characteristic of high quality media, were present for 25% of the failed discs.

Increases in mark length as measured by beta, asymmetry, or effect length, have been proposed as an indicator. Only 6% of the samples used in this study showed such changes, but indicated a decrease in mark length. The onset of E22, E32, or BURST errors clearly indicate end-of-life, but does not allow extrapolation over multiple storage time intervals to yield a failure time. Both total and peak E12 error rates as well as jitter may be appropriate for extrapolation purposes, but each manufacturer should determine the end-of-life measures appropriate to their processes.

[Return to Top of Page](#)

## **CONCLUSIONS**

CD-R discs are capable of excellent longevity, but achieving that potential requires diligence by both manufacturers and users. Manufacturers claims may be valid, or may be based upon flawed or non-existent data. Proper end-of-life indicators must be used to estimate longevity. This study has shown that BLER is not a universal indicator of media life, although most published longevity estimates have utilized BLER as the sole end-of-life indicator.

E22, E32, and BURST errors are valid end-of-life indicators. When present, they indicate a need for immediate duplication if a disc containing archival information is still readable. Such errors are not useful for estimating media life by extrapolating test results of discs that have been subjected to accelerated ageing. All quality indicators must be considered when selecting end-of-life indicators. This study suggests that total and peak E12 error rates as well as jitter may be useful indicators, provided that all other quality requirements are met.

High initial quality for each disc can only be achieved by managing variations in media and recording drive quality. Individual CD-R lot qualification should be employed where possible to confirm that manufacturing quality was high and was not degraded by subsequent packing, shipping, and storage events. Media handling and storage is very important. Both unrecorded and recorded disks should be archived in clean jewel cases in a stable storage environment of 10 C-15 C and 20%-50% RH, and protected from sunlight and other radiation sources.

CD-R media and drive manufacturers are responsible for product quality levels that support interchange and longevity. Not all manufacturers achieve this goal. Increasing demand may require new facilities or production lines that inevitably undergo growth pains. Technical advances can lead to new manufacturing processes that must be debugged. Price pressures may force compromises in quality that adversely affect baseline quality or increase fluctuations about that baseline. Identification of CD-R media and drives that support consistent, acceptable levels of interchange and longevity is the responsibility of the archivist. Proper vendor qualification and monitoring enables the user to confidently utilize the rich capabilities of CD-R, and rewards the manufacture with recognition of their diligent efforts to attain and maintain product quality.

Reliance upon brand names or upon readability of discs in one or a few drives cannot verify longevity. Confidence in longevity can only be achieved by initial testing of drives and media, through proper handling and storage, and by periodic resampling to confirm longevity or to identify a need for duplication while the original disc is still readable. Short cuts do not exist. The level of confidence will always be proportional to the amount of effort and expense incurred by the archivist in establishing and maintaining a high level of CD-R quality. Such methods are appropriate for all media types, and their proper application to CD-R information storage will satisfy the most critical requirements for interchange and longevity.

[Return to Top of Page](#)

---

**Media Sciences, Inc. — Dedicated to Quality**

# CD-R Media Survey

**Jerome L. Hartke, Media Sciences, Inc.**

**Published in *Software Fulfillment News*, May 18, 1998 and Updated May 22, 2000**

## **Contents of this Document**

[Purpose](#): Background of this study.

[Sample Selection](#): Compromises and definitions of test lots.

[Test Methods](#): Description of equipment and calibration procedures.

[Test Results](#): Pass/fail results by test type and by lot.

[Table of Test Results](#): Results for browsers that support tables.

[Summary](#): Conclusions of this CD-R media evaluation.

[Return to Home Page](#)

## **Purpose**

CD-R users experiencing problems may believe that all discs are defective, while others may feel that all discs are equally good. Neither expectations are accurate. Quantitative tests by Media Sciences show that the quality of CD-R discs from experienced manufacturers has improved from 30% defective discs in 1998 to 13% in 2000. Failures for other discs increased from 33% in 1998 to an alarming 60% in 2000, mostly for high radial tracking and jitter.

Falling prices, new suppliers, conflicting vendor claims, and silver - green - gold - blue alternatives present a bewildering matrix to CD-R buyers requiring both high quality and low cost. Information to guide procurement is often lacking, since readability in a few drives or the absence of coasters are not effective quality indicators. Costly and embarrassing field failures are often the result.

One laser beam in the read drive sustains multiple operations that are sensitive to CD-R disc quality. Beam intensity is bit detected in the data channel, and data bytes are then error corrected. Beam variations control the clock and spindle servos. One differential beam detector supports the track following servo. Another controls the focus servo. Since drive standards do not exist, readability is not an indicator of disc quality because various read drives can respond differently to media defects. Only conformance to disc standards provides confidence in readability and interchange.

CD-R recording is a sophisticated process involving many components that can result in highly variable quality. Discs, writers, system hardware and software, and recording software can degrade quality. Extensive error correction methods in read drives can mask severe flaws, therefore readability alone does not predict interchange and longevity. Quality can be established and

maintained only through in-depth testing using properly calibrated equipment and trained personnel.

This study provides an example of CD-R quality evaluation that properly forecasts interchange and longevity. Detailed results are reported for each of many tests instead of an ambiguous pass/fail conclusion for the disc. Results of this study are not reported by brand name nor are consumers provided with a list of approved or recommended suppliers, since brand quality may be inconsistent.

[Return to Top of Page](#)

## Sample Selection

Compromises were necessary in order to achieve timely and cost-effective results. The exact origin of each sample was not tracked, although suppliers may have various manufacturing plants or may resell discs from another source. Regular variations with time, or from lot-to-lot, were not explored. Some vendors were not included because they only resell privately labelled discs made by others. The following samples were selected to represent a significant and diverse part of industry capacity.

Lot A, tested in 1998, consisted of 23 samples of 74 minute discs from seven different experienced manufacturers. Cyanine dye (green with gold metal or blue with silver metal) samples from Denon, Verbatim, Taiyo Yuden, and TDK were evaluated. Phthalocyanine (gold) samples from Kodak, Mitsui Toatsu, and Ricoh were tested. Ten discs passed, six were marginal, and seven discs failed.

Lot B, also tested in 1998, contained twelve samples of 74 minute green discs from six manufacturers, generally from the Asia-Pacific region, including Anton, CMC, KAO, Lead Data, Mega Media, and Ritek. One disc passed, seven were marginal, and four discs failed.

Lot C used six samples of 63 minute discs manufactured by Mitsui Toatsu and TDK. These potentially had of higher quality because of lower data density associated with higher linear track velocity. All six discs passed.

Lot D tests during 1999 and 2000 used 87 samples from 17 different manufacturers, adding 80 min. discs, silver discs, and 8X recording. Fifteen samples were from manufacturers represented in Lot A; seven passed, six were marginal, and two failed, one for jitter and one for length deviation. Seventy-two were from manufacturers typical of Lot B; eleven passed, 18 were marginal, and 43 failed, mostly for high radial tracking and jitter.

[Return to Top of Page](#)

## Test Methods

Samples were recorded at 2X, 4X, and 8X in a high quality recording system of Media Sciences using selected components. The recording system was periodically retested using known media and provided consistent results. Each recorded sample was evaluated for parameters and errors in accordance with ISO/IEC 10149, with certain limits modified for CD-R. Philips calibration and

correlation discs were used to certify the test system.

A full suite of electrical tests was used to evaluate CD-R quality. Nominal results indicate high quality, while marginal or non-conforming results disclose problems that predict interchange and longevity failures. Reliance only upon pass/fail tests, or upon qualitative tests such as readability, often results in misleading results and erroneous conclusions.

Parametric tests were conducted for Itop (reflectance), radial tracking (push-pull), radial noise, I11/Itop, I3/Itop, asymmetry, jitter, length deviation, and radial contrast before. Local defects were evaluated using BLER, E22, E32, and burst error measurements.

Environmental tests were not conducted, although heat, humidity, sunlight, and chemicals may degrade CD-R quality. Experience indicates that longevity is best attained by assuring high initial quality. Discs can then be stored under the proper conditions with confidence that they will always be readable in any system of reasonable quality.

[Return to Top of Page](#)

### **Test Results - Itop**

Intensity of the reflected laser beam as measured by Itop is important to readability. Reflectivity is affected by both the type and thickness of metallization and by attenuation in the dye layer.

Lot A Results: No discs failed, two were marginal, and 21 discs passed.

Lot B Results: No discs failed, two were marginal, and ten discs passed.

Lot C Results: All six discs passed.

Lot D Results: Seven discs failed, four were marginal, and 76 discs passed.

### **Test Results - Radial Tracking**

This measure of sensitivity of the radial error signal predicts whether or not the radial servo of a read drive will accurately position the laser. Marginally high results are not uncommon, but very high values can result from pre-groove or dye problems, and such discs may be readable in some drives but not in others.

Lot A Results: No discs failed, seven were marginal, and sixteen discs passed.

Lot B Results: No discs failed, four were marginal, and eight discs passed.

Lot C Results: All six discs passed.

Lot D Results: Fifteen discs failed, 24 were marginal, and 48 discs passed.

### **Test Results - Radial Noise**

Excessive noise caused by defects can disrupt the radial servo of a write or read drive, resulting in unpredictable errors.

Lot A Results: No discs failed, two were marginal, and 21 discs passed.

Lot B Results: All twelve discs passed.

Lot C Results: All six discs passed.

Lot D Results: All 87 discs passed.

## **Test Results - I11/Itop and I3/Itop**

Measurement of the amplitude of the return laser beam at both the lowest and highest data rates provides an important quality indicator of the dye layer. Weak signals predict readability problems, especially if subsequent dust or scratches degrade the entrance surface of the disc.

Lot A Results: No discs failed, two were marginal, and 21 discs passed.

Lot B Results: No discs failed, two were marginal, and ten discs passed.

Lot C Results: All six discs passed.

Lot D Results: Two discs failed, three were marginal, and 82 discs passed.

## **Test Results - Asymmetry**

CD-R recorders perform optimum power calibration (OPC) prior to writing. Asymmetry evaluates the match between the dye layer and OPC, and also evaluates the radial uniformity of the dye.

Lot A Results: Three discs failed, none were marginal, and twenty discs passed.

Lot B Results: One disc failed, two were marginal, and nine discs passed.

Lot C Results: All six discs passed.

Lot D Results: Two discs failed, seven were marginal, and 78 discs passed.

## **Test Results - Jitter and Length Deviation**

Error-free data recovery requires that tolerances be maintained in the time intervals between mark-land transitions. Length deviation evaluates averages of eighteen different time intervals, while jitter measures random variations from those averages. Deficiencies usually result from either pre-groove or dye problems, and are a frequent cause of media failure.

Lot A Results: Five discs failed, four were marginal, and fourteen discs passed.

Lot B Results: Four discs failed, three were marginal, and five discs passed.

Lot C Results: All six discs passed.

Lot D Results: Twenty discs failed, seventeen were marginal, and 50 discs passed.

## **Test Results - Radial Contrast Before**

This important quality indicator for the unrecorded disc measures sensitivity to off-track position of the return laser beam. Radial position errors may be excessive during recording if it is low. Problems usually point to an incorrect pre-groove. No maximum limit exists in the standards. Samples did exhibit 3:1 variations that could cause problems in some recorders.

Lot A Results: All 23 discs passed.

Lot B Results: No discs failed, two were marginal, and ten discs passed.

Lot C Results: All six discs passed.

Lot D Results: All 87 discs passed.

## **Test Results - BLER**

BLER denotes frame error rate averaged over ten seconds. It accurately evaluates small errors caused by noise or microscopic manufacturing defects. The long averaging time masks large defects such as scratches, debris, or black spots.

Lot A Results: No discs failed, two were marginal, and 21 discs passed.

Lot B Results: All twelve discs passed.

Lot C Results: All six discs passed.

Lot D Results: Two discs failed, none were marginal, and 85 discs passed.

## Test Results - E22

Large defects generate errors that approach uncorrectable levels. Standards specify maximum defect sizes that indirectly forbid E22 errors, and directly prohibit them for archival storage. Since defects cause not only data errors but also adversely affect radial, focus, clock, and spindle servos, their effects are not predictable. Problems observable only as E22 errors in certain drives may generate more serious errors in other drives.

Lot A Results: Two discs failed, one was marginal, and twenty discs passed.

Lot B Results: No discs failed, three were marginal, and nine discs passed.

Lot C Results: All six discs passed.

Lot D Results: Eight discs failed, none were marginal, and 79 discs passed.

## Test Results - E32

Very large defects in the substrate, pre-groove, dye, or metallization generate E32 errors that are uncorrectable by C1/C2 circuitry in the read drive. All Standards require zero E32 errors that can cause read failure, low data rates resulting from read retries, or make the disc unsuitable for archival storage purposes.

Lot A Results: One disc failed, none were marginal, and 22 discs passed.

Lot B Results: All twelve discs passed.

Lot C Results: All six discs passed.

Lot D Results: Five discs failed, none were marginal, and 82 discs passed.

## Test Results - Burst

Defects or scratches along the track result in contiguous errors. Standards forbid burst errors that contain seven or more sequential error frames.

Lot A Results: Three discs failed, none were marginal, and twenty discs passed.

Lot B Results: All twelve discs passed.

Lot C Results: All six discs passed.

Lot D Results: Five discs failed, none were marginal, and 82 discs passed.

[Table of Test Results for Newer Browsers](#)

[Return to Top of Page](#)

## Summary

Test results clearly indicated that all discs were not alike, even if their colors were similar. Cost pressures have resulted in a broad matrix of stampers, dyes, metallizations, and processes. No correlation was observed between CD-R quality and dye type (cyanine or phthalocyanine), metallization (gold or silver), or recording speed (2X-8X). Quality is primarily determined by efforts at the manufacturing facility, and depends less on types of dyes or metallizations.

Good discs would be expected to satisfy all interchange and longevity requirements. Marginal or defective discs might be readable in high quality drives, but could fail in others. Degradation from handling or storage might cause poor quality discs to become unreadable while better discs could still

function.

Updated Lot D results from 1999-2000 indicate that the quality of discs from experienced manufacturers, similar to those evaluated in Lot A, has improved from 30% failures to 13%. Quality in Lot D from manufacturers comparable to previous Lot B has declined. Failures in this group increased from 33% to an alarming 60%. High radial tracking and jitter were the most commonly observed deficiencies. BLER was not a meaningful quality indicator.

Readability in a few drives does not confirm quality. Even reliance upon brand name can be ineffective unless each manufacturing location and product type is qualified and regularly monitored to assure consistency. Only in-depth testing can qualify media and assure interchange and longevity. The expense of such an effort is quickly repaid when recording processes flow smoothly and field failures are minimized. Establishment of a quality baseline enables further cost savings to be achieved by reducing the frequency of testing while maintaining a high level of confidence in the process.

This study will be expanded from time-to-time. Current results will be posted on this web site.

[Return to Top of Page](#)

---

**Media Sciences, Inc. — Dedicated to Quality**

# **NARA/Long-Term Usability of Optical Media**

## **The National Archives and Records Administration and the Long-Term Usability of Optical Media for Federal Records: Three Critical Problem Areas**

Since the early 1980s, staff at the National Archives and Records Administration (NARA) have monitored developments in optical media storage technology in order to understand how best to ensure the long-term usability of records of federal agencies stored digitally on optical media.

Three types of optical media can be used to store digital information--CD-ROM, WORM (Write Once, Read Many), and Rewritable. Because these three types of optical media use essentially the same technology to read digital information, the fundamental difference among them is how the information is written or stored. Typically, the production of CD-ROM (4.72 inches in diameter) involves a mastering process that produces multiple copies. In WORM technology, a laser beam and very powerful optics are used to record chunks of information on a single disk. A recent development in CD digital technology called CD-R, that uses the same approach as WORM technology, makes it possible to record on 4.72 inch disks that conform to international standards for physical and logical file characteristics. Information on CD-ROM, CD-R, and WORM media cannot be erased or revised, in contrast with rewritable optical media on which digital information can be changed and deleted almost without limit.

CD-ROM, CD-R, and WORM technologies inherently are attractive for long-term storage of digital information because they can not be erased or revised. Therefore, NARA's interest in optical media technologies focuses largely upon non-rewritable optical media, because of concerns about the long-term usability of digital records stored on CD-ROM, CD-R, and WORM optical media. An examination of the long-term usability of digital records stored on CD-ROM, CD-R, and WORM optical media involves consideration of three critical problem areas: (1) the life expectancy or longevity of the optical media, (2) the capacity of the computer system to measure and compensate for data degradation, and (3) a technology migration strategy that crosses information technology generations. Consequently, NARA initiated a three-part research project through its Technology Research Staff to address these three critical problem areas.

### **1. Life Expectancy**

Typically, WORM optical media manufacturers claim five years of shelf life for blank disks and twenty to thirty years of life after recording. These life expectancy claims are based upon test procedures that vary from one manufacturer to another. The absence of generally accepted test procedures for evaluating the life expectancy of WORM optical media means that comparing vendor claims for longevity is like comparing apples and oranges. In 1987 NARA contracted with the National Institute of Standards and Technology (NIST), then the National Bureau of Standards, to develop a standard test methodology for assessing the life expectancy of WORM optical media. The

results of this project are reported in the NIST study, *Development of a Testing Methodology to Predict Optical Disk Life Expectancy Values*, issued in February 1992. The study proposes a generalizable test methodology that can form the basis for a national and international standard.

## 2. Data Degradation

Although a standard test methodology for predicting the life expectancy of WORM optical media is very important, it does not address the equally important needs for users to have guidance on the care and handling of optical media. Consequently, in 1990, NARA commissioned an on-going project with NIST to produce a report on the care and handling of optical media. One goal of this study is to identify and develop standardized measurements to verify periodically any degradation in the quality of the recording. As part of this project and with support from other federal agencies, NIST has organized a working group composed of users and vendors that is focusing upon this problem.

Digital data errors can be introduced by the communications system transporting data from one place to another, by the mechanical systems writing and reading the data onto media, by deformations in the media such as spots or micro-level warping, and a host of other causes related to the storage media. From a narrow storage perspective, a primary factor influencing the number of data errors is the storage density of the medium. For example, current magnetic media generally have a storage density of about 50 to 60 million bits per square inch, while optical media store on the order of 150 to 400 million bits of data per square inch by utilizing a laser beam focused to approximately one micron to record and read digitized data. The close tolerances for spacing bits, tracks, and sectors on optical media place heavy constraints upon the positioning mechanisms of optical disk drives. A tracking error of one-half micron (approximately 1/50,000th of an inch) in an optical disk system is enough to cause a stored bit to be read incorrectly. In contrast, magnetic media and systems have much larger tolerances and the possibility of errors occurring when reading data is much lower.

Regardless of the medium, storage of digital information has always included some kind of error detection and error correction mechanism so that data can be retrieved error-free. A number of utility programs have been written for magnetic disk based systems to help users determine the location of these errors, to relocate data to other areas of the disk, and to reconstruct the data that has become partially damaged. Unfortunately, similar utilities suitable for the general user do not exist for optical media.

The close mechanical tolerances in optical media and systems require very powerful error detection and error correction schemes to ensure reliability of retrieved data. Optical systems typically provide a statistical probability of error of only one byte out of every one billion bytes. The application of error detection and error correction schemes to achieve this level of reliability is automatic and transparent to users. However, the error correction schemes are limited to handling error rates below five out of every 10,000 bytes. Once this limit has been exceeded, the error correction scheme can no longer compensate for or guarantee correction of all errors, and the optical medium essentially becomes useless.

One solution is to have drive electronics that are capable of providing access to error detection/correction data so that monitoring techniques can be used to monitor the gradual degradation of the

media before the level of errors becomes catastrophic. Utility programs could be written to capture this information on a periodic basis and provide the user with a profile of the optical media. The NIST working group mentioned earlier expects to produce a report in November 1992 that will identify this and other possible solutions. The report is expected to encourage industry cooperation in the development/modification of optical drive error reporting systems so that optical drives from different manufacturers will all have the capability of supporting a common set of error monitoring and reporting utilities. Eventually the industry agreement would become an international standard.

### 3. Technology Migration Strategies

The third crucial problem affecting the long-term usability of digital records created by federal agencies is the failure to develop a migration strategy for moving records to new media and technologies as older ones are displaced. The unavoidable fact is that digital records are technology dependent and therefore technology obsolescent is likely to be the most serious impediment to the long-term usability of digital records. Therefore, the development and implementation of a migration strategy to ensure that digital records created today can be both processed by computers and intelligible to humans in the 21st century is absolutely essential.

In the NARA study, *Digital Imaging and Optical Media Storage Systems: Guidelines for State and Local Government*, completed in late 1991, the broad characteristics of a viable migration strategy were outlined. A follow-on study of digital imaging and optical media storage systems with guidelines for federal agencies, scheduled for completion near the end of 1992, will explore in greater detail alternative migration strategies.

---

The research and investigation into these three critical problem areas for optical media is part of NARA's on-going information technology research activities to address problems of electronic records, a matter of great concern to the entire archival community. Consequently, the findings and conclusions reached through NARA sponsored information technology research are intended to be shared with the entire archival and professional community



---

[\[Search all CoOL documents\]](#)

[\[Feedback\]](#)

**This page last changed: August 24, 2005**

# Development of a Testing Methodology to Predict Optical Disk Life Expectancy Values

NIST Special Publication 500-200

Summary Prepared by

Technology Research Staff

National Archives and Records Administration

Special Publication 500-200, *Development of a Testing Methodology to Predict Optical Disk Life Expectancy Values* by Fernando L. Podio of the National Institute of Standards and Technology, is a technical research report that describes a methodology to predict the longevity of optical media. The National Archives and Records Administration supported this three-year project because there are no national/international standards for the longevity of optical disks that can assist managers in the federal government to select optical media for the storage of permanent records with a reasonable assurance of how long they may be stored on the media.

The study, which consists of five chapters, begins with a Program Overview that reviews the history of the NIST computer storage optical media research program and summarizes methodological issues involved in developing the report. Initially, the research plan called for testing several 300 mm WORM (Read Once Read Many) disks using vendor-supplied drives and other supporting equipment. When this became impossible due to costs, only Sony 300 mm WORM disks, which the National Archives was using in another research project, were tested. Other cost considerations also dictated the level of detailed information captured about the test disks.

A key methodological consideration was the Arrhenius Model, a set of mathematical procedures and computations for accelerated aging, which assumes that temperature and relative humidity are the crucial independent variables that over time affect the longevity of optical media. The NIST use of this model involved storing optical disks in three different high-stress environments (70° C, 80° C, and 90° C with a constant relative humidity of 90 percent) for an extended period of time (ranging from 4120 hours to 5711 hours). The disks were read periodically to monitor the effect of temperature and humidity on the error rate. A linear increase in error rates as harsh conditions of the stress environments increased from 70° C to 90° C over the test period made it possible to extrapolate from these errors and to predict error rates at nominal room temperatures.

Chapter Two focuses upon such technical matters as definition of the end of life of optical media, the measurement of the byte error rate (BER), the optimum number of bytes required for reliable statistical analysis, the pattern of test data (sequential, random, and high frequency), statistical analysis of data, measurement of signal degradation, and procedures for conducting the accelerated aging tests.

The end-of-life definition used in the study was  $5 \times 10^{-4}$ . This means an error rate of five bytes out of every 10,000 bytes, which exceeds the capacity of error correction codes to correct.

Of course, this does not necessarily mean catastrophic failure so that a disk is totally unreadable. Another important consideration was the selection of three different sections of the disk (inner, middle, and outer) of the disk surface area on which to record the test data.

Chapter Three presents the results of the tests, which in most instances are displayed in five tables and fifty graphs. The most significant findings that these tables and graphs convey are summarized below.

- The overall byte error rate (BER) increases as the temperature increases, which confirms the applicability of the Arrhenius Model.
- The byte error rate for the random and sequential test patterns is similar while that of the high-frequency test pattern is larger. The high-frequency test pattern represents the highest areal storage density.
- There is a significant difference in the byte error rate for the three different test sections, with the middle section showing the lowest byte error rate. Consequently, any test that predicts the life expectancy of optical media should calculate the byte error rate for each section and include this information in a report.
- Visual inspection of the three different test patterns revealed damage to the surface of the inner and outer edges, possibly caused by damage to the sealing on the inner and outer edges that may have occurred as a result during placement of the disks in ovens or from the very high stress conditions in the ovens.
- The extrapolated life expectancy with storage at nominal room temperature and 90 percent humidity for each of the three areas varies. The most conservative estimate is 57 years, while a more liberal estimate is 121 years. In either case, a relative humidity between 40 percent and 50 percent should lead to an even longer life expectancy.
- These test procedures are repeatable so that 300 mm optical disk manufacturers can employ them to predict the longevity of their media.

Chapter Four presents a number of conclusions, recommendations, and follow-on activities that flow out of the above discussion of the results of the tests. They include the following:

- Broad test parameters without precise specification can yield equally accurate but uncomparable results (e.g., the average byte error rate for the entire test surface area rather than the byte error rate for each of the surface areas).
- Vendor claims for life expectancy should be accompanied by detailed specification of the test parameters.
- A standardized test methodology for predicting the life expectancy should take into account at least five factors. The entries in boldface are the options used by NIST in their research program.

1. Test Method Used (substrate independent versus substrate dependent)
  - a. glass substrate
  - b. polycarbonate substrate

2. Quality Measurement Approach (byte error rate, bit error rate)
    - a. Data Patterns Used (random, sequential, high density)
    - b. Amount of Data Tested (900 sector blocks on each surface)
    - c. Area of Media Tested (outer, middle, inner)
  3. Mathematical Model Used for Extrapolation
    - a. Arrhenius Model
    - b. Eyring Model
  4. Criteria for Data Analysis (300 sector blocks for read/write)
  5. Experimental Stress Conditions on Media
    - a. Relative Humidity 90%
    - b. Test Temperatures
      - (1) 60°C
      - (2) 70°C
      - (3) 80°C
    - c. Temperature Ramp-Up/Ramp-Down Rates
- NIST will use the findings of this study to develop a common test framework for predicting the life expectancy of optical media that has the support of the optical media industry. This common test framework could lead to the development and approval of a national/international standard test methodology to predict the life expectancy of optical media.



---

[\[Search all CoOL documents\]](#)

[\[Feedback\]](#)

**This page last changed: August 03, 2004**

## Government Information Preservation Working Group

The Government Information Preservation Working Group web site has been move.

Please go to: <http://www.itl.nist.gov/div895/gipwog/index.html>

[Privacy Policy/Security Notice](#)  
[Disclaimer](#) | [FOIA](#)

[NIST](#) is an agency of the  
[U.S. Commerce Department's](#)  
[Technology Administration.](#)

Date created: 12/17/2003  
Last updated: 03/03/2005

# *Stability Comparison of Recordable Optical Discs—A Study of Error Rates in Harsh Conditions*

Volume 109

Number 5

September-October 2004

**Oliver Slattery, Richang Lu,  
Jian Zheng, Fred Byers, and  
Xiao Tang**

National Institute of Standards  
and Technology,  
Gaithersburg, MD 20899-8951

ollie@nist.gov  
richangl@yahoo.com  
jian.zheng@netzero.com  
byers@nist.gov  
xiao.tang@nist.gov

The reliability and longevity of any storage medium is a key issue for archivists and preservationists as well as for the creators of important information. This is particularly true in the case of digital media such as DVD and CD where a sufficient number of errors may render the disc unreadable. This paper describes an initial stability study of commercially available recordable DVD and CD media using accelerated aging tests under conditions of increased temperature and humidity. The effect of prolonged exposure to direct light is also investigated and shown to have an effect on the error rates of the media. Initial results show that high quality optical media have very stable characteristics and may be suitable for long-term storage applications. However,

results also indicate that significant differences exist in the stability of recordable optical media from different manufacturers.

**Key words:** archiving; CD-R; digital preservation; DVD-R; error rates; life expectancy.

**Accepted:** October 18, 2004

**Available online:** <http://www.nist.gov/jres>

## 1. Introduction

Recordable optical disc media contains an organic dye layer whose transparency can be altered either to absorb a laser beam or to allow the beam to pass through to a reflective layer behind the dye [1,2]. The nature of this organic dye is such that when the internal energies of its molecules reach a particular threshold, an irreversible chemical reaction occurs, and the dye layer loses its transparency. This property allows a high-energy beam to “write” data by burning “pits,” in the form of dark marks, to the disc during recording. A low powered laser reads the data by either passing through the transparent dye layer (without causing any molecular change) to the reflective layer or by being absorbed by the nontransparent marks in the dye.

Due to the organic nature of the dye, degradation and breakdown of the transparent portion of dye layer will occur over a long period of time as a natural process. This process, which has its roots in chemical kinetics, can take several years in normal environment conditions [3]. Higher temperatures and humidity will accelerate this process by increasing the thermal and kinetic energies of the dye molecules.

It is well known that temperature and humidity are among the most important factors affecting the life expectancy of optical discs. Yet, there is another important factor that has not been so well investigated. Light exposure can increase the rate of dye degradation precisely because the organic dye used in recordable media is light sensitive. This study also addresses this issue.

The effect of these processes can be modeled using various techniques including the Eyring model [4], which is derived from the study of chemical kinetics. The Eyring equation can model the effect of two stresses, such as temperature and relative humidity, on the rate of a reaction or degradation, which can be related to the time-to-failure of the optical disc.

The end of life of a disc can be defined as the time when an uncorrectable error occurs. Although the disc may still be readable after this point, some information has been lost. Consequently the life expectancy of a disc is the period of time in which the information recorded on the disc can be retrieved without loss. In an ideal case, the real time taken for actual failure to occur would be measured and used as the time to failure. However, this measurement is impractical to explore the degradation process, since a single end point cannot describe the complex process that led to failure. Instead we use the maximum value of some error rate monitor, whose gradual change can serve as an indicator of the media stability. In this study, the block error rate (BLER) [5] is used to monitor CDs and the parity inner (PI) [6] error rate, as summed over eight consecutive error correction blocks (PIE Sum8) [6], is used to monitor DVDs. A high BLER rate indicates a potential onset of uncorrectable errors (E32) [5] in CDs, and likewise a high PI error rate indicates a potential onset of uncorrectable errors (PO) [6] in DVDs. In both cases, these error rate monitors are used to characterize the extent of media deterioration.

## 2. Experimental Equipment and Procedural Overview

All testing occurred at the National Institute of Standards and Technology (NIST) as part of the digital data preservation program ongoing in the Information Access Division (IAD). Two types of environmental chambers were used for artificially aging the media. Both chambers were designed to allow aging of the media under a controlled environmental condition.

Temperature and humidity: A Blue M (model: FRM-256B)<sup>1</sup> environmental chamber was used to control the temperature and relative humidity through various settings of temperature (−18 °C to −93 °C) and relative humidity (5 % to 98 %). The specified control accuracy is ±0.5 °C for temperature and ±1 % for relative humidity (RH) respectively. The test stresses of aging used are given in Table 1.

A complete incubation cycle for temperature and RH accelerated testing is shown in Fig. 1. Once at the stress condition, the temperature and RH were held constant for a period of approximately 45 h followed by a gradual return to ambient conditions. Discs were analyzed after each incubation cycle. This cycle was repeated under the same stress condition until the error rate of most discs in the group increased to exceed an upper limit of the error rates (as indicated in the DVD and CD specifications) or until the disc became unreadable.

**Table 1.** Temperature and relative humidity stresses

Test stress	Incubation cycle duration	Minimum total time (multiple incubation cycles)
60 °C to 90 °C, 70 % to 90 % RH (various combinations)	Approximately 48 h including ramping	450 h to 850 h (approximately)

<sup>1</sup> Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

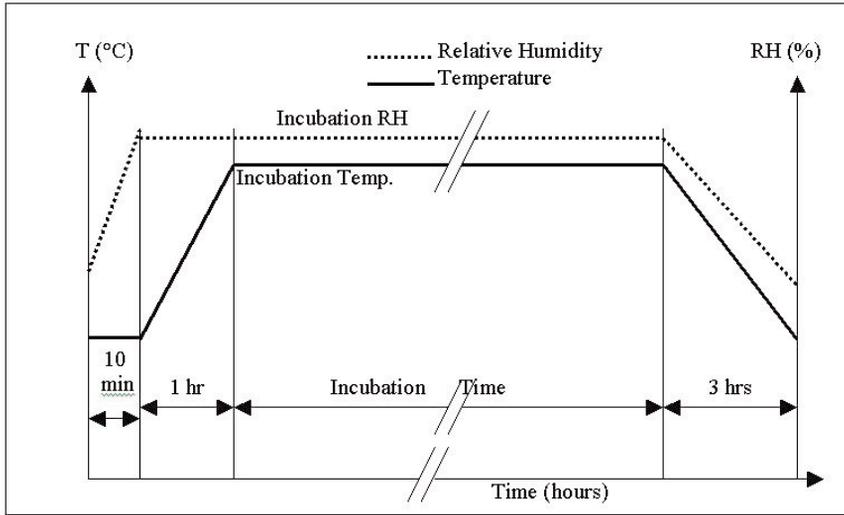


Fig. 1. Temperature and RH incubation cycle.

Light exposure: A light chamber was designed and built at NIST to meet the requirements for controlled light exposure (Fig. 2). Two cylindrical light bulbs were placed vertically in the center of the chamber, with up to twelve discs placed at equal distance from the light source. Intensity was measured at each disc



Fig. 2. Light Chamber.

location to check uniformity. The discs were installed with the recordable side facing the light source.

Two 150 W metal halide (M-H) [7] bulbs were used for the light source, giving a 47.5 mW/cm<sup>2</sup> light intensity at the disc surface. Light intensities were measured using a Scientech Victor S310 thermo-power meter with shield tube. The wavelength range of the metal halide lamps is similar to sunlight, centered at 500 nm, and partly extending to UV region.

*Disc Analyzers:* In order to monitor the change in the error rate during aging, discs were analyzed after each incubation cycle using disc analyzers. A CD-R analyzer capable of reading BLER (in the case of CD) and a DVD-R analyzer capable of reading PI error was used.

*DVD-R Analyzer:* The DVD 1000P analyzer conforms to DVD specifications and was capable of testing electrical, digital, and mechanical parameters in DVDs, including PI errors, PO errors and jitter.

*CD-R Analyzer:* The CD CATS SA3 Advanced allowed measurement of all relevant CD disc parameters including BLER, E32 errors and jitter. All measurements are performed according to optical disc industry standards.

Table 2. Light exposure stress conditions

Test stress	Incubation period duration	Minimum total time (multiple incubation periods)
Metal Halide	100 h (at controlled temperature)	1400 h (approximately)

*Test Specimens:* Test media were selected randomly from the commercial market. Efforts were made to include all the major dye technologies and many of the main commercial brands. The three dye types typically used in CD-Rs (phthalocyanine, cyanine and azo) were included. The dye types for DVD-R were unknown as no specific information had yet been released. Table 3 shows the CD-R test samples used in this experiment and indicates coating and dye type where possible. Similar information for the DVD-R test specimens was not available. Each sample set had several actual pieces of media to ensure that any particular result was representative of that entire sample.

**Table 3.** The CD-R specimens for light exposure test

Sample	Coating and Dye
S1	Unknown, Super Azo
S2	Unknown, Phthalocyanine
S3	Unknown, Super Azo
S4	Silver + Gold, Phthalocyanine
S5	Silver, Metal stabilized cyanine
S6	Silver, Phthalocyanine
S7	Silver, Phthalocyanine

## 2.1 Key Measured Parameters

*Jitter:* Jitter is the temporal variation or imprecision in a signal compared to an ideal reference clock. It is a measure of how well defined the pits and lands of a disc are. For CD discs, jitter is defined in nanometers (nm), and the CD specification states that jitter should not exceed 35 nm. For DVD recordable discs, jitter is defined in percentage points, and should not exceed 9%.

*BLER (CD only):* Block Error Rate is the number of blocks of data that have at least one occurrence of erroneous data. BLER is quantified as the rate of errors (total number of E11, E21, and E31 errors) [5] per second. According to the CD specifications, BLER may be a maximum of 220 per second. Maximum BLER is the maximum BLER measured anywhere on the disc.

*E32 (CD only):* E32 errors are errors that are uncorrectable by the C2-decoder in the CD error detection and correction system. E32 errors represent lost data and therefore no E32 errors are allowed for in the CD specification.

*PIE (DVD only):* Data is arranged in DVD discs in a two-dimensional array with appended parity check bits. Each 2-dimensional array is called an error correction code block. Parity Inner errors (PIE) is the number of parity inner rows with errors. According to the DVD specification, any eight consecutive ECC blocks (PI Sum8) may have a maximum of 280 PI errors.

*POE (DVD only):* Parity Outer errors (POE) are the number of uncorrectable parity outer columns in an ECC block. Since PO errors are uncorrectable by the DVD error detection and correction system, no PO errors are permitted by the DVD specification.

## 3. Results and Discussion

It should be noted that results presented in this paper represent continuous exposure to direct light and extreme temperature/humidity levels. The error rates are not representative of discs stored in typical, normal or ideal storage conditions. The results from these tests are to demonstrate, in terms of error rates, the ability of some DVD and CD media to maintain stability given these extreme conditions.

Also, as stated earlier, each sample set had several pieces of actual media to ensure that results were representative of the entire sample. While there may have been some differences in the results within each sample set, the range was small and thus the results presented here are considered representative of the entire set. Furthermore, particular media from any sample set was subjected to only one particular stress condition.

Results show that the key quality parameters of optical media are altered and error rates increase during exposure to temperature, relative humidity and/or continuous direct light. Since these conditions are key factors in the lifetime of optical media, an estimate for life expectancy can be achieved with a sufficient sample size using various statistical techniques. This investigation, however, was too small to make such an estimate.

The life expectancy of optical media will not be the same for all brands of discs. In a CD-R comparison (see Fig. 3), sample S4, which uses phthalocyanine as the dye and a silver and gold alloy as a reflective layer, is far more stable than any of the other samples during both the temperature/humidity and direct light exposure tests. In a DVD-R comparison (see Fig. 5), sample D2 showed the greatest stability to the temperature/humidity and light exposure tests.

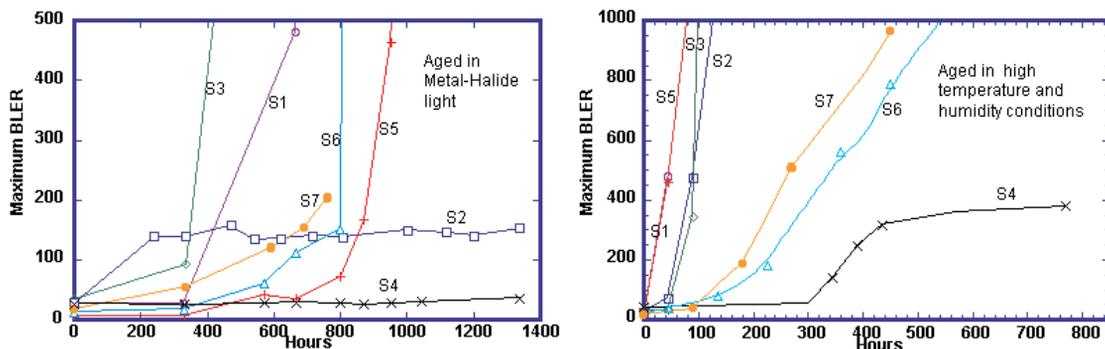


Fig. 3. Maximum BLER increase in CD-R when exposed to (A) M-H and (B) extreme temperature/humidity.

Phthalocyanine based samples S2 and S4 provide very good stability to prolonged direct light exposure as can be seen in the maximum BLER measurements in Fig. 3(A). Both maintain stable BLER levels (following an initial increase in the case of sample S2) beyond 1400 h of exposure to metal halide light whereas other samples have sharp BLER increase within 800 h.

Sample S4 also performs the best in temperature and humidity testing. It shows a BLER of less than 400 after 600 h of an extreme temperature and humidity stress condition while all other samples have BLER greater than 600 within 400 h of the same exposure. Some samples (including S1, S2, S3 and S5) show sharp BLER increases within 100 h. Higher stability for sample S4 is also shown for other key measurements including jitter and E32 under all of the accelerated aging stress conditions used (Fig. 4). According to these results, this disc is clearly more suitable for the long-term storage of important digital data.

Sample S2, however, performs poorly in the extreme temperature and humidity testing despite its good stability to direct light exposure. Within 150 h of extreme temperature/humidity aging, a BLER of over 1000 is observed. Sample S2 uses a darkened polycarbonate layer and this seems to have a limiting or filtering effect on the amount of harmful light reaching the data layer.

Samples S1 and S3, both of which use azo dye for the data layer, have higher error rates in both direct light exposure and extreme temperature/humidity testing. Both samples have sharp increases in BLER within 500 h of direct light exposure and within 100 h in extreme temperature/humidity conditions.

Other samples using phthalocyanine, samples S6 and S7, perform well in direct light exposure until approximately 600 h, but then a significant increase in BLER and errors in general is seen. These samples have low errors beyond 100 h of aging in extreme temperature/humidity conditions, but again have sharp BLER increase soon afterwards. Both of these discs have similar stability characteristics, which is not surprising since samples S6 and S7 are from the same manufacturer (although branded differently) and use the same dye and reflective layers.

Sample S5, which uses cyanine for its data layer, performs well under some conditions of direct light exposure but has problems in extreme temperature and humidity conditions. After 600 h of direct light exposure, sample S5 has a BLER of less than 50, second only to sample S4. After 900 h of exposure however, its BLER increases to more than 500. In extreme temperature and humidity testing, sample S5 has an instant and severe increase in BLER.

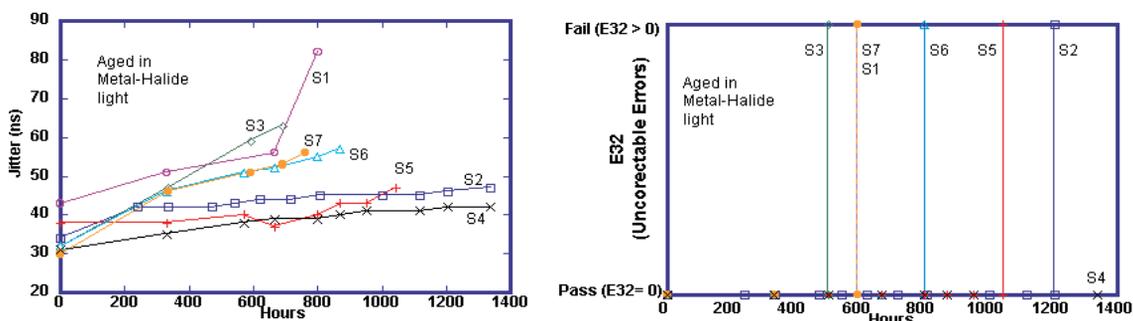


Fig. 4. Increase in (A) jitter and (B) E32 in CD-R exposed to M-H light.

Comparing the BLER for direct metal halide light exposure from Fig. 3(A) with the jitter and E32 errors from Fig. 4, it can be seen that a high level of correlation exists between the various error indicators of CD-R. It also demonstrates that jitter is a key factor in the quality and stability of CD-R media. As jitter increases, a clear correlation between maximum BLER and jitter emerges. In most results, sudden BLER increases or readability problems occur as jitter increases to approximately 50 nS. Fig. 4(B) also shows that BLER is a good predictor of data loss caused by uncorrectable E32 errors. In the example shown for metal halide light exposure, E32 errors occur for all discs in correlation with a sharp increase of BLER.

Many of the trends observed in the error rates of CD-R are also true for DVD-R. In particular, different samples of DVD-R media show different stabilities during exposure to direct light, temperature and relative humidity. Unfortunately, dye information for DVD-R is less accessible than for CD-R and it is therefore difficult to make a determination of stability based on dye type. However, most DVD-R discs tested are based on

a stabilized Cyanine dye. Since results from these samples of similar dye types are quite different, there appears to be varying proprietary modifications made to the dye formulations, and perhaps different manufacturing processes and quality control procedures.

Fig. 5 shows that sample D2 is the most stable of the three DVD-R media types tested. PIE for this sample remains low beyond 800 h of exposure to direct metal halide light compared with steady increases in PIE in samples D1 and D3 to approximately 1500 after 800 h.

The stability of sample D2 is further demonstrated when compared to the other samples during exposure to extreme temperature and humidity. Within 200 h, both samples D1 and D3 have reached PIE of approximately 1000 whereas sample D2 remains very low beyond 400 h of exposure.

Fig. 6 shows that there is a correlation between the key error rate monitor, PIE, and the onset of uncorrectable parity outer errors (POE), although with the small sample size, it is difficult to identify any clear value of PIE at which POE occurs. And as in the case of CD, jitter appears to have a good correlation with PI errors trends.

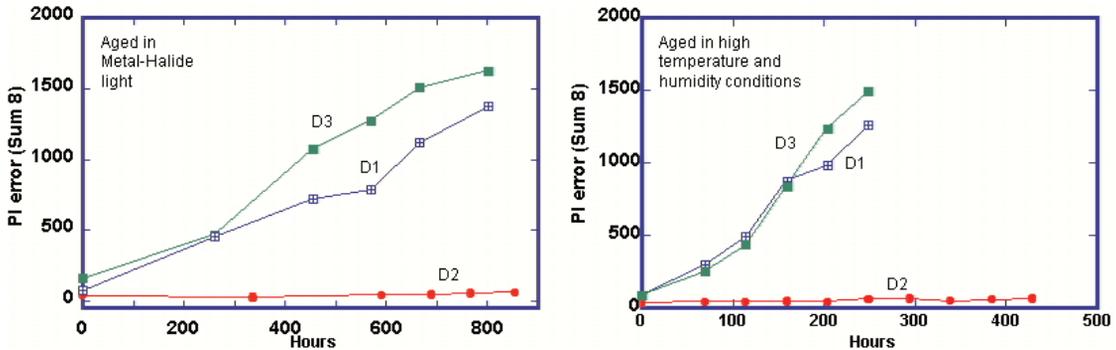


Fig. 5. PI (Sum 8) increase in DVD-R when exposed to (A) M-H and (B) extreme temperature/humidity.

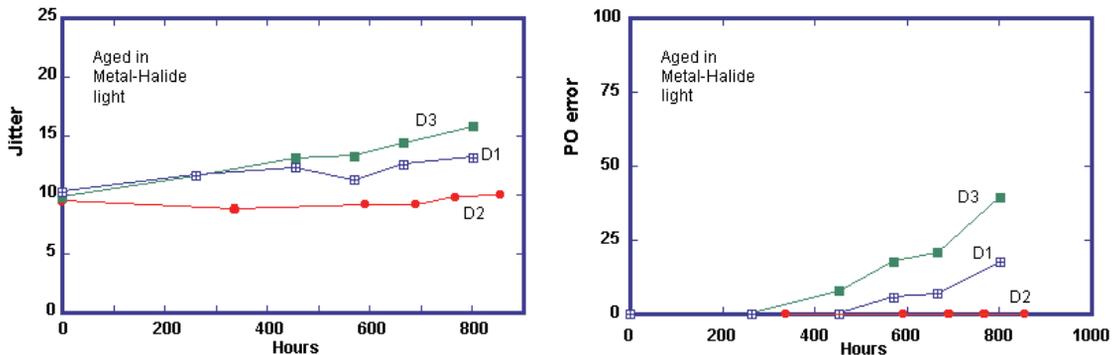


Fig. 6. Increase in (A) jitter and (B) POE in DVD-R when exposed to M-H light.

#### 4. Conclusions

The quantity of media tested in this study was too small to provide the statistical estimation for the life expectancy of the discs but shows relative error trends associated with the different media. A complete study to provide an estimate for the life expectancy is underway at NIST in collaboration with the Library of Congress (LoC).

While there are a number of factors that may contribute to the stability of the CD-R and DVD-R media, dye type is generally considered one of the more important ones. Based on the test results for CD-R media, this expectation appears to hold true, even with mixed results for the dye types. Samples containing phthalocyanine performed better than other dye types. In particular, phthalocyanine combined with a gold-silver alloy as a reflective layer was consistently more stable than all other types of CD-R media. Discs using azo dye as the data layer had less stability in light exposure and temperature/humidity stress testing. Media using cyanine dye performed well when exposed to light but had problems when under temperature/humidity stress conditions.

Although information is less accessible regarding the dye type used in DVD recordable media, it is believed that DVD-R media use a modified form of stabilized Cyanine dye for the recording layer. It is therefore difficult to make any determinations from these results based on dye types for DVD-R media. Furthermore, manufacturers make modifications to the dye to improve its stability or to make it less expensive. This process may result in similar dye types having considerably different qualities, which is shown to be the case in the DVD-R discs tested. And again, as in the case of the CD recordable media, the variation of stability among different brands of DVD recordable media is considerable.

Our results show that the effects of direct light exposure cannot be ignored. The spectral wavelength of metal halide is close to what may be expected within the higher spectrum of sunlight. Depending on the media type and intensity of the light, a disc may fail due to exposure to direct sunlight in as little as a few weeks. This will be especially true when coupled with the heating effect of exposure to sunlight or combined with any other heat source. For archival purposes, however, light is a less challenging issue since it is relatively simple to avoid direct light exposure or prolonged exposure to any damaging light source.

There are a number of physical disc parameters that will provide a good indication of the quality of the media. Based on these results, jitter is a key indicator of media quality in both CD and DVD recordable media. A dye's ability to maintain well-defined marks is crucial in maintaining low error rates. This also indicates that the dye layer is probably the most significant layer for media stability. Other layers, such as the polycarbonate layer, may also degrade but at a slower rate than the dye layer. Furthermore, a disc with a faded or damaged polycarbonate layer may still have all the data intact and therefore the data may be recovered and migrated to new media. If, however, the dye layer becomes damaged or has degraded, causing uncorrectable errors to occur, the uncorrectable data cannot be recovered. Uncorrectable data error may cause negligible, minimal, or up to catastrophic failure, depending on either the extent or the location of that uncorrectable error within the DVD data structure.

It is demonstrated here that CD-R and DVD-R media can be very stable (sample S4 for CD-R and sample D2 for DVD-R). Results suggest that these media types will ensure data is available for several tens of years and therefore may be suitable for archival uses. Unfortunately, it is very difficult for customers to identify these more stable media.

It is clear that an archive quality grade for media is necessary and should be based on a number of quality parameters rather than brand name or manufacturer. NIST has been leading this effort in consultation with other government agencies and has assisted in the formation of the "Government Information Preservation Working Group" to address this issue. This working group plans to clearly state their needs in regard to the longevity of optical media and work with the optical disc industry to develop requirements for an archival CD or DVD recordable media. A comprehensive study is underway in a collaboration between NIST and the Library of Congress (LoC) with two principle objectives: 1) to determine the life expectancy of DVD recordable media and 2) to develop a test which media manufacturers can use to assign an archive quality grade to their product.

#### Acknowledgments

The authors wish to thank the management of the NIST Information Technology Laboratory (ITL), the leadership of the Convergent Information Systems Division (CISD) and the Information Access Division (IAD) for their support in this important effort.

## 5. References

- [1] H. Bennett, Understanding CD-R and CD-RW, Optical Storage Technology Association, California, USA (2003).
- [2] H. Bennett, Understanding Recordable and Rewritable DVD, Optical Storage Technology Association, California, USA (2004).
- [3] D. Nikles and J. Wiest, Accelerated Aging Studies and the Prediction of the Archival Lifetime of Optical Disk Media, Center for Materials for Information Technology, University of Alabama, Alabama, USA (2000).
- [4] P. T. Kahan, A Study of the Eyring Model and its Application to Component Degradation, IBM Components Division, New York, USA (1970).
- [5] A. Svensson, CD-CATS SA3 Users Manual, AudioDev Inc., West Des Moines, Idaho, USA (2000).
- [6] CD Associates, Inc., The DVD1000P Analyzer Manual, CD Associates Inc., Irvine California, USA (CD Associates is now called DaTarius Inc.) (1998).
- [7] USHIO, General Lighting Catalog, USHIO America Inc., Cypress, California, USA.

### Further reading:

1. B. Mann and C. Shahani, Longevity of CD Media: Research at the Library of Congress,, Library of Congress, Washington DC, USA (2003).
2. F. Byers, Care and Handling for the Preservation of CDs and DVDs—A guide for Librarians and Archivists, NIST Special Publication 500-252, Gaithersburg, Maryland, USA (2003).
3. AES Standard for audio preservation and restoration—Method for estimating life expectancy of compact discs (CD-ROM), based on effects of temperature and relative humidity, Reference number: AES28-1997 (1997).
4. ISO International Standard, Imaging materials—Recordable compact disc system—Methods for estimating the life expectancy based on the effects of temperature and relative humidity, Reference number: ISO/FDIS 18927:2001 (2001).
5. E. Zwaneveld, Standards and Technology Strategies to Preserve Content on Magnetic and Disc Media, SMPTE Journal, New York, USA (2000).
6. K. Lee, O. Slattery, R. Lu, X. Tang, and V. McCrary, The State of Art and Practice in Digital Preservation, J. Research of National Institute of Standards and Technology Vol. 107, Gaithersburg Maryland, USA (2002).
7. X. Tang and J. Zheng, High-precision measurement of reflectance for films under substrates, Optical Engineering Vol. 41, No. 12., Amherst, New Hampshire, USA (2002).
8. W. Murray, Life Expectancy of Optical Systems, JTC SnSO IT9-5/AES (1990).
9. F. Akhavan and T. Milster, CD-R and CD-RW optical disk characterization in response to intense light sources, SPIE Conference on Recent Advances in Metrology, Denver, Colorado, USA (1999).
10. F. Podio, Development of a Testing Methodology to Predict Optical Disk Life Expectancy Values, NIST Special Publication 500-200, National Institute of Standards and Technology, Gaithersburg, Maryland, USA (1991).

*About the authors: Richang Lu was a guest researcher in the Convergent Information System and Advanced Networking Divisions; Jian Zheng was a guest researcher in the Convergent Information System and Information Access Divisions; Oliver Slattery and Fred Byers were technical staff members in the Convergent Information Systems Division and are currently members of the Information Access Division; Xiao Tang is the former leader of the Information Storage and Integrated System Group, within the Convergent Information Systems Division and is currently a member of the Advanced Networking Division. All Divisions are in the NIST Information Technology Laboratory. The National Institute of Standards and Technology is an agency of the Technology Administration, U.S. Department of Commerce.*

## DVD - Frequently Asked Questions

### What is DVD?

It is a type of high density compact disc that can hold from 7 to 14 times as much information as a conventional CD. It can hold video, audio, or computer data. At the present time, DVDs are used mostly for playing movies, although computer drives also use DVD-ROM discs.

### What do the letters DVD stand for?

Officially, nothing. But many people translate the letters as Digital Versatile Disc or Digital Video Disc.

### What are the DVD format types?

The most common ones are:

DVD-5	Stores 4.7 Gbytes	Data on 1 side in 1 layer
DVD-9	Stores 8.5 Gbytes	Data on 1 side in 2 layers
DVD-10	Stores 9.4 Gbytes	Data on 2 sides in 1 layer
DVD-18	Stores 17.0 Gbytes	Data on 2 sides in 2 layers

### Why are there different formats?

To accommodate different amounts of data. One side of one disc can hold about two hours' worth of video and audio, but of course some movies are longer than that. Some of the first long movies on DVD needed to be flipped partway through; new technology allows about four hours of seamless playback on one two-layer side. In some cases, publishers have put two versions of a movie –wide screen and so-called "pan and scan" – on the two sides of a disc, so purchasers can have a choice. Only about 5 percent to 10 percent of movies are too long to fit on one side of a DVD disc.

### Is a DVD disc made like a CD?

In many ways, the process is similar, but there are some big differences. The DVD disc is made of two layers, each half the thickness of a CD. So when they're bonded together they are as thick as a CD (1.2 mm). The pits of a DVD disc are half as big, and much more closely spaced, than those of a CD. That means the laser that "reads" the disc must be smaller and more sharply focused.

### Are all DVD discs compatible with all DVD players?

Generally, yes. For legal reasons, some publishers encode DVD discs so that they may play only in certain parts of the world. This is called "regional coding."

### Are DVD players compatible with my television now?

Yes. You just hook it up as you would any other accessory.

## **What about High Definition Television (HDTV)?**

At the present time, DVD won't work with HDTV because its data rates are so much higher.

## **Do DVD players play my present CDs too?**

Yes.

## **What can DVD do that VHS videotape can't?**

To begin with, the picture and sound quality of DVD are far superior to that of VHS. The discs won't wear out or degrade over time, as tape will. They can't be accidentally erased. Plus DVDs can have many extra features such as a choice of languages on the sound track, the ability to play back favorite sections instantly, special-effects playback (such as slow-motion), the choice of parental "locks" on objectionable scenes, and in some cases the ability to select different camera angles, such as in a filmed concert. The audio is better than that of a CD.

## **How about compared with laserdiscs?**

The quality of DVD video has generally been judged better than that of laserdiscs.

## **Can labels be printed on DVD discs?**

It depends on the type of disc. If it is single-sided, a label may be printed on the top side.

If a DVD disc has data on both sides silk-screening can not be used on the main surface. The so-called "mirror band" in the center of the disc may be printed on.

[Next](#) 

-  [FAQ-Authoring and Replication](#)
-  [FAQ-Packaging](#)
-  [DVD Overview](#)
-  [Glossary](#)

 [Top](#)

---

[Products](#) · [Services](#) · [Specifications](#) · [News](#) · [Order](#) · [Site Map](#) · [Home](#)

---

Optical Disc Solutions  
1767 Sheridan Street  
Richmond, Indiana 47374

BKlaine@odiscs.com  
Phone: 800.704.7648  
Fax: 765.935.0174

© 2005, Optical Disc Solutions, Inc., All Rights Reserved

# Optical Disk Formats: A Briefing

## ERIC Digest

Title: *Optical Disk Formats: A Briefing*. ERIC Digest.

Personal Author: Schamber, Linda

Clearinghouse Number: IR052626

Publication Date: May 88

Accession Number: ED303176

**Descriptors:** Information Retrieval; Information Sources; \*Information Storage; \*Information Systems; Interactive Video; \*Optical Data Disks; \*Technological Advancement; \*Videodisks

**Identifiers:** ERIC Digests

### Abstract

This digest begins with a brief description and review of the development of optical disks. Optical disk formats are then described by capability: Read Only Memory (ROM), Write Once, Read Many (WORM), Interactive (I), and Erasable (E); forms of information (audio, text or data, video or graphics, or a combination); and disk size (most often 12 or 4.72 inches in diameter). Some 12-inch formats are then briefly described: optical digital data disk, videodisc, digital video disk, and interactive video disk. Brief descriptions of compact disk formats cover compact audio disk, compact disk-read only memory (CD-ROM), and compact disk-write once, read many (CD-WORM). Future formats currently under development are also briefly described, including HDTV video disk (readable by high-definition television), compact disk-interactive (CD-I), digital video interactive (DVI), and compact disk-erasable magneto optic (CD-EMO). It is concluded that, although the new technology presents some problems, these problems will disappear within the next few years because of the new formats and systems being developed. It is suggested that interested users may keep up-to-date on new developments in this rapidly developing field by contacting manufacturers directly and by reading recent periodicals. (10 references) (CGD)

**Institution Name:** ERIC Clearinghouse on Information Resources, Syracuse, N.Y.

### Article Body

---

The first optical disks that became commercially available, around 1982, were compact audio disks. Since then, and during a particularly active marketing period from 1984 to 1986, at least a dozen other optical formats have emerged or are under development. The rapid proliferation of formats has led, understandably, to some confusion. This digest will briefly describe the most prominent formats (and their acronyms), and the contexts in which they are used.

Optical disks go by many names--OD, laser disk, and disc among them--all of which are acceptable. At first glance, they bear some resemblance to floppy disks: they may be about the same size, store information in digital form, and be used with microcomputers. But where information is encoded on floppy disks in the form of magnetic charges, it is encoded on optical disks in the form of microscopic pits. These pits are created and read with laser (light) technology.

The optical disks that are sold are actually "pressed," or molded from a glass master disk, in somewhat the same way as phonograph records. The copies are composed of clear hard plastic with a reflective metal backing, and protected by a tough lacquer. As the disk spins in the optical disk player (reader, drive), a reader head projects a beam from a low-power laser through the plastic onto the pitted data surface and interprets the reflected light. Optical disk players can stand alone, like portable tape players, or be connected to stereos, televisions, or microcomputers.

Optical disks lack the erasability and access speed of floppies, but these disadvantages are offset by their huge storage capacity, higher level of accuracy, and greater durability.

## Optical Disk Formats

Optical disk formats are described by capability, information form, and disk size. The most familiar capabilities are:

### Read Only Memory (ROM, RO)

permanent, unalterable storage, currently the only firmly established format.

### Write Once, Read Many (WORM, WO)

new data can be written to the disk, but existing data cannot be altered or erased. The drive head contains two lasers: one for reading, and a more powerful laser for etching new data pits. WORM capability may reduce storage capacity by a third. Both 5.25- and 12-inch versions were marketed around 1985.

### Interactive (I)

stores information in any or all forms, along with interactive programming that allows flexible access, animates visuals, runs games, etc.; similar to the capabilities of microcomputer software.

### Erasable (E)

information can be deleted and overwritten. This format requires complex technology and is still being developed.

These capabilities may or may not be available on disks containing different forms of information--audio, text or data, video or graphics, or a combination. And various forms of information can be found on different sizes of disk, most often 12 or 4.72 inches in diameter. Some 12-inch formats are:

- Optical Digital Data Disk (OD3, ODD): contains digitized information (bits), usually text, sometimes images, and used primarily in business and research settings. The disk may have ROM or WORM capabilities.
- Video Disk (videodisc, laser video disk, LV, reflective optical video disk) carries sound and moving images--movies--as analog (continuous) signals. The disk itself may be silvery, reflecting colors for a rainbow effect.
- Digital Video Disk (digitally encoded video disk) contains digital video, audio, and/or text. Still video is more common than motion video because it is more economical to store and easier to access. Digital motion video must be converted to analog signals for playing on standard home televisions. The disk may also include some interactivity programming.
- Interactive Video Disk (IVD) stores 54,000 still frames or 30 minutes of full-motion video with audio. A microcomputer interface allows IVD to gain some interactive capability, although interactivity is hampered by the analog video signal. This format has applications in education, business and industrial training, public information terminals, and games and other entertainment.

Most compact disks (CDs) are 4.72 inches across; a few are 5.25 inches. Audio, text, and still video are stored on CDs. The first two formats below are the most commercially successful to date:

#### Compact Audio Disk (digital audio disk; CD-audio)

contains digital stereo audio signals, generally 75 minutes of high-fidelity, static-free music.

#### Compact Disk-Read Only Memory (CD-ROM)

has tremendous storage capacity, used to store text databases such as bibliographic indexes. One compact disk holds 550 MB of data, the equivalent of 1,500 floppy disks. This amounts to 275,000 pages of information or 200 pounds of paper. Audio or still video may be integrated with text. This format is marketed for library, academic, and professional applications.

#### Compact Disk-Write Once, Read Many (CD-WORM)

often contains text or text with still images (see WORM). Several 5.25-inch formats with write and possibly erase capabilities are also under development.

## Future Formats

The next wave of optical disk formats is about to break. This year or next we are likely to see:

**HDTV Video Disk:** a digital format that is readable by high-definition television, with more than double the number of lines per screen. o **Compact Disk-Interactive (CD-I)**

a writable, very adaptable format that will store text, multichannel audio at several quality levels, still or motion video with low- or high-definition, and animated graphics. Signals may be all digital or include some analog video. CD-I players can stand alone or link to home entertainment systems. This disk and a similar product, compact disk video (CD-video, CD-V), are targeted for mass-market education and entertainment.

**Digital Video Interactive (DVI)**

similar to CD-I, but a highly integrated system that promises more. Compressed digital signals allow the storage capacity of CD-ROM and more motion video than CD-I and IVD: up to an hour of video, with multitrack audio. This disk also has the interactive graphics capabilities of a microcomputer program. Applications include simulations, video paint, special effects, sales tools, and scientific imaging. DVI's release is slated for 1990 or later.

**Compact Disk-Erasable Magneto Optic (CD-EMO)**

combines optic and magnetic technologies in a format that is expected to be easily erasable and reusable for 10 years. Like WORM, EMO requires a drive head with separate read and write lasers. In writing, heat from the laser works in conjunction with a change in magnetic polarity to change the shape of the pit. The erasable disk should also appear within two years.

## Conclusion

With the exception of audio CD, CD-ROM, and video disk, the optical disk formats described here are largely experimental: just released or about to be released. And because the technology is so new, there are very few standards for compatibility of hardware or software, or for encoding, accessing, or integrating data in various forms. But these problems will disappear within the next few years, with optical formats and optical systems that offer...

- players
- different wavelengths or colors
- networks
- organization

just to name a few. The optical technologies industry is moving so fast that the only way to keep up is to contact manufacturers directly, and to read periodicals. *CD-ROM Review*, *Electronic and Optical Publishing Review*, *Optical Information Systems, Database*, and *Journal of Information and Image Management* all focus on optical technologies; while *Byte Magazine*, *Information Today*, and *PC World* take a more general look at high-tech developments. Special issues of two journals focus on CD-ROM for libraries: *The Bulletin of the American Society for Information Science*, 14(1), October-November 1987; and *Wilson Library Bulletin*, December 1987.

## References

- Becker, Karen A. (1987, July/August). CD-ROM: A primer. *C&RL News*, 48(7), 388-399.
- Berg, Brian A. (1987, September-October). Critical considerations for WORM software development. *Optical Information Systems*, 7(5), 329-333.
- Brewer, B. (1987, March-April). Multimedia media mastery. *CD-ROM Review*, 2, 14-18.
- Brewer, B. (1987, April). Ready when you are, CD-I. *PC World*, 5, 252-255.
- Davies, David H. (1988, January). The CD-ROM medium. *Journal of the American Society of Information Science*. 39(1), 34-42.
- Herther, Nancy K. (1988, January). The next CD evolution: Compact Disk-Interactive (CD-I): An interview with Bert Gall." *Online*, 12(1), 68-74.
- Levine, R. (1987, February). Optical storage. *DEC Professional*, 6, 30-38.
- RCA Laboratories (David Sarnoff Research Center). (1987, November-December). New integrated video and graphics technology: Digital Video Interactive. *Optical Information Systems*, 7(6), 412-415.
- Ropiequet, S. (ed.). (1987). *CD-ROM Volume 2, Optical Publishing*. Redmond, WA: Microsoft Press, 1987.
- Saffady, William. (1987, September-October). Optical disks at the 1987 AIIM conference. *Optical Information Systems*, 7(5), 321-328.

Prepared for the ERIC Clearinghouse on Information Resources by Linda Schamber of the School of Information Studies, Syracuse University. May 1988.

This publication was prepared with funding from the Office of Educational Research and Improvement, U.S. Department of Education, under contract RI88062008. The opinions expressed in this report do not necessarily reflect the positions or policies of OERI or ED.

This digest was created by ERIC, the Educational Information Resources Center. For more information about ERIC, contact Access ERIC 1-800-USE-ERIC.



[\[Search all CoOL documents\]](#)

[\[Feedback\]](#)

**This page last changed: August 24, 2005**

# Ampex Virtual Museum and Mailing List

---



**Ampex Model 200A, Serial Number 003**

---

**Welcome to the home of the Ampex Virtual Museum and Ampex Mailing List!**

Entire contents of this site copyright © 2002-2005 by [Howard Sanner](#). All rights reserved. Reproduction is prohibited without written permission from the copyright owner.

[Search This Site](#)

[Manage your Ampex Mailing List subscription](#)

[Parts Sales and Equipment Prices](#)

[Search the Ampex Mailing List Archives](#)

The [Ampex Mailing List](#) provides an email discussion forum for those interested in Ampex products.

Dedicated to preserving the history of the most important manufacturer of magnetic recorders, the Ampex Virtual Museum provides online access to

[Manuals, Schematics, and Service Bulletins](#)

[Repair, Maintenance, and Modification Tips](#)

[Parts and Repair Sources](#)

[Catalogs, Sales Brochures, and Similar Literature](#)

[Pictures of Ampexes](#) (mostly)

[Pictures of Ampex Mailing List parties](#)

[Ampex history](#) & [Other Historical Information](#)

[Audio Clips](#) of Interviews with Former Ampex Employees

[Alignment Instructions](#)

[Links](#) of Interest to Ampex Users

[Obituaries](#)

[Illustrations](#) Accompanying Dale Manquen's Posts on Flutter

[Literature Scanning Project](#)

[FTP](#) Server for Uploading and Downloading Ampex-related Files

New material is added frequently. Check back often to see [what's new](#).

---

Stanford's [press release](#) regarding their acquisition of the collection of the former Ampex Museum of Magnetic Recording and other material from Ampex.

---

3M users should look into Matt Patoray's group for [3M professional audio and video recorders](#).

Check out the [Sony APR](#) group's web page.

Interested in Studers? Check out the [Studer List's](#) home page.

Looking for a forum to discuss reel-to-reel recorders in general? Check out the [Reel-to-Reel Mailing List](#).

---

Ampex is a registered trademark of Ampex Corporation.

The Ampex Virtual Museum and Ampex Mailing List are not affiliated with Ampex Corporation.

---

Questions or comments? Contact the [webmaster](#).

---

Created: 11 Nov 97; revised 16 Oct 2005.

## Schematics

---

[Ampexes](#) courtesy of Ampex Data Systems Corporation web server

[Compressors](#)

[Microphones](#)

[Non-Ampex equipment](#)

---

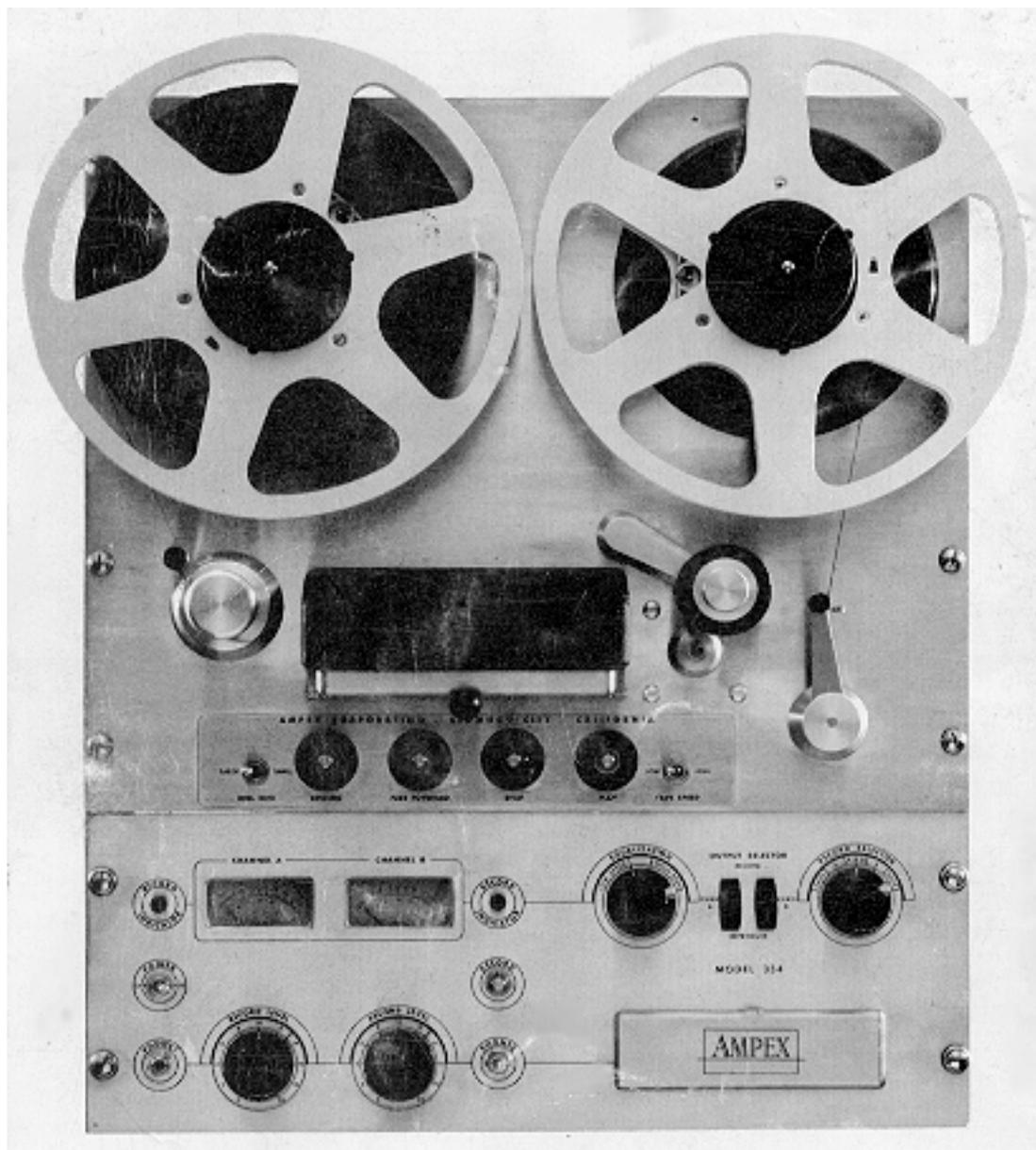
[Back to main page](#)

---

Created 22 Jan 2000; updated 11 May 2003.

# Ampex Repair, Maintenance, and Modification Tips

---



**Ampex 354 Rack Mounted**

---

These repair, maintenance, and modification suggestions are provided as a service to users of the Ampex Mailing List web site. **Listing here is not to be construed as an endorsement or recommendation of the procedure by anyone other than the document's author.** No one else in the universe makes any representations whatsoever about these tips. **It is the responsibility of users of this web page to determine whether a given procedure is applicable to their particular equipment and level of skill.** Anyone making use of information imparted herein does so at his

**own risk.**

---

## General Electronic Troubleshooting Hints

- [Novice's Guide to Electronics Troubleshooting](#) by Rick Chinn and Jay McKnight
- 

## Model-Specific Tips (in numerical order)

### 351

- [351 cosmetic restoration](#)
- [Capacitor & resistor list, with sources, for 30960 electronics restoration](#)
- [John Hughes' 351 restoration](#)

### AG-440

- [Michael Guthrie's 440 Play mod](#)

### MM-1100

- [Alastair Heaslett's procedure for improving MM-1100 reliability](#)
- [Jerry Hertel's further explanation of Heaslett's procedures](#)

### MM-1200

- [Doug McClement's MM-1200 maintenance log & troubleshooting guide](#)

### MX-10/MX35

Note: The MX-10 and MX-35 are identical electronically. Only the knobs and faceplate differ.

- [Lowell Cross's modifications for the MX-10 & MX-35.](#)
- 

## Tips Applicable to Many Models

### Bearings

- [Reel idler bearing replacement](#)

### Capstans

- [Ampex's recommended procedure for resurfacing capstan shafts](#)

### Heads

- [Nortronics Handylap head lapping manual](#)
- [Minneapolis Magnetics's instructions for installing heads on the 300, 350 series, 400, 450, and 3100](#)
- [John French's post on how to determine remaining head life](#)
- [Article on head life from the JRF Magnetics web site](#)
- [Introduction to tape heads from the JRF Magnetics web site](#)

### Motor Rebuilding

- [Beau/UMC capstan motor bearing replacement procedure](#)

### Oscilloscope Usage

- [How to use an oscilloscope](#) (from Tektronix web site)

### Recording Levels

- [Jay McKnight's explanation of how to determine the maximum flux level a recorder can produce](#)

## General Tape Recording Technology

- [Explanation of analog tape technology](#), including **alignment instructions** from Deutsche Welle
- [Art Shifrin's AG-440 wire playback machine](#)

---

[Back to main page](#)

---

Created 25 Oct 99; revised 5 Jun 2004.

# Ampex Parts



MM1200 Headblock

---

## Parts Sources

- [Dave Dintenfass's list of Ampex parts sources](#)
- [OC-11 \(turbine\) oil](#)
- [List of Ampex parts in catalog number order](#) (a work in progress)

## Repair Sources

- [Ampex Repair Technicians](#)
- 

[Back to main page](#)

---

Created 5 Mar 2001

Updated 20 Sep 2004

# Ampex Literature



Model 1260

Reproductions of printed material about Ampexes, from Ampex and other sources, **except** schematics, manuals, and service bulletins. See the separate [schematics page](#) for schematics, manuals, and service bulletins.

## Ampex Tape Recorder Brochures

### 200A

The following brochure for the Model 200A is from Jack Mullin's collection. It appears here through the courtesy of Eve Mullin Collier, his daughter. My thanks to her for making this rare, important document available.

- [200A brochure, p. 1](#)
- [200A brochure, p. 2](#)
- [200A brochure, p. 3](#)

- [200A brochure, p. 4](#)
- [200A brochure, p. 5](#)
- [200A brochure, p. 6](#)
- [200A brochure, p. 7](#)
- [200A brochure, p. 8](#)

#### **AG-300**

- [AG-300 brochure](#) in Acrobat (PDF) format
- [AG-300 brochure](#) in JPG format, p. 1
- [AG-300 brochure](#) in JPG format, p. 2
- [AG-300 brochure](#) in JPG format, p. 3
- [AG-300 brochure](#) in JPG format, p. 4
- [AG-300 brochure](#) in JPG format, p. 5
- [AG-300 brochure](#) in JPG format, p. 6
- [AG-300 brochure](#) in JPG format, p. 7
- [AG-300 brochure](#) in JPG format, p. 8

#### **AG-350**

- [AG-350 brochure](#)

#### **AG-440**

- [AG-440 brochure](#)
- [AG-440 price list A159](#), effective Feb. 20, 1967, p. 1
- [AG-440 price list A159](#), effective Feb. 20, 1967, p. 2

#### **AG-440-8**

- [AG-440-8 brochure](#), p. 1
- [AG-440-8 brochure](#), p. 2

#### **MR-70**

- [MR-70 brochure](#)

---

## **Ampex Accessories Brochures**

- [AM-10 mixer brochure](#), p. 1
  - [AM-10 mixer brochure](#), p. 2
  - [MX-10 mixer brochure](#)
  - [VS-10 vari-speed accessory brochure](#)
- 

## Ampex Catalogs

- [1970 Ampex catalog](#)
- 

## Advertisements

- [Audio Electronics ad showing Model 200A and \*prototype\* of Model 300](#)
  - [Audio Electronics ad showing \*production version\* of Model 300](#)
  - [Audio Electronics ad showing Model 300 and Model 400](#)
  - [Audio Electronics ad showing Model 400](#)
  - [Earliest Audio Electronics ad showing Model 350](#)
  - [Audio Electronics ad showing Model 450](#)
- 

## Ampex Parts & Price Lists

- [Ampex price list, June 1968.](#)
  - [Ampex Professional Audio Spare Parts Catalog](#), effective June 1, 1968  
(original landscape format--24 MB high-resolution scan in Acrobat format)
  - [Ampex Professional Audio Spare Parts Catalog](#), effective June 1, 1968  
(original landscape format--14 MB low-resolution scan in Acrobat format)
  - [Ampex Professional Audio Spare Parts Catalog](#), effective June 1, 1968  
(portrait format--22 MB high-resolution scan in Acrobat format)
  - [Ampex Professional Audio Spare Parts Catalog](#), effective June 1, 1968  
(portrait format--14 MB low-resolution scan in Acrobat format)
- 

## Ampex Tape & Tape Technical Topics

- [Ampex Magnetic Tape Trends](#)  
nos. 2 (July 1963), 4 (Oct. 1963), 7 (Apr. 1964) (46 MB high-resolution scan in Acrobat format)
  - [Ampex Magnetic Tape Trends](#)  
nos. 2 (July 1963), 4 (Oct. 1963), 7 (Apr. 1964) (30 MB low-resolution scan in Acrobat format)
-

## Miscellaneous

- [Magnetic Recording Theory for Instrumentation](#) (Book: 29 MB PDF file)
- [Ampex Ultra High Fidelity Tape Recorders](#) in Acrobat (PDF) format
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 1
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 2
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 3
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 4
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 5
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 6
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 7
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 8
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 9
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 10
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 11
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 12
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 13
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 14
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 15
- [Ampex Ultra High Fidelity Tape Recorders](#), in JPG format p. 16
- [Illustration of track widths](#) in Acrobat format
- [Illustration of track widths](#) in GIF format

---

[Back to main page](#)

---

Created 1 Nov 2000; revised 13 Oct 2002.

# Pictures



Ampex Model 601

---

## Pictures of Ampexes and other equipment

- [A Model 200A earns its keep at Capitol Records](#)
- [Another Model 200A at Capitol doing what it was made for](#)
- [Model 200A picture from the original manual](#)
- [Model 200A head block and tape path from the original manual](#)

- [Threading the model 200A from the original manual](#)
- [Model 200A capstan and motor assembly out of the machine from the original manual](#)
- [Model 200A capstan motor as installed from the original manual](#)
- [Model 200A brake assembly from the original manual \(One of the few problem areas in the design\)](#)
- [Model 200A electronics and wire gutter from the original manual](#)
- [Harold Lindsay and Alexander M. Poniatoff with Model 200A](#)
- [Early Model 300 serial number tag](#)
- [Early Model 300 deck plate. Note plywood in middle of sandwich.](#)
- [Early Model 300 seven-terminal barrier strip. Later ones have nine terminals.](#)
- [Early Model 300 head plug in](#)
- [Model 300 "bathtub" electronics](#)
- [Early Model 300 reel idler](#)
- [Three-track Model 300 with Sel-Sync](#)
- [Model 300 with "bathtub" electronics](#)
- [Model 400A brochure.](#)
- [Model 351 rackmounted](#)
- [Model 351 electronics. Note black dot under input transfer switch.](#)
- [Model 601](#)
- [Model 601-2](#)
- [Model 354 rackmounted](#)

- [AG-440C \(left\) and AG-440B \(right\)](#)
- [Four track MR-70](#)
- [Bottom of MR-70 transport](#)
- [Inside the MR-70's electronics](#)
- [An MM-1100-16 with the covers closed.](#)
- [An MM-1100-16 with the covers open.](#)
- [Top view of an MM-1200.](#)
- [MM-1200 headstack.](#)
- [TU-40 flutter meter](#)

## **Pictures taken at Peter Carli's house**

- [Peter Carli](#)
- [John French and Neil Muncy](#)
- [This is really Mike Rivers](#)
- [Mike Spitz, Mike Rivers, John French, Neil Muncy](#)

## **Pictures of Les Paul's studio**

- [Les Paul & Mary Ford, picture 1](#)
- [Les Paul & Mary Ford, picture 2](#)
- [Les Paul & Mary Ford, picture 3](#)
- [Les Paul in his 1949 studio](#)
- [Some of the electronics for the Octopus](#)
- [The big room of Les Paul's new studio](#)

- [The console in Les Paul's new studio](#)
- [Inside the console in Les Paul's new studio](#)
- [The control room of Les Paul's new studio](#)
- [The editing room of Les Paul's new studio](#)
- [Les Paul's first lathe](#)
- [Cover of the album Les Paul Now](#)

### **Picture from Ampex confab in Olympia, Wash.**

- [Karl Welty, Alex Kostelnick, and Dave Dintenfass in front of Karl's bathtub Model 300.](#)

### **Pictures from the Ampex Old Timers Picnic, 2000**

- [Jay McKnight and Walter Selsted](#)
- [Paul McManus](#)
- [Jim Wheeler](#)

### **Miscellaneous Pictures of List Subscribers**

- [Anna Heaslett](#)
- [Lonn Henrichsen](#)
- [Peter Carli on Nov. 16, 2000](#)

### **Ampex Buildings in October 2000**

- [1313 Laurel St., San Carlos \("Howard Ave. at Laurel"\), Ampex's first offices](#)
- [401 Broadway, view no. 1](#)
- [401 Broadway, view no. 2](#)

- [401 Broadway, view no. 3](#)
- [500 Broadway, Ampex's current office, across the street from 401](#)
- [Time capsule at 401 Broadway](#)
- [Ampex sign from southbound Bayshore Freeway](#)

## **Pictures Taken at ATR Services**

- [Mike Spitz shows the ATR-100 capstan motor to Ross Snyder](#)
- [Mike Spitz and Ross Snyder contemplate the ATR-102](#)
- [The workbench area at ATR Services](#)
- [Mike Spitz loads Doug McClement's MM-1200 using a forklift](#)
- [Mike Spitz, Peter Carli, Doug McClement with Doug's MM-1200s](#)

---

[Back to main page](#)

---

Revised 1 May 2003.

# Ampex History

---



**3200 Duplicator Line**

---

- [AG-440 identification guide](#) by Tom Fine
- [Chronology of Ampex products](#)
- [Scott Dorsey's canonical list of Ampex tape types](#)
- [Dave Dintenfass's canonical list of Ampex 351 electronics variants](#)
- [Dave Sarser's account of the first three-track Ampexes](#)

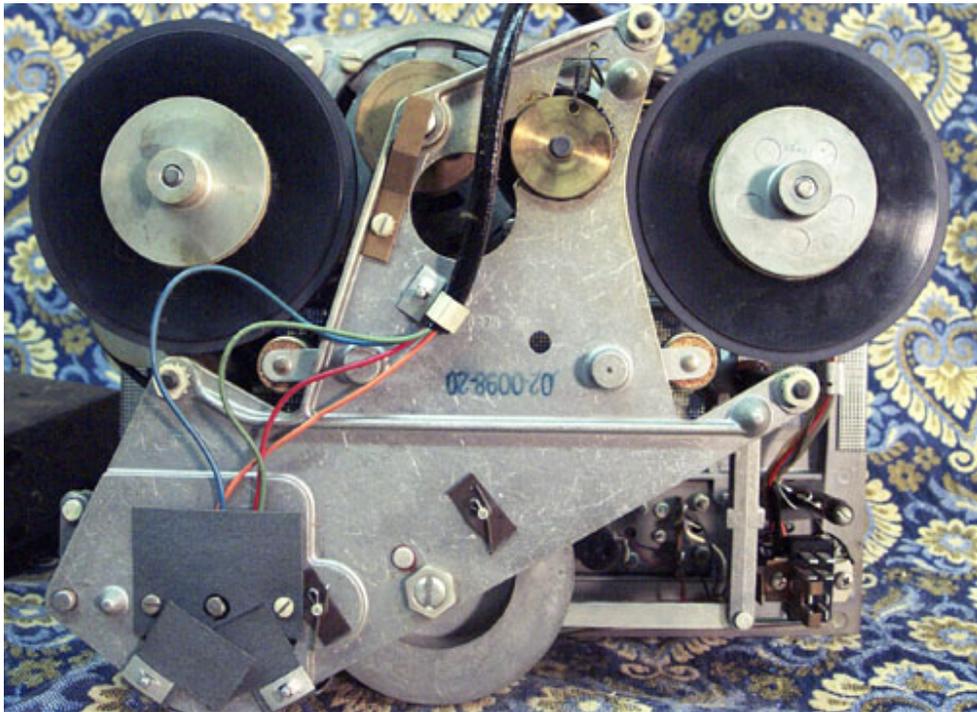
- [Ampex Collection at Stanford Inventory](#)
- 

[Back to main page](#)

---

Created 5 Mar 2001; rev. 18 Nov 2005.

# General Audio History

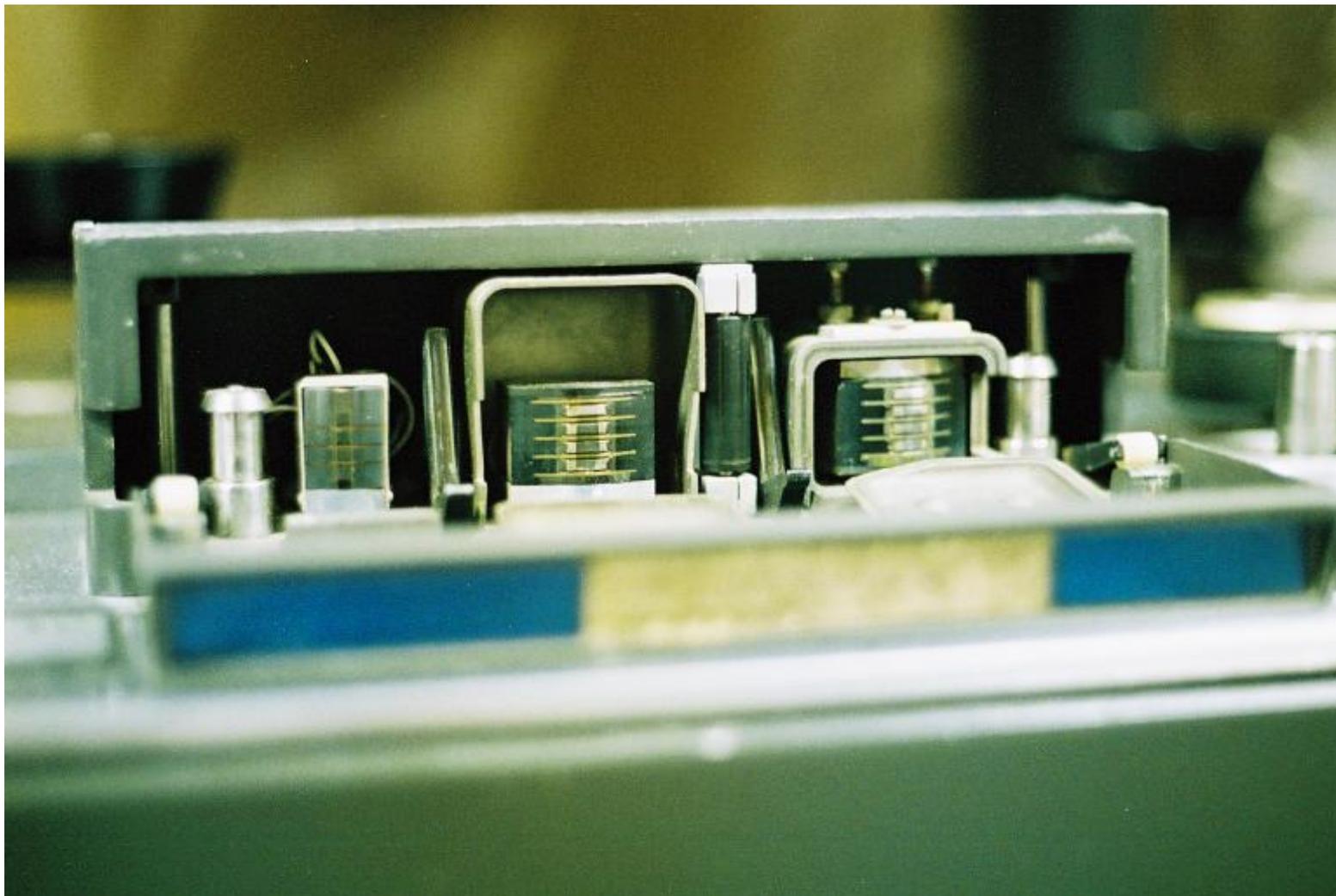


- [Introduction to list of 3M Reel to Reel Audio Tapes](#)
- [List of 3M Reel to Reel Audio Tapes](#)

[Back to main page](#)

Created: 5 Mar 2001  
Modified: 1 May 2003

# Alignment Information



**MR-70 Four-Track Headblock**

- 
- [Bill Vermillion's article on tape deck alignment](#)
  - [Form for recording alignment data, in ASCII](#)

---

[Back to main page](#)

---

Created 5 Mar 2001.

# The Ampex FTP Server

---

To access the FTP server, click [here](#) and change to the **ampex** (case is significant) directory once logged in:

- [FTP server](#)

To upload to the FTP server, log in to **ella.recordist.com** as anonymous with your email address as password. Change to the **ampex/incoming** directory, and begin uploading. Drop a note to the [listowner](#) to notify him that you have uploaded some files. Note that you will be unable to see the files you have uploaded if you attempt to list the directory's contents. This is a feature, not a bug.

---

[Back to main page](#)

---

Created 5 Mar 2001.



---

Welcome to the website of the Audio Engineering Society Historical Committee (AES HC, or just HC for short).

The AES Historical Committee is an international forum open to all who wish to investigate and learn about the achievements of those pioneers whose innovative ideas and inventions have contributed to audio's rich past. Dedicated to the preservation of over a century of audio history, the Committee is developing a broad-based history of audio engineering and the audio industry. **AES membership is encouraged but not required.**

If you have questions about the history of audio engineering, [search](#) this site or its links or for patents from all countries; check out the [Audio Engineering History](#) and [AES History](#). Join our [E-mail reflector](#) and post a historical question to it -- many of the people who made that history are here, and can answer your questions.

We would be glad to have you [volunteer](#) to help us with our projects -- AES membership is encouraged but not required.

[Jay McKnight](#)

Chair, AES Historical Committee

---

## [What's New](#) on this website

### Search

- [This Site and Its Links](#)
- [Patents of All Countries](#)

## [Practical Audio Preservation](#)

# Audio Engineering History

- [Links to Other Historical Sites, and Museum Coordination Project](#)
- [Timeline](#) of Audio Engineering, from cylinder to DVD
- AES Journal
  - [Book Listings and Reviews](#) in the AES Journal, 1990...2002, Index of
  - [Historical Issues](#)
  - [Obituaries](#), Index to
- [Audio Patents](#), the plan for a project to improve access to the patent literature.
- [Auditory Perspective](#), a scan of the classic Bell Telephone Labs paper from 1934 in Acrobat format (4.5 MB file)
- [Books & video tapes](#) now available on audio engineering history
- [The Thiele-Krause Archive for Audio-Visual Media Technology](#)
- [Digital Audio Engineering Standardization History at the AES](#): The history of the development of the AES digital audio standards, as given in the AES Journal minutes and reports of the Digital Audio Standards and Technical Committees between 1977 and 1984, and in Bart Locanthi's 1986 summary of the Committees' work.
- [Exhibits and Meetings of the Historical Committee at AES Conventions](#)
- [Historical Developments in Audio Engineering](#) 1877...1977, with 81 references to papers and patents
- [Forensic Audio Engineering](#): The complete "Watergate" report, formally called "The EOB Tape of June 20, 1972: Report on a Technical Investigation Conducted for the U.S. District Court for the District of Columbia, by the Advisory Panel on White House Tapes May 31, 1974", that is the basis for much of present-day forensic audio engineering.
- Magnetic Recording
  - [Ampex Museum & Historical Collection](#) at Stanford University
  - [Analog Magnetic Tape](#): 3M Tape types, Index to and scans of all 46 issues of 3M "Sound Talk," etc.
  - [Before 1943](#), History of Audio Engineering and Magnetic Recording
  - [Bibliography](#) of Histories of Magnetic Tape Sound Recording, Selected

- [Motion Pictures](#), History of Audio Engineering in
- [Standardization Activity of the AES](#)--the History Before 1982
- Personal Histories
  - Oral Interviews, including [Oral Histories](#)
  - Talks and Papers, including [Texts of Interviews and Talks](#)
- Company Histories
  - [Ampex History](#)

## AES History

- [AES, History of](#): Founding of the AES, Officers and Governors, Conventions and Conferences, and Awards
- [Japan Section of the AES](#), History of

## E-Mail Reflector

- [Join the AESHC E-mail Reflector](#)
- [Using the AESHC E-mail Reflector](#)
- Instructions for [configuring email clients](#) to send plain ASCII email and not HTML

## Committee Administration

- [Operating Structure](#): The Committee's charter
- [Guidelines](#): Committee purpose, officers, membership requirements, and meetings
- [Officers and Project Leaders](#)
- [Committee Reports](#): Committee activities
- [Agendas, etc.](#), of AES Historical Committee Meetings
- [Minutes of AES Historical Committee Meetings](#)

-

Last Revised 2005-05-27

# AN AUDIO TIMELINE

V--1999-10-17

*A selection of significant events, inventions, products and their purveyors, from cylinder to DVD*

The road that leads us from Edison's tin-foil cylinder to today's audio DVD is a fascinating avenue crammed with remarkable people, inventions and innovations. Our past accomplishments contribute to what we are today, and signpost the future as a never-ending quest to push the envelope of what is possible in audio.

In 1997 the Committee for the Fiftieth Anniversary of the AES was formed to increase the awareness of where we have been and what we have accomplished. Part of that effort was directed to the creation of an Audio Timeline, compiled by Jerry Bruck, Al Grundy and Irv Joel. It is intended to be a selection of significant events, inventions, products and their purveyors.

This Timeline is not complete, and probably never could be, given the wealth of discoveries, inventions and innovative products that did and do appear almost daily. Nor are the dates given always precise, depending as they often do on second hand documents or dim memories.

Its authors would welcome any substantiated corrections or additions to this timeline, for imperfect as it must be, it serves as both backbone and DNA in the evolution of our industry. [Please send your comments to us.](#)

We hope you will return to this page often to check on the evolution of our Timeline, as the first of many pieces to be generated by the Historical Committee as this Site expands.

## 1877

- Thomas Alva Edison, working in his lab, succeeds in recovering Mary's Little Lamb from a strip of tinfoil wrapped around a spinning cylinder.
- He demonstrates his invention in the offices of *Scientific American*, and the phonograph is born.

## 1878

- The first music is put on record: cornetist Jules Levy plays "Yankee Doodle."

## 1881

- Clement Ader, using carbon microphones and armature headphones, accidentally produces a

stereo effect when listeners outside the hall monitor adjacent telephone lines linked to stage mikes at the Paris Opera.

### ***1887***

- Emile Berliner is granted a patent on a flat-disc gramophone, making the production of multiple copies practical.

### ***1888***

- Edison introduces an electric motor-driven phonograph.

### ***1895***

- Marconi achieves wireless radio transmission from Italy to America.

### ***1898***

- Valdemar Poulsen patents his "Telegraphone," recording magnetically on steel wire.

### ***1900***

- Poulsen unveils his invention to the public at the Paris Exposition. Austria's Emperor Franz Josef records his congratulations.
- Boston's Symphony Hall opens with the benefit of Wallace Clement Sabine's acoustical advice.

### ***1901***

- The Victor Talking Machine Company is founded by Emile Berliner and Eldridge Johnson.
- Experimental optical recordings are made on motion picture film.

### ***1906***

- Lee DeForest invents the triode vacuum tube, the first electronic signal amplifier.

### ***1910***

- Enrico Caruso is heard in the first live broadcast from the Metropolitan Opera, NYC.

### ***1912***

- Major Edwin F. Armstrong is issued a patent for a regenerative circuit, making radio reception practical.

## **1913**

- The first "talking movie" is demonstrated by Edison using his Kinetophone process, a cylinder player mechanically synchronized to a film projector.

## **1916**

- A patent for the superheterodyne circuit is issued to Armstrong.
- The Society of Motion Picture Engineers (SMPE) is formed.
- Edison does live-versus-recorded demonstrations in Carnegie Hall, NYC.

## **1917**

- The Scully disk recording lathe is introduced.
- E. C. Wente of Bell Telephone Laboratories publishes a paper in *Physical Review* describing a "uniformly sensitive instrument for the absolute measurement of sound intensity" -- the condenser microphone.

## **1919**

- The Radio Corporation of America (RCA) is founded. It is owned in part by United Fruit.

## **1921**

- The first commercial AM radio broadcast is made by KDKA, Pittsburgh PA.

## **1925**

- Bell Labs develops a moving armature lateral cutting system for electrical recording on disk. Concurrently they introduce the Victor Orthophonic Victrola, "Credenza" model. This all-acoustic player -- with no electronics -- is considered a leap forward in phonograph design.
- The first electrically recorded 78 rpm disks appear.
- RCA works on the development of ribbon microphones.

## **1926**

- O'Neill patents iron oxide-coated paper tape.

## ***1927***

- "The Jazz Singer" is released as the first commercial talking picture, using Vitaphone sound on disks synchronized with film.
- The Columbia Broadcasting System (CBS) is formed.
- The Japan Victor Corporation (JVC) is formed as a subsidiary of the Victor Talking Machine Co.

## ***1928***

- Dr. Harold Black at Bell Labs applies for a patent on the principle of negative feedback. It is granted nine years later.
- Dr. Georg Neumann founds a company in Germany to manufacture his condenser microphones. Its first product is the Model CMV 3.

## ***1929***

- Harry Nyquist publishes the mathematical foundation for the sampling theorem basic to all digital audio processing, the "Nyquist Theorem."
- The "Blattnerphone" is developed for use as a magnetic recorder using steel tape.

## ***1931***

- Alan Blumlein, working for Electrical and Musical Industries (EMI) in London, in effect patents stereo. His seminal patent discusses the theory of stereo, both describing and picturing in the course of its 70-odd individual claims a coincident crossed-eights miking arrangement and a "45-45" cutting system for stereo disks.
- Arthur Keller and associates at Bell Labs in New York experiment with a vertical-lateral stereo disk cutter.

## ***1932***

- The first cardioid ribbon microphone is patented by Dr. Harry F. Olson of RCA, using a field coil instead of a permanent magnet.

## ***1933***

- Magnetic recording on steel wire is developed commercially.
- Snow, Fletcher, and Steinberg at Bell Labs transmit the first inter-city stereo audio program.

### **1935**

- AEG (Germany) exhibits its "Magnetophon" Model K-1 at the Berlin Radio Exposition.
- BASF prepares the first plastic-based magnetic tapes.

### **1936**

- BASF makes the first tape recording of a symphony concert during a visit by the touring London Philharmonic Orchestra. Sir Thomas Beecham conducts Mozart.
- Von Braunmühl and Weber apply for a patent on the cardioid condenser microphone.

### **1938**

- Benjamin B. Bauer of Shure Bros. engineers a single microphone element to produce a cardioid pickup pattern, called the Unidyne, Model 55. This later becomes the basis for the well known SM57 and SM58 microphones.
- Under the direction of Dr. Harry Olson, Leslie J. Anderson designs the 44B ribbon bidirectional microphone and the 77B ribbon unidirectional for RCA.
- RCA develops the first column loudspeaker array.

### **1939**

- Independently, engineers in Germany, Japan and the U.S. discover and develop AC biasing for magnetic recording.
- Western Electric designs the first motional feedback, vertical-cut disk recording head.
- Major Armstrong, the inventor of FM radio, makes the first experimental FM broadcast.
- The first of many attempts is made to define a standard for the VU meter.

### **1940**

- Walt Disney's "Fantasia" is released, with eight-track stereophonic sound.

### **1941**

- Commercial FM broadcasting begins in the U.S.
- Arthur Haddy of English Decca devises the first motional feedback, lateral-cut disk recording head, later used to cut their "ffrr" high-fidelity recordings.

## **1942**

- The RCA LC-1 loudspeaker is developed as a reference-standard control-room monitor.
- Dr. Olson patents a single-ribbon cardioid microphone (later developed as the RCA 77D and 77DX), and a "phased-array" directional microphone.
- The first stereo tape recordings are made by Helmut Kruger at German Radio in Berlin.

## **1943**

- Altec develops their Model 604 coaxial loudspeaker.

## **1944**

- Alexander M. Poniatoff forms Ampex Corporation to make electric motors for the military.

## **1945**

- Two Magnetophon tape decks are sent back to the U.S. In pieces in multiple mailbags by Army Signal Corps Major John T. (Jack) Mullin.

## **1946**

- Webster-Chicago manufactures wire recorders for the home market.
- Brush Development Corp. builds a semiprofessional tape recorder as its Model BK401 Soundmirror.
- 3M introduces Scotch No. 100, a black oxide paper tape.
- Jack Mullin demonstrates "hi-fi" tape recording with his reconstructed Magnetophon at an Institute of Radio Engineers (IRE) meeting in San Francisco.

## **1947**

- Colonel Richard Ranger begins to manufacture his version of a Magnetophon.

- Bing Crosby and his technical director, Murdo McKenzie, agree to audition tape recorders brought in by Jack Mullin and Richard Ranger. Mullin's is preferred, and he is brought back to record Crosby's Philco radio show.
- Ampex produces its first tape recorder, the Model 200.
- Major improvements are made in disk-cutting technology: the Presto 1D, Fairchild 542, and Cook feedback cutters.
- The Williamson high-fidelity power amplifier circuit is published.
- The first issue of *Audio Engineering* is published; its name is later shortened to *Audio*.

## 1948

- **The Audio Engineering Society (AES) is formed in New York City.**
- The microgroove 33-1/3 rpm long-play vinyl record (LP) is introduced by Columbia Records.
- Scotch types 111 and 112 acetate-base tapes are introduced.
- Magnecord introduces its PT-6, the first tape recorder in portable cases.

## 1949

- RCA introduces the microgroove 45 rpm, large-hole, 7-inch record and record changer/adaptor.
- Ampex introduces its Model 300 professional studio recorder.
- Magnecord produces the first U.S.-made stereo tape recorder, employing half-track staggered-head assemblies.
- A novel amplifier design is described by McIntosh and Gow.

## 1950

- Guitarist Les Paul modifies his Ampex 300 with an extra preview head for "Sound-on-Sound" overdubs.
- IBM develops a commercial magnetic drum memory.

## 1951

- The "hot stylus" technique is introduced to disk recording.
- An "Ultra-Linear" amplifier circuit is proposed by Hafler and Keroes.
- Pultec introduces the first active program equalizer, the EQP-1.
- The Germanium transistor is developed at Bell Laboratories.

## ***1952***

- Peter J. Baxandall publishes his (much-copied) tone control circuit.
- Emory Cook presses experimental dual-band left-right "binaural" disks.

## ***1953***

- Ampex engineers a 4-track, 35 mm magnetic film system for 20th-Century Fox's Christmas release of "The Robe" in CinemaScope with surround sound.
- Ampex introduces the first high speed reel-to-reel duplicator as its Model 3200.

## ***1954***

- EMT (Germany) introduces the electromechanical reverberation plate.
- Sony produces the first pocket transistor radios.
- Ampex produces its Model 600 portable tape recorder.
- G. A. Briggs stages a live-versus-recorded demonstration in London's Royal Festival Hall.
- RCA introduces its polydirectional ribbon microphone, the 77DX.
- Westrex introduces their Model 2B motional feedback lateral-cut disk recording head.
- The first commercial 2-track stereo tapes are released.

## ***1955***

- Ampex develops "Sel-Sync" (Selective Synchronous Recording), making audio overdubbing practical.

## ***1956***

- Les Paul makes the first 8-track recordings using the "Sel-Sync" method.
- The movie *Forbidden Planet* is released, with the first all-electronic film score, composed by Louis and Bebe Barron.

## 1957

- Westrex demonstrates the first commercial "45/45" stereo cutter head.

## 1958

- The first commercial stereo disk recordings appear.
- Stefan Kudelski introduces the Nagra III battery-operated transistorized field tape recorder, which with its "Neo-Pilot" sync system becomes the *de facto* standard of the film industry.

## 1959

- EMI fails to renew the Blumlein stereo patent. Hello - anybody home?

## 1961

- 3M introduces the first 2-track closed-loop capstan-drive recorder, the M-23.
- The FCC decides the FM stereo broadcast format.

## 1962

- The Society of Motion Picture and Television Engineers (SMPTE) sets the standard for the time code format.
- 3M introduces Scotch 201/202 "Dynarange," a black oxide low-noise mastering tape with a 4 dB improvement in s/n ratio over Scotch 111.

## 1963

- Philips introduces the Compact Cassette tape format, and offers licenses worldwide.
- Gerhard Sessler and James West, working at Bell Labs, patent the electret microphone.
- The Beach Boys contract Sunn Electronics to build the first large full-range sound system for their rock music concert tour.

## 1965

- The Dolby Type A noise reduction system is introduced.
- Robert Moog shows elements of his early music "synthesizers."
- Eltro (Germany) makes a pitch/tempo shifter, using a rotating head assembly to sample a moving magnetic tape.
- Herb Alpert and the Tijuana Brass tour with a Harry McCune Custom Sound System.

## 1967

- Richard C. Heyser devises the "TDS" (Time Delay Spectrometry) acoustical measurement scheme, which paves the way for the revolutionary "TEF" (Time Energy Frequency) technology.
- Altec-Lansing introduces "Acousta-Voicing," a concept of room equalization utilizing variable multiband filters.
- Elektra releases the first electronic music recording: Morton Subotnick's *Silver Apples of the Moon*.
- The Monterey International Pop Festival becomes the first large rock music festival.
- The Broadway musical *Hair* opens with a high-powered sound system.
- The first operational amplifiers are used in professional audio equipment, notably as summing devices for multichannel consoles.

## 1968

- CBS releases "Switched-On Bach," Walter (Wendy) Carlos's polyphonic multitracking of Moog's early music synthesizer.

## 1969

- Dr. Thomas Stockham begins to experiment with digital tape recording.
- Bill Hanley and Company designs and builds the sound system for the Woodstock Music Festival.
- 3M introduces Scotch 206 and 207 magnetic tape, with a s/n ratio 7 dB better than Scotch 111.

## ***1970***

- The first digital delay line, the Lexicon Delta-T 101, is introduced and is widely used in sound reinforcement installations.
- Ampex introduces 406 mastering tape.

## ***1971***

- Denon demonstrates 18-bit PCM stereo recording using a helical-scan video recorder.
- RMS and VCA circuit modules introduced by David Blackmer of dbx.

## ***1972***

- Electro-Voice and CBS are licensed by Peter Scheiber to produce quadraphonic decoders using his patented matrixes.

## ***1974***

- D. B. Keele pioneers the design of "constant-directivity" high-frequency horns.
- The Grateful Dead produce the "Wall of Sound" at the San Francisco Cow Palace, incorporating separate systems for vocals, each of the guitars, piano and drums.
- 3M introduces Scotch 250 mastering tape with an increase in output level of over 10 dB compared to Scotch 111.
- DuPont introduces chromium dioxide (CrO<sub>2</sub>) cassette tape.

## ***1975***

- Digital tape recording begins to take hold in professional audio studios.
- Michael Gerzon conceives of and Calrec (England) builds the "Soundfield Microphone," a coincident 4-capsule cluster with matrixed "B-format" outputs and decoded steerable 2- and 4-channel discrete outputs.
- EMT produces the first digital reverberation unit as its Model 250.
- Ampex introduces 456 high-output mastering tape.

## ***1976***

- Dr. Stockham of Soundstream makes the first 16-bit digital recording in the U.S. at the Santa Fe Opera.

## ***1978***

- The first EIAJ standard for the use of 14-bit PCM adaptors with VCR decks is embodied in Sony's PCM-1 consumer VCR adaptor.
- A patent is issued to Blackmer for an adaptive filter (the basis of dbx Types I and II noise reduction).
- 3M introduces metal-particle cassette tape.

## ***1980***

- 3M, Mitsubishi, Sony and Studer each introduces a multitrack digital recorder.
- EMT introduces its Model 450 hard-disk digital recorder.
- Sony introduces a palm-sized stereo cassette tape player called a "Walkman."

## ***1981***

- Philips demonstrates the Compact Disc (CD).
- MIDI is standardized as the universal synthesizer interface.
- IBM introduces a 16-bit personal computer.

## ***1982***

- Sony introduces the PCM-F1, intended for the consumer market, the first 14- and 16-bit digital adaptor for VCRs. It is eagerly snapped up by professionals, sparking the digital revolution in recording equipment.
- Sony releases the first CD player, the Model CDP-101.

## ***1983***

- Fiber-optic cable is used for long-distance digital audio transmission, linking New York and Washington, D.C.

## ***1984***

- The Apple Corporation markets the Macintosh computer.

## ***1985***

- Dolby introduces the "SR" Spectral Recording system.

## ***1986***

- The first digital consoles appear.
- R-DAT recorders are introduced in Japan.
- Dr. Gunther Theile describes a novel stereo "sphere microphone."

## ***1987***

- Digidesign markets "Sound Tools," a Macintosh-based digital workstation using DAT as its source and storage medium.

## ***1990***

- ISDN telephone links are offered for high-end studio use.
- Dolby proposes a 5-channel surround-sound scheme for home theater systems.
- The write-once CD-R becomes a commercial reality.
- 3M introduces 996 mastering tape, a 13 dB improvement over Scotch 111.

## ***1991***

- Wolfgang Ahnert presents, in a binaural simulation, the first digitally enhanced modeling of an acoustic space.
- Alesis unveils the ADAT, the first "affordable" digital multitrack recorder.
- Apple debuts the "QuickTime" multimedia format.
- Ampex introduces 499 mastering tape.

## ***1992***

- The Philips DCC and Sony's MiniDisc, using digital audio data-reduction, are offered to

consumers as record/play hardware and software.

- The Nagra D is introduced as a self-contained battery-operated field recorder using Nagra's own 4-channel 24-bit open-reel format.

### ***1993***

- In the first extensive use of "distance recording" via ISDN, producer Phil Ramone records the "Duets" album with Frank Sinatra.
- Mackie unveils the first "affordable" 8-bus analog console.

### ***1994***

- Yamaha unveils the ProMix 01, the first "affordable" digital multitrack console.

### ***1995***

- The first "solid-state" audio recorder, the Nagra ARES-C, is introduced. It is a battery-operated field unit recording on PCMCIA cards using MPEG-2 audio compression.
- Iomega debuts high-capacity "Jaz" and "Zip" drives, useful as removable storage media for hard-disk recording.

### ***1996***

- Record labels begin to add multimedia files to new releases, calling them "enhanced CDs."
- Experimental digital recordings are made at 24 bits and 96 kHz.

### ***1997***

- DVD videodiscs and players are introduced. An audio version with 6-channel surround sound is expected to eventually supplant the CD as the chosen playback medium in the home.

### ***1998***

- The Winter Olympics open with a performance of Beethoven's "Ode to Joy," played and sung by synchronizing live audio feeds from five continents with an orchestra and conductor at the Olympic stadium in Nagano, Japan, using satellite and ISDN technology.
- Golden Anniversary celebration held in New York on March 11, the exact date of the first AES meeting in 1948, with ten of the original members present.

- MP-3 players for downloaded Internet audio appear.

## ***1999***

- Audio DVD Standard 1.0 agreed upon by manufacturers.

---

[Back to Documents](#)



---

## Historical Interviews, Talks, and Articles

---

One of the charges of the AESHC is to "record oral histories (sound only, sound with video, and/or sound with still photographs) of important figures in the history of audio engineering." We encourage all of our members to record oral histories, and submit them (or transcripts of them) to the Chair of the AESHC [John G. \(Jay\) McKnight](#) for consideration for inclusion on this website. Currently, only a few are available, and these are the following papers and talk transcripts:

- [Ray Dolby's personal tour of the audio industry and how it changed from analog to digital](#), a transcript of his 1992 Heyser lecture
- [Cyril Francis, Recording Engineer \(Acoustic and Electrical Recordings\) with Parlophone in England, 1926 to 1936](#)
- [Sel-Sync and the Octopus](#), Ross Snyder's article from the [ARSC Journal](#), v. 34, no. 2 (fall 2004): 209-213.
- [The Education and Tribulations of a Precursory Disk Recording Engineer](#) by Robert J Callen -- Audio engineering in 1925...1928, AES Preprint 794 from the year 1971

---

[Back to the main page](#)



# 3M Analog Magnetic Tape Technology

---

## Canonical List of 3M Tape Types

- [Introduction to 3M Reel to Reel Audio Tape Types](#)
- [List of 3M Reel to Reel Audio Tape Types](#)
- [List of 3M Reel to Reel Audio Tape Types in Microsoft Excel Format](#)

## Other 3M Tape Documents

- [Analog Audio Mastering Tape Print-Through \(Technical Bulletin A011194\)](#)
- [Splicing and recording tips inside of a 'Scotch' recording tape box, around 1970](#)

## 3M Sound Talk Index and PDF Scans of all 46 Bulletins

- [Sound Talk index in HTML format](#) and links to the PDF scans of the Bulletins themselves
- [Sound Talk index in ASCII format](#)

---

[Back to the main page](#)

# Introduction to 3M Audio Open Reel Tape List

Copyright © 2000 by Delos A. Eilers. Reproduction in any form prohibited without the written permission of Delos A. Eilers.

During the winding down of the magnetic tape business at 3M, it struck me that audio archivists and historians would probably appreciate a listing of the products 3M had made over the years, and their basic characteristics. There have been a number of serious aging problems with several magnetic tapes of 3M and other brands, which had led to concern about all magnetic tape records. However, most products sold have not had aging problems when properly stored.

This chart was devised to help the archivist determine which 3M Audio Open Reel product they have. Then, knowing the tape types they possess, they can monitor their libraries by product types. When a given product type of a given vintage is identified with a problem, those recordings in the library of the same binder type and vintage can be more closely checked for aging problems.

The parameters chosen for this database were tape type/number, estimated year of introduction, binder type, oxide coating color, base material, base caliper, oxide caliper, total caliper, back treated (yes or no), remanence, coercivity, retentivity, and any special feature worth noting. These parameters were chosen because they can either be seen or measured to help determine which 3M product a tape might be when it is not labeled or identified. Electro-magnetic (recording performance) data was not shown since it is not absolute, but dependent upon a recording machine operating conditions and a reference tape. Frankly, a common set of test conditions and reference tapes was not used over the nearly 50 years of tape design, thus technical data sheet electro-magnetic data from one era can not be compared to another. Even the basic magnetic properties are not exactly comparable from the old technical data sheets to new as refinements and improved accuracy in measurement changed the data shown over the years. This list is our best guess of the physical and basic magnetic data that we could assemble from technical data sheets kept over the years.

A word about binder type shown. The binder chemistries used to make magnetic tapes went through many changes over the years. We also used different identification systems over the years to describe them. Sometimes they were reflections of the binder chemistry make-up. Sometimes they were the initials of the developer(s). Sometimes they were just a code name. Rather than use this coded jargon, I chose to merely group them into their basic types and use a letter designation starting with A going through the alphabet till I ran out of group types. The binder groups were never referred to as A, B, C, etc. within 3M.

Lastly the earliest tape data is very "sketchy" as there is little historical documentation in existence today. Talking recently with one of the chemists from that era, he admitted that the changes made to a given tape type in those days often were significant. As we learned more about making tapes and found improvements, they were incorporated in the products. We didn't really appreciate what a standardized product type/name/number meant. In some of our early Sound Talk bulletins we identify products as the same tape type but also show different internal numbers (3RBA, 4RBA, 5RBA, 6RBA). These were all product "improvements"/changes made to the same basic tape type. In some cases these changes were quite significant on performance. By the early 50's this practice was stopped, so that significant performance changes were accompanied by a new product number

identification.

We hope this database is of interest and is useful to you. While I believe it to be accurate, there may be some minor errors. Some of these can be "found" looking at different versions of technical data sheets for the same product. These "errors" are really slight changes in measurement accuracy or in product "fine tuning" that occurred over the years. For this database I had to pick one number to represent the product characteristic.

The data file, "3M Audio Open Reel Tapes", is available for you to download in two formats:

First, as an [HTML file](#) that you can read on your browser. (On the Netscape 4 browser, "ctrl-]" will make the screen text larger.) You can print out this file in "landscape" orientation on a letter-size page.

Second, as a [Microsoft Excel \(xls\) file](#) that you can import into many spreadsheets ("MS Excel", a part of the MS Office package; or "Quatro Pro", a part of the Corel "WordPerfect" package; etc). From these spreadsheets, you can also use "save as" to save the data in other formats such as Dbase 3 or 4, or tab- or comma-delimited text files. By putting the data into a database program, you can do various useful things such as re-sorting the data (which is now in Product Number sequence) by introduction date, or by binder type, etc; or add you own columns with the data in SI units, etc. Finally, you can import the xls file directly into some word processors (e.g., WordPerfect).

Delos (Del) A. Eilers  
Senior Technical Service Specialist  
3M Company

2000-02-01

---

Technical Definitions and SI Units (note by J. McKnight):

"Caliper, mils", in this sense, means "thickness". A "mil" is one-thousandth of an inch; for thickness in micrometers [ $\mu\text{m}$ ], multiply the number of mils by 25.4 .

"Remanance, flux lines" is the saturation remanance flux from a quarter-inch wide tape. A "line" is an alternate name for a maxwell, equal to 10 nanowebers. For saturation fluxivity in nanowebers per meter [ $\text{nWb/m}$ ], multiply the number of flux lines per quarter inch by 1600.

For coercivity in kiloamperes per meter [ $\text{kA/m}$ ], multiply the number of oersteds by 0.08 .

"Retentivity, gauss" is the saturation remanance flux density; for retentivity in millitesla [ $\text{mT}$ ], divide the number of gauss by 10.

---

[Back to AES Historical Committee documents page](#)

# 3M Audio Open Reel Tapes

Copyright © 2000 by Delos A. Eilers. Reproduction in any form prohibited without the permission of Delos A. Eilers.

3M												
Audio Reel												
Tapes												
Product Number	Est. Yr. Intro'd	Binder Type	Oxide Ctg Color	Base Material	Base Caliper	Oxide Caliper	Total Caliper	Back Treated	Remanence	Coercivity	Retentivity	Special Feature
					mils	mils	mils		flux lines	Oersteds	Gauss	
100	1947	A	black	paper	1.5	0.7	2.2	no	0.6	320	550	
101	1948	B	brown	paper	1.5	0.5	2	no	0.64	270	900	
102 (111AM)	1953	B	brown	polyester	1.45	0.5	1.95	no	0.64	270	900	
111 (111A)	1948	B	brown	acetate	1.42	0.5	1.92	no	0.64	270	900	
112	1948	B	brown	acetate	1.42	0.7	2.12	no	0.45	170	450	tape for Ampex 200A
120 (120A)	1953	B	dark green	acetate	1.42	0.65	2.07	no	1.1	260	1170	
122 (120AM)	1954	B	dark green	polyester	1.45	0.65	2.1	no	1.1	260	1170	
131	1958	B	brown	acetate	1.42	0.4	1.82	no	0.64	270	1000	
138	1958	B	brown	polyester	1.45	0.4	1.85	no	0.64	270	1000	
139	1958	B	brown	polyester	0.92	0.4	1.32	no	0.64	270	1000	
140	1960	B	brown	acetate	1.42	0.35	1.77	no	0.64	260	1100	
141	1960	B	brown	polyester	1.45	0.35	1.8	no	0.64	260	1100	
142	1960	B	brown	polyester	0.92	0.35	1.27	no	0.64	260	1100	
144	1960	B	brown	polyester	0.5	0.35	0.85	no	0.64	260	1100	
150	1954	B	red/brown	polyester	0.92	0.35	1.27	no	0.64	260	1100	
151	1964	B	brown	polyester	0.92	0.43	1.45	*	0.64	270	950	back lubricated tape
152	1964	B	brown	polyester	0.92	0.43	1.92	*	0.64	270	950	double ctd lube tape
153	1965	D	black	polyester	0.92	0.43	1.43	*	0.64	265	940	back lubricated tape
154	1972	E	brown	polyester	0.92	0.27	1.3	*	0.5	280	1150	back lubricated tape
156	1972	E	brown	polyester	0.92	0.27	1.3	*	0.5	280	1150	back lubricated tape
157	1973	E	brown	polyester	0.75	0.38	1.23	*	0.64	320	1050	back lubricated tape

## 3M Audio Open Reel Tapes

158	1973	E	brown	polyester	0.92	0.27	1.3	*	0.47	320	1050	back lubricated tape
175	1965	C	black	polyester	1.42	0.5	1.92	no	0.72	275	900	
176 (211, 228)	1972	E	brown	polyester	1.42	0.4	1.82	no	0.64	325	1025	
177 (212, 229)	1972	E	brown	polyester	0.92	0.4	1.32	no	0.64	325	1025	
178 (213)	1972	E	brown	polyester	0.5	0.4	0.9	no	0.64	325	1025	
179 (214)	1972	E	brown	polyester	0.43	0.24	0.67	no	0.39	325	1025	
186	1989	L	brown	polyester	1.46	0.36	1.82	no	0.68	365	1190	
190	1954	B	red/brown	acetate	0.95	0.35	1.3	no	0.64	260	1100	
200	1957	B	red/brown	polyester	0.5	0.35	0.85	no	0.64	260	1100	
201	1962	C	black	acetate	1.42	0.51	1.93	no	0.64	315	790	
202	1962	C	black	polyester	1.42	0.51	1.93	no	0.64	315	790	
203	1962	C	black	polyester	0.92	0.51	1.43	no	0.64	315	790	
206	1969	D	black	polyester	1.42	0.56	2.08	yes	0.93	320	1050	
207	1969	D	black	polyester	0.85	0.56	1.51	yes	0.93	320	1050	
208	1971	D	black	polyester	1.42	0.4	1.9	yes	0.64	330	1000	
209	1971	D	black	polyester	0.92	0.4	1.4	yes	0.64	330	1000	
217	1982	H	brown	polyester	0.56	0.34	0.98	*	0.75	360	1400	back lubricated tape
219	1985	H	brown	polyester	0.79	0.38	1.25	*	0.75	360	1250	back lubricated tape
226	1979	H	brown	polyester	1.3	0.56	1.94	yes	1.25	360	1400	
227	1980	H	brown	polyester	0.79	0.56	1.43	yes	1.25	360	1400	
250	1974	F	brown	polyester	1.3	0.68	2.06	yes	1.3	380	1200	
265	1977	K	brown	polyester	0.79	0.15	1.04	yes	0.32	700	1350	digital tape
275	1984	L	dark brown	polyester	0.79	0.19	1.04	yes	0.45	710	1400	digital tape
282	1960	B	red/brn,blk/brn	polyester	1.42	0.39	1.85	no	0.64	260	1030	mag layer overcoat
283	1960	B	red/brown	polyester	0.92	0.39	1.35	no	0.64	260	1030	mag layer overcoat
290	1964	C	black	polyester	0.5	0.17	0.67	no	0.25	265	900	
294	1965	C	black	polyester	0.6	0.43	1.03	no	0.64	265	940	
295	1974	E	brown	polyester	0.43	0.18	0.68	yes	0.36	315	1260	logging tape
296	1974	E	brown	polyester	0.85	0.18	1.1	yes	0.36	315	1260	logging tape
311	1960	B	brown	pvc	1.42	0.38	1.8	no	0.63	260	1050	
806	1986	H	brown	polyester	1.42	0.4	1.9	yes	0.9	370	1400	
807	1986	H	brown	polyester	0.92	0.4	1.4	yes	0.9	370	1400	
808	1986	H	brown	polyester	1.46	0.32	1.86	yes	0.67	360	1320	

## 3M Audio Open Reel Tapes

809	1986	H	brown	polyester	0.92	0.32	1.32	yes	0.67	360	1320	
908	1993	I	brown	polyester	1.38	0.36	1.82	yes	0.9	365	1600	
966/986	1992	I	brown	polyester	1.38	0.49	1.95	yes	1.25	365	1600	
967	1992	I	brown	polyester	0.92	0.49	1.49	yes	1.25	365	1600	
996	1990	I	brown	polyester	1.42	0.63	2.13	yes	1.7	360	1700	
8135	1972	E	brown	polyester	0.43	0.2	0.7	yes	0.38	315	1250	logging tape
8136	1972	E	brown	polyester	0.85	0.2	1.12	yes	0.38	315	1250	logging tape
8206	1977	J	dark brown	polyester	0.79	0.2	1.05	yes	0.34	345	1050	logging tape
8207	1977	J	dark brown	polyester	0.43	0.2	0.69	yes	0.34	345	1050	logging tape
8265	1981	K	brown	polyester	0.92	0.16	1.14	yes	0.32	650	1250	digital tape
8614	1978	J	dark brown	polyester	0.79	0.2	0.99	no	0.34	345	1050	logging tape
Classic DP	1975	H	brown	polyester	0.42	0.48	0.98	yes	0.9	350	1200	
Classic LP	1975	H	brown	polyester	0.8	0.48	1.36	yes	0.9	350	1200	
Classic SP	1975	H	brown	polyester	1.42	0.48	1.98	yes	0.9	350	1200	
Master	1978	H	brown	polyester	0.8	0.48	1.36	yes	0.9	350	1200	
Master XS	1980	H	brown	polyester	0.79	0.56	1.43	yes	1.25	360	1400	
rev.3, 2/9/00												
scotch												

# A Selected Bibliography of Histories of Magnetic Tape Sound Recording

Compiled by Jay McKnight, Magnetic Reference Lab, [mrltapes@flash.net](mailto:mrltapes@flash.net)

The history of magnetic sound recording and magnetic tape has been told in numerous articles in journals and magazines, and in several books. I have listed below the articles that I have in my collection that I consider to be primary references -- that is, those either written by "those who were there", or written by those who did extensive research and interviews with them. These articles and books in turn have many further references to other articles on this subject. This bibliography makes no pretense at being "complete" -- it is just my answer to the question "what books and papers have you read that tell the story of magnetic tape recording". This bibliography does not include books and papers dealing primarily with technical aspects of magnetic recording.

In the present context, I take "History" to mean a tale or story, primarily from the viewpoint of its author. This is to say that I make no judgment if two authors have conflicting views of the same event, because if I excluded every historical paper that might have some minor difference of opinion, we would have no history at all!

The papers in the AES Journal, and the books and video tapes (unless otherwise stated) are available for purchase from the [AES](#).

## Papers (in chronological order)

**John T Mullin**, "Creating the Craft of Tape Recording", High Fidelity Magazine, Vol ?, Nr ?, pp 62...67 (1976 April). Mullin's own telling of the story of his finding the Magnetophons in Germany, sending them back home to the US, showing them to the IRE meeting, working with Bing Crosby and also working with Ampex.

**Harold Lindsay**, "Magnetic Recording, Part 1", db Magazine, Vol 11, Nr 12, pp 38...44 (1977 December). History and development of the Ampex Model 200A. [5 references]

**Harold Lindsay**, "Magnetic Recording, Part 2", db Magazine, Vol 12, Nr 1, pp 40...44 (1978 January). More on the Ampex Model 200A. History and development of the Models 300, 3200 (duplicator), Todd-AO, consumer products, 4-track stereo tapes, multichannel recorders, acquisition of "Irish tape", and ATR-100. [18 references]

**Peter Hammar and Don Ososke**, "The Birth of the German Magnetophon Tape Recorder 1928...1945", db Magazine, Vol 16, Nr 3, Cover photo, pp 47...52 (1982 March).

**Peter Hammar**, "In Memoriam: Harold Lindsay", AES Journal Vol 30, Nr 9, pp 691, 692 (1982 Sept).

**Friedrich Karl Engel**, "1888-1988: A Hundred Years of Magnetic Sound Recording", AES Journal, Vol 36, Issue 3, pp 170...178 (1988) [18 references]

Abstract: In the past, the essay "Some Possible Forms of Phonograph" by the American engineer Oberlin Smith, dating from 1888, has been regarded merely as a first indication of the possibility of electromagnetic sound recording. A recently discovered reader's letter proves that Smith constructed a unit with functional transducers, which could at least be used for experimental purpose, and is therefore the inventor of magnetic sound recording technique.

**Heinz H. K. Thiele**, "Magnetic Sound Recording in Europe up to 1945", AES Journal, Vol 36, Issue 5, pp: 396...408 (1988) [42 references]

Abstract: The 50th Anniversary of the Magnetophon. This jubilee was an occasion for audio engineers to look back on how the tape recorder was born and to see what has become of it. Its evolution during the years from 1888 to almost 1945 is discussed.

**Friedrich Karl Engel**, "Magnetic Tape -- From the Early Days to the Present", AES Journal, Vol 36, Issue 7/8, pp 606...616 (1988) [6 references]

Abstract: Cooperation between AEG and BASF -- then IG Farben, Ludwigshafen works -- concerning the Magnetophon sound-recording system began in the fall of 1932. Formulation of the first tapes and their production are described: first carbonyl iron (1933), then iron oxide  $Fe(3)O(4)$  (1936), and finally iron oxide  $Fe(2)O(3)$  (1939) on a cellulose acetate base, the latter oxide mixed in since 1943 and, since 1945, coated on polyvinyl chloride film. A concise survey from the mid-1950s traces development leading up to modern analog tape technology.

**Rudolph Müller**, "On Improvements of Magnetic Tape Shown by Measurements on Early and Newer Tapes", AES Journal, Vol 36, Issue 10, pp 802...820 (1988)

Abstract: Magnetic tapes from the early days of magnetic tape recording are investigated using today's measuring methods. Comparing mechanical, magnetic, and electroacoustic properties of studio tapes, the steps in development that had to be taken to reach the high-quality standard required for modern analog recording are outlined. In addition this investigation shows possible areas for the further development of the magnetic tape system.

**Peter Hammar**, "Jack Mullin: The Man and His Machines", AES Journal, Vol 37, Issue 6, pp 490...512 (1989)

An illustrated review of "The John T Mullin Collection: The History of Sound Recording", as exhibited at the 85th AES Convention in Los Angeles, in 1988 November. The early history of magnetic recording is summarized, then Mullin's involvement with bringing the German Magnetophon to the US, demonstrating it to US engineers, and working with Bing Crosby and with Ampex.

## Books

**S.J. Begun**, "Magnetic Recording", Rinehart Books Inc, 1949, now out of print. The very first

magnetic recording book (at least that I know of). The history, theory, equipment and applications of magnetic recording as it was in 1949. Steel-wire, steel-tape, homogeneous and coated tape media and recorders. Many photos of historic and "modern" (in 1949) equipment, and many references to the literature up to 1949.

**Marvin Camras** (ed), "Magnetic Tape Recording", Van Nostrand Reinhold Co, (1985), now out of print. Camras includes interesting original historical introductions ("Editor's Comments") to these historically-important *technical* (as opposed to historical) papers on magnetic tape recording.

**Marvin Camras**, "Magnetic Recording Handbook", Van Nostrand Reinhold Co, (1988), now out of print. Altho this is primarily a technical book, it begins with a chapter "Magnetic Recording History and Early Recorders". The main chapters have historical introductions and include products of historical interest. And it ends with "Bibliography and References" (13 pages); appendix A "Highlights of Magnetic Recording Development" (41 pages, including 28 illustrations); and appendix B with photos of 39 individuals who contributed to magnetic recording.

**Daniel, Mee and Clark** (editors), [Magnetic Recording, The First 100 Years](#) (1999) (available now from the IEEE Press)

**S. J. Begun** (M. Clark, ed), [Magnetic Recording: The Ups and Downs of a Pioneer](#) The memoirs of Semi Joseph Begun (2000) (available now from the AES)

## Video Tapes

[An Afternoon with Jack Mullin](#) (available now from the AES)

**Dale Manquen** (producer), [A Chronology of American Tape Recording](#) (8 hours) (available now from the AES)

Draft of 2001-06-19 12:48

Revised 2001-07-24 16:22

Revised 2002-09-25: Added Begun 1949 book; pages for Hammar and Ososke, 1982.

Revised 2005-08-11: Changed URL for ordering Daniel, Mee, and Clark book.



## ORAL HISTORY PROJECT

Project Leader: Irv Joel

The Guidelines of the AES Historical Committee include:

2.13 Record oral histories (sound only, sound with video, and/or sound with still photographs) of important figures in the history of audio engineering.

The Project is responsible for coordinating and recording interviews which will preserve a first hand recording and/or text transcription of the recollections of persons involved with important audio engineering inventions and/or events, for future generations.

Our main goal at this moment is to record the interviews. Because we have only a small volunteer staff, we will edit these recordings as time permits into a form that we can make available. At this time, none of the interviews are available.

Here is the [List of Completed Interviews and Interview Candidates](#).

The [Interview Guidelines](#), based on our experience, recommend procedures for recording oral interviews. The [Release Form](#) can be printed from this pdf file. Comments and suggestions for improvements should be sent to the project leader, [Irv Joel](#).

revised 2001-02-16 IJ, jm

2001-04-25 jm

2003-03-03 jm

2003-03-04 jm

2004-10-12 jkc

2001-11-14

## Stanford Acquires Ampex History Collections

The Stanford University Libraries have acquired historical and archival collections of Ampex Corporation, one of Silicon Valley's pioneering technology companies and for more than five decades an industrial leader in magnetic recording and data storage. These historical collections include the artifact collection of the former Ampex Museum of Magnetic Recording, an extensive photographic archive of more than 200,000 images, documentation and product files, and Ampex publications. These materials will provide scholars with a major resource in the history of audio and video recording technology and the early development of Silicon Valley.

According to Michael A. Keller, Stanford University Librarian, "This is a brilliant addition to our holdings supporting the study of the history of technology in the second half of the 20<sup>th</sup> century. It also provides a wonderful counterpoint to our Archives of Recorded Sound. Between them, we have a wealth of material covering both recorded content and technology." Keller, a former music librarian, added that Ampex has been the pivotal influence in recording technology since its creation in 1944.

The collections were given to Stanford by Ampex Data Systems Corporation, of Redwood City. Acquisitions costs, including shipping, storage and preliminary processing, have been underwritten by a gift from Dolby Laboratories, Inc., of San Francisco. (Recording innovator Ray Dolby began his career in recording technology at Ampex.)

The Ampex collections, currently in storage, will require several years of curatorial effort to organize, stabilize, describe and re-house before the collection is fully accessible for research. Henry Lowood, Curator for History of Science & Technology Collections and head of the Stanford and the Silicon Valley Archives commented, "We are still digesting the collection, which we received in several parts over the summer, and right now our efforts are going to assessing what we have. Before long we will provide a website for the collection, which will be a good place to look for more information about access to it." [As of 2002-01-24, the website is not yet available.] Until the collection has been processed, the Libraries will be unable to provide public access to the collection or fulfill information requests about it. [In, 2002 January, NPR aired a [story about the collection](#) by Peter Jon Shuler (KQED).]

The Ampex Electric and Manufacturing Company was founded in 1944 by Alexander M. Poniatoff in San Carlos, Calif. Four years later, the American Broadcasting Company (ABC) used an Ampex Model 200 audio recorder for the "The Bing Crosby Show," the first tape-delayed radio broadcast in the United States. In March, 1956, Ampex demonstrated the first videotape recorder, the VRX-1000, at a meeting of the National Association of Radio and Television Broadcasters in Chicago. Over the next decades, Ampex introduced many innovations in audio, video, and data recording and storage. The museum collection includes recording devices and media, dating back to the 1940s, and spans the critical years of development of video recording.

The photographic archives and audio recordings date back to the 1940s and document not only the development of technology and products, but also trace close links between Ampex and the emerging television networks, electronic media and the entertainment industry. Stanford hopes eventually to display some of the key artifacts in the Libraries' exhibit spaces, though specific plans are not firm yet.

The Stanford and Silicon Valley Archives <[svarchive.stanford.edu](http://svarchive.stanford.edu)> seek to identify, preserve, and make available the documentary record of science, technology, and related business and cultural activities in Silicon Valley. It has been a dynamic and strong component of the Stanford Libraries' collecting program since 1983 and has preserved the papers of Douglas Engelbart, William Shockley, Frederick Terman, Donald Knuth, George Forsythe, John McCarthy, Edward Feigenbaum, Charles A. Rosen, and many others active or influential in the development of computers and computing or other aspects of the Silicon Valley economy. The archives also have preserved the Stephen Cabrinety Collection in the History of Microcomputing and the historical collections of Apple Computer, as well as historical records of Fairchild Semiconductor, Interval Research Corporation and the American Association for Artificial Intelligence, to name only a few examples. Oral histories are also an important component of the Stanford and Silicon Valley Archives Project, such as "Silicon Genesis: Interviews with Semiconductor Pioneers."

Stanford University Libraries & Academic Information Resources supports the teaching, learning and research mandates of the university through delivery of bibliographic and other information resources and services to faculty, student and staff. It is tackling the challenges of the digital age, especially pertaining to scholarly communication and research libraries, while continuing the development, preservation and conservation of its extensive print, media, manuscript and technological artifact collections.

CONTACT: Andrew Herkovic, University Libraries (650) 224-3711; [andrew.herkovic@stanford.edu](mailto:andrew.herkovic@stanford.edu)

COMMENT: Henry Lowood, Curator for History of Science and Technology Collections (650) 723-4602; [lowood@stanford.edu](mailto:lowood@stanford.edu)

**Relevant Web URLs:**

[svarchive.stanford.edu](http://svarchive.stanford.edu)

2002-01-24: html format, corrections per H Lowood, add Shuler story about collection. J Mcknight

**Volume 20, Number 6**  
**Nov 1996**



## **NARA Conference on Preserving Tapes & Disks, March 1996**

### **Facts and Advice from the Speakers**

by Ellen McCrady

With apologies to the speakers, I will use an unconventional but time-saving method of reporting this conference: simply transcribing from my notes the most important or interesting facts and advice, without attribution. If I get anything wrong, or if anyone needs to know who made a given statement, please call me at the office, 512/929-3992, so that I can correct or supply the information.

The conference was held March 14. Speakers were Peter Z. Adelstein (on standardization activities), Fred Layne (tape), John Van Bogart (tape storage), John Powers (tape restoration), Fynette Eaton (The Archives' Center for Electronic Records), Douglas Stinson (CD lifetimes), Barry Roginski (optical disk migration) and Chris Cain (backup tape).

The optical disk people are doing a good job of writing standards.... The Tape Head Interest Committee (THIC) is producing documents on the permanence of tape.

Tape binders are mainly polyurethane, which have evolved significantly, and are more lasting now. Tape pack is the most important factor in archivability.... Tape 456 is Quantegy Corporation's most permanent tape.... Archive the software and generational protocol with the tape.... Japanese competitors of American tape manufacturers are very secretive, which is an obstacle to cooperative formulation of standards.

"Sticky shed" syndrome is caused by breakdown of the binder. Ligamers migrate to the surface, the tape squeaks and stops. When this was discovered in 1986 or '87, a heat treatment was developed in a crash program: bake them at 125°F for eight hours, which recombines the ligamers. The effect lasts 30 days at most, but the treatment can be repeated at least 20 times. Before the ligamers break down again, copy the tape. [Note: Most specialists recommend against use of this method by untrained people.]

Oxide systems last longer than metal particles under adverse conditions.... Analog media probably

last longer than digital.... In studies comparing different recording media, life expectancy needs to be qualified by temperature, RH, failure criteria, percent surviving, and confidence interval.... Until we have standards for magnetic tape, write specs into your purchase order; ask for the life expectancy of the product and how they calculate it.... Keep the environment clean and dustfree. Bit densities are getting higher, so smaller debris does more damage.... Store tails out to preserve a good tape pack. Rewinding periodically is not as important now that big reels aren't used as much.

In the LBJ Library, there is a wide variety of tape formats of varying quality. They have lots of spare parts, equipment and technicians. It took them six months to find a machine that would play a databelt. They finally got one on loan from IBM's museum.

Make sure your machine works before you put an old tape in it... Know what's on the tape. The more descriptive information on the tape, the better. Transfer the data when you rerecord. A lot of the tapes in the LBJ Library are acetate, which is OK, but as soon as it starts to deteriorate, it goes rapidly. If you smell it, copy it. Store tapes with vinegar syndrome separately from the others.

In the 1970s, the LBJ Library had its obsolete tape copied, using a low-bid service company that copied them onto tapes that are now completely obsolete. Now the Library is doing things right, recording first on multiple DAT (digital audio tape) copies, then doing an analog master.

The National Archives' Center for Electronic Records has been in existence for 30 years.... The Archival Preservation System (APS) was formed with several goals in mind, including: to retain control of the media and do their own preservation, and to widen the choice of formats for receiving and sending records. They will get nine 586 Intel processors, to handle all formats.

Archival Electronic Records Inspection and Control (AERIC) was organized in 1992.

**NEVER ACCEPT BACKUP TAPES AS ARCHIVAL MEDIA.**

The Infoguard Protection System is Kodak's scratch-tolerant top coating for CD-ROMs. "This is the first archival storage system that's made money for anyone."

Only 1 to 3% of government records are retained forever. All media are accepted by the Archives. The holdings of the Archives are growing exponentially. In 1994, NARA started accepting CD-ROMs (data files) on a trial basis. Optical *tape* is new, and could be an archival medium; but there is no commercial product yet.

In 1991, the Archives issued the first guidelines for state and local governments on long-term access to electronic data. It emphasized that this was the users' responsibility, not the vendors', who say "Don't worry-we'll take care of that," and "This will save you money." A survey of state and local government records systems showed that many systems were proprietary. They were advised to put the software code in the bank.

NARA learned 13 lessons during a large conversion project (to get copies of the Bush tapes so as to make them available to the public). Some of the lessons were: Do not postpone conversion. Expect it

to be resource-intensive. Upgrade system components at the same time. Select system components that meet standards. Get any available hardware or software ever used with the files. Clean up files before you convert. Perform routine backups as you convert.



[Copyright 2004 Abbey Publications, Inc. All rights reserved.](#)



---

[\[Search all CoOL documents\]](#) [\[Feedback\]](#)

**Page last changed: August 03, 2004 11:43**

[Data Recovery](#)[Disk Recovery](#)[RAID Data Recovery](#)[Tape Data Recovery](#)[Tape Recovery](#)[Data Recovery Service](#)[Data Recovery Software](#)[Hard Drive Recovery](#)[Lost Data](#)[Windows Recovery](#)[UNIX Data Recovery](#)[F A Q](#)

## Quick data recovery .com

Welcome to Quick Data Recovery .com, your 100% free informational resource designed to answer all your questions regarding the recovery of lost data. Every day we ask more of our computer technology. We use it to store our addresses, our phone numbers, our financial records, our work, etc. Basically, we rely on our computers a great deal; probably more than we should. (Online fraud is one of the fastest growing criminal enterprises, due to the prevalence of Internet hackers and the seemingly infinitesimal number of credit card purchases that are made on a daily basis.)

Anyone who has ever suffered through the experience of losing some kind of needed data from their computer knows all too well the need for creating up-to-date backups of all important information that is stored on their computer(s). As it is, however, only the most expensive, cutting edge systems have the capability to provide up-to-the-second data backups. The reality for the rest of us is that if we even bother to make a backup, it will generally be a one-time thing. Fortunately, technology is at a point where it boasts that virtually no lost data is irretrievable.

Professional data recovery companies boast an ability to retrieve lost data resulting from:

- Virus/Worm attack
- Human error
- Fire/Smoke/Chemical/Water damage
- After fdisk
- Formatting errors
- Power failure
- Software/Hardware failure
- Overwritten data

For more information on how data recovery specialists go about the process of retrieval, or to get more information on the specialists themselves, take some time to go through our informational site and comprehensive list of partners.

What's the first thing I should know about [DATA RECOVERY ?](#)

Learn more about the specialists who provide [DATA RECOVERY SERVICES](#)



[Back to Home](#) ◀

▶ [Products](#)

▶ [Support & Downloads](#)

▶ [Where to Buy](#)

▶ [About Imation](#)

◀ [Back to Home](#)

## Welcome to Imation

**We are currently in the process of improving our site.**

You may need to update any bookmarks you have to pages within the Imation site. Some sections are still under construction, so please check back next week. In the meantime, [click here](#) to be directed to the home page.



◀ [Back to Home](#)

## Welcome to Imation

**We are currently in the process of improving our site.**

You may need to update any bookmarks you have to pages within the Imation site. Some sections are still under construction, so please check back next week. In the meantime, [click here](#) to be directed to the home page.

## National Media Lab - Overview of Archival Stability of Recording Media



Tech Stuff

Based on Magnetic Properties		Magnetic Pigment Type				MP	ME	MP
Effects of Temperature	$\gamma$ -Fe <sub>2</sub> O <sub>3</sub>	BaFe	CrO <sub>2</sub>	Co- $\gamma$ -Fe <sub>2</sub> O <sub>3</sub>	(data)	(video)	(video)	
T= 20 C / RH = 50%								
T= 30 C / RH = 50%								
T= 40 C / RH = 50%								
Effects of Humidity	$\gamma$ -Fe <sub>2</sub> O <sub>3</sub> is incompatible for High Density Recording							
T= 20 C / RH = 85%								
T= 30 C / RH = 85%								
Effects of Pollution	$\gamma$ -Fe <sub>2</sub> O <sub>3</sub> is incompatible for High Density Recording							
T= 20 C / RH = 50% >= Battelle Class II								
<b>Product:</b>	9-Track Computer Tape Low Density Diskette	4 MB Diskette Floptical Hi8 video	3480 3490	D-1 High Density Diskette	D-2 R-DAT 8mm Data 4mm DDS	Hi8 video	Hi8 video	

**KEY:**

- GOOD No corrosion or signal problems expected
- FAIR May be suitable if some signal loss and/or bit errors can be tolerated
- POOR Unsuitable for storage under these conditions

Best Performer  
 Worst Performer

Chart recommendations are specific to digital recording media. Based on changes observed in tape magnetic remanence and coercivity upon temperature/humidity accelerated aging and Battelle FMG laboratory and field tests conducted by the NML and others as of December, 1993. The above chart does not consider effects associated with binder and substrate instabilities.

P.O. Box 33015 St. Paul, MN 55133-3015  
 (612) 733-0468, Fax (612) 733-4340

[Magnetic Stability](#)

[Barium Ferrite \(BaFe\) Technical Data](#)

[Amplitude Comparison of BaFe vs. Metal Particle Tape](#)



# Magnetic Tape Storage and Handling

## A Guide for Libraries and Archives

**Dr. John W.C. Van Bogart**  
**National Media Laboratory**  
**June 1995**

[Table of Contents](#) | [Glossary](#)

---

Published by:

[The Commission on Preservation and Access](#)

**1400 16th Street, NW, Suite 740**  
**Washington, DC 20036-2217**

and the:

[National Media Laboratory](#)

**Building 235-1N-17**  
**St. Paul, MN 55144-1000**

This report is a joint project of the Commission on Preservation and Access and the National Media Laboratory, developed within the Commission's Preservation Science Research initiative. The initiative encourages new techniques and technologies to manage chemical deterioration in library and archival collections and to extend their useful life.

---

## Table of Contents

[Preface](#)

[Author's Acknowledgments](#)

### [1. Introduction](#)

- 1.1 Magnetic Media Compared to Paper and Film
- 1.2 The Scope of the Report

### [2. What Can Go Wrong with Magnetic Media?](#)

- Figure 1. Diagram of a Tape Reel
- Figure 2. Cross Section of a Magnetic Tape
- 2.1 Binder Degradation
- Lubricant Loss
- 2.2 Magnetic Particle Instabilities

## 2.3 Substrate Deformation

## 2.4 Format Issues

### Helical versus Longitudinal Scan Recording

#### Figure 3. Helical Scan Recording

#### Figure 4. Types of Mistracking for Helical Scan Recording

#### Figure 5. Longitudinal Recording

### Analog versus Digital Storage

## 2.5 Magnetic Tape Recorders

## [3. Preventing Information Loss: Multiple Tape Copies](#)

## [4. Life Expectancy: How Long Will Magnetic Media Last?](#)

### 4.1 Tape Costs and Longevity

### 4.2 Practical Life Expectancies

## [5. How Can You Prevent Magnetic Tape from Degrading Prematurely?](#)

### 5.1 Care and Handling

#### Frequent Access

#### Transportation of Magnetic Tape

### 5.2 Storage Conditions and Standards

#### Temperature and Relative Humidity

#### Figure 6. Temperature and Humidity Conditions and Risk of Hydrolysis Variations in Temperature and Humidity

#### Figure 7. Bad Tape Wind Examples

#### Dust and Debris

#### Figure 8. Size of Tape Debris Relative to the Tape/Head Spacing

#### Corrosive Gases

#### Storage Recommendations

#### Table 1. Current Recommendations for Magnetic Tape Storage

#### Table 2. Key Features of Access and Archival Storage of Magnetic Tape

#### Removal of Magnetic Tapes from Archival Storage

#### Table 3. Acclimation Times for Magnetic Media

#### Removed from Archival Storage

### 5.3 Refreshing of Tapes

## Appendix

### [Ampex Guide to the Care and Handling of Magnetic Tape](#)

#### Recommended practices

#### Figure 9. Tape Debris

#### Tape handling

#### Rotary Head Recorders

### [Estimation of Magnetic Tape Life Expectancies \(LEs\)](#)

#### Figure 10. Life Expectancies for a Hi Grade VHS Tape

### [Further Reading](#)

#### Resources for Transfer and Restoration of Video and Audio Tape

### [Glossary](#)



# Magnetic Tape Storage and Handling

## A Guide for Libraries and Archives

Dr. John W.C. Van Bogart  
National Media Laboratory  
June 1995

[Table of Contents](#) | [Glossary](#)

Proceed to: [3. Preventing Information Loss: Multiple Tape Copies](#)

Go back to: [1. Introduction](#)

## 2. What Can Go Wrong with Magnetic Media?

Magnetic tape consists of a thin layer capable of recording a magnetic signal supported by a thicker film backing. The magnetic layer, or top coat, consists of a magnetic pigment suspended within a polymer binder. As its name implies, the binder holds the magnetic particles together and to the tape backing. The structure of the top coat of a magnetic tape is similar to the structure of Jell-O that contains fruit - the pigment (fruit) is suspended in and held together by the binder (Jell-O). The top coat, or magnetic layer, is responsible for recording and storing the magnetic signals written to it.

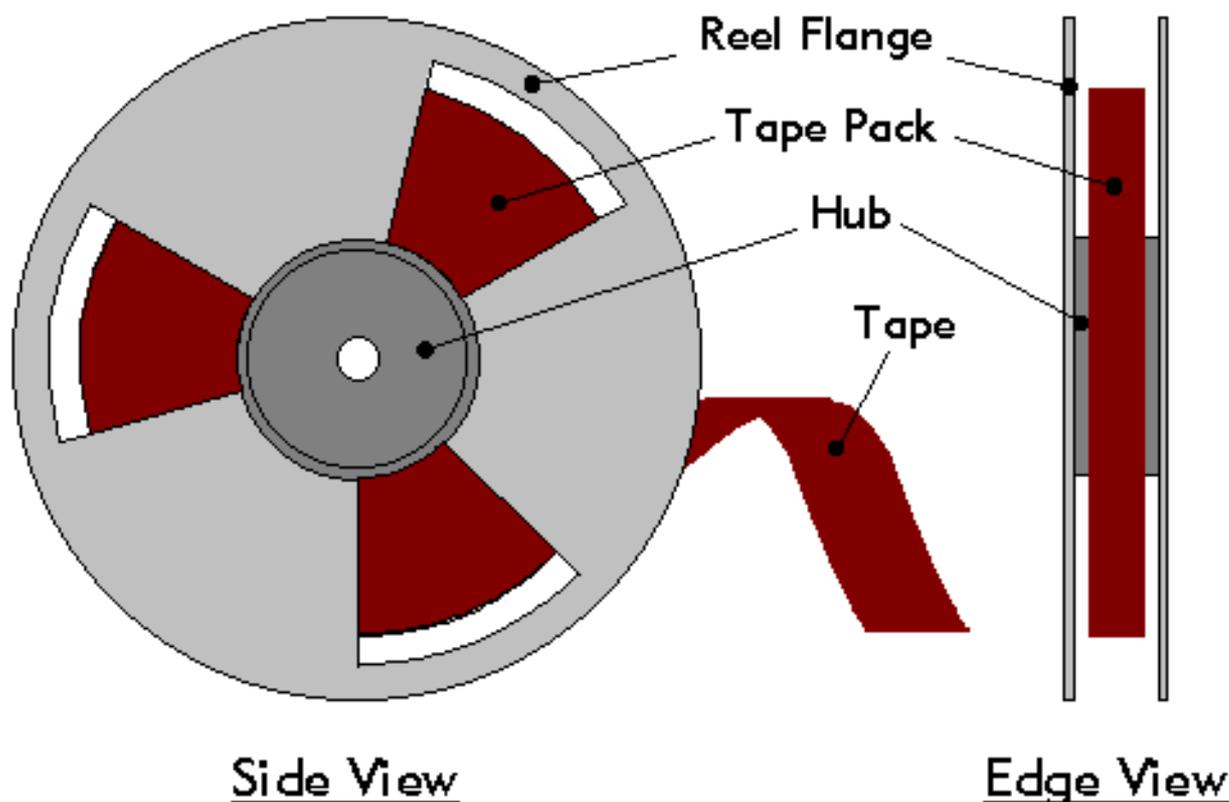


Figure 1. Diagram of a Tape Reel A schematic of a tape reel showing the principal components. Tape is

wound around the hub of a tape reel forming a tape pack. The tape pack is protected from damage and disruption by flanges on the reel.

The binder also has the function of providing a smooth surface to facilitate transportation of the tape through the recording system during the record and playback processes. Without the binder, the tape surface would be very rough, like sandpaper. Other components are added to the binder to help transport the tape and facilitate information playback. A lubricant is added to the binder to reduce friction, which reduces the tension needed to transport the tape through the recorder and also reduces tape wear. A head cleaning agent is added to the binder to reduce the occurrence of head clogs that result in dropouts. Carbon black is also added to reduce static charges, which attract debris to the tape.

The backing film, or substrate, is needed to support the magnetic recording layer, which is too thin and weak to be a stand-alone film layer. In some tape systems, a back coat is applied to the backside of the tape substrate layer. A back coat reduces tape friction, dissipates static charge, and reduces tape distortion by providing a more uniform tape pack wind on the tape reel (Figure 1). A schematic diagram of a magnetic tape construction is shown in Figure 2.

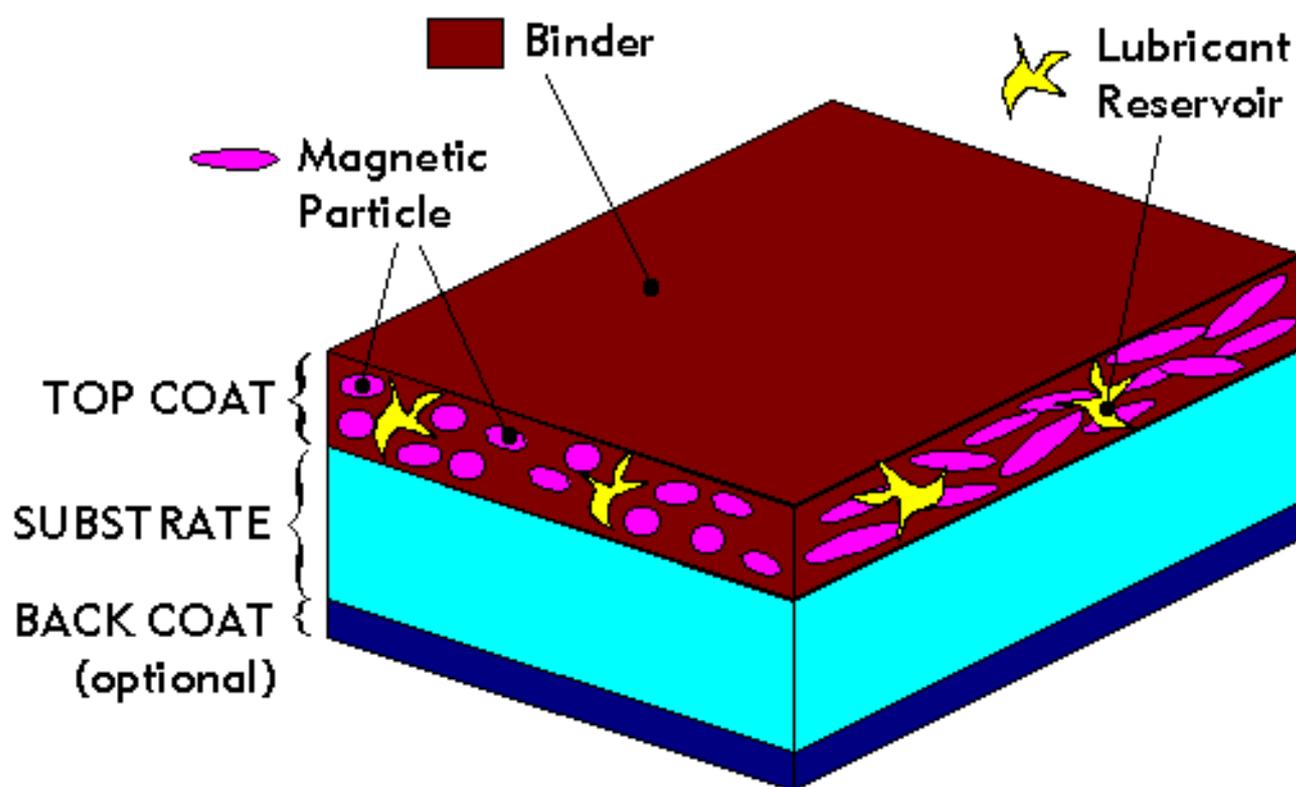


Figure 2. Cross Section of Magnetic Tape Magnetic particles are held together with a binder coated on a film substrate. Lubricant and other agents (not shown) may also be included in the top coat layer. A back coat may also be added to control friction and static charges. The structure of the top coat is analogous to that of Jell-O filled with grapes where the grapes represented the magnetic particles and the Jell-O represented the binder.

All three tape components - magnetic particle, binder, and backing - are potential sources of failure for a magnetic tape medium. The Magnetic-Media Industries Association of Japan (MIAJ) has concluded that the shelf life of magnetic tape under normal conditions is controlled by the binder rather than the magnetic particles ("DDS Specs Drive DAT Reliability," *Computer Technology Review*, 13 (5), May 1993: 30). In this instance, the shelf life would refer both to the life of recorded as well as unrecorded

media; the life of the binder is independent of whether or not the tape has ever been recorded.

## 2.1 Binder Degradation

The binder is responsible for holding the magnetic particles on the tape and facilitating tape transport. If the binder loses integrity - through softening, embrittlement, loss of cohesiveness, or loss of lubrication - the tape may become unplayable. Sticky tape and sticky shed are commonly used terms to describe the phenomenon associated with deterioration of the magnetic tape binder.

The binder polymers used in magnetic tape constructions are subject to a chemical process known as hydrolysis. In this process, long molecules are broken apart by a reaction with water to produce shorter molecules. The shorter molecules do not impart the same degree of integrity to the binder system as do the longer molecules. As in a wool sweater, if enough individual yarns are cut, the sweater will eventually fall apart.

Specifically, it is the polyester linkages in the commonly used polyester polyurethane-based binder systems that undergo scission (are broken) by water molecules. Water must be present for the hydrolysis reaction to occur. Furthermore, the more water that is present, the more likely it is that polyester chains will be broken. The binder polymer will absorb water from the air. It will absorb more water in a high humidity environment than a low humidity one. This process is analogous to that observed for open bags of crackers, potato chips, and breakfast cereals: They will lose their crunch quickly on humid, summer days (80 to 90% RH) as they absorb high amounts of moisture from the air. In the winter, however, indoor humidities generally can be lower (10 to 20% RH), less moisture is absorbed from the air, and the snacks never seem to get as stale.

Binder hydrolysis can lead to a sticky tape phenomenon characterized by a softer than normal binder coating, higher friction, and/or gummy tape surface residues. A sticky tape can exhibit sticky shed, produce head clogs, result in stick slip playback, and in extreme cases, seize and stop in the tape transport. Tape binder debris resulting from binder deterioration will result in head clogs that will produce dropouts on a VHS tape when played back. The sticky tape syndrome will result in the squealing of audio tapes as the tape very rapidly sticks to and releases from the playback head.

Procedures such as tape baking can temporarily improve binder integrity, allowing sticky tapes to be played and data recovered. The Ampex Recording Media Corporation reports that treating a sticky tape at 122° F (50° C) for three days will sufficiently firm up the binder coating so that the tape can be played. The effect of the treatment is temporary, and it is recommended that the information on the treated tape be transcribed to new tape within one to two weeks. Tape baking should not be considered a universal panacea for the treatment of sticky tapes. The tape baking procedure was developed for a specific type of degradation phenomenon on specific tape types - hydrolysis of reel-to-reel audio tapes and computer tapes. For other kinds of degradation on other tape types, tape baking may actually cause more damage. Expert advice is recommended.

### Lubricant Loss

Lubricants are normally added to the binder to reduce the friction of the magnetic topcoat layer of the tape. Lower friction will facilitate tape transport through the recorder and reduce tape wear. In a VHS recorder, where the tape is wrapped around a rapidly rotating head, low friction is also important as it

prevents overheating of the tape. The surface of a magnetic tape is actually quite porous. In some tapes, a liquid lubricant is added to the binder and will reside in these pores, similar to water absorbed in a wet sponge. When the tape passes over a head or a tape guide, lubricant is squeezed out onto the tape surface, providing a slippery interface between the tape and the guide pin. After passing by the guide pin, the excess lubricant on the surface of the tape is absorbed back into the surface of the tape. The phenomenon is similar to that observed when the surface of a wet sponge is gently pressed and released - water is exuded to the surface when the sponge is pressed and is reabsorbed when the pressure is released.

Over time, the level of lubricant in the tape decreases. Lubricants are partially consumed every time the tape is played. This is all part of their job as lubricants - they are consumed and worn down sacrificially to protect the tape. Some of the lubricant will migrate from the tape to the guide pins and heads of the recorder each time the tape is played.

Lubricant levels decrease over time even in unplayed, archived tape as a result of evaporation and degradation. The lubricants used in some tapes are oily liquids that are volatile and slowly evaporate away over time. Some lubricants are also subject to degradation by hydrolysis and oxidation, just like the binder polymer, and will lose their essential lubrication properties with time.

The information stored on severely degraded magnetic tapes can be recovered, in specific instances, after relubrication of the tapes. By significantly reducing the friction of the magnetic coating with the addition of lubricant, tapes can be made to play back. Prior to relubrication, the tape may have seized in the tape transport as a result of high friction, or the magnetic coating may have been readily torn off the tape backing by a high speed tape head. Relubrication of tapes must be done carefully by experienced individuals. If a tape is over-lubricated, the excess lubricant on the surface of the tape will act as debris and increase the head-to-tape spacing, causing signal losses and dropouts.

## **2.2 Magnetic Particle Instabilities**

The magnetic particle, or pigment (the terminology is a carryover from paint and coatings technology), is responsible for storing recorded information magnetically as changes in the direction of the magnetism of local particles. If there is any change in the magnetic properties of the pigment, recorded signals can be irretrievably lost. The magnetic remanence characterizes the pigment's ability to retain a magnetic field. It refers to the amount of signal that remains after a recording process. The strength of the signal recorded on a tape magnetically is directly related to the magnetic remanence of the pigment. Thus, a decrease in the magnetic remanence of the pigment over time can result in a lowered output signal and potential information loss.

The coercivity characterizes the pigment's ability to resist demagnetization. It refers to the strength of the magnetic field that must be applied to a magnetic particle in order to coerce it to change the direction of its magnetic field. Demagnetization of a tape can result from an externally applied field, such as that produced by a hand-held metal detector at an airport security check point. A magnetic tape with a lower coercivity is more susceptible to demagnetization and signal loss.

Magnetic pigments differ in their stability - some particles retain their magnetic properties longer than others. Thus, some tapes will retain information, which is stored magnetically, longer than others. Iron oxide and cobalt-modified iron oxide pigments are the most stable pigment types of those used in audio and videotapes. These pigments are generally used in the lower grade audio and low to high grade VHS/

Beta videotape formulations. The low coercivity of these pigments disallows their use in high grade audio formulations.

Metal particulate (MP) and chromium dioxide ( $\text{CrO}_2$ ) pigments provide a higher tape signal output and permit higher recording frequencies than the iron oxide pigments, but are not as stable as the iron oxide pigments. A decrease in signal output of two decibel (dB) may be observed over the lifetime of metal particle and chromium dioxide based tapes. However, even with these losses, the output signal will still be better than a comparable iron oxide based tape. A loss in signal will manifest itself as a reduction in the clarity and volume of a sound recording and in the loss of hue and reduction in saturation for a video recording. Chromium dioxide is used in medium to high grade audio tape and some high grade VHS/Beta video tape. Metal particulate is used in high grade audio and 8mm video tape. Metal particulate is also used in most digital audio and digital video tape formulations. The type of pigment used in the audio or video tape formulations is normally indicated in the product literature that comes with the tape. This information can also be obtained from the manufacturer via the toll-free number provided on the literature that accompanies the tape cassette or reel.

There is not much that can be done to prevent the magnetic deterioration that is inherent in the metal particulate and chromium dioxide pigment types. However, the rate of deterioration can be slowed by storing the tapes in cooler temperatures. The level of humidity has little direct effect on the deterioration of magnetic pigments. However, by-products of binder deterioration can accelerate the rate of pigment deterioration, so lower humidity would also be preferred to minimize the degradation of the magnetic pigment.

Metal evaporated (ME) video tapes are prevalent in the 8mm video formats. These tapes require no binder polymer, as the entire magnetic layer consists of a single, homogeneous metal alloy layer that is evaporated onto the tape substrate. These tapes have chemical stabilities similar to those of metal particle tapes. However, because the magnetic coating on an ME tape is much thinner than the corresponding layer on an MP tape, they are generally not as durable and do not hold up well in repeated play or freeze-frame video applications.

## 2.3 Substrate Deformation

The tape backing, or substrate, supports the magnetic layer for transportation through the recorder. Since the early 1960s, audio tapes and videotapes have used an oriented polyester (also known as polyethylene terephthalate, PET, or DuPont Mylar®) film as a tape substrate material. Polyester has been shown, both experimentally and in practice, to be chemically stable. Polyester films are highly resistant to oxidation and hydrolysis. In archival situations, the polyester tape backing will chemically outlast the binder polymer. The problem with polyester backed videotapes is that excessive tape pack stresses, aging, and poor wind quality can result in distortions and subsequent mistracking when the tapes are played.

The best way to reduce the degree of tape backing distortion is to store magnetic media in an environment that does not vary much in temperature or humidity. Each time the temperature or humidity changes, the tape pack will undergo expansion or contraction. These dimensional changes can increase the stresses in the tape pack that can cause permanent distortion of the tape backing. Distortion of a VHS tape backing will show up as mistracking when the tape is played.

Tape backing deformation can also arise if the tape experiences nonlinear deformation as a result of

nonuniform tape pack stresses. This normally results if the tape pack wind quality is poor as indicated by popped strands of tape - one to several strands of tape protruding from the edge of a wound roll of tape. Methods of controlling the quality of the tape pack wind are discussed in the Ampex Guide to the Care and Handling of Magnetic Tape that appears in the Appendix.

Older tapes used other backing materials. In the 1940s and 1950s, acetate (cellulose acetate, cellulose triacetate) film was used as an audio tape backing. This is the same material used in some older movie film. In general, if light can be seen coming through the tape windings when the reel is held up to a light, it is an acetate based magnetic tape. This substrate is subject to hydrolysis and is not as stable as polyester film. However, more stable vinyl binder systems were used during this time period. Thus, the life of tapes produced during this period can be limited by the degradation of the backing rather than the binder. Degradation of the backing in these tapes is indicated by the presence of the vinegar syndrome, where a faint odor of vinegar (acetic acid) can be detected coming from the tapes. In the advanced stages of degradation, the magnetic tape will become brittle and break easily if bent too sharply or tugged. The backing also shrinks as it decomposes, resulting in a change in the length of the recording. Any tape on an acetate backing should be stored in a low-temperature, low-humidity archive to reduce the rate of deterioration of the acetate tape backing.

Acetate film has also been used as a base film for photographic film, cinema film, and microfilm. The "IPI Storage Guide for Acetate Film" has been prepared by the Image Permanence Institute, Rochester Institute of Technology, Post Office Box 9887, Rochester, New York, 14623-0887, Phone: 716-475-5199, as an aid in preserving still and motion picture film collections on acetate base films. The comments in that guide are equally appropriate for acetate based magnetic recording tape. In general, lower storage temperatures and relative humidities are recommended to increase the time to onset of the vinegar syndrome. Tapes having the vinegar syndrome should be stored separately to prevent the contamination of other archive materials by acetic acid. After the onset of the vinegar syndrome, acetate films degrade at an accelerated rate. Tapes that have been stable for fifty years may degrade to the point of being unplayable in just a few years. Any valuable tape showing vinegar syndrome should be transcribed as soon as possible.

Prior to cellulose acetate, paper was used as a tape backing material. Audio recordings of this type are very rare and should be stored in a tape archive. Although generally stable, these backings are very fragile and subject to tearing or breaking on playback. For this reason, particular care should be taken to ensure that the playback recorder is very well maintained.

## **2.4 Format Issues**

### **Helical versus Longitudinal Scan Recording**

The susceptibility of the recording to loss as a result of dimensional changes in the backing is dependent on recording format. Videotape, which uses a helical scan recording format, is more sensitive to disproportionate dimensional changes in the backing than analog audio tape, which uses longitudinal recording.

Helical (Figure 3). Tracks are recorded diagonally on a helical scan tape at small scan angles. When the dimensions of the backing change disproportionately, the track angle will change for a helical scan recording. The scan angle for the record/playback head is fixed. If the angle that the recorded tracks make

to the edge of the tape do not correspond with the scan angle of the head, mistracking and information loss can occur.

Distortion of a helical scan videotape can result in two types of mistracking - trapezoidal and curvature (Figure 4). In trapezoidal mistracking, the tracks remain linear, but the track angle changes so that the playback head, which is always at a fixed angle to the tape, cannot follow them. Curvature mistracking can be a more serious type of deformation where the recorded tracks become curved as a result of nonlinear deformation of the tape backing. Mistracking will result in a video image where some or all of the screen is snowy or distorted. For example, in the case of trapezoidal mistracking, the upper portion of the TV screen may appear normal, whereas the lower portion of the screen may be all static. The appearance on the screen will be similar to the playback of a good tape where the tracking adjustment control has been purposely misadjusted.

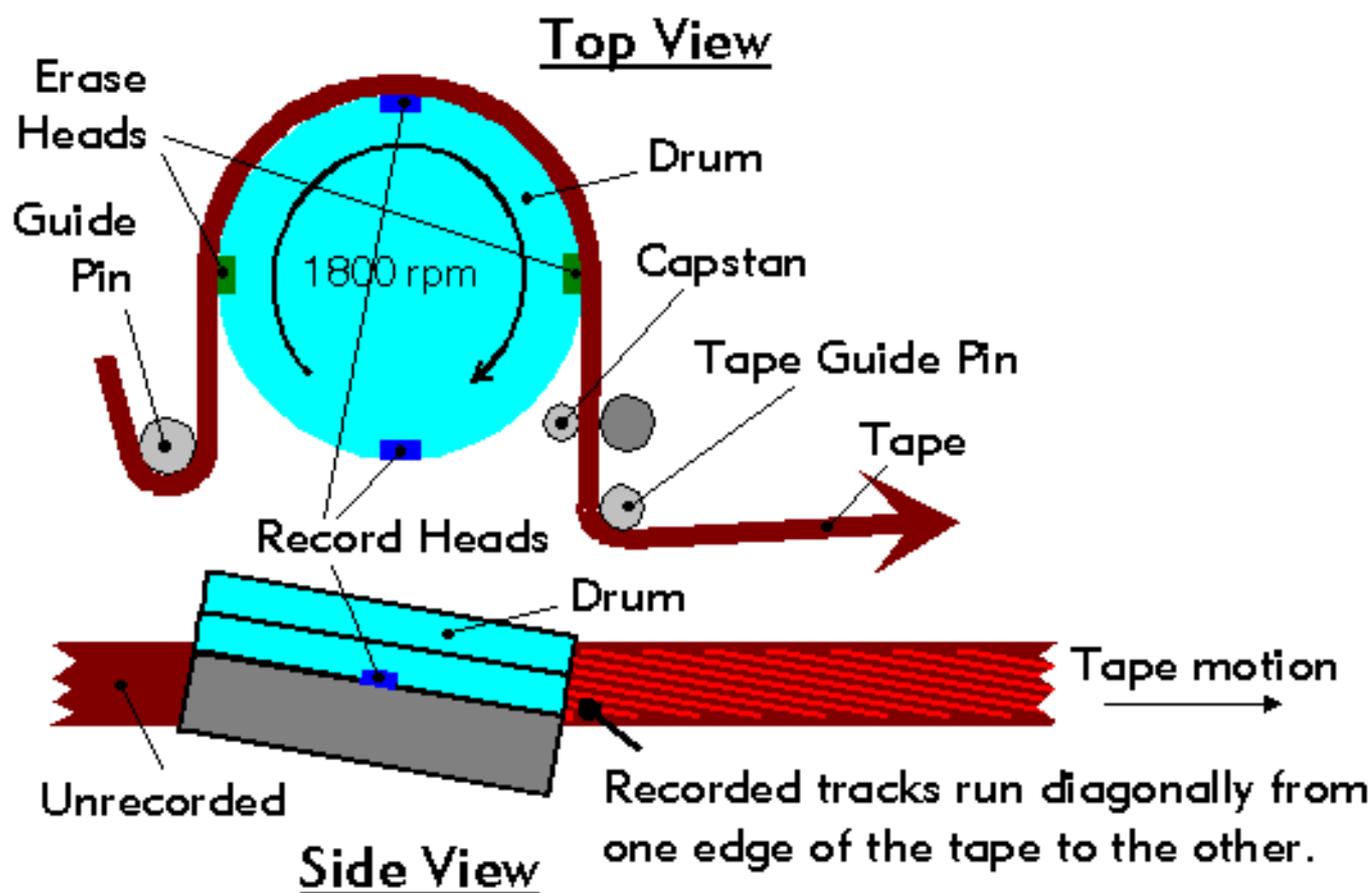


Figure 3. Helical Scan Recording A moving tape wraps 180° around a cylindrical drum rotating at high speeds; the rotating head is oriented at a slight angle to the tape so that the tracks written by the tiny record head embedded in the surface of the rotating drum run diagonally across the tape from one side to the other.

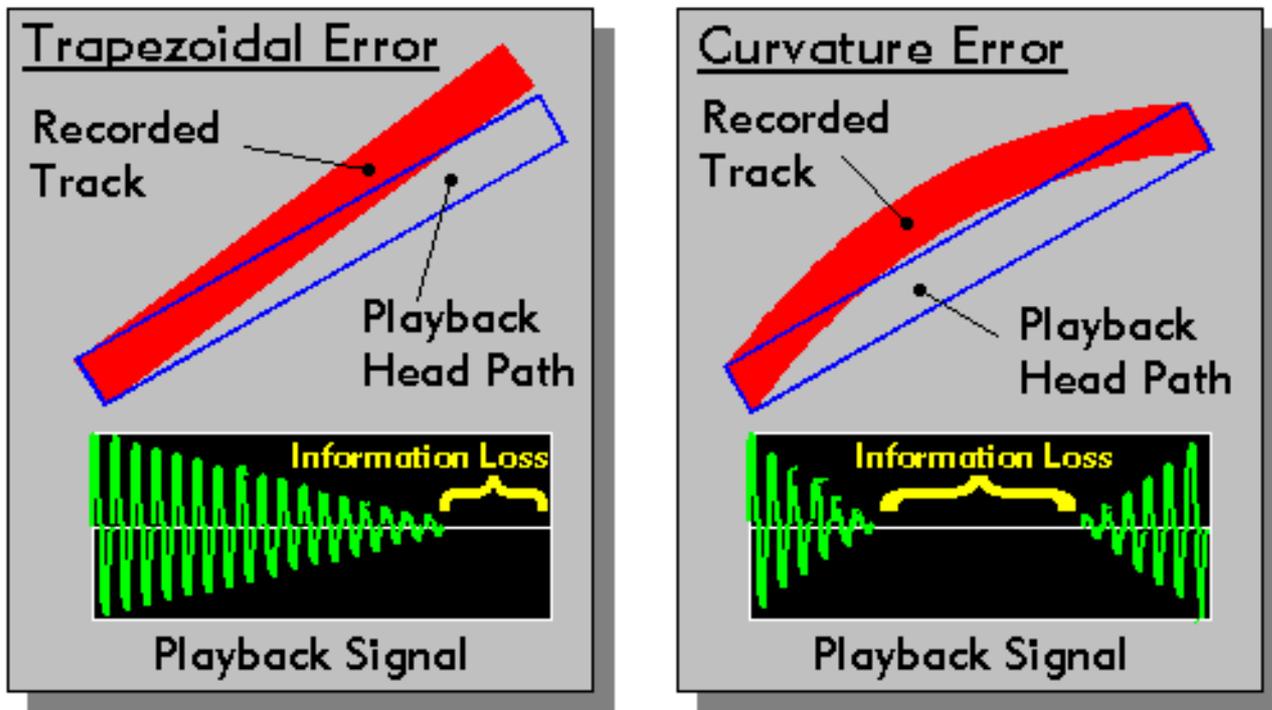


Figure 4. Types of Mistracking for Helical Scan Recording Trapezoidal error occurs when the angle of the recorded track does not agree with the scan angle of the playback head. Curvature error occurs when the tape has deformed nonlinearly. The playback signal corresponds to that for a single helical scan.

Longitudinal (Figure 5). In a longitudinal tape system, the heads are arranged along a fixed head stack - one head per track - and the tracks will always remain parallel to the edge of the tape. Mistracking is not as great a problem in longitudinal recording for this reason.

Distortion of a longitudinal audio recording tape will appear as a temporary muffling, change in pitch, or loss of the audio track. Distortion of the tape backing can impart a slight curve to the normally linear tape. When the distorted portion of tape passes over the playback head, the recorded tracks can move out of alignment with the head gap, causing a temporary reduction in sound volume and quality.

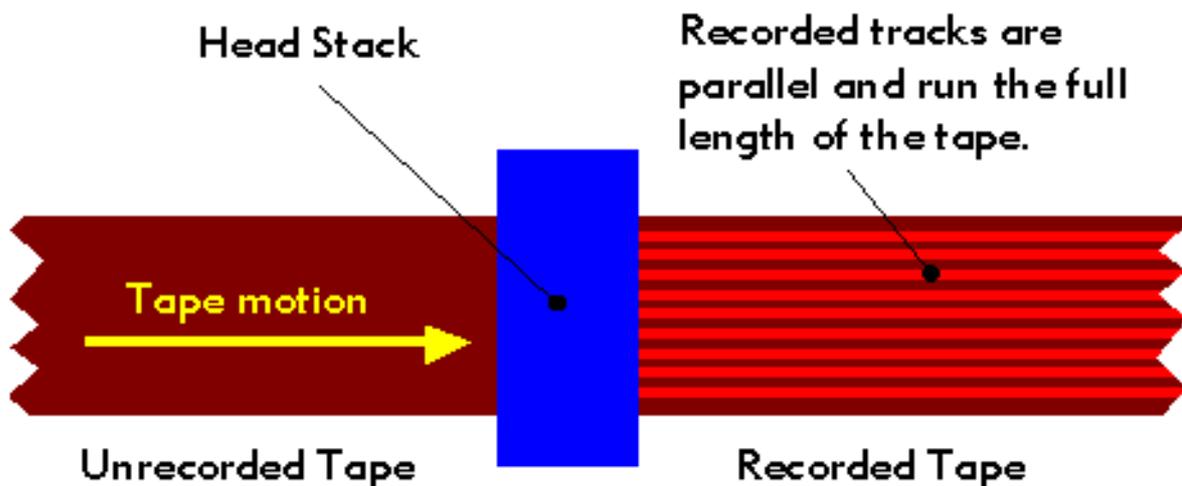


Figure 5. Longitudinal Recording A moving tape passes across a stationary record head. The recorded tracks are parallel to the edge of the tape and run the full length of the tape. A nine-track tape is shown.

## Analog versus Digital Storage

Some comments concerning the archival stability of analog versus digital materials may be instructive. In an analog recording, the signal recorded on the audio or videotape is a representation of the signal originally heard or seen by the microphone or video camera. The volume of a sound recording or the intensity of the color of a video image is directly related to the strength of the magnetic signal recorded on the tape. In a digital recording the audio or video source signal is digitized - the signal is sampled at specific points in time and converted to a number that reflects the intensity of the signal at the time of sampling (analog-to-digital conversion). These numbers, in binary form, are written to the tape, rather than the analog signal. On playback, the numbers are read and used to reconstruct a signal that is representative of the original signal (digital-to-analog conversion).

The chief advantage of an analog recording for archival purposes is that the deterioration over time is gradual and discernible. This allows the tape to be transcribed before it reaches a point where the recording quality has degraded to an unusable level. Even in instances of severe tape degradation, where sound or video quality is severely compromised by tape squealing or a high rate of dropouts, some portion of the original recording will still be perceptible. A digitally recorded tape will show little, if any, deterioration in quality up to the time of catastrophic failure when large sections of recorded information will be completely missing. None of the original material will be detectable in these missing sections.

The chief advantage of a digital recording is that copies of the original tape can be made without any loss in recording quality. A copy of a digital tape can be made that is truly identical to the original source tape. When an analog tape is copied, the original information signal is actually copied along with any tape noise inherent in the tape and any electronic noise inherent in the recording device. This will be written to a new tape that also has its own level of inherent tape noise. Therefore, the noise level on the dubbed copy will always be greater than that on the original tape, or the sound quality of the original recording will be altered as it is filtered to reduce noise. The presence of noise in the recording will make the recorded information less distinct to see or hear. (Recording engineers refer to a signal-to-noise ratio, which defines the quality of the recording with a higher value being better.) Digital tape recordings are virtually unaffected by tape noise, even though digital tapes are not noise free. In digital recording, binary numbers (comprised entirely of ones and zeros) are read from and written to the tape. The ones and zeros are easily distinguished from the background noise. In an analog recording, the recorder cannot distinguish between the recorded signal and the tape noise so that both are read and reproduced on playback. In addition, digital recordings usually have an error correction system that uses redundant bits to reconstruct areas of lost signal.

Analog recording continuously records the complete signal heard or seen by the recording microphone or video camera. However, distortion in both recording and playback will vary with the quality of the electronic components used. In digital recording, the source signal is quantized to a fixed number of allowed signal levels. For example, a video image quantized at 8-bits/color would only allow for 256 distinct colors to be reproduced, whereas an analog image would allow an infinite number of colors. By increasing the number of bits/color used, the number of color levels that can be reproduced will increase (see bit in the Glossary for more detail). For example, an image quantized at 24-bits/color will allow 16,777,216 distinct colors. With digital recording, higher quality video images require greater storage volumes. Some audiophiles with highly trained ears claim that they can hear limitations in a digital CD audio recording (16-bit quantization permitting 65,536 distinct sound levels and a maximum frequency of 22 kHz) when compared to an analog recording of the same sound source.

Analog tape recordings do not require expensive equipment for recording and playing. Digital audio and

video equipment which records high frequencies at high speeds and performs the complex tasks of analog-to-digital and digital-to-analog conversion and error correction is relatively expensive.

## 2.5 Magnetic Tape Recorders

This document is primarily concerned with tape media, not tape recorders. However, in discussing what can go wrong with media, recorders must be mentioned. Audio and video recorders must be maintained in excellent condition in order to produce high quality recordings and to prevent damage to tapes on playback. Dirty recorders can ruin tape by distributing debris across the surface of the tape and scratching the tape. Recorders that are not mechanically aligned can tear and stretch tape, produce poor tape packs, and write poorly placed tracks. Recorders that are poorly aligned electrically can cause signal problems that will result in inferior playback. Follow the manufacturer's instructions for good recorder maintenance in order to protect recordings.

---

Proceed to: [3. Preventing Information Loss: Multiple Tape Copies](#)

Go back to: [1. Introduction](#)



# Magnetic Tape Storage and Handling

## A Guide for Libraries and Archives

Dr. John W.C. Van Bogart  
National Media Laboratory  
June 1995

[Table of Contents](#) | [Glossary](#)

---

Proceed to: [4. Life Expectancy: How Long Will Magnetic Media Last?](#)

Go back to: [2. What Can Go Wrong with Magnetic Media?](#)

---

## 3. Preventing Information Loss: Multiple Tape Copies

As already discussed, this report is primarily concerned with preventing magnetic tape from degrading prematurely. However, it is worth mentioning the use of multiple copies as another strategy for preventing loss of information. Recorded information can be lost because the medium on which it is recorded has deteriorated to the point of being unplayable. Information can also be lost if the tape on which it is recorded disappears (misplaced, stolen, destroyed by fire or flooding, and so forth). Both types of loss can be prevented by maintaining more than one copy of the information and storing all copies in separate locations.

If funds are available, it is preferable to maintain both access storage and archival storage of important information. As the names imply, the access environment keeps the recording readily available for playback. Archive storage involves a separate environment designed to maximize the longevity of the tape. Refer to Section 5.2: Storage Conditions and Standards for a more detailed discussion of these storage conditions.

The quality of care a magnetic tape receives should be commensurate with the perceived value of the information contained on the tape. Refer to Section 4.1: Tape Costs and Longevity for more information. In reality, a library or archive may not have the budget, the personnel, the time, or the space to maintain two copies of all of the recordings in a video or audio tape collection. In this case, the value and use requirements of individual tapes in the collection should be assessed and prioritized. Those tapes considered the most valuable and most likely to be used should be duplicated and the originals should be placed in an archive environment. If duplicates of information are disallowed, some or all of the collection could be placed in an archive, but this would greatly limit access to the information. In instances where the information is considered extremely valuable, it may be worthwhile to maintain several copies of the original in the archive along with the original tape.

---

Proceed to: [4. Life Expectancy: How Long Will Magnetic Media Last?](#)

Go back to: [2. What Can Go Wrong with Magnetic Media?](#)





# Magnetic Tape Storage and Handling

## A Guide for Libraries and Archives

Dr. John W.C. Van Bogart  
National Media Laboratory  
June 1995

[Table of Contents](#) | [Glossary](#)

---

Proceed to: [5. How Can You Prevent Magnetic Tape from Degrading Prematurely?](#)

Go back to: [3. Preventing Information Loss: Multiple Tape Copies](#)

---

## 4. Life Expectancy: How Long Will Magnetic Media Last?

Unfortunately, media life expectancy (LE) information is largely undocumented, and a standard method for determining magnetic media lifetimes has yet to be established. The need for this information fuels the ongoing NML media stability studies, which have incorporated accelerated temperature/humidity and corrosion environments to measure performance over time and to develop models to forecast extended media lifetimes. A simple example as to how LEs can be determined is provided in the Appendix under Estimation of Magnetic Tape Life Expectancies (LEs).

According to manufacturers' data sheets and other technical literature, thirty years appears to be the upper limit for magnetic tape products, including video and audio tapes. LE values for storage media, however, are similar to miles per gallon ratings for automobiles. Your actual mileage may vary.

Recently, articles have been appearing which suggest that the life expectancy of magnetic media is much shorter than originally thought. For example, an article in the January 1995 Scientific American ( Jeff Rothenberg, "Ensuring the Longevity of Digital Documents") conservatively estimated the physical lifetime of digital magnetic recording tape at one year. Because of the confusion that can result from such a statement, NML officially responded with a letter to the editor that appeared in the June 1995 issue of Scientific American. The letter states that the "physical lifetimes for digital magnetic tape are at least 10 to 20 years."

### 4.1 Tape Costs and Longevity

Some people assess storage media solely in terms of media cost. This view assumes that the sound, images, or information stored on the media have no intrinsic value. However, a storage medium should be evaluated in terms of the cost of losing the recorded information in the event that the storage medium degrades irreversibly.

The value of the tape cassette must be equated with the cost of preserving the data. When the cost of losing the information is considered, it may be economically justified to invest more in a medium/system

of proven reliability. It may also warrant the cost of making and keeping replicated copies of original data and stockpiling systems to play back the data at future times.

When purchasing media of a specific format, some archivists are required to deal with a procurement bidding process. In most cases, the archivist will end up with the lowest bidder's media, which may not be the best media. Tape manufacturers' products differ in coating thickness, magnetic particle stability, and durability. Procurement specifications should exclude the poorer media. The vendor should be asked for experimental proof of the stability of the media if the tape is to be used for archival storage.

## 4.2 Practical Life Expectancies

Those accustomed to storing paper and microfilm may be annoyed by the relatively short life expectancies (ten to thirty years) of magnetic tape materials. Some gold plated/glass substrate digital optical disc technologies promise 100-year lifetimes. However, a 100-year life expectancy is irrelevant when the system technology may be in use for no more than ten or twenty years (or less).

Audio and video recording technologies are advancing at a much faster rate than printing and microfilming technologies. We are fortunate if a recording technology stays current for more than twenty years. In the case of a magnetic recording media with a fifty-year life expectancy, the media would undoubtedly outlive the recording system technology. To truly achieve a fifty-year archival life, recording systems, sufficient spare parts, and technical manuals would need to be archived along with the recorded media.

In the case of audio and video archives, transcription is inevitable. Rather than trying to preserve old, outdated recording formats and technologies, it may be more practical to transcribe on a regular basis - every ten to twenty years or even more frequently. The old copy could be preserved until the new copy is transcribed to the next generation of recording system. In this fashion, at least two copies of the material are always in existence.

---

Proceed to: [5. How Can You Prevent Magnetic Tape from Degrading Prematurely?](#)

Go back to: [3. Preventing Information Loss: Multiple Tape Copies](#)



# Magnetic Tape Storage and Handling

## A Guide for Libraries and Archives

Dr. John W.C. Van Bogart  
National Media Laboratory  
June 1995

[Table of Contents](#) | [Glossary](#)

---

Proceed to: [Appendix: Ampex Guide to the Care and Handling of Magnetic Tape](#)

Go back to: [4. Life Expectancy: How Long Will Magnetic Media Last?](#)

---

## 5. How Can You Prevent Magnetic Tape from Degrading Prematurely?

The remainder of this document answers this question. Some of the factors to be discussed are more controllable than others. For example, you can normally decide the storage conditions and level of access to an archive collection. However, you do not always have control over the quality of the tape wind, or the brand, type, and format of the tape media on which the information is stored.

Factors affecting the life of the tape over which you have some control are:

- The care with which it is handled and shipped, discussed in Section 5.1: Care and Handling.
- The quality of the conditions in which it is stored, discussed in Section 5.2: Storage Conditions and Standards.
- The number of times the tape is accessed during its lifetime, discussed in Section 5.1: Care and Handling: Frequent Access.

Other factors that affect media over which you have less control are:

- The physical components of the tape, discussed in Section 2: What Can Go Wrong with Magnetic Media?
- The quality of the tape being purchased; for example, standard grade versus high grade VHS.
- Variation in the quality of the manufacturer; for example, a name brand versus a bargain brand.
- Future availability of system technology to play back the tape. For example, quadruplex videotapes still exist in archives; however, the equipment to play them back is considered obsolete, and it is difficult to find working recorders.

### 5.1 Care and Handling

Magnetic tape should receive the same kind of care that you would give to a valuable book or important

photograph. In general, handle the tapes with care, keep them clean, and apply common sense:

- Use and store magnetic tape reels and cassettes in a clean environment.
- Avoid contamination of the tapes by dirt, dust, fingerprints, food, cigarette smoke and ash, and airborne pollutants.
- Take care not to drop tapes or cartridges.
- Keep tapes out of strong sunlight and avoid contact with water.
- Do not store tapes on radiators, window sills, televisions, electronic equipment, or machinery.
- When the tapes are not in use, they should be placed back on the storage shelf, and stored on end. They should not be allowed to lay flat (reel flanges parallel with the table top) for extended periods of time.

Refer to the Ampex Guide in the Appendix for more information. Magnetic tapes do require some unique care and handling precautions. Because they are a magnetic form of storage, exposure to strong magnetic fields must be avoided to prevent information loss. This is generally not a problem, unless the materials need to be transported or shipped.

## **Frequent Access**

Tapes that are frequently accessed may have a reduced life expectancy due to wear and tear. The life of the media may not be determined by data error rates, but by the life of the media housing. In one instance, the life of a tape cassette was limited by failure of the cassette door, not because of any fault of the tape media. How many insert and eject cycles will your media be required to handle? This may limit the life of the cassette.

The more a tape or cassette is handled, the more it is contaminated with fingerprints and debris. It is also exposed to less than ideal conditions, especially if the materials are removed from the building in which they are normally stored.

Every time a VHS cassette is loaded into a recorder, the recorder mechanism pulls tape from the cassette. This mechanism can damage the tape if the guide pins are not properly aligned. Debris on the loading mechanism can scratch the surface of the tape. Also, when a tape is removed from a recorder, the tape must properly retract into the cassette, otherwise it will be damaged when the cassette doors close and the tape cassette is ejected from the recorder. Most of us have probably had experience with a VHS deck that has eaten a tape.

Because of potential damage to the tape, it is important that the tapes be inserted and ejected at areas of the tape that contain no recorded information. A tape should NEVER be ejected in the middle of an important recording.

## **Transportation of Magnetic Tape**

Care must be exercised to ensure that tape collections are not harmed when they are transported. When magnetic media are transported, temperatures should not exceed 110° F (43° C). Collections should be transported in the spring or the fall when outdoor temperatures are moderate, if possible. Properly wound tape reels can survive greater variations in temperature and humidity without permanent damage than can poorly wound tape packs.

Tapes and cassettes should be shipped in the same orientation as they are stored - on edge - with the weight of the tape pack being supported by the reel hub. Tapes that are shipped in the flat position are particularly subject to damage from dropping and other forms of shock. This is especially true of tapes that experience large changes in temperature during shipment or tapes that are poorly wound.

Media should be protected from damage due to shock by packing them in materials that will absorb shock (special packages, bubble wrap), using special labeling, and transporting them in appropriate vehicles. Shock-absorbing packaging will often have the added advantage of providing insulation that helps protect the media from large swings in temperature and humidity.

Exposure to strong magnetic fields must also be avoided to prevent information loss. Some of the detectors used to screen luggage in overseas airports have been known to partially erase tapes. Walk through metal detectors and X-ray scanners do not pose a threat to recorded information. Some hand-held metal detectors can cause problems since they use strong magnetic fields. Refer to the section on Stray Magnetism in the Ampex Guide in the Appendix.

## **5.2. Storage Conditions and Standards**

Storing magnetic tape in a clean, controlled environment is the most important precaution you can take to extend the life of the media. High temperatures, high humidity, and the presence of dust and corrosive elements in the air all affect the physical components that make up magnetic tape and can result in loss of readable data through decreased magnetic capability and deterioration of the binder or backing of the tape. Too low temperatures should also be avoided. In some cases, temperatures lower than 32° F (0° C) may actually harm the media and shorten, rather than extend, life expectancies by risking exudation of the lubricant from the binder, which may clog heads. Rapid temperature changes are also undesirable as they introduce stresses in the wound tape pack. Tapes that are to be played in an environment different from the storage environment should be allowed to acclimate to the new temperature.

### **Temperature and Relative Humidity**

For years tape manufacturers have recommended that you store your tapes in a cool, dry place. In Section 2: What Can Go Wrong with Magnetic Tape?, the reasons behind this dictum were discussed in terms of the chemistries of the tape components: Binder hydrolysis is dependent on the moisture content of the tape, and lower humidity results in lower rates of hydrolysis. Furthermore, this reaction will proceed more slowly at lower temperatures. The latter is also true for the magnetic pigments - they will degrade more slowly at lower temperatures. Finally, to reduce unnecessary stresses on the wound tape that could result in deformation of the backing, a limited variation in temperatures and humidities is recommended. (See Figure 6.)

Storage at high temperatures (> 74° F; > 23° C) increases tape pack tightness. This results in distortion of the tape backing and an increase in permanent dropouts as wound-in debris is forced into the tape magnetic layer. Many layers of tape before and after the debris can be affected by impressions of the debris. Layer to layer adhesion, known as tape blocking, also can result after long term storage at elevated temperatures.

Storage at high humidity (> 70% RH) results in increased degradation of the binder as a result of the

higher moisture content of the tape pack. High humidities will also cause increased tape pack stresses as the tape absorbs moisture from the air and expands, causing distortion of the tape backing and an increase in permanent dropouts.

Fungal growth is also possible at high humidities and temperatures. Molds can live off the binder polymer and added components. This is yet another cause of binder breakdown in high humidities. Hairy growths at the edges of the tape are a sign of mold. The spores that are produced on this fuzz can get onto the tape surface and cause many dropouts.

Changes in both temperature and humidity can also cause mistracking problems on helical scan recordings (See Section 2.4: Format Issues: Helical versus Longitudinal Scan Recording). Substrates will expand or shrink with changing temperature and humidity just as metals do in heat or cold. The substrate films are not completely balanced in their reaction to these changes in temperature and humidity. In other words, they stretch and shrink differently in length and width directions. This causes a change in the angle of the recorded helical scan tracks. Most of these changes are recoverable by returning to a temperature and humidity close to the one at which the tape was recorded. However, heat can also cause premature aging of the substrate in the form of nonrecoverable shrinking and stretching.

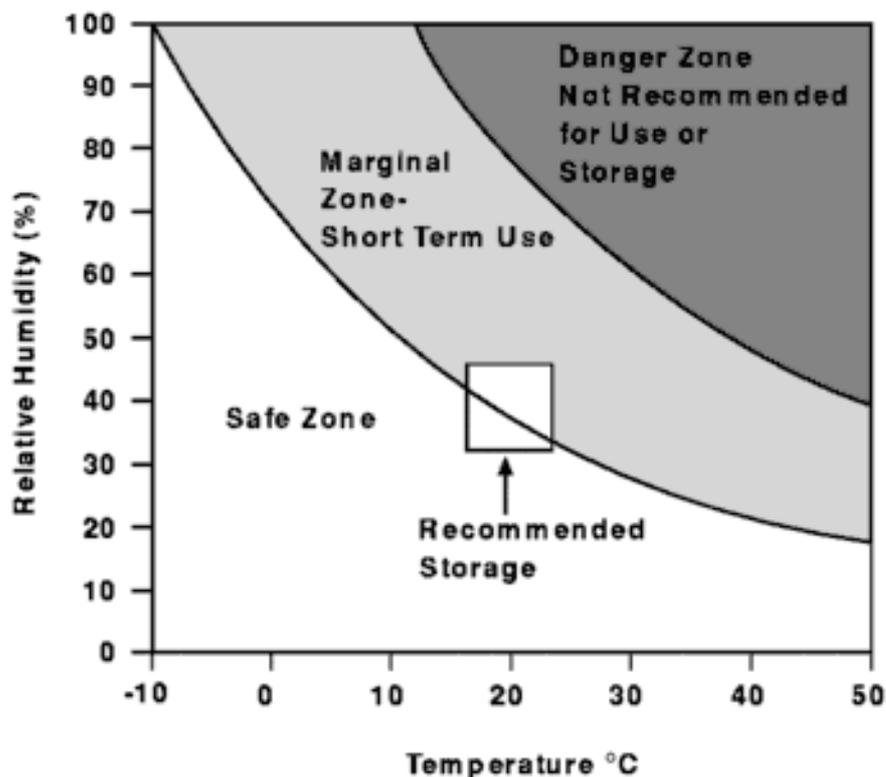


Figure 6. Temperature and Humidity Conditions and Risk of Hydrolysis This figure depicts the effects of humidity and temperature and shows that  $15 \pm 3^\circ \text{C}$  ( $59 \pm 5^\circ \text{F}$ ) and 40% maximum relative humidity (RH) are safe practical storage conditions. A similar diagram appears in ISO TR 6371-1989 that suggests even more stringent conditions (RH 20% max.) for long-term storage of instrumentation tapes. (Source: Ampex. Reprinted with permission.)

## Variations in Temperature and Humidity

Generally, the temperature and humidity in a tape storage facility are set to specific values, or set points,

and infrequently varied or adjusted. This does not mean that the temperature and humidity in the facility are invariant. Changes in the outdoor temperature and humidity will cause the temperature in the tape storage facility to vary slightly.

If the temperature outdoors is higher than the set point temperature in the facility, the actual temperature in the facility will be slightly higher than the set point. If the outdoor temperature is lower than the set temperature, the actual facility temperature will be lower than the set point. The variations in temperature experienced will be larger at larger distances from the thermostat in the facility. The same logic applies to the humidity level in the facility. Larger discrepancies in the set point and the actual temperature will be observed if one of the walls of the facility is an exterior wall, or if the heating/cooling capacity of the environmental controller is less than that required to properly control the tape archive.

The set point in a tape archive may be constant, but the archive will still experience some degree of daily and seasonal variations in temperature and humidity. A tape archivist must have knowledge of the set points in the archive as well as the variations in temperature and humidity to ensure that the archive complies with recommended storage conditions.

Variations in temperature and humidity can cause tape problems. Tape packs are wound under a considerable amount of tension. This is necessary to maintain the shape of the tape pack. A reel of tape can be permanently damaged if the tape pack tension is too high or too low. If the tension is too high, the tape backing can stretch. If the tension gets too low, tape layers can slip past each other, resulting in pack slip, cinching, or popped strands on playback (see Figure 7). Relaxation of the tape backing can also occur if the tape pack tension is not properly maintained. Relaxation, stretching, and deformation of the tape backing can cause mistracking of a videotape or sound distortion on an audio tape. Every time a tape pack is heated or cooled, the tape pack tension will increase or decrease, respectively. The best way to reduce the degree of tape backing distortion is to store magnetic media in an environment that does not vary much in temperature or humidity.

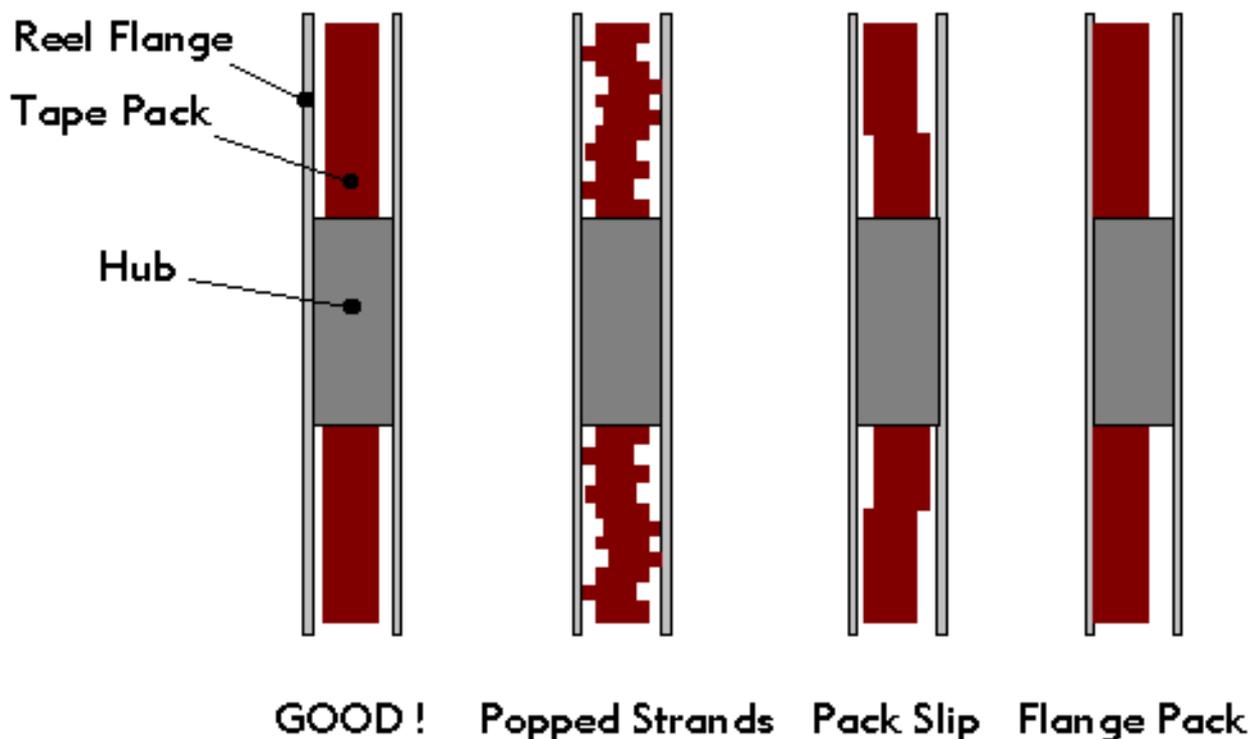


Figure 7. Bad Tape Wind Examples This figure shows schematic examples of popped strands, pack slip,

and a flange pack. The illustrations show a cross-section slice of the tape pack through the hub.

## Dust and Debris

Dust, smoke particles, and tape debris present in the environment can get wound into the tape pack as the tape is played, resulting in dropouts when the tape is subsequently played. The lost signal is generally greater than expected from the size of the particle. The record and read heads must maintain very close contact with the tape. A particle of dust on the tape causes the head to ride up over the particle and lose contact with the tape. For perspective on the size of various debris particles compared to the normal head to tape spacing, see Figure 8.

### Debris Perspective on High Density Digital Recording Tape

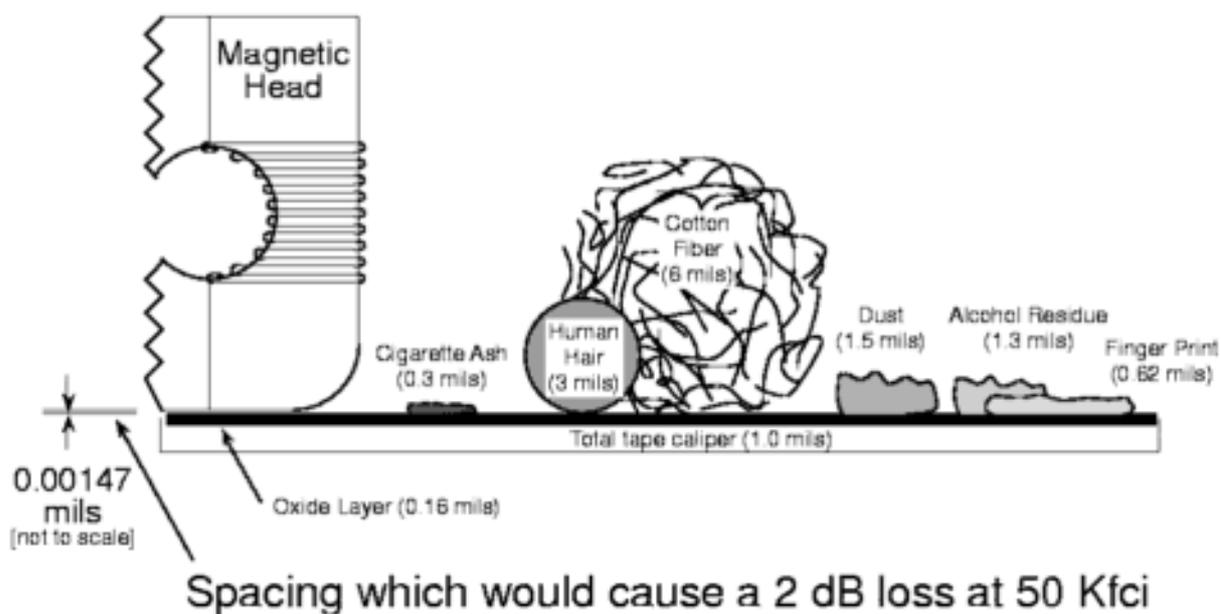


Figure 8. Size of Tape Debris Relative to the Tape/Head Spacing This figure shows the relative size of debris commonly found on tapes and on recorders relative to the tape-head spacing. It is clear from this diagram that even the smallest airborne particles can result in a dropout if the debris gets between the head and the tape.

## Corrosive Gases

Polluted air is known to cause problems with books, photographs, and works of art. Airborne sulfides, ozone, and nitrous oxides can cause accelerated deterioration of these objects. Silverware and black and white photographs are blackened by airborne sulfides produced by the degradation of wool fibers, the burning of coal, and bioeffluents. Magnetic tapes are no exception. They, too, are susceptible to corrosive gases in the environment.

Exposure to very low levels of corrosive gases representative of urban office environments has been known to cause corrosion on bare metal particle (MP) and metal evaporated (ME) tapes. In general, these tapes are contained in cassettes, and the cassette shells have been shown to be an effective armor against pollutants in the environment. This corrosion problem is limited to the metal based MP and ME tapes and

is not a significant factor in the deterioration of oxide tapes (iron oxide, chromium dioxide, barium ferrite).

If a tape archive is known to contain MP or ME based magnetic tapes, and the tape archive is situated in an environment characterized by high levels of pollutants (e.g., downtown Los Angeles), some precautions may be necessary to ensure that the level of chlorine and sulfides in the archive are at a sufficiently low level. Air conditioning systems may require special filters to remove pollutants if the archive is located in an urban environment.

## Storage Recommendations

Current industry standards recommend that materials be stored around 65 - 70° F (18 - 21° C) and 40 - 50% relative humidity (RH) (Table 1). Unfortunately, these recommendations are based, in part, on what is best for recording and playback, and what has historically proven to be good for film and paper storage. They may not be the best conditions for the long-term storage of magnetic media. Standards committees are beginning to consider storage conditions specific to magnetic tape and are recognizing that magnetic tapes benefit from storage at temperatures and humidities lower than those recommended in the past.

Agency/Researcher	Date	Temperature	Relative Humidity
Cuddihy	1982	65°F ± 3°F 18°C ± 2°C	40% ± 5%
SMPTE (RP-103)	1982	70°F ± 4°F 21°C ± 2°C	50% ± 20%
NARA	1990	65°F ± 3°F 18°C ± 2°C	40% ± 5%

Table 1. Current Recommendations for Magnetic Tape Storage Note: These are general recommendations that were being made in the 1980s. Standards committees are beginning to recognize the benefits of lower humidities and temperatures for the long term storage of magnetic tape. The above conditions may not be optimal for preserving magnetic tape for as long as is physically possible.

AES, ANSI, NARA, and SMPTE standards committees are coming to recognize that organizations have different storage needs and requirements. In some cases, information older than five years is considered obsolete. In other cases, information needs to be preserved in perpetuity. The optimal storage conditions for each of these requirements differs (Table 2). In the case of short-lived information, storage conditions can be at or near the room ambient conditions of the facility in which the tape collection is housed. No special storage facilities would be required, assuming that temperatures stayed between 68 - 76° F (20 - 24° C) year round and humidity never exceeded 55% RH. For the indefinite storage of information, special storage facilities would be required to maximize the lifetime of the media. No medium lasts forever, so transcription of information from old, deteriorating media to new media would eventually be required; however, storage conditions can be optimized to preserve the current media copy of the information for as long as possible.

Information stored at room ambient conditions would be readily accessible and playable. On the other hand, information stored in deep archive conditions would require a period of time to acclimate to the conditions of the facility in which the information would be played back. As such, the storage condition

recommendations are generally referred to as access storage and archive, or preservation, storage.

<b>Key Feature</b>	<b>Access Storage</b>	<b>Archival Storage</b>
Function	To provide storage for media that allows immediate access and playback.	To provide storage that preserves the media for as long as possible.
Acclimation required prior to playback?	No.	Yes.
Media Life Expectancy	At least 10 years when stored at the indicated temperature and humidity conditions.	The maximum allowed for the particular media type.
Temperature set point.	At or near room ambient. In the range: 60 to 74°F (15 to 23°C).	Significantly lower than room ambient. As low as 40°F(5°C).
Humidity set point	At or near room ambient. In the range: 25 to 55%RH.	Significantly lower than room ambient. As low as 20%RH.
Temperature variations	Difference between maximum and minimum value should not exceed 7°F (4°C).	Difference between maximum and minimum value should not exceed 7°F (4°C).
Humidity variations	Difference between maximum and minimum value should not exceed 20%RH.	Difference between maximum and minimum value should not exceed 10%RH.

Table 2. Key Features of Access and Archival Storage of Magnetic Tape Information represents a general summary of conditions being proposed in drafts of storage recommendations by SMPTE, ANSI, AES, and others.

Access storage conditions are recommended for those materials that need immediate access for playback purposes and for information that has a functional lifetime of ten years or less. Access storage conditions are close to the temperature and humidity conditions of the playback facility - generally room ambient conditions. The single, one-size-fits-all storage condition recommended for magnetic tape in the 1980s and early 1990s generally fit the category of access storage.

Archival storage conditions are recommended for materials that need to be preserved as long as possible. The conditions are specifically designed to reduce the rate of media deterioration through a lowering of the temperature and humidity content of the media. The temperature and humidity are also tightly controlled to reduce the deformation of the tape pack as a result of thermal and hygroscopic expansion/contraction.

Considerable cost is normally involved in maintaining a temperature/humidity controlled archive. However, as mentioned elsewhere in this report, the quality of care a magnetic tape receives should be commensurate with the perceived value of the information contained on the tape. If the information stored on the tape is of great value and must be preserved indefinitely, this could justify the cost of purchasing and maintaining the recommended archive facility. See Section 4.1: Tape Costs and Longevity for more information.

## Removal of Magnetic Tapes from Archival Storage

Tapes cannot be immediately removed from archival storage conditions and played on a recorder. Time must be allowed for the tapes to equilibrate to the temperature and humidity of the recorder environment prior to playback. This allows the stresses in the pack to equalize and the track shapes (helical scan) to return to normal. In the case of very low temperature storage, it may be necessary to place the tapes in an intermediate storage environment first to prevent condensation of moisture on the tapes and reduce stresses on the tape pack that would be introduced by rapid temperature changes.

In general, it is the width of the tape that determines how rapidly it will come to equilibrium. A tape that is twice as wide will take four times as long to stabilize to the new environment. Table 3 indicates the amount of time that should be allowed for the tapes to come to equilibrium after significant changes in temperature and/or humidity ("Heat and Moisture Diffusion in Magnetic Tape Packs," IEEE Transactions on Magnetics, 30 (2), March 1994: 237).

Tape Format	Time for Temperature Acclimation	Time for Humidity
Compact audio cassette	1 hour	6 hours
1/4-inch reel-to-reel	1 hour	1 day
2-inch reel-to-reel	16 hours	50 days
VHS/Beta cassette	2 hours	4 days
8mm video cassette	1 hour	2 days
U-matic cassette	4 hours	8 days

Table 3. Acclimation Times for Magnetic Media Removed from Archival Storage

A tape that is stored at a temperature or humidity that is significantly below that of room ambient conditions must be allowed to acclimatize prior to playback.

### 5.3 Refreshing of Tapes

In order to maximize their useful life, tapes may require periodic refreshing. This is a nonstandard term in the tape recording trade that can refer to the retensioning or rerecording of the tape, depending on the community of tape users. To avoid confusion, the terms retensioning and rerecording are preferred to refreshing.

Retensioning is normally recommended where prolonged tape pack stresses could cause damage to the tape. Some manufacturers have recommended that tapes be unspooled and rewound at regular intervals (often three years) to redistribute tape stress and prevent tape pack slip, cinching, and tape backing deformation. For example, retensioning was often recommended for large diameter tape reels, such as the old twelve-inch quadruplex videotape reels, so that tape stresses near the hub of the reel could be relieved. Some tape user communities refer to the process of retensioning as exercising the tape.

Rerecording requires that data be read from and written to the same tape periodically to refresh the magnetic signal and prevent data loss. Rerecording was employed primarily with some older nine-track computer tapes used in the 1960s and 1970s that were susceptible to print through.

Transcription, the copying of one tape to another, has also been referred to as refreshing. Transcription is the preferred term for this process. Tapes purchased today generally utilize small diameter tape reels and high coercivity magnetic pigments so that they often do not require retensioning or rerecording on a periodic basis. In some specific instances, tape manufacturers may still recommend the periodic retensioning of tape (see Ampex Guide in the Appendix, for example). It is best to check with the manufacturer to determine if tape retensioning is necessary.

Finally, refreshing should not be confused with restoration. Refreshing is a preventative maintenance procedure. Restoration refers to the reconditioning of a damaged or degraded tape in order to allow playback. Restoration is a repair or damage recovery procedure.

---

Proceed to: [Appendix: Ampex Guide to the Care and Handling of Magnetic Tape](#)

Go back to: [4. Life Expectancy: How Long Will Magnetic Media Last?](#)



# Magnetic Tape Storage and Handling

## A Guide for Libraries and Archives

Dr. John W.C. Van Bogart  
National Media Laboratory  
June 1995

[Table of Contents](#) | [Glossary](#)

---

Proceed to: [Estimation of Magnetic Tape Life Expectancies \(Les\)](#)

Go back to: [5. How Can You Prevent Magnetic Tape from Degrading Prematurely?](#)

---

## Ampex Guide to the Care and Handling of Magnetic Tape

The Ampex Recording Media Corporation, a U.S. magnetic tape manufacturer, has developed many informational and training materials about magnetic tape. The "Guide to the Care and Handling of Magnetic Tape" is reproduced here with the permission of the Ampex Recording Media Corporation. Additions, changes, and comments by NML are shown in square brackets [ ]. Some of the sections of this document deal with recorder aspects that may be beyond your control, such as wind speed and tension, if you are using a simple VHS, cassette, or reel-to-reel audio deck. However, these sections still contain useful information on what to look for as signs that the tape is damaged or needs to be copied. All sections of the original document are included for completeness, but not all sections may be appropriate for your particular tape collection.

### Recommended practices

- Tape should be handled only in no smoking, no food, clean areas.
- Do not let tape or leader ends trail on the floor.
- [Do not drop or subject to sudden shock.]
- Keep tape away from magnetic fields. Don't stack tapes on top of equipment.
- Tape storage areas should be cool and dry. Never leave open reel or cassette tapes exposed to the sun.
- Avoid subjecting tapes to rapid temperature changes. If storage and operating area temperatures differ by more than 15° F (8° C), allow an acclimatization time within the operating area of four hours for every 18° F (10° C) difference.
- Store open reel and cassette tapes with the reels or tape packs vertical. Reels should be supported by the hub. [Tapes should be stored like books on a library shelf - on end. They should not be stored laying flat.]
- Use high quality reels or cassettes, boxes/containers, and accessories.
- Return tapes to their containers when they are not in use.
- Cut off damaged tape or leader/trailer ends from open reel tapes.
- For open-reel tapes, use protecting collars if available.
- Do not use general purpose adhesive tapes to secure the tape end or for splicing. If necessary, use

adhesive products designed for the purpose.

- Minimize tape handling.
- Do not touch the tape surface or the edge of the tape pack unless absolutely necessary and then wear lint-free gloves.
- Clean the recorder tape path thoroughly at the recommended intervals.
- Discard tapes with scratches or any other surface damage, which causes significant debris to be left in the recorder tape path.
- Ensure tapes to be reused are thoroughly bulk-erased before they are put back into service.

## Cleanliness

Cleanliness is important because minute debris can cause loss of reproduced signal by disturbing the intimate contact necessary between the tape surface and the reproducing head. Figure 9 shows typical dimensions of common contaminants in the context of significant tape to head separation. A separation less than 1/10th of the diameter of a smoke particle will cause a 12 dB loss, reducing the signal to 1/4 of the proper amplitude.

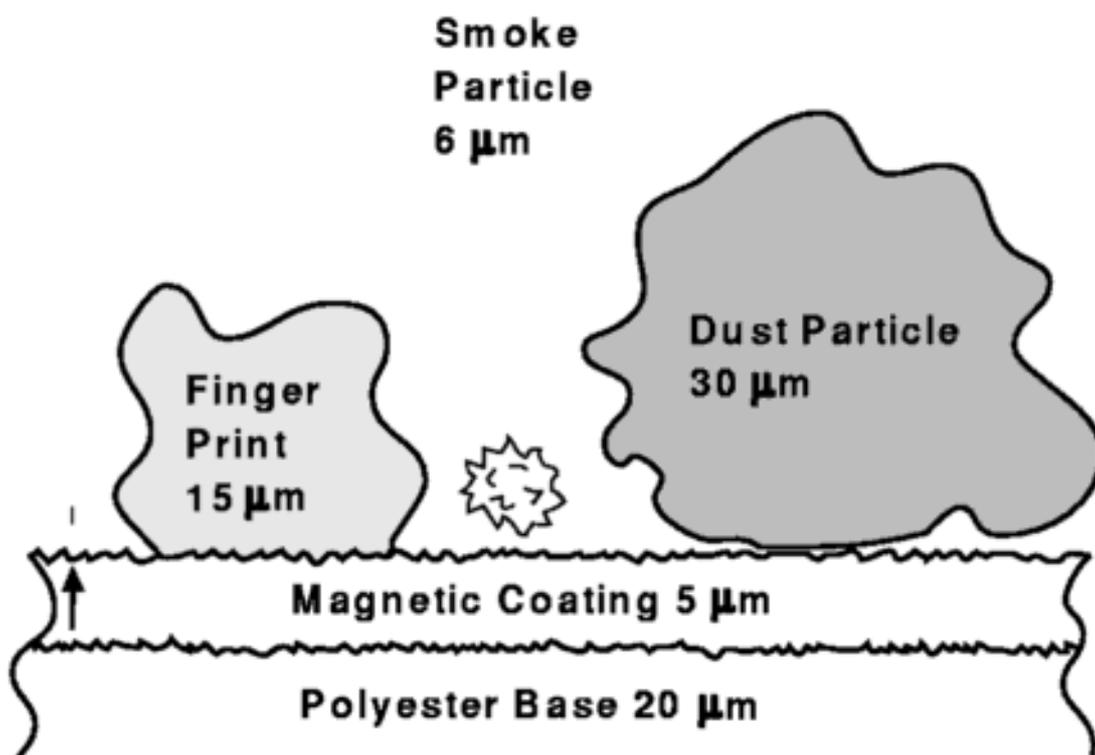


Figure 9. Tape Debris. (Source: Ampex. Reprinted with Permission.)

For analogue recording, especially audio recording, the effects of dirt and debris are much less important than for high density digital recording and video recording. Relatively severe dropouts will be unnoticed in analogue hi-fi reproduction and even worse dropouts will not impair the intelligibility of speech.

Dropouts are much more important in instrumentation data recording and any form of high density digital recording. If the signal losses are sufficiently severe to overwhelm the error correction, data errors may result.

In video recording, very short duration dropouts appear as irritating flashes in the picture, and in this case, perhaps unusually, the eye is more critical than the ear. For any type of recording, things are not as difficult as they appear because spacing due to debris is confined to only a small part of the track width, but the message for tape care is clear. However, most physical tape damage occurs when tapes are being loaded on a recorder or during handling before or after loading. It is, therefore, preferable that tapes be kept clean to avoid the need for special cleaning that involves extra handling and passage through additional mechanisms. For general purpose tapes, a class 10,000 clean room environment (less than 10,000 0.5 mm particles per foot) is a good aim. High density digital recording may benefit from cleaner conditions.

The worst contaminants, which should never arise, are sticky residues from improper tape end fixing tabs or elsewhere. Special end retaining tape or tabs have non-oozing clean peeling adhesive.

## Stray Magnetism

This is less of a problem than often thought. Devices such as walk-through metal detectors use small fields that have absolutely no effect. Hand-held detectors are best avoided as high local fields may be present. X-rays have no effect on unrecorded or recorded tapes. Similarly, radiation from radar antennas can be disregarded, unless the field strengths are sufficient to injure people. [Some detectors used to screen luggage in airports use powerful magnetic fields that may partially erase recorded information on tapes. These devices are used in some European airports.]

It is prudent to keep tape away from transformers, heavy electrical machinery, [and other very strong magnets]. Magnetizing forces of the order of 500 A/m and above can cause partial erasure and/or increase print through in the case of recorded tape. Such fields may put low frequency (LF) noise on unrecorded tape. This can be removed by bulk-erasure. The risk of increased print through applies to alternating fields that can act as a bias, encouraging layer-to-layer printing.

Magnetic field problems are very rare, even for tapes shipped internationally without special precautions. The best protection for shipping is a minimum of 50 mm [2 inches] of nonmagnetic material all round. The inverse square law ensures that the fields from even heavy electrical equipment will not affect tape at 50 mm [2 inch] distance. Metallic boxes and foils offer no useful protection against stray fields but may help exclude adverse environments.

## Tape Handling

### General

Cassettes provide good protection for the tape inside. Cassettes should be returned to their library boxes for additional protection when not in use.

The protection offered by reels can be improved if wrap-around collars fitting around or between the flanges are used. Such collars prevent the flanges deflecting and pressing against the edge of the tape; they also help exclude dust and retain the tape end, avoiding the risk of contamination with glue from unsatisfactory retaining tabs.

[Shock, such as dropping the tapes, should be avoided.]

## **Tape edge quality**

Tape is slit to precise widths with smooth straight edges. These qualities must be preserved if the tape is to perform well, [since most recorders edge guide the tape].

Modern recorders use narrow tracks. [If a tape edge is nicked, dented, bent or stretched] the recorder head [will not properly track over the recorded signal (mistracking)]. Bent or nicked reels, therefore, should be promptly discarded before significant tape edge damage results.

If an uneven tape pack is noted within a cassette, it may be appropriate to copy any valuable data for the same reason.

## **Tape pack/wind quality**

Tape is least vulnerable to external damage when wound in a smooth, even pack. Popped strands, where a few turns of tape stand away from the majority, are very easily damaged and should be avoided by using good quality tape and properly adjusted recorders.

Wound tape packs tend to loosen at low temperature (the tape thickness shrinks faster than the length). [This can also occur if the tape has reached high temperature and/or humidity and is brought back down to access conditions]. Vertical storage prevents pack slip under such conditions. Supporting reels by their hubs ensures the flanges are not deflected. In the ideal case, the flanges will then not contact the tape.

[Flange packing is a condition that occurs when the tape is either wound up against one flange by a poorly aligned recorder, or has fallen against the flange due to a loose wind and flat storage. Flange packing often leads to damaged edges from the tape scraping against the edge of the flange as it unwinds through the recorder or winds back to the reel. When a poor wind with popped strands is also present, the strands that stick out of the pack can be severely bent when the tape is flange packed.]

## **Embossing**

Reels should have smooth tape take-up surfaces. Even small bumps close to the hub will produce impressions in the tape repeating for several tens of meters. This embossing effect applies for lumps as small as 30 mm [1.2 mil; 0.0012 inch] high, and the impressions produce measurable tape-from-head separation. Note that even well-made splices stand higher than 30 mm so the embossing effect applies.

A wrinkled tape end on the hub can cause similar problems. A wrinkled or frayed end at the beginning of a tape is likely to deposit debris in the recorder tape path before embossing the tape as it winds onto the take-up reel.

## **Winding speed and tension**

As indicated above, a smoothly wound pack is always desirable. A nominal winding tension in the region

of 2.2 N [8 ounces] is appropriate for 25.4 mm [1 inch] wide tape with nominal thickness 25  $\mu$ m [1 mil; 0.001 inch]. For other widths and/or thicknesses, the tension may be adjusted pro-rata. At slow winding speeds (< 381 mm/s [15 inches/sec]), very little air is trapped in the pack as it is wound, and there is a negligible air lubrication effect. In these conditions, lower tension may be desirable.

Excessive tension (at any speed) leads to a tape pack showing radial lines known as spokes. These radial lines result from the pressure from outer layers in the pack compressing the inner layers so that the turns develop a small kink. These kinks align radially and appear as a spoke [when you look through the flange at the edge of the tape].

In severe cases, the periphery of the tape pack may lose its smooth round form and become lumpy. A tape showing any such signs of distress should be rewound immediately, ideally at a low speed (e.g., 760 mm/s [30 inches/sec]) and any valuable data copied. The tape may return to normal, but there is a risk of the edges having been stretched more than the center, which leads to wrinkled edges and subsequent tracking and tape-to-head contact problems.

Several different winding tension control systems are popular. Most tape leaving the factory is wound with constant torque. Many recorders wind with constant tension. There is also the so-called programmed winding tension advocated by several U.S. Government agencies. In this case, the tape is wound with low tension close to the hub. Increased tension is applied midtape and then the tension reduces again as the outer diameter is approached. A plot of tension (vertical axis) versus tape length (horizontal axis) gives rise to another name for this technique, which is the bathtub curve approach.

This special technique yields a pack with certain types of tape that survives a particular sequence of temperature and humidity cycles very well, but either constant tension or constant torque winding is perfectly satisfactory for normal applications and storage conditions.

## **Periodic rewinding**

For long-term storage, it is helpful to rewind tapes at an interval of not more than three years. This relieves tape pack stresses and provides early warning of any problems.

## **Rotary Head Recorders**

### **Tape scratches and head clogging**

All the foregoing considerations apply equally to stationary head and rotary head [VHS; 8mm] recorders; however, the much greater head-to-tape speed associated with the latter can lead to special problems if the tape becomes scratched. Tape scratches may be inflicted by damaged heads or a sharp surface somewhere along the tape path.

Scratches can also be caused by mobile debris reaching the spinning head area. In such cases, high temperatures can result at the head-to-tape interface, and a blob of molten debris can become welded to the head. This solidifies and, as it spins on the head, inflicts more damage on the tape. A head with such a damaging attachment neither records nor reproduces properly and is said to be clogged. It is, therefore, very important to be scrupulous in following the cleaning procedure recommended by the recorder

manufacturer.

If there is any suspicion of tape scratching, the recorder tape path and heads should be cleaned immediately to avoid risk of damage to other tapes. Similarly, a scratched tape should be taken out of use as soon as possible to avoid the risk of clogged heads and damage to other tapes. Once a tape is scratched, its surface integrity is lost, and it will tend to clog on even the most perfect recorder.

---

Proceed to: [Estimation of Magnetic Tape Life Expectancies \(Les\)](#)

Go back to: [5. How Can You Prevent Magnetic Tape from Degrading Prematurely?](#)



# Magnetic Tape Storage and Handling

A Guide for Libraries and Archives

Dr. John W.C. Van Bogart  
National Media Laboratory  
June 1995

[Table of Contents](#) | [Glossary](#)

---

Proceed to: [Further Reading](#)

Go back to: [Ampex Guide to the Care and Handling of Magnetic Tape](#)

---

## Estimation of Magnetic Tape Life Expectancies (LEs)

Magnetic tape degrades by known chemical processes. When the kinetics of these processes is fully understood, the degradation mechanisms can be modeled and the life expectancy (LE) of tapes can be estimated. The binder systems used in today's audio and videotapes are generally based on polyester polyurethanes. These polymers degrade by a process known as hydrolysis - where the polyester linkage is broken by a reaction with water. One of the by-products of this degradation is organic acids. These organic acids accelerate the rate of hydrolytic decomposition. Furthermore, the acids can attack and degrade the magnetic particles.

The lifetime of a tape is defined as the length of time a tape can be archived until it will fail to perform. Tape failure in terms of a change in tape properties will be a characteristic of the particular system on which the tape is intended for play. An end-of-life criterion is a key property and a value which, if exhibited by the storage medium, would indicate a situation where significant data loss is expected. For example, the degree of hydrolysis of a tape binder system is a critical property that may determine the lifetime of a magnetic tape. Figure 10 shows the life expectancy for a Hi Grade VHS tape assuming that the tape will fail when 12% of the binder polymer has hydrolyzed.

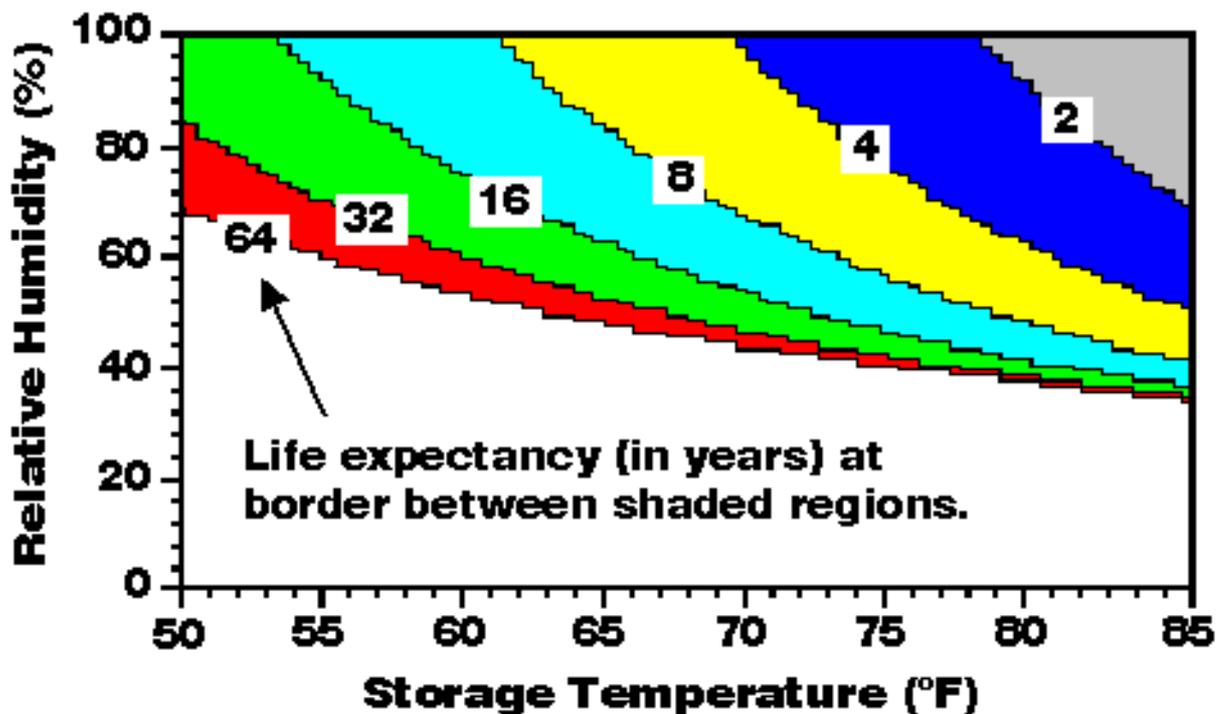


Figure 10. Life Expectancies for a Hi Grade VHS Tape Estimated by the degree of binder hydrolysis using an end-of-life criteria of 12%. LE values are indicated as a function of storage conditions.

Note that from the above chart, humidity is more important in determining the lifetime of the VHS tape than the storage temperature. At 20° C (68° F) and 50% RH, an estimated LE value of ~30 years is indicated. If the storage temperature is raised to 25° C (76° F) at 50% RH, the LE is reduced to ~10 years. However, if the humidity is raised to 80% at 20° C (68° F), the LE is reduced to ~5 years.

The life expectancy chart above was generated solely on the basis of a specific degree of hydrolytic degradation of the binder polymer. Tapes can fail for several reasons, however. Tapes can become too sticky to play as a result of an increase in the coefficient of friction or an overabundance of hydrolysis products. They can fail due to a loss in the magnetic signal as a result of a decrease in magnetic remanence or coercivity. They can fail because the magnetic coating has failed to adhere to the tape backing. They can fail due to irreversible shrinkage of the tape substrate.

The above information was provided to show how estimates of life expectancies can be made. The LE method outlined above is a simple explanation of a much more complicated issue. Standards committees such as the ANSI IT 9-5/AES Joint Technical Commission are endeavoring to determine procedures by which the life expectancy of magnetic tape materials can be determined.

#### **The Commission on Preservation and Access**

1400 16th Street, NW, Suite 740

Washington, DC 20036-2217

Phone (202) 939-3400

FAX (202) 939-3407

#### **National Media Lab**

Building 235-1N-17

St. Paul MN 55144-1000

**Phone (612) 733-3670**

**FAX (612) 773-4340**

---

Proceed to: [Further Reading](#)

Go back to: [Ampex Guide to the Care and Handling of Magnetic Tape](#)



# Magnetic Tape Storage and Handling

## A Guide for Libraries and Archives

**Dr. John W.C. Van Bogart**  
**National Media Laboratory**  
**June 1995**

[Table of Contents](#) | [Glossary](#)

---

Proceed to: [Glossary](#)

Go back to: [Estimation of Magnetic Tape Life Expectancies \(LEs\)](#)

---

## Further Reading

3M Technical Bulletin, 84-9811-2085-4, "Magnetic Tape Recording: Forever?"

3M Technical Bulletin, 84-9811-2075-5, "Temperature and Humidity Recommendation for VTR Facilities."

AMIA Newsletter, The Newsletter of the Association of Moving Image Archivists, c/o National Center for Film and Video Preservation, The American Film Institute, PO Box 27999, 2021 North Western Avenue, Los Angeles, CA 90027.

De Lancie P., "Sticky Shed Syndrome - Tips on Saving Your Damaged Master Tapes," *Mix*, May 1990, p. 148.

Ford, H., "Handling and Storage of Tape," *Studio Sound*, December 1984.

Geller, Sidney B., *Care and Handling of Computer Magnetic Storage Media*, NBS Special Publication 500-101, June 1983.

Jenkinson, B., "Long Term Storage of Videotape," *BKSTS Journal*, March 1982.

Kalil, F., "Care, Handling, and Management of Magnetic Tape," *Magnetic Tape Recording for the Eighties*, NASA Reference Publication 1075, April 1982.

Krones, F., "Guidelines for the Conservation of Magnetic Tape Recordings - Preservation and Restoration of Moving Images and Sound," *International Federation of Film Archives*, 1986.

Reilly, J., "IPI Storage Guide for Acetate Film," *Image Permanence Institute*, 1993.

Ritter, N., "Magnetic Recording Media: Part 1: Care and Handling of Magnetic Tape," 3M Company, 1985.

SMPTE Recommended Practice RP 103, "Care and Handling of Magnetic Recording Tape," 1982.

St.-Laurent, G., "Preservation of Recorded Sound Materials," ARSC Journal (Association for Recorded Sound Collections, PO Box 10162, Silver Spring, MD 20914), V.23, n.2, Fall 1992.

Wheeler, J., "Long-Term Storage of Videotape," SMPTE Journal, June 1983.

## **Resources for Transfer and Restoration of Video and Audio Tape**

[Note--The original printed document contained a list of persons identified as resources for audio and videotape recovery. This static listing has been replaced with the dynamic link below.]

[Resources for Transfer and Restoration of Video and Audio Tape](#)

---

Proceed to: [Glossary](#)

Go back to: [Estimation of Magnetic Tape Life Expectancies \(LEs\)](#)



# Magnetic Tape Storage and Handling

## A Guide for Libraries and Archives

Dr. John W.C. Van Bogart  
National Media Laboratory  
June 1995

[Table of Contents](#) | [Glossary](#)

---

Go back to: [Further Reading](#)

---

## Glossary

**Access storage:** Storage conditions at or near room ambient conditions that allow tape collections to be readily accessed for immediate playback.

**AES:** Abbreviation for Audio Engineering Society.

**Analog recording:** A recording in which continuous magnetic signals are written to the tape that are representations of the voltage signals coming from the recording microphone or the video camera.

**Analog-to-digital:** The process in which a continuous analog signal is quantized and converted to a series of binary integers.

**ANSI:** Abbreviation for American National Standards Institute.

**Archival storage:** Storage conditions specifically designed to extend or maximize the lifetime of stored media. Generally involves the use of temperatures and humidities lower than access storage conditions. Temperatures and humidities are also tightly controlled within a narrow range, and access by personnel is limited.

**Backing:** See substrate.

**Binary number:** A number that can be represented using only two numeric symbols - 0 and 1. A number in base 2.

Decimal Number	Binary Equivalent
0	0
1	1
2	10

4	100
12	1100
100	1100100
1995	11111001011

Binary numbers are used by computers because they can easily be represented and stored by device hardware that utilizes switches, magnetic fields, or charge polarities that are normally in one of two states. The on or off, north or south, or positive or negative states can easily represent the 1s and 0s of a binary number, respectively.

**Binder:** The polymer used to bind magnetic particles together and adhere them to the tape substrate. Generally, a polyester or polyether polyurethane based system. See polymer.

**Bit:** A single numeric character. Each bit of a binary number can either be 0 or 1. An n-bit number is composed of exactly n numeric characters. An n-bit binary number can have  $2^n$  distinct values. For example, an 8-bit binary number has  $2^8 = 256$  distinct values, namely all the numbers between 00000000 (0 in decimal) and 11111111 (255 in decimal), inclusive. 8-bit quantization would discretely sample a signal and assign each sampling a value between 0 and 255, permitting 256 possible values.

**Blocking:** The sticking together or adhesion of successive windings in a tape pack. Blocking can result from (1) deterioration of the binder, (2) storage of tape reels at high temperatures, and/or (3) excessive tape pack stresses.

**Cinching:** The wrinkling, or folding over, of tape on itself in a loose tape pack. Normally occurs when a loose tape pack is stopped suddenly, causing outer tape layers to slip past inner layers, which in turn causes a buckling of tape in the region of slip. Results in large dropouts or high error rates.

**Coercivity:** The level of demagnetizing force that would need to be applied to a tape or magnetic particle to reduce the remanent magnetization to zero. A demagnetizing field of a level in excess of the coercivity must be applied to a magnetic particle in order to coerce it to change the direction of its magnetization. Coercivity is the property of a tape that indicates its resistance to demagnetization and determines the maximum signal frequency that can be recorded by a tape. Hc is the common abbreviation for coercivity.

**Cohesive force:** The force that holds a material together. The force that holds a material to itself.

**Cohesiveness:** See cohesive force.

**Curvature error:** A change in track shape that results in a bowed or S-shaped track. This becomes a problem if the playback head is not able to follow the track closely enough to capture the information.

**dB:** See decibel.

**Decibel:** The unit of measure used to indicate relative changes in signal intensity or sound volume. The actual expression for calculating the difference in decibels between signal A and signal B is:

decibel (dB) =  $20 \cdot \log_{\text{base}10} (\text{signal A amplitude}/\text{signal B amplitude})$

+6 dB represents a doubling of the signal or a 100% increase

+5 dB represents a 78% increase

+4 dB represents a 58% increase

+3 dB represents a 41% increase

+2 dB represents a 26% increase

+1 dB represents a 12% increase

+0 dB represents no change-signals are equal

-1 dB represents a 11% decrease

-2 dB represents a 21% decrease

-3 dB represents a 29% decrease

-4 dB represents a 37% decrease

-5 dB represents a 44% decrease

-6 dB represents a halving of the signal or a 50% decrease

**Digital recording:** A recording in which binary numbers are written to the tape that represent quantized versions of the voltage signals from the recording microphone or the video camera. On playback, the numbers are read and processed by a digital-to-analog converter to produce an analog output signal.

**Digital-to-analog:** The process in which a series of discrete binary integers is converted to a continuous analog signal.

**Dropout:** Brief signal loss caused by a tape head clog, defect in the tape, debris, or other feature that causes an increase in the head-to-tape spacing. A dropout can also be caused by missing magnetic material. A video dropout generally appears as a white spot or streak on the video monitor. When several video dropouts occur per frame, the TV monitor will appear snowy. The frequent appearance of dropouts on playback is an indication that the tape or recorder is contaminated with debris and/or that the tape binder is deteriorating.

**Flange pack:** A condition where the tape pack is wound up against one of the flanges of the tape reel.

**Format:** The arrangement of information tracks on a tape as prescribed by a standard. The two most common categories of recording formats are longitudinal and helical scan.

**Head clog:** Debris trapped in the playback head of a video recorder. Clogging of the playback head with debris causes dropouts.

**Helical scan recording:** The recording format in which a slow moving tape is helically wrapped 180° around a rapidly rotating drum with a small embedded record head. The tape is positioned at a slight angle to the equatorial plane of the drum. This results in a recording format in which recorded tracks run diagonally across the tape from one edge to the other. Recorded tracks are parallel to each other but are at an angle to the edge of the tape.

**Hydrolysis:** The chemical process in which scission of a chemical bond occurs via reaction with water. The polyester chemical bonds in tape binder polymers are subject to hydrolysis, producing alcohol and acid end groups. Hydrolysis is a reversible reaction, meaning that the alcohol and acid groups can react

with each other to produce a polyester bond and water as a by-product. In practice, however, a severely degraded tape binder layer will never fully reconstruct back to its original integrity when placed in a very low-humidity environment.

**Hygroscopic:** The tendency of a material to absorb water. An effect related to changes in moisture content or relative humidity. The hygroscopic expansion coefficient of a tape refers to its change in length as it takes up water upon an increase in the ambient relative humidity.

**Longitudinal recording:** A recording format in which a slow or fast moving tape is passed by a stationary recording head. The recorded tracks are parallel to the edge of the tape and run the full length of the tape.

**Lubricant:** A component added to the magnetic layer of a tape to decrease the friction between the head and the tape.

**Magnetic particles:** The magnetic particles incorporated in the binder to form the magnetic layer on a magnetic tape. Iron oxide, chromium dioxide, barium ferrite, and metal particulate are various examples of magnetic pigment used in commercial tapes. The term pigment is a carry over of terminology from paint and coating technology - the magnetic coating on a tape is analogous to a coat of paint in which the magnetic particle is the paint pigment.

**Magnetic pigment:** See magnetic particles.

**Magnetic remanence:** The strength of the magnetic field that remains in a tape or magnetic particle after it is (1) exposed to a strong, external magnetic field and (2) the external field is removed. The property of a tape that determines its ability to record and store a magnetic signal. Mr is the common abbreviation for magnetic remanence. Magnetic remanence, Mr, and magnetic retentivity, Br, both refer to the ability of the tape to retain a magnetic field; however the latter is expressed in units of magnetic flux density.

**Magnetic retentivity:** See magnetic remanence.

**Mistracking:** The phenomenon that occurs when the path followed by the read head of the recorder does not correspond to the location of the recorded track on the magnetic tape. Mistracking can occur in both longitudinal and helical scan recording systems. The read head must capture a given percentage of the track in order to produce a playback signal. If the head is too far off the track, recorded information will not be played back.

**NARA:** The abbreviation for National Archives and Records Administration.

**Pack slip:** A lateral slip of selected tape windings causing high or low spots (when viewed with tape reel laying flat on one side) in an otherwise smooth tape pack. Pack slip can cause subsequent edge damage when the tape is played, as it will unwind unevenly and may make contact with the tape reel flange.

**PET:** Abbreviation for polyethylene terephthalate. The polymeric substrate material used for most magnetic tapes.

**Polymer:** A long organic molecule made up of small, repeating units (literally, many mers). Analogous to

a freight train, where each individual unit is represented by a freight car. At very high magnification, a chunk of polymer would resemble a bowl of cooked spaghetti. Plastic materials are polymers. The strength and toughness of plastics is due, in part, to the length of its polymer molecules. If the chains (links in the freight train) are broken by hydrolysis, the shorter chains will impart less strength to the plastic. If enough polymer chains are broken, the plastic will become weak, powdery, or gooey. See binder.

**Popped strand:** A strand of tape protruding from the edge of a wound tape pack. **Print through:** The condition where low frequency signals on one tape winding imprint themselves on the immediately adjacent tape windings. It is most noticeable on audio tapes where a ghost of the recording can be heard slightly before the playback of the actual recording.

**Quantization:** A process in which a continuous signal is converted to a series of points at discrete levels. The quantized version of a ramp, a continuum of levels, would be a staircase, where only certain distinct levels are allowed.

**Refreshing:** This term can refer to periodic retensioning of tape, or the rerecording of recorded information onto the same tape (or different tape) to refresh the magnetic signal. In the audio/video tape community, refreshing generally refers to retensioning of the tape, but it can also refer to the copying of one tape to another. See transcription.

**Relative humidity (RH):** The amount of water in the air relative to the maximum amount of water that the air can hold at a given temperature.

**Restoration:** The process where a tape degraded by age is temporarily or permanently restored to a playable condition. The tape backing procedure is an example of a tape restoration procedure.

**Retensioning:** The process where a tape is unspooled onto a take-up reel and then rewound at a controlled tension and speed. In performing this procedure, tape pack stresses are redistributed and, thus, the tape is retensioned. This has sometimes been referred to as refreshing (or exercising the tape).

**RH:** The abbreviation for relative humidity.

**Room ambient conditions:** The temperature, humidity, and air quality of the surrounding conditions. Those conditions generally found in a library, resource, studio, or office facility with a controlled environment (heating and air conditioning), which should range between 66 to 78° F (19 To 26° C) and 30 to 70% relative humidity year round. Analogous to room temperature conditions, except that this term only refers to the temperature of the room.

**Scission:** The process in which a chemical bond in a molecule is broken either by reaction with another molecule, such as water, or by the absorption of a high energy photon.

**Signal-to-noise ratio:** The ratio of the recorded signal level to the tape noise level normally expressed in decibels. Commonly abbreviated as S/N.

**SMPTE:** Abbreviation for the Society of Motion Pictures and Television Engineers.

**Stick slip:** The process in which (1) the tape sticks to the recording head because of high friction; (2) the tape tension builds because the tape is not moving at the head; (3) the tape tension reaches a critical level, causing the tape to release from and briefly slip past the read head at high speed; (4) the tape slows to normal speed and once again sticks to the recording head; (5) this process is repeated indefinitely. Characterized by jittery movement of the tape in the transport and/or audible squealing of the tape.

**Sticky shed:** The gummy deposits left on tape path guides and heads after a sticky tape has been played. The phenomenon whereby a tape binder has deteriorated to such a degree that it lacks sufficient cohesive strength so that the magnetic coating sheds on playback. The shedding of particles by the tape as a result of binder deterioration that causes dropouts on VHS tapes.

**Sticky tape:** Tape characterized by a soft, gummy, or tacky tape surface. Tape that has experienced a significant level of hydrolysis so that the magnetic coating is softer than normal. Tape characterized by resinous or oily deposits on the surface of the magnetic tape.

**Stress:** Force per unit area, such as pounds per square inch (psi). A tape wound on a reel with high tension results in a tape pack with a high interwinding stress. See tension.

**Substrate:** Backing film layer that supports the magnetic layer in a magnetic tape. PET is currently the most commonly used tape substrate.

**Tape baking:** A process in which a magnetic tape is placed at an elevated temperature for a brief time in order to firm up the tape binder. This procedure is recommended as a temporary cure for the sticky shed or sticky tape syndrome. The tape baking procedure is discussed in the reference, "Sticky Shed Syndrome - Tips on Saving Your Damaged Master Tapes," Mix, May 1990, p. 148.

**Tape noise:** A magnetic signal on the tape resulting from the finite size and nonuniform distribution of magnetic particles in the magnetic layer of the tape. Tape noise is inherent in any magnetic tape but can be reduced by using smaller pigment sizes in tape formulations. The iron oxide pigments found in less expensive tapes have the largest tape noise level. Ranked in size: iron oxide > chromium dioxide > metal particulate > barium ferrite. Therefore, ranked in order of tape noise: iron oxide > chromium dioxide > metal particulate > barium ferrite.

**Tape pack:** The structure formed by and comprised solely of tape wound on a hub or spindle; a tape reel consists of a tape pack, the metal, plastic, or glass hub, and flanges. Tape transport: The mechanics used to guide and move the tape through the recording system and past the read and write heads of the recorder. The tape transport consists of the tape guide pins, capstan, rollers, tension controllers, etc.

**Tension:** Force, or force per tape width. The force on a tape as it is transported through a recorder. A tape wound on a reel with high tension results in a tape pack with a high interwinding stress. See stress.

**Thermal:** An effect related to changes in temperature. The thermal expansion coefficient of a tape refers to its change in length upon a change in the ambient temperature.

**Track angle:** The angle that the track of a helical scan recording makes to the edge of the tape. This should correspond with the scan angle of the helical recorder - the angle that the tape makes to the equatorial plane of the rotating drum head. If the track angle and scan angle do not correspond,

mistracking will occur.

**Transcription:** The process of copying all of the information on one tape to another tape of the same or different format. The term refreshing is commonly used by some archivists and librarians to refer to the process of copying information from one tape to a newer tape of the same format (e.g., VHS to VHS). When the information is copied to a different format (e.g., BetaMax to VHS), the terms reformatting and converting have been used.

**Trapezoidal error:** A change in the angle of a recorded helical scan track. Can result in mistracking.

**Vinegar syndrome:** Characteristic of the decomposition of acetate based magnetic tape where acetic acid is a substantial by-product that gives the tape a vinegar-like odor. After the onset of the vinegar syndrome, acetate tape backings degrade at an accelerated rate - the hydrolysis of the acetate is catalyzed further by the presence of acetic acid by product.

---

Go back to: [Further Reading](#)

Return to: [Table of Contents](#)

Atomic Switch

Atomic Holographic Optical  
Nanostorage Drive

Spintronics

Display\_n\_Stor



Programmable  
Refracting OLED Lens

Holographic Imager

Homepage

Other Applications



## Dense holographic storage promises fast access

*John H. Hong and Demetri Psaltis*

Adapted from **Laser Focus World** April 1996 p. 119.

Scanned, filtered by optical character recognition. For further details please see the original article. This resource is intended for engineering students.

Science Center, Thousand Oaks, CA 91360. DEMETRI PSALTIS is professor of electrical engineering and executive officer of Computational and Neural Systems at the California Institute of Technology, Pasadena, CA 91125.

JOHN H. HONG is manager of optical information processing at Rockwell

With three-dimensional recording and parallel data readout, holographic memories can outperform existing optical storage techniques.

In its basic form, a hologram is the photographic record of the spatial interference pattern created by the mixing of two coherent laser beams. One of the beams usually carries spatial information and is labeled the "object" beam. The other is distinguished by its particular direction of travel and is labeled the "reference" beam. Illuminating the recorded hologram with the reference beam will yield or reconstruct the object beam and vice versa. As the holographic material becomes thicker, the reconstruction becomes very sensitive to the particular angle of incidence of the reference beam, which allows multiple objects to be recorded in the same volume and accessed independently by using an appropriate set of associated reference beams. Such holograms would be recorded sequentially, each object beam illuminating the holographic material simultaneously with its unique reference beam.

The angularly selective property of holograms recorded in thick materials enables a unique form of high-capacity data storage distinguished by its parallel data access capability. A holographic data storage system is fundamentally page-oriented, with each block of data defined by the number of data bits that can be spatially impressed onto the object beam. The total storage capacity of the system is then equal to the product of the page size (in bits) and the number of pages that can be recorded.

A holographic data storage system can be constructed to exploit this principle by using a spatial light modulator to properly shape the object beam, an optical beam scanner to point the reference beam, a detector array to convert the reconstructed output object data into an electronic bit stream, electronics to control the entire process and condition the input/output electronic information, and a sufficiently powerful laser to overcome the optical losses of the system (see Fig. 1).

Holographic data storage system developed by Rockwell for avionics applications is based on an acousto-optic addressing scheme and contains no moving parts.

In practice, the number of holograms that can be stored and reliably retrieved from a common volume

of material is limited to less than 10,000 so that spatial multiplexing techniques must be used. Although solid-state designs are possible, it is easiest to envision a storage material formed as a volume disk in which holograms in a particular cell are stored and retrieved by angular multiplexing and where random access to arbitrary cells is enabled by rotation of the disk (see Fig. 2).

## Demonstration Systems

The holographic 3-D disk being developed at the California Institute of Technology (Caltech) is a practical example of such a system. The 3-D disk shown in Fig. 2 has a 100-micron-thick photopolymer laminated onto a glass disk substrate. By superimposing 32 holograms each hologram consisting of 590,000 bits-in an area  $1.77 \text{ mm}^2$ , an areal density greater than  $10 \text{ bit}/\mu\text{m}^2$  has been demonstrated. No errors were detected in the reconstructions. Thicker recording media can yield densities in excess of  $100 \text{ bit}/\mu\text{m}^2$ . Another example is a system demonstrated at Rockwell that has no moving parts and is based on an acousto-optic addressing scheme (see photo on p. 119). Rockwell is developing the technology for avionics applications in which resistance to vibration and shock must be provided while maintaining the beneficial features of holographic storage. In addition to such technology demonstrations, Holoplex (Pasadena, CA) has recently developed and delivered a commercial holographic memory product that stores up to 1000 images, each consisting of  $640 \times 480$  pixels, and is capable of reading out its entire contents in one second (see Fig. 3).

Holographic data-storage systems use devices that are currently being refined by display and electronic imaging applications, so cost reductions and performance improvements can be expected in some proportion to the large volumes expected in those markets. Some of the key components, such as the spatial light modulator (used to compose a given data page) and the electronic detector array (which converts the holographic data reconstruction into an electrical signal) (see *Laser Focus World*, March 1996, p.38).

## Mass memory applications

Mass memory systems serve computer needs by providing archival data storage; emerging applications also involve network data and multimedia services. In general, such systems require high capacity and low cost.

New compact visible wavelength laser sources such as high-power semiconductor and frequency-doubled solid-state lasers are also available for use in holographic systems. Further optical and system engineering, including error correction and other issues, however, is needed to integrate such devices into demonstration systems. Several entities,

board random access memory (RAM). A variety of hardware approaches are currently available and can be classified with respect to two important performance measures: storage capacity and effective data transfer time (see Fig. 4).

The effective data transfer time is a measure of the time required to fully retrieve an arbitrarily located data block from the system and is a combination of the data transfer rate (the rate at which data are transferred from the mass memory to the user or CPU) and the data latency (the time lag

between the address set-up and the appearance of valid data at the output). The effective data transfer time is given by the sum of the data latency and the data block size in bytes divided by the data transfer rate in bytes per second. For specific comparisons, the block size must be chosen with care depending on the application of interest. No matter how fast the data transfer may be, a fundamental limitation exists for all approaches involving mechanical motion of either the read/write head or the storage medium.

Magnetic and optical tape storage systems are cost-effective for archival storage, when data access time is less critical. At the other extreme, flash memory, which is a solid-state semiconductor approach, offers extremely fast data access time at relatively low packing density but at high cost. Although incremental improvements to existing disk-based systems may be sufficient to address certain new applications, such as digital video CD-ROM, they will fall short of addressing applications in which both high capacity and short effective data transfer time are featured simultaneously, as is the case with network servers and image databases.

A new storage technology must displace the incumbent in all other portions of the storage spectrum. To do this, the demands of a mature market must be met, as well as technical challenges. New technologies that attempt to increase only the achievable storage capacity must compete with magnetic-and optical-disk-based systems for which tremendous resources are constantly brought to bear. In particular, the areal density of magnetic disk systems can be increased by use of optical tracking and better read/write heads, while for an optical system the areal density can be increased using multiple-layer CDs and variable pit depth recording. New data storage techniques must provide a hardware solution while conforming to cost/price constraints that have been imposed by existing technology.

We believe that holographic mass memory systems will have distinctive appeal for applications that require both high capacity and short data access time. This double technical goal is relevant in light of the numerous applications that exist in both commercial and military sectors for maintenance and usage of large image databases and digital video information. The low likelihood of incremental improvements to current storage technology fully meeting requirements for both high capacity and short access time presents an important window of opportunity for holographic mass memory systems and possibly other contending technologies.

In contrast to the currently available storage strategies, holographic mass memory simultaneously offers high data capacity and short data access time. Consider, for example, a system in which each page of data to be recorded and retrieved contains 1 Mbit of data. The storage of 500,000 such pages using a combination of spatial and common volume multiplexing will yield a total... To reach the desired capacity equivalent to magnetic systems, 50 such units must be spatially multiplexed, which is feasible. Moreover, because each page of data can potentially contain upwards of ~ Mbit of information, the parallel retrieval of a single page in a time interval as long as 100  $\mu$ s (the detector array response time) yields a total data transfer rate of ~0 Gbit/s 1.25 Gbyte/s.

The electronic data can be scanned in parallel using a multiple-tap CCD array. Magnetic disks, by comparison, typically feature 10 Mbyte/s data rates, RAID systems feature greater than 100 Mbyte/s, and CD-ROMs achieve about 1 Mbyte/s.

FIGURE 1. Basic components of a holographic memory system are a spatial light modulator, beam-steering device, and detector array.

FIGURE 2. Volume holographic disk stores a stack of holograms in a particular cell. Data are stored and retrieved by angular multiplexing, and random access to arbitrary cells is made possible by rotation of the disk.

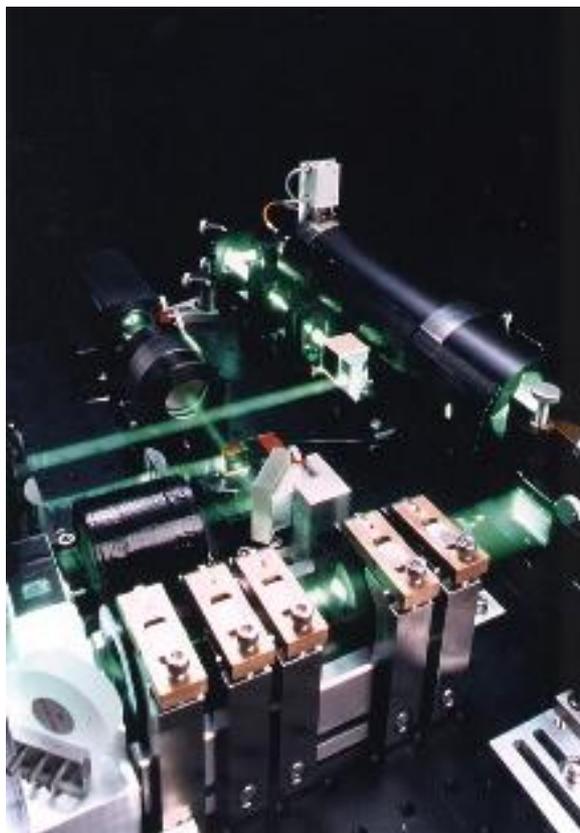
FIGURE 3. MH100 is the first commercially available holographic memory product.

FIGURE 4. Different hardware approaches to data storage are classified with respect to storage capacity and effective data transfer, two important measures of performance.

[Home](#) | [Products & services](#) | [Support & downloads](#) | [My account](#)→ **Select a country**[← IBM Almaden Home](#)[← IBM S&T Research](#)**Science and Technology**[Chemistry](#)[Computational Science](#)[Data Storage](#)[· Magnetic Materials and Phenomena](#)[· Novel Recording Technology](#)[Exploratory Technology](#)[Magnetism](#)[Nanoscale Science](#)[Quantum Information](#)[Scientific Services](#)→ [All Projects](#)**Related Links**[About S&T](#)[Student Opportunities](#)[Science Colloquium](#)[Collaborations](#)[Research S&T](#)[Career Opportunities](#)[Feedback](#)[Worldwide Labs](#)

# IBM Almaden Research Center

## Holographic Storage | Demon

**Overview**

The DEMON holographic storage demo is designed to implement as many key elements of a holographic storage device as possible with available components. It uses angular multiplexing in a single stack of holograms in a lithium niobate single crystal. Features include high resolution, pixel-matched imaging between the SLM and CCD arrays, near-video frame rates during readout, and the use of prototype modulation and error-correction codes developed at Almaden. Data processing is implemented via software running on two PCs. A short MPEG digital video (about 30 seconds of video, 6.7 MB of data) was stored and retrieved error-free using 1200 superimposed holograms. The DEMON was featured on the cover on the November 1996 issue of Laser Focus World.

**Novel Recording Technology Projects**[Holographic Storage](#)**Holographic Storage | Demon Related Links****Related subjects:**

- [Associative Retrieval](#)
- [Coding](#)

- [Predistortion](#)
- [Study of noise sources](#)

## Publications

### Journal publications

- G. W. Burr, J. Ashley, H. Coufal, R. K. Grygier, J. A. Hoffnagle, C. M. Jefferson, and B. Marcus "Modulation coding for pixel-matched holographic data storage," *Optics Letters*, **22**, 9, 639--641 (1997).
- G. W. Burr, H. Coufal, R. K. Grygier, J. A. Hoffnagle, and C. M. Jefferson, "Noise reduction of page-oriented data storage by inverse filtering during recording," *Optics Letters*, **23**, 289--291 (1998).
- G. W. Burr, W. C. Chou, M. A. Neifeld, H. Coufal, J. A. Hoffnagle, and C. M. Jefferson, "Experimental evaluation of user capacity in holographic data storage systems," *Applied Optics*, **37**, 5431-5443 (1998).
- G. W. Burr, M. A. Neifeld, G. Barking, H. Coufal, J. A. Hoffnagle, and C. M. Jefferson, "Gray-scale data pages for holographic data storage," *Applied Optics*, **23**, 1218-1220 (1998).

### Trade press

- J. Ashley, M.--P. Bernal, M. Blaum, G. W. Burr, H. Coufal, R. K. Grygier, H. Guenther, J. A. Hoffnagle, C. M. Jefferson, R. M. Macfarlane, B. Marcus, R. M. Shelby, G. T. Sincerbox, and G. Wittmann, "Holographic optical data storage: progress and promise," *Laser Focus World*, **32**, 11, 81--93 (1996).

### Conference presentations

- G. W. Burr, "Holographic storage," *IEEE Computer Elements Workshop*, June 1996.

- G. T. Sincerbox, M.--P. Bernal, G. W. Burr, H. Coufal, R.K. Grygier, J. A. Hoffnagle, C. M. Jefferson, R. M. Macfarlane, and R. M. Shelby, ``Evaluation of holographic recording materials from a storage systems perspective," *CLEO 1997*, May 1997, paper CMB3.
- G. W. Burr, J. Ashley, B. Marcus, C. M. Jefferson, J. A. Hoffnagle, and H. Coufal, ``Optimizing the holographic digital data storage channel," invited talk at *SPIE 1998*.

G. W. Burr, ``High-density holographic data storage," invited talk at *OSA 1998*

Search

Advanced search



Bell Labs Home

About Bell Labs

- ▶ History
- ▶ Awards
- ▶ People
- ▶ News & Features
  - 1999 Archive

Research Areas

Employment

Software Downloads

FAQs

## Lucent, Imation Developing Bell Labs Holographic Storage Technology

MURRAY HILL, N.J. (Aug. 11, 1999) -- Lucent and Imation Corp. today announced a joint agreement to collaborate on developing Bell Labs holographic data storage technology for the enterprise storage market.

Recent research advances at Bell Labs, the research and development arm of Lucent, in both disk drive and storage materials technologies indicate that the commercialization of holographic storage is possible. Imation, a leading provider of data storage media, will work jointly with Bell Labs to develop holographic disks.

Holographic storage provides dramatic advances in both data storage density and transfer rates required by the latest Internet applications and data warehousing.

Unlike other storage methods, which record only on the surface of a disk, holographic digital data storage allows recording through the entire thickness of the material, which allows for a huge increase in storage density. In addition, much higher transfer rates are achievable because the data is stored and recalled in "page format," which can be accessed one million bits at a time.

Based on the experimental advances, first generation drives would have the potential to store 125 gigabytes of user data on a removable 5.25-inch disk. This single disk capacity would be equivalent to that of 27 current 4.7 gigabyte DVD (digital versatile disk) disks. The transfer rates would be around 25 times faster than that of DVD.

"With this capacity, the information in a typical large university library could be stored on about 10 holographic disks," said Alastair Glass, director of Bell Labs Photonics Research Lab. "Future generations of devices are expected to store around a terabyte on a single disk with about 150 times the transfer rates of current DVDs."

"This joint development agreement is another example of how we seek strong partners to help us commercialize Bell Labs technology beyond Lucent's core businesses," said Steve Socolof, Lucent New Ventures Group Vice President.

"Imation has an impressive record in the development of optical storage technologies and information management solutions. We look forward to a productive relationship," he said.

"We are pleased to work with Lucent Technologies - a company with a well-deserved reputation for innovation across many fields - to co-develop this exciting new technology," said Steve Ladwig, President, Imation Data Storage and Information Management. "As Imation continues to expand its broad portfolio of data storage offerings to help customers manage and protect their ever-increasing amounts of data, we expect that this collaboration will fully leverage our respective strengths."

"We will be discussing our plans for developing this technology into a prototype data storage disk and device with systems companies, and we are evaluating this technology as a possible venture," Socolof said. "We plan to push this opportunity forward as quickly as possible."

Imation supplies a variety of products and services worldwide for the information and image management industry, specializing in data storage and information management, color management and imaging solutions. In 1998, the company reported revenues of approximately \$2.0 billion. As of June 30, 1999, Imation employed approximately 4,900 people worldwide.

Additional information about Imation is available on the company's Web site at [www.imation.com](http://www.imation.com) or by calling Imation at 1-888-466-3456. To receive recent earnings and news releases, corporate information and related shareholder services for Imation, call its shareholder information line at 1-888-IMN-NYSE (1-888-466-6973).

This information is based on a press release written by [David Bikle](#) of Lucent Technologies Media Relations.

[Terms of use](#) | [Privacy statement](#) | [Agere](#)

Copyright © 2002 Lucent Technologies. All rights reserved. \*



United Business Media

SEARCH

[Advanced Search](#)[Newsletters](#) | [ACE Awards](#)[Print Subscription](#) | [ProductCasts](#)
[HOME](#) [LATEST NEWS](#) [SEMI NEWS](#) [EDA NEWS](#) [LOCAL LANGUAGE](#) [DESIGN ARTICLES](#) [NEW PRODUCTS](#) [ABOUT](#) [FEEDBACK](#) [MEDIA KIT](#) [RSS](#) [CONTACT](#)

EE Times:

## Holographic storage nears debut

Margaret Quan

[EE Times](#)

(04/26/2001 10:59 AM EDT)

PRINT THIS STORY  
 SEND AS EMAIL  
 DISCUSS THIS STORY



**F**or more than 20 years researchers worldwide have pursued the Holy Grail of holographic data storage, an optical method of storing massive amounts of data in small areas by writing data as light patterns in three dimensions on a filmlike medium.

During that time, hardware advances carried out independent of holography have made holographic storage more achievable. These include such improvements as CMOS sensor technology, development of spatial light modulators using ferroelectric liquid crystals and mirror arrays, and reduction in the cost and size of shorter wavelength green lasers. Yet the biggest challenge has been to find the right material for the recording medium, one that works and is inexpensive enough to produce commercially.

In the last year, some research groups at universities, corporations, government labs and startups claimed to have found the material that will propel the technology forward and enable its adoption in commercial storage media products in two to three years.

However, storage industry analysts have a different view of when holographic storage will become reality.

Jim Porter, president of DiskTrend Inc., a storage industry research firm in Mountain View, Calif., said the holographic storage announcements he's heard have been mostly smoke and no fire, none of them having discussed specific products, and few offering realistic timelines for commercial availability.

In fact, Porter disagreed with the characterization of holographic storage as an "emerging" market, saying it's more of a nonexistent market.

Analysts such as Porter may be cynical about near-term expectations for the technology, but researchers in the field can't be more enthused about its prospects.

Rob Hermes, chief scientist working on holographic data storage at HoloStor, a division of Manhattan Scientific Inc. in Los Alamos, N.M., described research activity in this field as very hot and called it a "race" to create a good, inexpensive photopolymer.

Hermes is a polymer scientist who joined HoloStor a year ago

### Related Products

- [Railway supply powers signalling](#)
- [Low-cost UPSes include automatic voltage regulation](#)
- [High-speed cable suits design and debug apps](#)
- [Elan pocket-size oscilloscope has 100MHz analogue bandwidth](#)
- [100-V power modules extend to motor control, UPS apps](#)

### New White Papers

- [FREE White Paper on >> the 10 Essential Technologies](#)
- [PXI Measurement Suites for WLAN & GSM/EDGE](#)
- [Rapid SOC Design >> Using Configurable Processors](#)

[All White Papers >](#)

### MICROSITES

#### FEATURED TOPIC

#### ADDITIONAL TOPICS

- ▶ [Huge customer requirements, missed ship dates. Learn how DSO helps.](#)
- ▶ [Build high-speed designs using FPGAs with transceivers](#)
- ▶ [Maximize content delivery to grow future networks](#)
- ▶ [Learn what engineers want from memory products.](#)
- ▶ [FPGA design constraints? Use Platform ASICs/ Structured ASICs](#)

#### Sponsored Products

### Site Features

[Calendar Events](#)  
[Conference Coverage](#)  
[Forums](#)  
[Job Postings](#)  
[Multimedia](#)

[Print Edition](#)  
[Column Archive](#)  
[Special Reports](#)  
[Subscriptions](#)  
[Print | Digital](#)

after working at Los Alamos National Laboratories. He is perfecting a photopolymer that was originally developed by the government-funded MCC consortium in Austin, Texas in the 1980s that set out to develop systems for holographic data storage. The intellectual property was sold to Tamarack Storage Devices and finally ended up in the hands of businessman/executive Marvin Maslow, who founded Manhattan Scientific, and formed the HoloStor unit in 1998.

The firm owns more than 20 patents, including one for a photopolymer formulation that would form the basis for holographic media. This formulation is a proprietary mixture of nine components that Hermes claims has the required sensitivity or film speed, and the dynamic range (capacity) to be useful for storing several pages of data containing 1 million bits on each page.

To make holographic storage competitive, though, Hermes said, the medium must store a minimum of 100, 1 million-bit page images and eventually 1,000 page-images to rival existing technologies.

Today, Hermes works to enhance the formulation so that it has improved image quality (sharpness) and subsequent readability. HoloStor is targeting image lifetimes of 10 to 20 years, and proof that it can multiplex many images in the same spot or location (the 3-D part of holographic data storage).

Hermes said tests performed on the material at a facility he declined to name showed HoloStor's formulation to be good, but not exceptional—a requirement for commercial development—something he said is still three to five years away.

#### **Spin-off at work**

Meanwhile, researchers at Aprilis Inc. (Cambridge, Mass.), a Polaroid-spin-off that is developing holographic data storage systems, licensed a polymer patented by Polaroid that is based on epoxy-modified silicones. The material is based on Crop (cationic ring opening polymerization) chemistry.

According to Aprilis' vice president of research and development David Waldman, Crop solves the problem of volume shrinkage associated with conventional photopolymers formed by free-radical polymerization chemistry.

Volume shrinkage occurs when a photopolymer activated by laser light shrinks in volume as it moves from liquid to solid state and optically throws off the pixels recorded in the hologram to impair the image's fidelity. It is one of the bugaboos that's hindered the development of holographic media.

Aprilis' material was chosen to be the write-once holographic recording media in three public demonstrations of holographic data storage systems, including one at Stanford University in November 2000. That demonstration involved several companies in the Holographic Data Storage System (HDSS), a DARPA-funded group that includes IBM Corp.'s Almaden Research Center (San Jose, Calif), Rockwell Science Center (Thousand Oaks, Calif.) and Stanford University.

The demo achieved data rates of 6 Gbits per second with megapixel data pages, a good result.

Aprilis recently received a multimillion-dollar investment from venture firm Zero Capital (Cambridge) and is pursuing collaborative development of holographic storage systems by licensing technology from other players in the United States

and abroad.

Waldman said the company intends to commercialize both the holographic recording media and a holographic storage technology system in approximately two years.

Another startup in the race to provide a solution to the holographic media conundrum is InPhase Technologies, a Longmont, Colo., group launched out of Lucent Technologies in January. InPhase aims to commercialize technology based on a photopolymer material developed at Bell Laboratories, and storage media and manufacturing technology co-developed by Lucent and Imation Corp. (Oakdale, Minn.) under a 1999 agreement. InPhase gains access to these technologies via a license from Lucent.

The company's technology rests on the invention of a photopolymer with bit-storage characteristics and environmental ruggedness suitable for extreme temperature and humidity. In a 1999 interview with EE Times, InPhase Technologies' chief technology officer, Kevin Curtis, said the material possessed "more sensitivity and better dynamic range than lithium niobate," the preferred material at the time. No time line had been set for the introduction of products, and it was not clear which storage markets the business would target.

But two years later, in a January 2001 interview, the company's president and chief executive, Nelson Diaz, said he believes the technology will enable "point-of-sale kiosks" where consumers would purchase movies stored on very cheap media, or be incorporated into information systems for data archiving and retrieval. Development of commercial products is under way, but at least two years off.

#### **Down to earth**

As enthusiastic as the researchers are about holographic storage media, analysts are quick to bring their claims down to earth.

DiskTrend's Jim Porter said holographic storage media will have limitations at the get-go because most researchers say the first products will be write-once, read-many-times, rather than rewritable media. "Because the first products will be write-once, it means the first holographic storage media won't affect what you're doing today with magnetic-tape or magnetic-disk technology," Porter explained.

Write-once, read-many optical drives have a limited and specialized use in government and financial applications. These applications involve data being stored on media in robotic-driven libraries such as government storage of its huge Social Security database, where having three-dimensional storage would be more cost-effective than existing technologies. Porter said he could also envision the technology being deployed in banks or financial institutions, where the holographic media would enable a storage system that would let bank employees access images of checks on terminals instantaneously in offices worldwide. For the long run, Porter said he'd have to "wait a few years to see. I go by results, not promises."

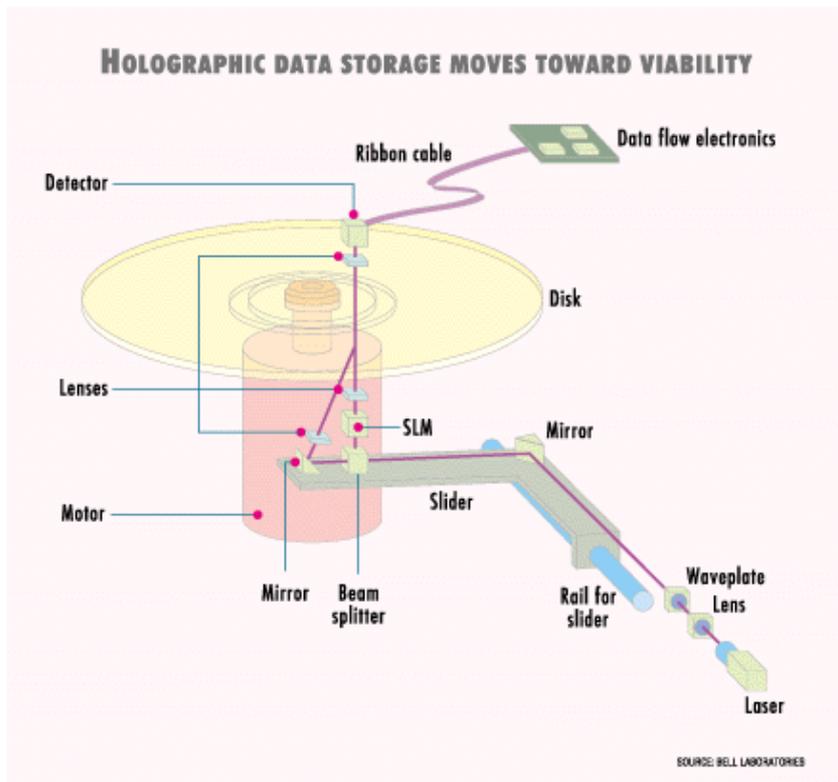
In the meantime, Constellation3D Inc. (New York), a developer of a new optical-storage media called Fluorescent Multilayer Disc (FMD), is creating technology it claims could enable large storage densities that go beyond compact disk and DVD, but fall short of holographic storage's promises. The company's technology is based on a fluorescent dipolymer as a coating for each storage layer on a CD-sized disk.

### Company secret

The fluorescent material remains a company secret, but was described to EE Times as an inexpensive material commonly used in food processing. A Russian chemist and expert in photochemistry invented the technology along with several colleagues who founded the firm in 1995. With researchers in Israel, Florida and Russia, the company is betting the technology will capture the interests of Hollywood and high tech.

John Ellis, the company's vice president of marketing, said his organization would demonstrate a DVD pit-density ROM disk with a 20-Mbit per second transfer rate in June 2001. In February 2001, Constellation3D inked a deal with Plasmon plc of London, a provider of recordable optical-disk technology, to develop production processes for mass production of FMD media. The company said it also signed letters of intent to work with manufacturers of CD-ROM drives, as well as equipment providers and a chemical film company.

Wolfgang Schlichting, an analyst covering removable storage at International Data Corp. (Framingham, Mass), said FMD technology has significant potential, "but the company will need partners to commercialize it in a big way." He said the company faces engineering challenges that it won't know about until it begins manufacturing. For that reason, the technology "has a long way to go."



**SPEC SEARCH**

eeProductCenter Launches SpecSearch®, New Parametric Parts Search Engine  
 In our continuing effort to enhance our site, eeProductCenter introduces SpecSearch® powered by GlobalSpec. [Click here.](#)

**Free Subscription to EE Times**

First Name	Last Name
Company Name	Title
Business Address	City
State	Zip
Email address	

**Electronics Marketplace**

• **[Measurement Computing- leader in low-cost USB DAQ](#)**

Analog I/O modules from \$99, including 8 channels, simultaneous sampling for \$399. Digital I/O modules from \$149. New temperature measurement modules USB-TEMP (all common sensor types, 8 channels, \$499) and USB-TC (thermocouple, 8 channels, \$299)!

• **[Measurement Computing: \\$99 for USB 24 DIO bits](#)**

The best value in DAQ. USB-1024LS offers 24 bits for just \$99. USB-DIO96H offers 96 DIO bits (with high current 64mA sink, 24mA source outputs) for \$299. USB-PDIS08 offers 8 electromechanical relays, 8 isolated digital inputs for \$299.

• **[Membrane Switches and Membrane Keyboards](#)**

Pannam Imaging, with its ISO 9001:2000 certification is the worldwide leader in the design and manufacture of custom membrane switch assemblies. Our digital printing capabilities allow for prototypes in less than 2 weeks.

• **[PXI Measurement Suites for WLAN & GSM/EDGE](#)**

As cellular and wireless data devices converge, Aeroflex addresses the test requirements with new measurement suites for WLAN and GSM/EDGE for use in conjunction with the 3000 series PXI Modular RF test platform.

• **[C Algorithm to Hardware RTL In Less Than a Day](#)**

Tensilicays XPRES Compiler automatically generates customized RTL engines from standard ANSI C/C++ . Graphically compare different performance/gate-count trade offs in minutes. Read the Microprocessor Report review.

[Buy a link NOW:](#)

**[HOME](#) | [ABOUT](#) | [EDITORIAL CALENDAR](#) | [FEEDBACK](#) | [SUBSCRIPTIONS](#) | [NEWSLETTER](#) | [MEDIA KIT](#) | [CONTACT](#) | [REPRINTS](#)**

**NETWORK WEBSITES**

[CommsDesign](#) | [DeepChip.com](#) | [Design & Reuse](#) | [Embedded.com](#) | [Embedded Edge Magazine](#) | [Embedded Computing Solutions](#) | [Planet Analog](#) | [eeProductCenter](#) | [Electronics Supply & Manufacturing](#) | [Inside \[DSP\]](#) | [Automotive DesignLine](#) | [Power Management DesignLine](#) | [Wireless Net DesignLine](#) | [Video/Imaging DesignLine](#) | [Green SupplyLine](#) | [Industrial Control DesignLine](#) | [Network Systems DesignLine](#) | [Digital TV DesignLine](#) | [Programmable Logic DesignLine](#) | [Audio DesignLine](#) | [Mobile Handset DesignLine](#) | [TechOnLine](#)

**INTERNATIONAL**

[EE Times JAPAN](#) | [EE Times Asia](#) | [EE Times CHINA](#) | [EE Times FRANCE](#) | [EE Times GERMANY](#) | [EE Times Korea](#) | [EE Times Taiwan](#) | [EE Times UK](#) | [Electronics Express](#) | [Elektronik i Norden](#) | [Electronics Supply & Manufacturing - China](#) | [Microwave Engineering Europe](#)

**NETWORK FEATURES**

[Career Center](#) | [Conference/Events](#) | [Custom Magazines](#) | [EE Times Info/Reader Service](#) | [GlobalSpec](#) | [NetSeminar Services](#) | [Sponsor Products](#) | [Subscribe to Print](#) | [Global Supply Chain Summit](#) | [Product Shopper](#) | [ProductCasts](#) | [Reprints](#)

All material on this site [Copyright © 2005 CMP Media LLC](#). All rights reserved.  
[Privacy Statement](#) | [Your California Privacy Rights](#) | [Terms of Service](#)





**Archives**

- Columns
- Features
- Print Archives
- 1994-1998

**Special**

- BYTE Digest
- Michael Abrash's Graphics Programming Black Book
- 101 Perl Articles

**About Us**

- How to Access BYTE.com
- Write to BYTE.com

**Newsletter**

**Free E-mail Newsletter from BYTE.com**

# BYTE.com

► **SEARCH:**

- HOME
- ABOUT US
- ARCHIVES
- CONTACT US
- REGISTER

- BYTE**
- ARTICLES
- BYTEMARKS
- FACTS
- HOTBYTES
- VPR
- TALK



## Creating Holographic Storage

### April 1996 / Cover Story / When Silicon Hits Its Limits, What's Next? / Creating Holographic Storage

A research team at IBM's Almaden Research Center has built a precision Photorefractive Information Storage Materials (PRISM) test stand for evaluating photosensitive samples. It also illustrates the fundamental components of a holographic storage system, as shown in the figure.

The device first splits a blue-green argon laser beam into separate reference and object beams. The object beam, which carries the data, gets expanded so that it fully illuminates a spatial light modulator (SLM). An SLM is simply an LCD panel that displays a page of raw binary data as an array of clear or dark pixels.

The object beam finally interacts with the reference beam inside a photosensitive crystal. The ensuing interference pattern--the substance of the hologram--gets stored

**Flexible C++**

*Matthew Wilson*  
My approach to software engineering is far more pragmatic than it is theoretical--and no language better exemplifies this than C++.

[more...](#)

**BYTE Digest**



as a web of varying optical characteristics inside this crystal. To read out the data, the reference beam again illuminates the crystal. The stored interference pattern diffracts the reference beam's light so that it reconstructs the checkerboard image of the light or dark pixels. The image is directed upon a charge-coupled device (CCD) sensor array, and it instantly captures the entire digital page.

When reading out the data, the reference beam has to hit the crystal at the same angle that's used in recording the page. The beam's angle is crucial, and it can't vary by more than a fraction of a degree.

This apparent flaw in the recording process is actually an asset. It's how holographic storage achieves its high data densities. By changing either the angle of the reference beam or its frequency, you can write additional data pages in to the same volume of crystal.

However, all the holograms appear dimmer because their patterns must share the material's finite dynamic range. In other words, the additional holograms alter a material that can support only a fixed amount of change. Ultimately, the images become so dim that noise creeps into the read-out operation, thus limiting the material's storage capacity.

The dynamic range of the medium determines how many pages it can hold reliably; therefore, the PRISM project examines the limitations in a variety of photosensitive materials. Current work uses iron-doped lithium niobate, strontium barium niobate, or barium titanate crystals. "We're also looking into polymers and other organic materials," says Glenn T. Sincerbox, the principal investigator from IBM.

Because the interference patterns are spread uniformly throughout the material, it endows holographic storage with another useful capability: high reliability. "While a defect in the medium for disk or tape storage might garble critical data, a defect in a holographic medium doesn't wipe out information. Instead, it only makes the hologram dimmer," he says.

The PRISM consortium has stored up to 200 holograms composed of 37.5-KB data pages (640 by 480 bits) into a crystal with less than 1 centimeter on a side, achieving a storage density of 48 MB per cubic cm. This is far short of the goal of a practical storage density of 10 GB per cubic cm, but it's sufficient to pursue the development of Holographic Data Storage System (HDSS) hardware.

Sincerbox believes that it will take several more years to refine the technology enough to build small desktop HDSS units. Such devices might be ready by about the year 2003.

Because HDSS hardware uses an acousto-optical light deflector (i.e., a crystal whose

*BYTE Digest* editors every month analyze and evaluate the best articles from *Information Week*, *EE Times*, *Dr. Dobbs' Journal*, *Network Computing*, *Sys Admin*, and dozens of other CMP publications—bringing you critical news and information about wireless communication, computer security, software development, embedded systems, and more!

[Find out more](#)

### BYTE.com Store



BYTE CD-ROM

NOW, on one CD-ROM, you can instantly access more than 8 years of BYTE.

refractive properties change according to sound waves traveling through it) to modify the beam angle, Sincerbox estimates that an HDSS system can retrieve adjacent data pages in under 100 microseconds. "Any conventional optical or magnetic storage unit will require some sort of mechanical means to access different data tracks, which takes on the order of milliseconds to accomplish," he explains. "A gigabit-per-second data rate appears reasonable for holographic storage, and this should make it a cost-competitive leader with whatever exists."

While holographic storage appears to be a radically new technology, actually it's not. The basic concepts were worked out almost 30 years ago. What's changed, according to Sincerbox, is the availability of key low-cost components. "Consumer electronics has played a large part in making holographic storage feasible today," he says. "Thirty years ago, lasers were made of glass tubes that were 6 feet long and had unreliable output. Now they consist of small, reliable, semiconductor junctions, similar to those mass-produced for CD players. The SLM is the result of fabrication techniques that make LCD screens for laptop computers and calculators. The CCD sensor array comes straight from a digital video camera. Neither of these were available 30 years ago--perhaps not even 10 years ago."

### How Holographic Storage Works

[illustration\\_link](#) (48 Kbytes)



*-- To read data out, the reference beam illuminates the crystal, and an image of the pattern gets projected onto a CCD array (e).*

*-- The object beam (a) passes through an LCD (b) that displays the data pattern. The object beam interferes with the reference beam (c) inside a crystal to make a hologram of the pattern (d).*



Up Level



Previous



Next



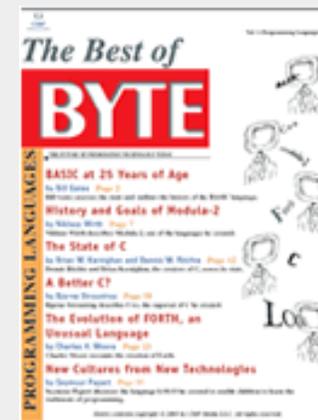
Search



Comment



Subscribe



### The Best of BYTE

### Volume 1: Programming Languages

In this issue of *Best of BYTE*, we bring together some of the leading programming language designers and implementors...

[Copyright © 2005 CMP Media LLC](#), [Privacy Policy](#), [Your California Privacy rights](#), [Terms of Service](#)

Site comments: [webmaster@byte.com](mailto:webmaster@byte.com)

SDMG Web Sites: [BYTE.com](#), [C/C++ Users Journal](#), [Dr. Dobb's Journal](#), [MSDN Magazine](#), [New Architect](#), [SD Expo](#), [SD Magazine](#), [Sys Admin](#), [The Perl Journal](#), [UnixReview.com](#), [Windows Developer Network](#)



**Archives**

- [Columns](#)
- [Features](#)
- [Print Archives](#)
- [1994-1998](#)

**Special**

- [BYTE Digest](#)
- [Michael Abrash's Graphics Programming Black Book](#)
- [101 Perl Articles](#)

**About Us**

- [How to Access BYTE.com](#)
- [Write to BYTE.com](#)

**Newsletter**

Free E-mail Newsletter from BYTE.com



► **SEARCH:**

- [HOME](#)
- [ABOUT US](#)
- [ARCHIVES](#)
- [CONTACT US](#)
- [REGISTER](#)

- BYTE**
- ARTICLES**
- BYTEMARKS**
- FACTS**



**When Silicon Hits Its Limits, What's Next?**

**[April 1996](#) / [Cover Story](#) / **When Silicon Hits Its Limits, What's Next?****

***A glimpse at three technologies that could be the subsystems of tomorrow's desktop computers***

***Tom Thompson***

In 1987, BYTE reported that the International Electronic Devices Meeting in Los Angeles had decreed that VLSI technology was on the verge of obsolescence. Only a year later, no less a personage than Jack St. Clair Kilby, inventor of the IC in 1958, philosophically told BYTE: "Nothing goes on forever. There may not be another five orders of magnitude of improvements to be made."

Today, Kilby's creation is 38 years old, and there are no signs that its influence will wane in the near future. Incredibly resourceful engineers have managed to push the

**Flexible C++**

*Matthew Wilson*  
My approach to software engineering is far more pragmatic than it is theoretical--and no language better exemplifies this than C++.

[more...](#)

**BYTE Digest**



bounds of fabrication techniques so that chips with submicron features are a common staple in today's desktop computers. For example, the 200-MHz Pentium Pro and PowerPC 604e have circuit features measuring only 0.35 micron across. The delivery of devices composed of 0.25- and 0.18-micron features is virtually assured; such chips are in the development phase and will ship in the next several years.

But there are signs that this technology is reaching its limits. While the features on the chip die have shrunk, the cost of the equipment necessary to fabricate these devices has ballooned. Intel alone has spent over a billion dollars apiece for the construction of several new "fabs" (the manufacturing plants that fabricate the chips) located in Oregon, New Mexico, and Arizona. Both IBM and Motorola have also broken ground on new high-price fabs.

The soaring costs of these facilities may eventually slow or halt the development of chips sporting ever-smaller features before the technological limits do. Once that happens, what does the microcomputer industry do next?

As small as these chip features are, they are still made up of huge aggregates of atoms. New computing technologies might operate on smaller scales, possibly at the molecular or even the atomic level. Or fundamentally new ways to handle information might be the answer, such as storing binary data as a holographic pattern whose data can be written or read in parallel.

This month, let's look to the future--specifically at two new storage media and one new CPU technology that may one day supplant silicon. But to do that, we must first examine the technology already in place.

### **It's Not Just a Good Idea, It's Moore's Law**

Since the IC was developed, the number of transistors that designers can pack on a chip has increased at a phenomenal rate. This rate, where the transistor count doubles approximately every 18 months, has become an axiom known as Moore's law. It's named after Gordon Moore, who first noticed this trend in the early 1960s. Within the span of 10 years, for example, the logic density in the x86 processor has increased 20 times, as shown in the figure "x86 Transistor Counts" .

The basis of these ever-higher logic densities is *photolithography* -- the same technology that etches the plates that print this magazine, only more complex. Here's how it works: Companies make ICs by layering patterns of metal or chemically treated (i.e., doped) silicon, one atop another, onto a die of silicon. The layout of these patterns, composed of either conductive or insulating material, builds the transistors that make up the IC's logic gates.

*BYTE Digest* editors every month analyze and evaluate the best articles from *Information Week*, *EE Times*, *Dr. Dobbs' Journal*, *Network Computing*, *Sys Admin*, and dozens of other CMP publications—bringing you critical news and information about wireless communication, computer security, software development, embedded systems, and more!

[Find out more](#)

### **BYTE.com Store**



**BYTE CD-ROM**

NOW, on one CD-ROM, you can instantly access more than 8 years of BYTE.

Adding a new layer first involves covering the die with a photosensitive coating. A mask in the shape of the desired pattern blocks light from reaching the coating, as shown in the figure "[The Limits of Silicon Fabrication](#)". Chemical processing etches off those sections of the coating that are exposed to the light. Logic gates thus get built, step by step, in a cycle where another doped layer gets applied, followed by another coating, another mask exposure, and more etching.

To accurately reproduce features onto the die, the wavelength of the light must be at least as small as the features themselves. Current lithographic processes employ a mercury light source whose 0.365-micron wavelength creates the 0.35-micron features. Successfully achieving the smaller 0.25-micron feature size requires the utilization of a krypton-fluoride ultraviolet laser that has a 0.248-micron wavelength.

Still-smaller features will be handled in the future by the use of argon-fluoride lasers with a 0.193-micron wavelength. But achieving 0.1-micron feature size requires optical trickery involving masks that phase-shift the light to improve the resolution. Building even-smaller chip features requires using light sources with even shorter wavelengths. In doing so, chip designers have traversed the electromagnetic spectrum from visible light, to ultraviolet light, and finally into X-ray territory.

But using X rays for the photolithographic process introduces a whole new set of production problems. With visible and ultraviolet light, masks are typically four to five times larger than the feature size. When the fab machinery projects the masks onto the die, lenses perform a reduction operation. With X rays, the masks must be the size of the features themselves, since X rays can't be focused with optical lenses. In short, making defect-free masks is as difficult as making the chip itself. Also, materials that are opaque to light aren't necessarily opaque to X rays.

Finally, there's the issue of having a reliable X-ray source. Mark Bohr, an Intel Fellow, hints at the scope of that problem by joking, "Part of the price tag of a future fab, if X-ray lithography is used, might very well be for the construction and operation of an on-site synchrotron."

John E. Kelly III, vice president of systems, technology, and science at the T. J. Watson Research Center, says that his group has fabricated logic gates as small as 0.07 micron using X-ray lithography. "They work--they switch--but there are still manufacturing challenges to be addressed," he admits.

Despite these hurdles, Intel and IBM say that current CMOS technology still has a lot of life in it. Says Bohr: "There's no sign of the technology slowing down. If we're going to run into a wall, it's more than 10 years out." Kelly agrees. "With CMOS technology and a lot of hard work, in a decade we'll use X-ray lithography and other techniques to deliver a processor that has 50 million to 100 million transistors and operates at 1 GHz," he predicts.



### The Best of BYTE

#### Volume 1: Programming Languages

In this issue of *Best of BYTE*, we bring together some of the leading programming language designers and implementors...

## Light Storage

Future compute-intensive jobs will present technical challenges in other areas besides the development of new processors. Whether they're made of CMOS or a fundamentally new technology, the quantity of data that these processors demand will tax the capabilities of other subsystems in a computer.

The capacities of today's mass-storage devices are indicative of this trend. Today, CD-ROMs are a common staple for distributing software, multimedia, and games. That's because they store up to 650 MB of error-corrected data on a single side of a platter. Magnetic-storage techniques are advancing rapidly as well. Within the last year or so, the typical storage capacity of the hard drive in a desktop computer jumped to more than a gigabyte.

Still, future computers will routinely handle hundreds of gigabytes or terabytes of information--orders of magnitude larger than the capacity of any existing CD-ROM or disk drive. Managing such vast quantities of data and delivering it in a torrent to an ultrahigh-speed processor requires a radically different type of storage system.

An optical recording technology known as *holography* shows great promise because it achieves the necessary high storage densities as well as fast access times. This capability occurs because a holographic image, or hologram, encodes a large block of data as a single entity in a single write operation. Conversely, the process of reading a hologram retrieves the entire data block simultaneously. (For more on the fundamentals behind holographic recording, see the sidebar "*Creating Holographic Storage*".)

Holographic data storage uses lasers for both reading and writing blocks of data, or "pages," into the photosensitive material. Theoretically, thousands of such digital pages, each containing a million bits, can be stored within the volume of a sugar cube. This is a storage density of 1 TB per  $\text{cm}^3$ . Practically, researchers expect to achieve storage densities of 10 GB per  $\text{cm}^3$ --still impressive compared to today's magnetic-storage densities, which are around 100 Kb per  $\text{cm}^2$  (not including the drive mechanism).

At this density, a block of optical media roughly the size of a deck of playing cards would house a terabyte of data. Because such a system can have no moving parts and its data pages are accessed in parallel, it's estimated that data throughput on such a storage device can hit 1 Gbps or higher.

The extraordinary capabilities of holographic storage have attracted the attention of universities, industry research labs, and the government. This interest has sparked two research projects. One is the Photorefractive Information Storage Materials (PRISM) program, a 2-1/2-year project jointly funded by the U.S. Department of

Defense's Advanced Research Projects Agency (ARPA) and other project members, [such as IBM's Almaden Research Center](#) (the principal investigator), GTE, and Rockwell International. The purpose of PRISM is to research optimal photosensitive materials for storing holograms and to understand their potential for storage.

The second research project is called the Holographic Data Storage System (HDSS). It has the same principal investigators as the PRISM project and includes such participants as IBM's Watson Research Center, Rockwell International, and GTE.

While PRISM investigates media, HDSS is developing the hardware technologies necessary to implement a practical holographic data-storage system. HDSS concentrates on building several key system components: a high-speed data-input mechanism, a sensor array to recover the data, and a high-powered red-light semiconductor laser (required for holographic I/O). These components will be integrated with the PRISM medium into prototype storage platforms to demonstrate the potential of this technology.

### **Molecules as Bits**

Even smaller objects might serve as storage devices or replace conventional semiconductor memory. Professor Robert R. Birge, director of the W. M. Keck Center for Molecular Electronics, has implemented a prototype memory subsystem that uses molecules to store digital bits.

The molecule in question is a protein called *bacteriorhodopsin*. This purple, light-harvesting protein is present in the membrane of a microorganism called halobacterium halobium, which thrives in salt marshes, where temperatures can hit 150. It uses the protein for photosynthesis when the oxygen levels in the environment are too low for using respiration to obtain energy.

Birge selected bacteriorhodopsin because its *photocycle*, a sequence of structural changes that the molecule undergoes in reaction to light, makes it an ideal AND data-storage gate, or flip-flop (see the figure "[Storing Bits in a Molecule](#)"). According to Birge, the *bR* (where the state is 0) and the *Q* (where the state is 1) intermediates are both stable for many years. This situation is due, in part, to the remarkable stability of the protein, which appears to have evolved to survive the harsh conditions of a salt marsh.

He estimates that data recorded on a bacteriorhodopsin storage device would be stable for approximately five years. "We have lab samples that have held information reliably for two years," he says. Another important feature of bacteriorhodopsin is that these two states have widely different absorption spectra. This makes it easy to determine a molecule's current state using a laser tuned to the proper frequency.

Birge has built a prototype memory system where bacteriorhodopsin stores data in a 3-D matrix. He builds this matrix by placing the protein into a cuvette (a transparent vessel) filled with a polyacrylamide gel. The cuvette is oblong and 1 by 1 by 2 inches in size. The protein, which is in the *bR* state, gets fixed in place by the polymerization of the gel. A battery of krypton lasers and a charge-injection device (CID) array surround the cuvette and are used to write and read data.

To write data, first a yellow "paging" laser fires to pump up the molecules to the *O* state. A spatial light modulator (SLM), which is an LCD array, slices this beam so that it excites a 2-D plane of material inside the cuvette. This energized plane of material is a data page that has the ability to hold an array of 4096 by 4096 bits. (See the figure "[How Molecular Memory Works](#)".)

Before the protein can return to its resting state, a red data-write laser, located at right angles to the paging laser, fires. Another SLM displays the binary data, and it sections up this beam so that certain spots on the page are irradiated. Molecules at these locations convert to the *Q* state and represent binary 1s on the page. The remainder of the page returns to the rest state and represents binary 0s.

To read data, the paging laser fires again, which excites the targeted page into the *O* state. This is done to further widen the absorption spectra differences between the digital 0s and 1s (the *Q* state). Two milliseconds later, a low-intensity red laser bathes the page. The low intensity is required to prevent the molecules from flipping into a *Q* state. Molecules representing 0s absorb the red light, while those in the binary 1 state let the beam pass through. This creates a checkerboard pattern of light and dark spots on the CID array, which captures the image as a page of digital information.

To erase data, a brief pulse from a blue laser returns molecules in the *Q* state back to the rest state. The blue light doesn't necessarily have to be a laser; you can bulk-erase the cuvette by exposing it to an incandescent light with ultraviolet output.

To ensure data integrity during selective page-erase operations, Birge caches several adjacent data pages. The read/write operations also use 2 additional parity bits to guard against errors. A page of data can be read nondestructively about 5000 times. Each page is monitored by a counter, and after 1024 reads, the page is refreshed via a new write operation.

How fast can data be accessed with this design? While a molecule changes states within microseconds, the combined steps to perform a read or write operation take about 10 milliseconds. However, like the holographic storage system, this device obtains data pages in parallel, so a 10-MBps rate is possible. This speed is similar to that of slow semiconductor memory.

By ganging up eight storage cells so that entire bytes can be accessed in parallel, Birge believes an 80-MBps data rate is possible. Maintaining this throughput depends on how you implement the memory subsystem. In some versions, the SLM does page addressing. Less-expensive designs use galvanometric mirrors that slew the beam to the correct page. While the SLM offers a millisecond response time, it also costs four times as much.

Says Birge: "Such a system would operate nearly as fast as semiconductor RAM until a page fault occurs. Then we have to reposition the laser beam to access pages on the other side of the container. Depending on the design, we can keep the page-fault access time in the milliseconds so that the memory system behaves like a hard drive during paging. Page caching appears to solve the access-time problem, but it's expensive because of the large page size [about 1.7 MB per page]. If you're willing to spend the money to cache about 10 pages, then you can eliminate the paging effects."

Theoretically, the cuvette described could hold 1 TB. Practically, Birge has stored about 800 MB on the cuvette, and he hopes to achieve a storage capacity of approximately 1.3 GB. Problems with the lens system and protein quality limit the system to this amount for now.

However, the merits of molecular storage have garnered sufficient interest that three of NASA's Space Shuttle missions explored methods to improve the manufacture of the data cubes by using microgravity. The resulting material was more homogeneous and provided an enhanced storage density. It remains to be seen, however, whether microgravity manufacturing will be sufficiently cost-effective to justify the observed factor-of-four improvement.

Birge's system, which he categorizes as a level-I prototype (i.e., a proof of concept), sits on a lab bench. He has received additional funding from the U.S. Air Force, Syracuse University (Syracuse, NY), and the W. M. Keck Foundation to develop a level-II prototype. Such a prototype would fit and operate within a desktop personal computer. "We're a year or two away from doing internal testing on a level-II prototype," says Birge. "Within three to five years, we could have a level-III beta-test prototype ready, which would be a commercial product."

Can molecular storage compete with traditional semiconductor memory? The design certainly has its merits. First, it's based on a protein that's inexpensive to produce in quantity. In fact, genetic engineering is being used to boost the output of the protein by the bacterium. Second, the system has the ability to operate over a wider range of temperatures than semiconductor memory.

Third, the data is stable. If you turn off the memory system's power, the bacteriorhodopsin molecules retain their information. This makes for an energy-efficient computer that can be powered down yet still be ready to work with

immediately because the contents of its memory are preserved.

Finally, you can remove the small data cubes and ship gigabytes of data around for storage or backups. Because the cubes contain no moving parts, it's safer than using a small hard drive or cartridge for this task.

### Quantum Computing

The scale of the function of new mass-storage and memory subsystems has grown progressively smaller. Holographic storage imprints data on crystalline lattices, and the rhodopsin memory system operates on batches of molecules.

But what about the processor itself? Is there a way to replace its machinery? Perhaps with something even smaller: individual atoms. For years, physicists have manipulated individual atoms in the lab. Now they're trying to coax computations out of them. But this work is like nothing you can imagine. At this scale, you get a whole new set of rules: The normal physical behavior that you expect even for minuscule CMOS logic gates no longer applies. Instead, quantum mechanics dictates the manner in which subatomic particles behave.

Quantum mechanics has every atom act as either a particle or a wave (the so-called wave-particle duality). This means that when subatomic particles behave as particles, they occupy only discrete energy states, called *quanta*.

When particles behave as waves, they exhibit strange counterintuitive behaviors. As the quantum wave that represents, say, an electron spreads out over time, its location becomes vague, and the laws of probability reign supreme. (The situation is analogous to throwing a rock into a pond: The wave centers around the point of impact the moment the rock hits the water. Over time, the wave spreads out over the surface of the pond and is everywhere.) The electron, in a sense, can be everywhere at once.

This fuzzy state of affairs continues until the electron interacts with another particle or photon that reveals its position, at which point the spread-out wave "collapses" into several localized waves (the electron and the other particle). As an example of this bizarre action, suppose a minute junction holds an electron. Its presence can be represented as a wave. This wave function has a certain probability that the particle can also be *outside* of the junction. Under the right conditions, the electron escapes from one junction to another by "tunneling" through the junction's walls, simply because the electron's wave function makes it probable to do so.

In the 1960s and 1970s, Rolf Landauer and Charles H. Bennett at the IBM Thomas J. Watson Research Center did research that investigated the basic physics of computing, which laid the groundwork for quantum computing. Notably, Bennett

demonstrated abstractly that you could build a molecular computer that implemented a Turing machine.

Around 1980, Paul Benioff of Argonne National Laboratory showed that computing could be done on a system that exactly obeys the laws of quantum mechanics. David Deutsch at the University of Oxford pointed out in 1985 that such a system could do quantum parallelism. While this research was still in the abstract stage, it indicated that a quantum computer could have greater capabilities than a classic digital computer.

In 1993, Seth Lloyd, who was then at Los Alamos National Lab, showed that many quantum systems, including an ordinary grain of salt, could function as quantum computers. That same year, Peter W. Shor of AT&T Bell Labs demonstrated that a quantum-mechanical computer could execute a practical task faster than any digital computer--factoring large numbers. All these findings have triggered a renaissance in quantum-computing research, where various groups are working on the construction of prototype components that represent quantum-computer "circuits."

The theoretical proposals to implement a basic quantum "gate" vary as widely as the number of research teams that are currently working on the problem. However, two groups have taken some important steps in demonstrations of actual laboratory implementations. This work has been carried out by David J. Wineland's group at the National Institute of Standards and Technology (NIST), which has built an XOR gate using an atomic ion held in a trap, and by H. Jeff Kimble's team at CalTech, which uses an optical cavity with a trapped atom to build a quantum phase gate (QPG). This latter gate's output, which modifies the phase shift of input laser beams, might be used to implement a variety of functions.

Constructing these building blocks isn't easy. NIST's logic gate involves a vacuum chamber with four electrodes, as shown in the figure "[How a Quantum-Logic Gate Works](#)".

Although the NIST group built a logic gate that implements the truth table of a classic electronic gate, it's important to note that quantum logic doesn't have to function that way. As mentioned earlier, quantum computing can exploit a kind of parallel processing because of that fuzzy-wave behavior of particles, and even the NIST gate exhibits this feature. "The state space of a quantum-computing system is far larger than the state space of a classic computer system, because the quantum system can exist in exponentially many states all at once," says Kimble.

Because of this, quantum bits are termed *qubits* to distinguish them from conventional bits. "A 3-bit register holds only one number, but a 3-qubit register can hold all eight possible numbers until you read it out," according to Chris Monroe, a member of the NIST team.

In theory, this quantum parallelism allows you to perform complex tasks quickly. For example, factoring a large number normally requires a computer to perform numerous divide operations, which can quickly reach an exponential amount of computations for large numbers. "A quantum computer would attack the problem by raising a smaller number to all different powers at once," explains Bennett. "A repeat period for a particular power function tells you how to factor the original number."

Furthermore, a quantum computer does not have to perform digital computations. The late Richard Feynman proposed that quantum computers could simulate other quantum-mechanical systems--in other words, operate as analog computers.

This idea is championed by Seth Lloyd, who's now with the department of mechanical engineering at MIT. As an example, Lloyd wants to simulate the time evolution of 40 particles that make up the matter at the core of an exploding star. Performing these calculations digitally would require setting up and working on  $2^{40}$  by  $2^{40}$  matrices that would accurately describe all the quantum characteristics of these particles, such as their spin.

"It would take  $10^{24}$  digital operations to compute the result," says Lloyd. "A TFLOPS system would require a trillion seconds--31,709 years--to compute the outcome. However, by using lasers to program the behavior of 40 ions in an ion trap, a quantum computer would have to operate for only a hundred quantum interactions." Such a quantum analog computer would use the very quantum properties of these particles, such as the spin, to compute the quantum effects of the simulation. Most of the purposes of quantum analog computing are similarly specialized.

Although quantum computing has lots of potential, there are still many problems yet to be solved. According to Landauer, there's the formidable issue of maintaining a coherent quantum system. "A quantum computer has to operate under two conditions that are hard to reconcile," he explains. "The qubits must interact strongly with one another to perform the computations. Yet they must do so without interacting with the environment itself. That's very difficult to do, especially if you're trying to perform computations over any length of time. For example, the thermal vibrations of the frame that holds the bits in their proper positions will cause the quantum logic to lose its coherence. Another problem is that flaws in the equipment cause errors to build up--unlike with digital computation, where at every stage the system is pushed back to a level of 0 or 1."

Monroe admits that "nobody's really studied these issues. Even the XOR gate loses coherence after 10 or 20 operations, perhaps due to minute instabilities in the laser." Bennett and others have investigated the use of error-correcting quantum codes to tackle the problem. According to Bennett: "Peter Shor discovered promising leads in quantum data storage for correcting errors. He proved that we could use 9 qubits to maintain an error-correcting code. It's not efficient, but it works. However, these

codes require reliable quantum processing to function. Unfortunately, it looks like we're going to need a breakthrough just to achieve reliable quantum processing."

Although the picture appears bleak, remember that quantum computing as a technology is still in its infancy. The situation is similar to when Bell Labs built the first transistor in 1947. Researchers are just starting to cast some of quantum computing's decades-old theories into real-world components that can do something. Says Kimble of the situation: "Implementing the quantum analog of classical circuits probably isn't the optimum strategy. Quantum physics is a rich and unexplored land where we're still discovering how to do things."

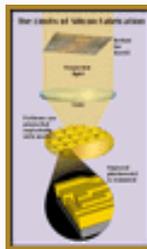
Even if quantum computing's problems are intractable, future processors will be built-- somehow. "Between the limits of conventional lithography and moving atoms around, there's a lot of space to build logic gates," says Kelly.

History is littered with technologies that showed great promise but failed to live up to expectations or usability. (See the sidebar "*Whatever Happened to Josephson Junctions?*" as a case in point.) This applies to all the technologies described here, not just quantum computing. Any one of them might founder due to unforeseen technical problems or because of cost issues. However, it's equally possible that offshoots from other disciplines might usher in a breakthrough, just as an eighteenth-century technology -- photolithography -- did for digital electronics.

---

## The Limits of Silicon Fabrication

[illustration\\_link \(36 Kbytes\)](#)



*Chip vendors use photolithography to etch patterns onto doped silicon layers. The smallness of the features is limited by the frequency of the light beam and the resolution of the lens.*

---

## **x86 Transistor Counts**

---

[illustration\\_link \(10 Kbytes\)](#)

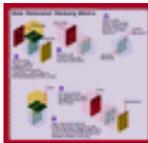


*The x86 processor's transistor count has increased by 20 times in 10 years.*

---

### How Molecular Memory Works

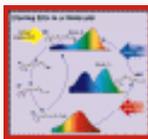
[illustration\\_link \(72 Kbytes\)](#)



- 1. An LCD array steers a yellow paging laser so that the beam excites a layer (i.e., a data page) of bacteriorhodopsin to the O state.*
  - 2. Another LCD array sections up a red data-write laser.*
  - 3. Spots on the page struck by this beam switch to the Q state, encoding binary 1s.*
  - 4. To read out data, the yellow laser fires, pushing the page into the O state.*
  - 5. Now a low-level red laser bathes the page. Molecules in the O state (0s) absorb light, while those in the Q state (1s) let the beam pass through, striking a CID array.*
- 

### Storing Bits in a Molecule

[illustration\\_link \(42 Kbytes\)](#)



*A photocycle is the sequence of structural changes that a molecule undergoes in reaction to light. The molecule remains at a resting state, known as bR. Yellow light starts the photocycle, where the molecule goes through several intermediate states,*

known as *K*, *M*, and *O*. If left alone, the molecule returns to the *bR* state. If the molecule is illuminated with red light during the *O* state, the photocycle detours into a *P* state, and then *Q*. The molecule remains at the *Q* state until irradiated by blue light, at which point it returns to the *bR* state. Both *bR* and *Q* are stable configurations and represent a binary 0 or 1, respectively.

---

### How a Quantum-Logic Gate Works

[illustration\\_link \(50 Kbytes\)](#)



An *XOR* gate built by the NIST research team. The chamber produces an electromagnetic field that suspends one beryllium atom. Two tuned ultraviolet lasers shine through quartz windows and manipulate the atom's state, namely its oscillation and spin. These two characteristics are used to implement a 2-bit register that behaves as an *XOR* gate. Another laser measures the atom's current state. If the atom fluoresces in response to the read-out beam, it's called a 0. If it doesn't, it's a 1.

---

### Holograms Could Be Mass Storage in the Future

[photo\\_link \(11 Kbytes\)](#)



Future mass-storage devices might use holograms to record digital information on a doped crystal, in a way similar to that of the test apparatus shown here at IBM's Almaden Research Center.

Commercial-scale equipment would be much smaller, have no moving parts, and use a high-powered semiconductor red laser. A crystal the size of a pack of playing cards would hold a terabyte of data.

---

*Tom Thompson is a BYTE senior technical editor at large with a B.S.E.E. degree from the University of Memphis. He writes extensively on Mac-related and general computing issues. You can reach him by sending E-mail to [tom\\_thompson@bix.com](mailto:tom_thompson@bix.com).*



**Up Level**



**Next**

[Copyright © 2005 CMP Media LLC](#), [Privacy Policy](#), [Your California Privacy rights](#), [Terms of Service](#)

Site comments: [webmaster@byte.com](mailto:webmaster@byte.com)

SDMG Web Sites: [BYTE.com](#), [C/C++ Users Journal](#), [Dr. Dobb's Journal](#), [MSDN Magazine](#), [New Architect](#), [SD Expo](#), [SD Magazine](#), [Sys Admin](#), [The Perl Journal](#), [UnixReview.com](#), [Windows Developer Network](#)

# Demetri Psaltis

## OPTICAL INFORMATION PROCESSING

California Institute of Technology

---

[Home](#)

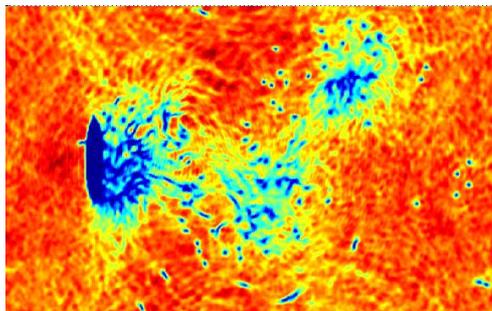
[Research](#)

[Courses](#)

[Publications](#)

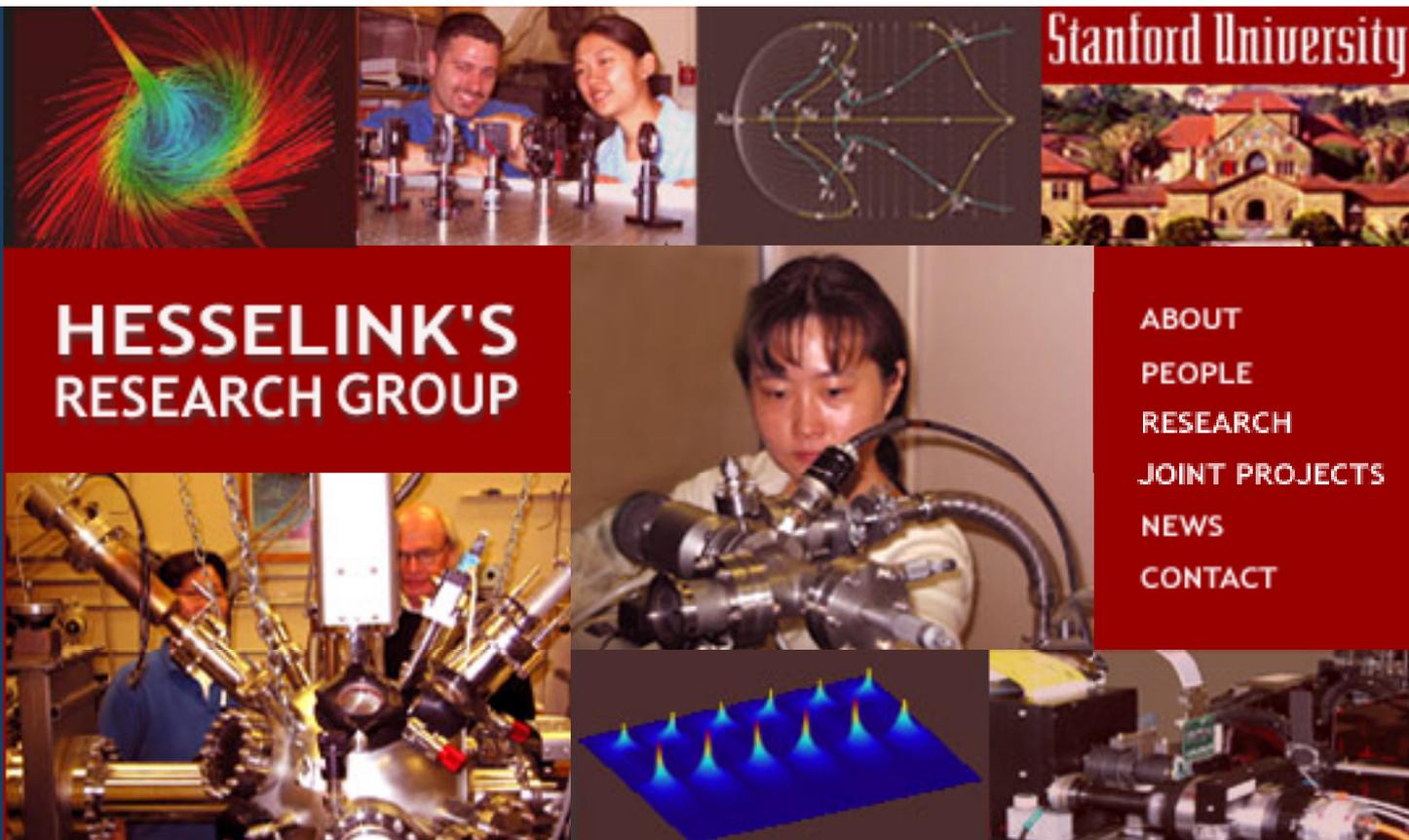
[People](#)

[Contact](#)



---

© California Institute of Technology | Last Update: April 29, 2005



Stanford University



# HESSELINK'S RESEARCH GROUP

- ABOUT
- PEOPLE
- RESEARCH
- JOINT PROJECTS
- NEWS
- CONTACT

© 2002 Hesselink's Research Group, Stanford University. All rights reserved. Email [webmaster](mailto:webmaster).

# Unscientific American

Date: Fri, 3 Feb 1995 10:15:12 -0500  
Sender: Association for Moving Image Archivists  
<AMIA-L@UKCC.UKY.EDU>  
From: Jim Wheeler <Jimwheeler@AOL.COM>

*The January Scientific American article "Ensuring the Longevity of Digital Documents" is very unprofessional and several people and organizations are writing letters to Scientific American about it. My letter is too long to be published in Scientific American but I want the Editor to understand the issue. I also enclosed a copy of my paper "Videotape Preservation".*

Jan 30, 1995  
Scientific American  
415 Madison Ave  
New York, NY 0017-1111

Dear Editor,

I have been a continuous subscriber to *Scientific American* for the past 40 years, and I have always considered your articles to be both scientific and understandable. I was shocked to read an article in the January edition that has several gross scientific errors in it. The author of the article "Ensuring the Longevity of Digital Documents" may be an expert on digital processing, but he certainly is not knowledgeable about the permanence of magnetic media.

Mr. Rothenberg lists the lifetime of magnetic tape as one year. I have been an archival tape engineer for over 30 years and I can say that is absolutely false! The National Institute of Standards and Technology (previously known as the National Bureau of Standards), the National Media Lab, the Battelle Institute, Ampex, and Sony have performed life tests on tapes and a 20 year life expectancy is considered reasonable for magnetic tape. Personally, I have family tapes 47 years old that play today without any special treatment or handling.

I suspect that the author is confusing failure to play back data properly with the failure of the tape itself. These are two completely separate issues! To explain the difference, I will use the common VHS videotape format as an example. JVC introduced the VHS format in 1976 and it is now the most used videotape format in the world. The term format is used to define the width of the tape, the speed of the tape, the width of the video tracks, the method used to record the signal on the tape, etc. In short, all of the specifications needed for someone to build a tape machine to play that particular tape. Videotape formats are documented and controlled by the Society of Motion Picture and Television Engineers (SMPTE). In the case of VHS, SMPTE has documented VHS as the "H" format.

What has happened with the cheap and popular VHS format, is that some companies are using the consumer VHS machines to record digital data. These consumer machines are mass-produced at the

rate of about a million per month and they are not built for durability or for high-quality recording/playback.

Using VHS machines incorrectly and for a purpose for which it was not designed has given videotape (and digital) a bad reputation. A dropout for video occurs in a flash and the eye may not even notice it. But, for data, a dropout might cause a multi-million dollar mistake. Consumers who use VHS machines for video will tolerate slightly degraded playback quality and occasional dropouts in the picture. That's what you get for a \$500 videotape recorder. On the other hand, professional tape recorders are precision machines built to a very high standard of performance, but the price is \$40,000 to \$100,000.

There are other errors in the article:

1. The idea of copying a tape every year is absurd. If the problem is with the format, then I suggest copying the information to a proven format. That will require only one copy, not one every year!
2. There are no common magnetic fields strong enough to erase magnetic tapes and discs, unless the media is purposely placed next to a magnet. This concern keeps being repeated by people who do not understand magnetics.
3. Digital is better than analog for long-term storage of data. This is because analog recordings lose some information each time they are copied, and there is no way of quantifying the quality of an analog recording. With digital (recorded on a professional tape recorder), the copy is a clone of the original. Also, with digital, the raw error rate can be monitored and a copy can be made when the uncorrected error rate exceeds an undesirable level. The copy is made using error correction.

The magnetic information recorded on the tape will last hundreds of years.

The common reasons for tape deterioration are:

1. Physical damage to the tape--usually caused by a faulty tape recorder.
2. Storing the tape in a high humidity and/or high temperature environment.

Magnetic tapes and discs should not be exposed to a temperature over about 80 & 176;F or a humidity over about 50 % RH for very long because this can cause the binder to degrade.

I do agree with the author on one thing, and that is the problem of format obsolescence. ALL high density media have the problem of becoming obsolete after a few years because someone will develop a new method of packing more data into less space and at a lower cost. Today, there are data tape recorders that can record the entire *Encyclopedia Britannica* on one cassette!

I should point out that ALL forms of high-density data storage media have a limited physical life as well as a limited format life. Gold-plated CD's should last hundreds of years, but finding a working CD player in the year 2050 will be extremely difficult. So, just as you must copy your 45's and LP's because record players are now obsolete, all archival information must eventually be copied to a new

format.

Also, *ALL* forms of data storage (not just magnetic media) must be stored in a *COOL & DRY* environment to prevent it from deteriorating. This is true for film, books, and CD's, as well as magnetic media.

I am a member of a American National Standards Institute (ANSI) committee and also a SMPTE committee which are developing standards for long-term storage of magnetic media.

Sincerely,

Jim Wheeler



---

[\[Search all CoOL documents\]](#)

[\[Feedback\]](#)

**This page last changed: October 18, 2004**

# Mag Tape Life Expectancy 10-30 years

Dr. John W. C. Van Bogart  
National Media Lab

---

Date: Mon, 13 Mar 1995 12:57:28 -0600  
Sender: Data Recording System and Media Information Exchange  
<DATA\_RECORDING@NML.ORG>  
From: "Wetzel, Peg" <pewetzel@MSMAIL.MMMG.COM>  
Subject: Mag Tape Life Expectancy 10-30 years

*Confusion and controversy over the expected lifetime of magnetic tape may have begun as the result of an article in the January 1995 Scientific American which cited magnetic tape life expectancy as 1-2 years. The NML refutes this figure. Years of research, industry, and operations support experience at the NML show magnetic tape life expectancy to be 10-30 years.*

*Dr. John W. C. Van Bogart, principal investigator of media stability studies at the NML, sent the following response to the editor of the Scientific American. Please feel free to comment.*

John Rennie, Editor in Chief  
Scientific American, Inc.  
415 Madison Avenue  
New York, NY 10017-1111

A Letter to the Editor of the Scientific American:

I am writing in regard to the article, "Ensuring the Longevity of Digital Documents," which appeared in the January 1995 issue of *Scientific American*. I am the Principal Investigator for the Magnetic Media Stability Program at the National Media Laboratory (NML), an industry resource supporting the U.S. Government in the evaluation of storage media and systems.

My NML colleagues and I agree with the key point made in this article-that the technological obsolescence of digital recording systems is a challenge for those individuals tasked with preserving digital archives. Digital archives should be transcribed every 10 to 20 years to ensure that they will not become technologically obsolete. To realize lifetimes greater than this, one would be required to archive the recording system, system software, operating system, computer hardware, operations manuals, and ample spare parts along with the recorded media.

My main contention is that the author has severely underestimated the physical lifetimes of digital magnetic tape. A chart in the article indicates that the physical lifetimes of magnetic tape are only one to two years. He states "digital magnetic tape should be copied once a year to guarantee that none of the information is lost," and "media with increased longevity are not on the horizon." Both of these statements are grossly inaccurate for current digital tape formats. Experience indicates that physical

lifetimes for digital magnetic tape are at least 10 to 20 years, a value commensurate with the practical life of the digital recording technology. One government agency responsible for maintaining meteorological data archives recently transcribed approximately 20,000 ten-year-old 3480 tape cartridges, of which only two cartridges had unrecoverable errors. Properly cared for reel-to-reel, 9-track computer tapes recorded in the 1970's can still be played back in the 90's, even though the 9-track format became obsolescent in the 80's. The NML has investigated the stability of several forms of digital storage media over the last six years. Life expectancies for magnetic media can be estimated by modeling the deterioration of tape properties induced experimentally in accelerated aging environments. Life expectancy estimates of 10 to 30 years for magnetic tapes are common. Given the fact that digital recording technologies can be supplanted by a newer format every 5 to 10 years, the bigger problem facing archivists is the lifetime of the technology, not the lifetime of the medium.

Of course, media life expectancies are like miles per gallon ratings on automobiles-"your actual mileage may vary." They are highly dependent on media storage conditions. In general, a controlled range of storage temperatures and humidities will increase media life expectancies. The National Bureau of Standards publication, *Care and Handling of Computer Magnetic Storage Media*, recommends that magnetic tape be stored at 65 +/- 3 degrees Fahrenheit and 40% +/- 5% Relative Humidity.

In conclusion, I believe that the author has been unduly pessimistic in his estimation of the physical life of digital magnetic media. Studies by the NML indicate that magnetic media, properly cared for, should have a lifetime which equals or exceeds that of the recording technology (10 to 20 years).

Sincerely,

Dr. John W. C. Van Bogart  
Principal Investigator, Media Stability Studies  
(612) 733-1918



---

[\[Search all CoOL documents\]](#)

[\[Feedback\]](#)

**This page last changed: August 03, 2004**



SEARCH:

[HOME CATALOGUE](#) [ASK US](#) [GUIDES](#) [INDEXES & DATABASES](#)

[FIND](#) ▼ [FOR](#) ▼ [HELP](#) ▼ [ABOUT US](#) ▼ [VISIT US](#) ▼ [NEWS & EVENTS](#) ▼ [SHOP](#) ▼

Presented by Ross Harvey, Division of Information Studies, [Nanyang Technological University](#) (Singapore) at the [2nd National Preservation Office Conference: Multimedia Preservation - Capturing the Rainbow](#), in Brisbane, 28-30 November 1995.

- [Abstract](#)
- [Introduction](#)
- [Rothenberg et al.](#)
- [Magnetic Tapes](#)
- [Optical Disks](#)
- [Other Media](#)
- [Training Needs](#)
- [Conclusion](#)
- [References](#)

---

## Abstract

A critical issue for the preservation of interactive multimedia is how to ensure that the digital data of which it consists maintains its integrity and remains usable. There are two possible approaches to take when considering the preservation of digital data: to preserve the artefact on which it is stored, or to direct efforts towards migrating the data (or digital 'object') to new systems as they are introduced. This paper describes the first approach and examines its implications for organisations which are committed to maintaining digital data for any length of time.

The digital storage media examined are magnetic tapes and optical disks (including magneto-optical disks). For each medium the claims of manufacturers about their longevity, results of accelerated aging tests, and observations from field sites are presented. Recent research, including that presented by Jeff Rothenberg and by the National Media Laboratory, St. Paul, Minnesota, is noted.

The paper concludes that there are at present too many unknowns to commit digital data to currently-available artefacts for anything other than short-term storage. The preferred option is to direct preservation efforts towards solutions which preserve the information content - the digital 'object' -

rather than the digital 'artefact'.

---

## Introduction

A critical issue for the preservation of interactive multimedia is how to ensure that the digital data of which it consists maintains its integrity and remains usable. There are two possible approaches to take when considering the preservation of digital data:

- to preserve the artefact on which it is stored; or
- to direct efforts towards migrating the data (or digital 'object') to new systems as they are introduced.

My task in this paper is to describe the first approach and examine its implications for organisations which are committed to maintaining digital data for any length of time. This task has been made significantly easier by the publication in June 1995 of John Van Bogart's report *Magnetic Tape Storage and Handling: A Guide for Libraries and Archives* [1]. It would be churlish of me to lament the fact that an equivalent report for optical disks does not exist, but one can at least hope that this will soon appear.

### *The Issues*

First, some working definitions:

- *multimedia* are physical formats in which information in more than one medium is stored: because multimedia in use today are recorded digitally, I am concerned here with digital data.
- *archival* appears to have many meanings. Archivists and librarians mean life spans of several hundred years. Manufacturers of compact discs talk in terms of decades. Computing people may talk of up to two years. Note a recent statement: 'Archivability' is defined as 'How long the media will last on the shelf and still be playable . . . [it] is normally thought of as being a medium problem, but it also becomes a machine problem after the machines cease to be manufactured [2]. This last point - machine obsolescence (and software obsolescence too) - are other significant parts of the equation.

Others more eloquent than I have captured the essential issues in words. Don Waters, [Yale University Library](#), in his position paper *Some Considerations on the Archiving of Digital Information* notes:

Preserving the media on which information is electronically recorded is now well understood to be a relatively short-term and partial solution to the general problem of preserving digital information. Even if the media could be physically well-preserved, *rapid changes* in the means of recording, in the formats for storage, and in the software for use *threaten to render the life of information in the digital age as, to borrow a phrase from another arena of discourse on civil society, 'nasty, brutish and short.'* [3]

Archivists have been forced to develop strategies to preserve electronic records because they have had in their care for many years significant quantities of such records. In 1988 this thinking was put into print by the American David Bearman:

*We must begin by accepting the information life of specific recording formats as a fact of physics. While we can influence the production of new media and formats and encourage current information recorders to use formats with longer lives, the 'format life' of any given format is the outside boundary beyond which we cannot rationally plan to retain the information without transforming the medium.* [4]

The Commission on Preservation and Access's *Annual Report July 1, 1991-June 30, 1992* summarises the issues for libraries. Its President, Patricia Battin, argues that 'we must remove the burden of *archival copy* from the . . . artefact' and refocus on 'the concept of managing continuing access to information stored on a variety of media and requiring a variety of ever-changing access hardware and software.' To this end she believes that 'preservation policies must now focus, not on the permanency of the medium, but on the management of permanency in the digital environment.' [5]

So, in relation to electronic records, we are rapidly moving away from the conventional preservation approach - attempting to preserve the artefact. With the rapid rise of electronic records, it is now obvious that the notion of 'saving object X for Y years' will become obsolete, or perhaps applied only to specific categories of information-carrying artefacts such as the book. The primary questions become those of **what is worth keeping**, and **for how long**; only when we have answered these can we look at the question of what medium to convert to. The issues here are fragility of each medium, rapid rate of obsolescence of the operating apparatus (software, operating system, etc.), the ease of altering the data, ownership of information (copyright, etc.), and who takes responsibility for its preservation.

As earlier indicated, my role in this paper is to concentrate on the matter of what medium to convert to. To address this question requires knowing about the physical and chemical makeup of the media and about optimum conditions for their storage and handling. Information about these is available: but it is not readily accessible to the consumers (by which I mean librarians, archivists and indeed anyone without the scientific and technical education to understand this information). This point is made by the writers of another report initiated by the Commission on Preservation and Access, this one entitled *Research on Magnetic Media-Phase 1*:

*While there is ongoing research and data available on durable or 'robust' magnetic media, archivists and librarians do not have ready access to this information and are generally unable to interpret the technical data resultant from the research. We know or understand little about the nature of the media, how they react to ambient conditions in storage and during use, and how these properties relate to the long-term preservation and use of information recorded on magnetic media. . . . In order to make decisions regarding migration of recorded information . . . archivists and librarians need to be able to predict the life expectancy of magnetic media when stored under a diverse and*

*variable set of environmental conditions'*. [\[6\]](#)

In short, the key points are:

- we need to decide how long we want to keep digital information
- there is a choice between keeping the physical objects themselves (the 'digital artefacts') in usable condition, or keeping the data contained in the physical object (the 'digital object') in a state in which it can be used. [\[7\]](#)

To these must be added a third point:

- if we choose to preserve the digital artefact, then we must be aware that it is a short-term expedient.

## **Rothenberg *et al.***

Jeff Rothenberg's article in the January 1995 *Scientific American* has succeeded in focusing popular attention on this question of the short times we can expect our digital artefacts to last. His conservative estimates ('expected lifetimes are estimated conservatively to guarantee that none of the data are lost', he notes) are [\[8\]](#):

Magnetic tape years	1 year	equipment obsolescence	5
Videotape years	1-2 years	equipment obsolescence	5
Magnetic disk years	5-10 years	equipment obsolescence	5
Optical disk years.	30 years	equipment obsolescence	10

Rothenberg's estimates caused a flurry of refutation, including a letter from staff of the National Media Laboratory to *Scientific American* which indicated that 'the physical lifetimes for digital magnetic tape are at least 10 to 20 years.' [\[9\]](#)

Current advertising in the press presents another story. We cannot blame the readers of popular computing magazines or the computer supplements of daily newspapers if they appear to be confused. A completely unscientific sample of current advertising provides the following:

- Reports in *The Straits Times* (Singapore) variously state that the life-span of CD-ROM is approximately 100 years (1 March 1995), almost indefinite (7 March 1995), and for CD-R (Compact Disc-Recordable) a 'long shelf life . . . 10 years' (8 February 1995)
- Digitised images stored on CD-ROM will be 'yours to keep without deterioration in quality. Hey! It will still be in mint condition for posterity's sake when your descendants see it!' (*HomePC* (Singapore, October 1995): p.71)

- By comparison, an advertisement by Pinnacle in *PC Magazine* (21 November 1995, p.43) claims a shelf life for CD-R of one hundred years:

The advertisement is split into two columns. The left column is titled 'TAPE IS OUT.' and features a black magnetic tape cartridge with a red prohibition sign (a circle with a diagonal slash) over it. Below the tape are five bullet points: '• Tape is slow', '• No random access', '• Five-year shelf life (Avg.)', '• Too many different formats', and '• Reliable?'. The right column is titled 'OPTICAL IS IN.' and features a colorful CD-ROM. Below the CD are five bullet points: '• Recordable CD is fast', '• Random access', '• One hundred-year shelf life', '• CD-ROM standard format', and '• Very reliable'.

(Source: *PC Magazine* 21 November 1995, p.43)

Let me again note (although it is not within the province of this paper to dwell on this aspect) that even where the life expectancy of the digital artefact can be estimated to be in decades, the equipment's life-cycle is only at the most ten years. Even given the possibility of developing software emulators address this issue, there is still a major problem.

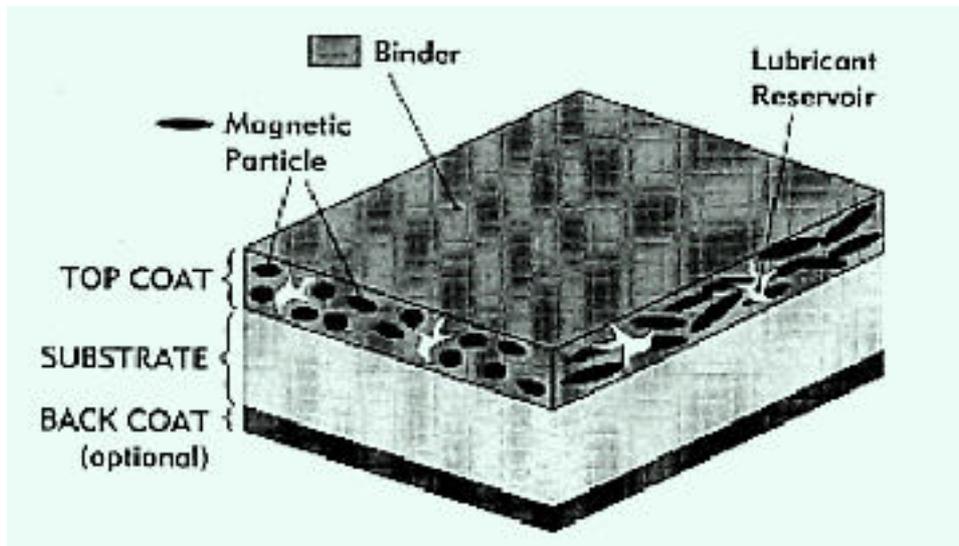
This paper examines primarily the two main digital storage media, magnetic tape and optical disks.

## Magnetic Tapes

The only magnetic medium I will note here is magnetic tape, as it is the primary magnetic medium in serious use for storing non-current digital data. This section is a summary of Van Bogart's already mentioned report *Magnetic Tape Storage and Handling: A Guide for Libraries and Archives*. While his report refers to video- and audiotapes, the author has advised me that it can be extrapolated to data tape, whose structure and composition is almost identical. There are, however, a few differences. [\[10\]](#)

### *Physical Structure*

Some basic knowledge about the structure of magnetic tape allows a better understanding of its preservation problems. The information is stored in the alignment of magnetic particles which are suspended within a polymer binder. This binder adheres the magnetic information carrying layer to a base or substrate; it also provides a smooth surface to ensure that the tape passes smoothly through the tape heads. Other substances are added, for example a lubricant to reduce friction, and a head cleaning agent.



**Cross Section Of Magnetic Tape**  
(Source: Van Bogart (1995) Figure 2)

### *Effect of Physical Structure on Longevity*

The **binder** is the most significant factor which determines the longevity of magnetic tapes. It may soften or become embrittled through hydrolysis, a chemical reaction which requires water to be present for it to occur. The polyester linkages in the binder break - the more moisture in the air, the more likely hydrolysis is to occur. Hydrolysis leads to the 'sticky tape' phenomenon where the binder can stick to the recorder heads and lead to clogging of the heads, dropout and other problems. Another problem is lubricant loss, which occurs with the passage of time, even if the tape is unplayed. Less lubricant means increased friction and the 'sticky tape' phenomenon can occur.

The **magnetic particles** store data in the form of changes in the direction of the magnetism in the particles. These particles (or pigments) vary in their ability to remain magnetically stable (which directly affects the quality of data recorded) according to what they consist of. Although the most stable are iron oxide and cobalt-modified iron oxide, these are not used for high quality tapes where metal particulate (MP) and chromium dioxide (CrO<sub>2</sub>) pigments are used because of their superior characteristics for recording higher frequencies and allowing higher signal outputs. Van Bogart notes 'there is not much that can be done to prevent the magnetic deterioration that is inherent in the metal particulate and chromium dioxide pigment types' [11] but indicates that storing the tapes in lower temperatures slows the rate of deterioration.

The **substrate** is most commonly made of polyester film (Mylar, polyethylene terephthalate, or PET). [12] Polyester film is well known to be chemically stable and will easily outlast the binder; rather, its problems arise from mechanical problems such as stresses on the tape caused by fluctuations in temperature and humidity levels in storage areas. These can result in mistracking during playback. Deformation of the substrate can also arise if the tape is not appropriately stressed when it is wound or rewound.

Other factors which affect data loss include whether the recording is helical scan (for example videotapes) or longitudinal scan (for example analog audio tapes) and, of course, the quality and

maintenance of the tape recording device itself. [13]

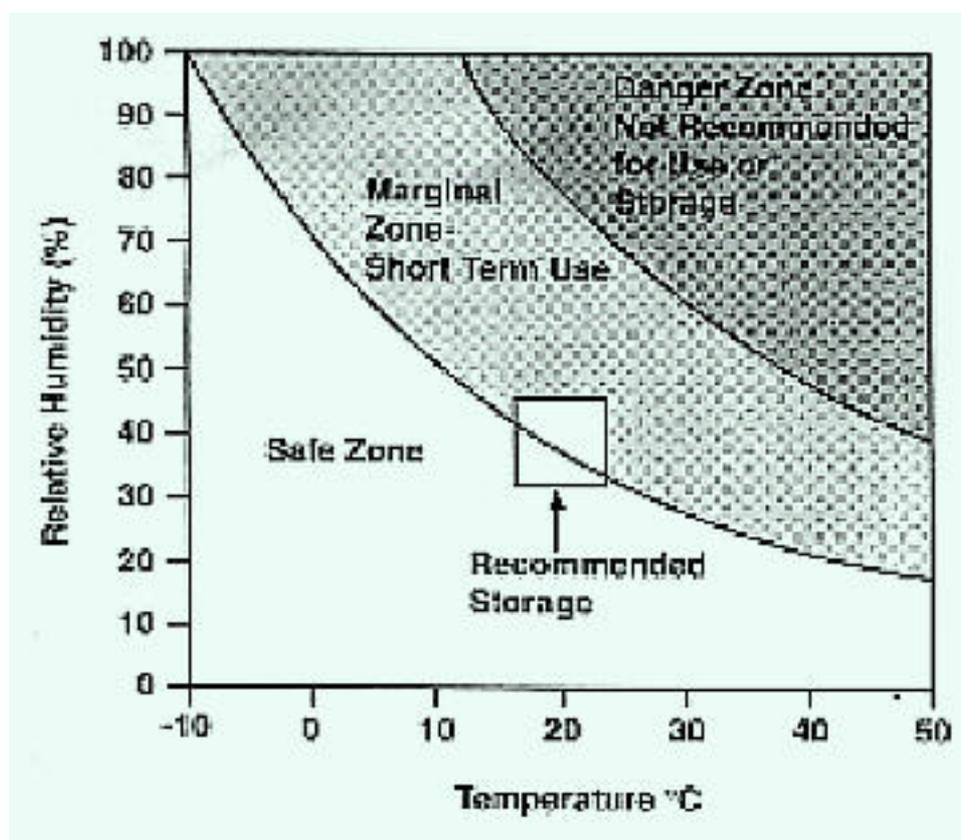
### *Improving Longevity*

It follows from the above description of physical structure of magnetic tape that ways to improve their longevity are based around:

- care and handling: quality of storage conditions, care in handling, number of times the tape is accessed
- quality of the tape
- future availability of the technology to play back the tape.

Only over the first of these can we exercise any real control as custodians of digital data.

I refer the readers to Van Bogart's report [14] for detailed descriptions of the care and handling required, and continue here with storage conditions. Because binder hydrolysis is the key factor in tape deterioration, and as this depends on the moisture content of the tape, lowering humidity levels means reduced rates of hydrolysis and lower temperatures slow down the rate of hydrolysis. Similarly, the magnetic pigments degrade more slowly at lower temperatures. Reducing the variation of fluctuation of temperature and humidity levels also assist. Storage at high temperatures (indicated by Van Bogart to be higher than 23°C) increases the tightness with which the tape is packed, thereby increasing the distortion of the tape backing and resulting in an increase in permanent dropouts. Storage at relative humidities higher than 70% can also result in increased tape pack stresses as the tape absorbs moisture and expands. Fungal growth may also occur at high temperatures and humidities.



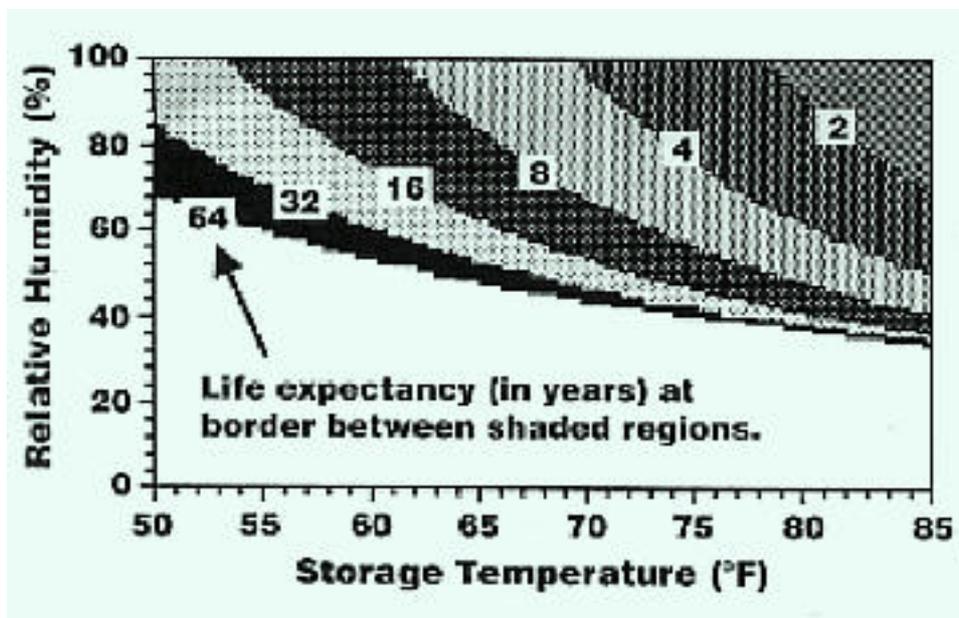
**Temperature and Humidity Conditions and Risk of Hydrolysis  
(Source: Van Bogart (1995) Figure 6)**

Attention also needs to be paid to minimising variations in temperature and relative humidity in the facility, to maintaining air quality at a high level, and to reducing dust and debris, and again I refer the reader to Van Bogart's report. Conditioning (acclimatisation) is also required if the tape is stored in a different environment from that in which it is used. Van Bogart summarises current conditions being proposed in drafts of storage recommendations by various standards organisations. [\[15\]](#) Here is my summary of his summary:

<b>KEY FEATURE</b>	<b>ACCESS STORAGE (storage for media that allows immediate access and playback)</b>	<b>ARCHIVAL STORAGE (storage that preserves the media for as long as possible)</b>
<i>Acclimatisation required before playback?</i>	No	Yes
<i>Media life expectancy</i>	At least 10 years	The maximum possible for the media type
<i>Temperature</i>	Room ambient (15-23oC) Maximum variation 4oC	As low as 5oC Maximum variation 4oC
<i>Humidity</i>	Room ambient (25-75% RH) Maximum variation 20% RH	As low as 20% RH Maximum variation 10% RH

### ***Life Expectancies***

I still have not addressed the question of 'how long'? Taking the key factor of binder hydrolysis (and he is at pains to point out that there are other reasons why tapes can fail, and that his estimate is capable of qualification in many areas) Van Bogart provides this illustration:



**Life Expectancies for a Hi Grade VHS Tape**  
 Estimated by the degree of binder hydrolysis using an end-of-life criteria of 12%  
 (Source: Van Bogart (1995) Figure 10)

For the kinds of ambient room temperatures in Melbourne, say - 25°C and 50% RH, and assuming an unlikely low level of fluctuation - this table suggests an estimated life expectancy of about 10 years. Living in Singapore - 30°C and 80% RH - I must reconcile myself to something more like 1-2 years.

## Optical Disks

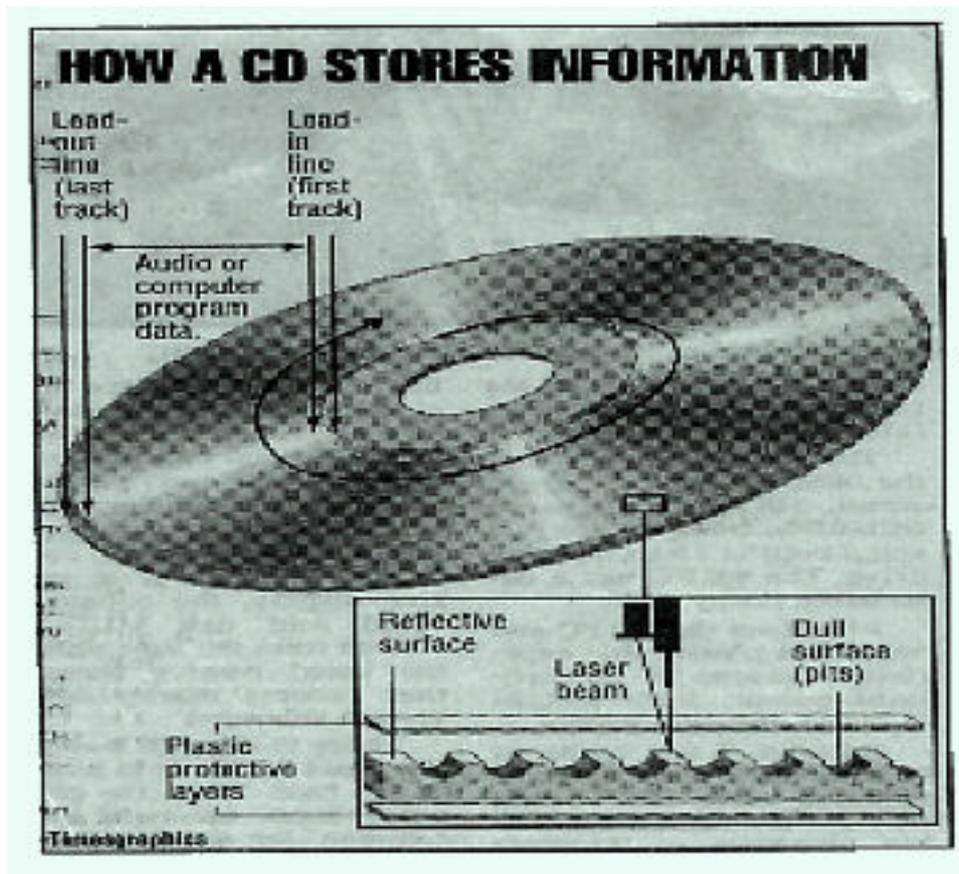
Unfortunately I know of no source for optical disks as current and concise as Van Bogart's is for magnetic tapes. [16] An almost bewildering variety of types can be included under the name 'optical disk', but I am concerned here with 'optical storage products which use light - specifically, the light from lasers - to record and retrieve [information from] ... light-scattering holes, bumps, or bubbles.' [17] These fall into two general categories: *read/write* and *read-only*. Read/write is further divided into two categories: *write-once* and *rewritable*. The first optical disks to appear were WORM (Write Once Read Many) disks at the end of the nineteen-seventies, so we have almost twenty years of practical and anecdotal experience to draw on. One clear implication to be drawn from this evidence is the importance of standards, with numerous stories about institutions committing themselves to one format only to find that it is no longer manufactured after a year or two - but this is not the main concern of this paper.

### *Physical Structure*

All optical disks use basically the same structure, the main difference being the way in which the data is recorded. [18] My examples will concentrate on CD-ROM, but the same physical structure and consequent problems apply more generally.

Videodisks are made by a laser which burns minute holes into a glass master, which is then used to make a metal master from which plastic discs are stamped. An acrylic protective coating is applied. Compact discs are similarly produced by a laser which burns pits into a coating on a glass master

from which a metal master is produced. This stamps a plastic base or substrate which is next coated with a thin layer of metal, usually aluminium, and is then covered with a protective polymer (or sometimes lacquer) layer.



(Source: *Straits Times*)

### *Effect of Physical Structure on Longevity*

The **metal reflecting layer** is considered to be the most susceptible of the factors, largely because the aluminium usually used is more vulnerable to oxidation than other metals or alloys. Oxidation leads over time to corrosion which obscures the distinction between pit and surface (that is, between 0 and 1, the way in which digital data is stored), and the data becomes unreadable. Some manufacturers have used other metals or alloys (platinum or gold, for instance) but the manufacturing costs are consequently significantly higher. The **polymer base**, whose primary function is to support the metal substrate, can itself be permeable to oxygen, thus affording incomplete protection against oxidation. It can also contain rough spots or other defects which promote localised corrosion. Although optical disk manufacturers use alloys which are more resistant to oxidation, this can only retard, not halt, deterioration. The protective **polymer coating** can also fail and again allow oxidation to occur. [19] The **bonding materials** need further investigation: how are the layers bonded? What is known about the stability of the process? And the **ink** used for printing onto the disc has been noted as causing oxidation of the metal layer because it caused breakdown of the polymer coating.

### *Improving Longevity*

As for magnetic tapes, ways to improve the longevity of optical disks are based around:

- care and handling: quality of storage conditions, care in handling
- quality of the disk
- future availability of the technology to play back the disk.

And again as with magnetic tapes, we can only exercise any real control over the first of these.

It is generally assumed that optical disks are less vulnerable to damage caused by poor handling than are magnetic tapes, although we are probably all familiar with the temporary unreadability resulting from fingerprints on a CD-ROM. All information storage media need careful and respectful handling, and optical disks are no exception.

A 3M employee makes the comment that the protective coating on a CD is very thin indeed and 'where there's a fault, normally it is that the seal coat doesn't cover everything and something gets in'. [20] It follows then that anything which can minimise the possibility of a fault occurring is worth pursuing. As one example, it is inadvisable to apply adhesive labels to CD-ROMs. Storage at extremes of temperature and humidity can also, clearly, affect the physical structures: for example, as plastic substrates can absorb moisture and oxidation of the metal layer can occur as a result, then high humidity conditions should be avoided. For the same reasons storage areas in which temperature and humidity fluctuates, resulting in condensation, should be avoided.

### *Life Expectancies*

Peter Adelstein noted in 1993 that while some excellent studies had been carried out on optical disk longevity, there were at that date no national or international specifications. [21] This still appears to be the case today. At one end of the spectrum is an estimate of three to five years for CD-ROMs reported as the view of [NARA \(National Archives and Records Administration\)](#) in 1992. The main problem 'according to Ken Thibodeau of NARA, is that the aluminum substrate on which the data is recorded is vulnerable to oxidation . . . The plastic that protects the substrate is oxygen-permeable, so it provides no protections against the oxidation process.' [22] In late 1994 NARA was still not considering CD-ROMs as an acceptable archival medium:

NARA views CD-ROM as an acceptable transfer medium for permanent records, but has not yet sanctioned it as an archival medium. This means that federal government records that have long-term or permanent value . . . may be transferred to NARA on CD-ROM media but will not be stored permanently on CD-ROM. Once NARA receives such records on CD-ROM, NARA will copy them onto 3480 class magnetic tape cartridges - the only currently acceptable electronic archival medium for permanent storage. [23]

Of more general applicability is a report of accelerated aging studies of CDs carried out at 3M, with input from the [National Media Laboratory](#), in 1992. These resulted in 'a 25-year warranty that assures 100 year life-time at room temperature': that is, the lower estimate takes account of 'general storage fluctuation, as long as it's non-condensing', [24] and 100 years is more like the lifetime to be expected

from high quality storage conditions.

Studies are still continuing on the life expectancies of optical disks and will certainly need to continue, especially as new kinds become available and enter into common use. An announcement was made recently that the National Media Laboratory is setting up stability studies for CD-R. [25] As yet no useful results are being reported in the non-scientific, library or archival studies literature.

Saffady [26] summarises manufacturers' lifetime estimates for read/write optical disks:

<b>RECORDING TECHNOLOGY</b>	<b>LIFETIME ESTIMATE</b>
Ablative technology	10-40 years
Thermal bubble	10-50 years
Dual alloy	100 years
Dye-based	15 years
Magneto-optical	10-30 years
WORM phase change	15 years
Rewritable phase change	10 years

Although these estimates are subject to change as more accurate tests are devised and applied, their implication is clear: the lifetime of optical disks of all kinds, and especially CD-Rs, is greater than the technological obsolescence factor of their recording and playback technology.

## **Other Media**

Many other media exist for the storage of digital data and new media are being developed and promoted on a regular basis. (My favourite is the promising write-once medium called Digital Paper whose 'effectiveness continues to be hampered by its perplexing nonexistence.' [27]) I have conveniently ignored these newly-developing removable storage media - magneto-optical, Zip drives, Sysquest, for example. Those which establish themselves commercially will clearly need to be tested to determine their life expectancies.

## **Training Needs**

What are the implications of the above for training needs in multimedia preservation?

The primary requirement is that of ensuring that the trainers (using this term to include educators at all levels, from on-the-ground library staff to teachers at tertiary institutions) have ready access to the kind of interpretation of the scientific and technical data which Van Bogart has supplied in his report on magnetic tape. Although trainers have an obligation to ensure that their knowledge and skills base are kept up to date, their task is not always easy, especially if the required data is effectively lost to them because they lack the scientific background to interpret it. While all trainers in the preservation area should have a modicum of scientific knowledge, they often do not: humanities and social science backgrounds prevail amongst librarians and archivists. The Commission on Preservation and Access provides the kind of interpretation which is required, but perhaps there is also a more local role to be

played? One envisages the National Preservation Office taking responsibility for a series of papers aimed at trainers, regularly updated and widely disseminated, which summarise current findings and recommendations.

On a more popular level, the misconceptions that we saw in the 'Tape is out. Optical is in' advertisement need to be countered more effectively than is currently the case. If the person in the street is convinced that their Kodak photo CD-ROM will last for ever (whereas their colour film, they know from direct experience, will not) because advertising has consistently and loudly told them that CDs are forever, so, then the more informed but considerably quieter voice of the librarian or archivist will not readily change their mind. This suggests that a different kind of publicity campaign is needed. Perhaps the computer equipment manufacturers, with a significant interest in more sales of their products, can assist with funding to promote the idea that migration of data is the key to this issue?

## Conclusion

### *Horses for courses?*

All of the above ineluctably points to the need to pose (and answer satisfactorily) three questions:

- how do we decide what information we want to retain?
- why do we want to retain that information?
- how long do we want to retain it?

These questions are not easy to answer. The best thinking on the matter has been carried out by archivists, for their profession has developed criteria to apply in selecting categories of data to retain. If we assume here that we have answered the 'what' and 'why', these answers allow us to then decide how long we need to retain each category, and leading on from that the appropriate preservation strategy to apply. Some hypothetical examples:

1. *'in-house' staff training multimedia in CD format: short-term retention (say five years) - conserving the artefact itself and storage at room ambient conditions (assuming that they are within acceptable workroom limits, say 20-24oC and <55% RH [28]);*
2. *business records, say a combination of images and machine-readable text on optical disk, which need to be retained for legal reasons for a minimum of ten years - conserving the artefact itself may be sufficient, as long as care is taken to ensure that the data remains readable and the equipment and software to access it remains in working order; special storage facilities would be required;*
3. *medical records on optical disk which need to be retained for the shorter of the patient's life-time or for twenty years - conserving the artefact itself may be sufficient, as long as care is taken to ensure that the data remains readable and the equipment and software to access it remain in working order; special storage facilities would be required; but refreshing of data and migrating it to new systems as they are introduced will probably also be required;*
4. *national heritage material such as multimedia published in Australia in a variety of formats: 'permanent' retention - the solution is NOT to preserve the digital artefact, but rather to*

concentrate attention on the digital object (for instance by refreshing of data and migrating it to new systems as they are introduced).

In fact this last point is surely the inescapable conclusion to be drawn from this paper: that the equipment (and software) obsolescence factor is where our efforts must be concentrated. In one sense the precise life expectancies of the digital artefacts do not matter. What matters is that they are for periods less than the effective life span (obsolescence period) of the equipment. Although it is platitudinous to say this, it is worth repeating here: equipment costs are decreasing rapidly. [29] What seem unlikely now may be easily affordable in only a brief period of time.

---

## References

Sections of this paper first appeared in the proceedings of the LAS/PPM Multimedia Workshop, 29-30 March 1995, Singapore.

1. Van Bogart, John. *Magnetic Tape Storage and Handling: A Guide for Libraries and Archives* (Washington, DC: The Commission on Preservation and Access and National Media Laboratory, 1995); also available on the World Wide Web at URL [http://www.nml.org/resources/misc/commission\\_report/contents.html](http://www.nml.org/resources/misc/commission_report/contents.html).
2. Email from Jim Wheeler to Data Recording list (DATA\_RECORDING@NML.ORG), 23 August 1995.
3. Don Waters, Yale University Library, January 1995; available on the IFLA World Wide Web site; italics added for emphasis.
4. Bearman, David. "Archival Methods", *Archives and Museums Informatics Technical Report*, 9 (Pittsburgh: Archives and Records Informatics, 1988, reprinted 1991), p.24; italics added for emphasis.
5. The Commission on Preservation and Access *Annual Report July 1, 1993-June 30, 1994*, p.1.
6. "Research on Magnetic Media-Phase 1" made by Chris Ward *et al* in conjunction with Commission on Preservation and Access, January 1994; available on the CoOL (Conservation Online) World Wide Web site; italics added for emphasis.
7. I am aware of, but am conveniently ignoring, other aspects here. One is of the necessity to preserve metadata, or as Peter S. Graham calls it 'intellectual preservation', that is, information about the integrity and authenticity of the information as originally recorded. See URL <http://aultnis.rutgers.edu/texts/dps.html> for Graham's text 'Long-term Intellectual Preservation' (version dated 18 July 1995 seen).
8. Rothenberg, Jeff. 'Ensuring the Longevity of Digital Documents', *Scientific American* (Jan 1995): 24-29.
9. Available on the CoOL (Conservation Online) World Wide Web site.
10. I must again note here my indebtedness in this paper to John Van Bogart. The reader should understand that I use his words here at times without direct acknowledgment. The provisos are noted in an email message to me from John Van Bogart, 21 November 1995. Summarised, the differences are: 1) *Cleanliness*: higher levels of cleanliness of storage environments and of tape drives are required, as missing information on a data tape caused by dropouts (missing signal caused by dust or debris) are not compensated for by the brain, as they are on audio- or

- video-tape; 2) *Tape acclimatisation*: if a tape is not fully acclimatised before play, mistracking may occur; the result of this in a data tape may be that data files cannot be read, whereas in a videotape it may only be an annoying band on the screen ; 3) *Loading/Unloading from drive*: because tape drive mechanisms wear out, the tape may become caught. Care is required when ejecting tapes: 'NEVER attempt to eject a tape while it is in a read/write operation'; 4) *Periodic retensioning*: data tapes require this more frequently than audio- or videotapes.
11. Van Bogart (1995), p.5.
  12. Older tapes were made from other materials such as acetate, which is less stable as a substrate. See Van Bogart (1995), p.6, for further details.
  13. See Van Bogart (1995) pp.7-8 for more information; a recent email gives in detail some of the mechanical problems which can arise with 8mm drives (Patricia Adams to Data Recording List (DATA\_RECORDING @NML.ORG), 20 November 1995).
  14. Van Bogart (1995), pp. 13-14, 23-27.
  15. Van Bogart (1995), p.18.
  16. I am very conscious that most of the sources I have used are not as current as I would like. Although this may be due to my literature-searching abilities, it could also suggest that current interests are focused elsewhere than on the life expectancy of optical disks.
  17. Saffady, William. *Electronic Document Imaging Systems: Design, Evaluation, and Implementation* (Westport, CT: Meckler, 1993), p.63.
  18. See Lieberman, Paula. 'Taking Measure of Magnetic, Optical, and Magneto-Optical Media and Drives', *CD-ROM Professional* 8, 7 (July 1995) and Saffady (1993) for more information about the various types of recording methods used by CD-Rs such as phase change, magneto-optical, dye.
  19. Saffady (1993), p.116.
  20. Arps, Mark. 'CD-ROM: Some Archival Considerations' in *Preservation of Electronic Formats & Electronic Formats for Preservation*, ed. Janice Mohlhenrich (Fort Atkinson, Wisc.: Highsmith, 1993), p.96.
  21. Adelstein, Peter Z. 'The Stability of Optical Disks: A Science-Standards Review', *The Commission on Preservation and Access Newsletter* 58 (July 1993): 3-4.
  22. Publishing E-Journals: Publishing, Archiving and Access List (VPJEJ-L@VTVM1.BITNET), ca March 1992.
  23. Email from Gene Hickock to Data Recording List (DATA\_RECORDING@NML.ORG), 27 April 1995.
  24. Arps (1993), p.102.
  25. Email from Robert D. Lorentz, 'Dye Stability of CD-R', 24 Aug 1995, to Data Recording List (DATA\_RECORDING@NML.ORG): 'Studies of the stability of CD-R under different conditions (light, heat, humidity) are being established in the National Media Lab. Your experiences with any particular problems encountered would be valuable input to this task. Once results are obtained, they will be available through NML. To find more about NML, try our home page at <http://www.nml.org/>.'
  26. Saffady (1993), p.118, Table 5-1.
  27. Lieberman (1995), p.72. Another example is the HD-ROM: 'The High-Density Read-Only Memory, or HD-ROM, uses a unique ion beam to inscribe information on pins of stainless steel, iridium or other materials that are built to last. An HD-ROM holds about 180 times more information than a comparably sized Compact Disc Read-Only Memory, or CD-ROM, today's cheapest data storage medium. Storage costs of HD-ROM are roughly one-half percent of CD-

ROM costs. . . . "The HD-ROM marks a complete departure from existing data storage technologies . . . For the first time, a non-magnetic, non-optical data storage system can be made from truly robust materials." HD-ROM materials are hard, non-malleable, non-flammable and don't react easily with chemicals. Since the medium isn't magnetic, electromagnetic fields can't destroy the data on HD-ROMs, unlike computer hard drives. . . . 'HD-ROM is virtually impervious to the ravages of time whether from material degradation due to thermal or mechanical shock or from the electromagnetic fields that are so destructive to other storage media.'" News release distributed in HPCWire, 23 June 1995..

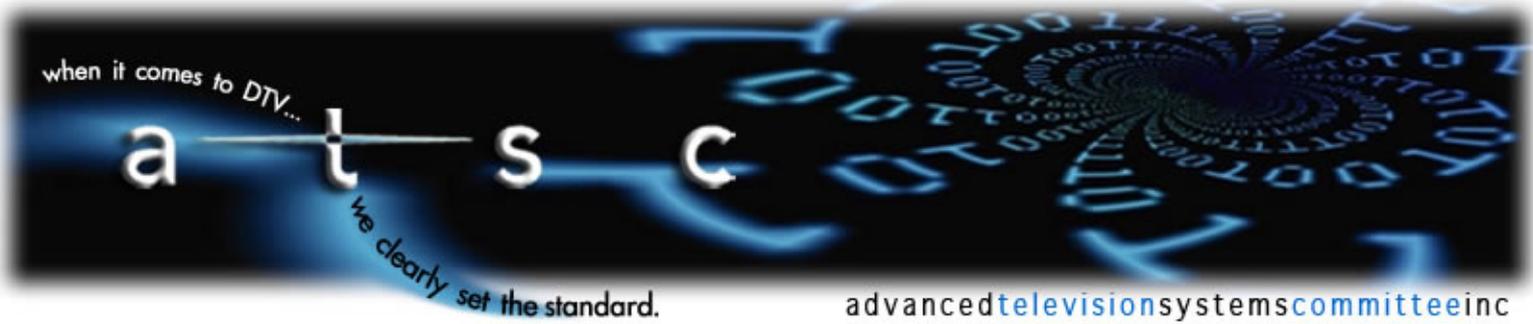
28. Van Bogart (1995), p.17.

29. One example: Lesk gives the equipment cost in 1990 for storing one gigabyte on magnetic disk as \$4,000 (Lesk, Michael. *Image Formats for Preservation and Access* (CPA, 1990), reprinted in *Information Technology and Libraries* 9, 4 (1990): 300-308). Today the cost is less than \$400.

---

[Return to 1995 NPO Conference](#)

**Last updated 27 February 1998**



Not a member?  
Join the ATSC today!

Mail or Fax?  
Membership Form

The Advanced Television Systems Committee in cooperation with the Society of Broadcast Engineers ([SBE](#)) is featuring a half-day seminar focusing on digital electronic newsgathering ([ENG](#)) on February 21, 2006 at the 11th Annual Hollywood Post Alliance ([HPA](#)) Technology Retreat in [Palm Springs, CA.](#) [CLICK HERE for Registration Information](#)

in\_the\_news

- Digital Electronic Newsgathering (ENG) [Read the Press Release](#)
- Digital Television Emerges from the Shadows...and Consumers See the Light [Read all about it in the November 2005 Special Edition of "The Standard"](#)



Copyright © 2005 Advanced Television Systems Committee, Inc. All rights reserved.

**> ABOUT FIAT**

- What is FIAT
- FIAT policy
- Statutes
- Join FIAT/IFTA
- FIAT organisation
- Members details
- News
- Calendar of events

**> CONFERENCES**

- Conferences
- Seminars & workshops

**> PROJECTS & STANDARDS**

- Projects information
- Professional standards, guidelines

**> SERVICES**

- Training opportunities

**> AWARDS**

- Short presentation
- Competition 2005
- Competition 2004
- Competition 2003
- Hall of Fame
- Nomination form

**> LINKS**

- Links

The **International Federation of Television Archives** is an international professional association established to provide a means for co-operation amongst broadcast and national audiovisual archives and libraries concerned with the collection, preservation and exploitation of moving image and recorded sound materials and associated documentation.

[> More](#)**> FOCUS**

- [New-York Conference programme : updated version](#)
- [The Parker is filling up. Book your room at the nearby Holiday Inn](#)
- [Vienna seminar](#)
- [FIAT @ IBC 2005 : Archive day](#)
- New members : [TWI](#), [Framepool AG](#)
- [Announcing the Extraordinary General Assembly](#)
- [Register now for the New-York Conference](#)
- [FIAT Award shortlisted programmes](#)
- [Latest News - May 2005](#)
- [New-York conference : announcement](#)
- [New member : University of Ghent \(Belgium\)](#)
- [Access the database listing all FIAT IFTA members](#)
- [Visit the RAI Archives](#)
- [Visit the Paris Conference picture gallery !](#)
- [Paris papers](#)

[> More](#)

The International Federation of Television Archives acts as a forum for media archive experts :

- Setting international standards and best practices
- Facilitating training opportunities
- Providing practical advice
- Offering strong partnership to media, IT and hardware manufacturers

The International Federation of Television Archives has solutions for :



© Ina

- Archive managers
- Broadcasting companies
- Media libraries

**> CALENDAR OF EVENTS**

- FIAT at IBC 2005 - Archive Day : 10th September, Amsterdam
- FIAT IFTA Annual Conference : 16-20th September, New York
- FIAT / IASA Southern Africa Conference : 10-15th October, Johannesburg

[> More](#)**> CONTACT**

- Content : [office@fiatifta.org](mailto:office@fiatifta.org)
- Technical : [webmaster@fiatifta.org](mailto:webmaster@fiatifta.org)

[> More](#)

# A Letter From QUANTEGY Recording Solutions to Our Valued Customers



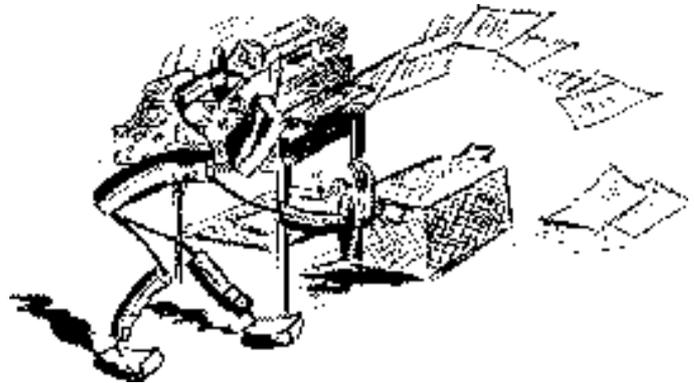
- PRODUCTS
- WHERE TO BUY
- NEW PRODUCTS
- NEWS & INFO
- ABOUT QUANTEGY
- SUPPORT
- CONTACT US



## the Dead Media Project

The Dead Media Project consists of a database of field Notes written and researched by members of the Project's mailing list.

The [Dead Media List](#) consists of occasional email to that stout band of souls who have declared some willingness to engage in this recherche field of study.



For more information on the purpose of the project, please read Bruce Sterling's [Dead Media Manifesto](#). For more information on the mailing list, including how to join, please read the [Frequently asked questions](#).

### The Dead Media Working Notes



The collection of [dead media working notes](#) is the purpose of the mailing list; the accumulated, archived, (and now collated!) collection of submissions from list members; an ad hoc database of the deceased, the slowly-rotting, the undead, and the never-lived media.

Other Dead Media websites

- The Dead Media Project site at [Griffith Multimedia](#). One of the authors told me that he and others listed in the credits for the site, are no longer actively involved in its maintenance. Quite lovely nonetheless.
- Last, but hardly least, [Bruce Sterling's home page](#) contains many dead-media related things, including >>gasp<< gopherspace.



-- Tom Jennings, moderator <tomj@deadmedia.org>

---

Andy McFadden's  
**CD-Recordable FAQ**

Last-modified: **2005/11/13**

Version: **2.63**

Send corrections and updates to Andy McFadden. If you have a question you'd like to see answered in here, either post it to one of the comp.publish.cdrom newsgroups (if you don't have the answer), or send it to fadden@fadden.com (if you do).

This FAQ is updated every couple of months. If you are looking at a version that is more than two or three months old, it may be an out-of-date copy. The most current version is always at <http://www.cdrfaq.org/>.

This was originally developed (and is still maintained) as a Usenet newsgroup FAQ. If you'd like to check out the newsgroups, point your news reader at the following (or go to <http://groups.google.com/> and read them with your web browser):

- [comp.publish.cdrom.hardware](http://groups.google.com/group/comp.publish.cdrom.hardware)
- [comp.publish.cdrom.software](http://groups.google.com/group/comp.publish.cdrom.software)
- [comp.publish.cdrom.multimedia](http://groups.google.com/group/comp.publish.cdrom.multimedia)
- [alt.comp.periphs.cdr](http://groups.google.com/group/alt.comp.periphs.cdr)

The "canonical" FAQ is available from <http://www.cdrfaq.org/> in HTML format and from [the MIT FAQ archives](#) in plain text format. You can get an [all-in-one-file](#) version of the HTML in a .ZIP file, suitable for printing. Some translations are available:

- [Turkish](#)
- [Hungarian](#)
- [Italian](#)
- [French](#)
- [Russian](#)
- [Spanish](#)

If you're maintaining a translation, or just really want to know what has changed since the last update, you can get a set of [context diffs](#) in a .ZIP file.

Please DO NOT post copies of the HTML version on your web site unless you plan to keep them up to date automatically. I have on several occasions received e-mail from people reading versions that are several months old. Just use a link to the [www.cdrfaq.org](http://www.cdrfaq.org) site instead.



## **[2] CD Encoding**

[\[2-1\] How is the information physically stored?](#)

[\[2-2\] What is XA? CDPLUS? CD-i? MODE1 vs MODE2? Red/yellow/blue book?](#)

[\[2-3\] How do I know what format a disc is in?](#)

[\[2-4\] How does copy protection work?](#)

[\[2-4-1\] ...on a data CD-ROM?](#)

[\[2-4-2\] ...on an audio CD?](#)

[\[2-4-3\] ...on an audio CD \(Macrovision - SafeAudio\)](#)

[\[2-4-4\] ...on an audio CD \(SunnComm - MediaCloQ and MediaMax CD3\)](#)

[\[2-4-5\] ...on an audio CD \(Midbar Tech - Cactus Data Shield\)](#)

[\[2-4-6\] ...on an audio CD \(Key2Audio / Sony DADC\)](#)

[\[2-4-7\] ...on an audio CD \(BayView Systems - Duolizer\)](#)

[\[2-4-8\] ...on an audio CD \(Sanyo\)](#)

[\[2-4-9\] How does the Doc-Witness OpSecure CD-ROM work?](#)

[\[2-4-10\] What's the Sony BMG rootkit \(First 4 Internet XCP\)?](#)

[\[2-5\] What's a multisession disc?](#)

[\[2-6\] What are subcode channels?](#)

[\[2-7\] Are the CD Identifier fields widely used?](#)

[\[2-8\] How long does it take to burn a CD-R?](#)

[\[2-9\] What's the difference between disc-at-once and track-at-once?](#)

[\[2-10\] Differences between recording from an image and on-the-fly?](#)

[\[2-11\] How does an audio CD player know to skip data tracks?](#)

[\[2-12\] How does CD-RW compare to CD-R?](#)

[\[2-13\] Can DVD players read CD-Rs?](#)

[\[2-14\] Should I buy a DVD recorder instead?](#)

[\[2-15\] What are "jitter" and "jitter correction"?](#)

[\[2-16\] Where can I learn more about the history of CD and CD-R?](#)

[\[2-17\] Why don't audio CDs use error correction?](#)

[\[2-18\] How does CD-R compare to MiniDisc?](#)

[\[2-19\] What does finalizing \(and closing and fixating\) do?](#)

[\[2-20\] How are WAV/AIFF files converted into Red Book CD audio?](#)

[\[2-21\] What does MultiRead mean? MultiPlay?](#)

[\[2-22\] If recording fails, is the disc usable?](#)

[\[2-23\] Why do recorders insert "00" bytes at the start of audio tracks?](#)

[\[2-24\] How many tracks can I have? How many files?](#)

[\[2-25\] Will SCMS prevent me from making copies?](#)

[\[2-26\] Is a serial number placed on the disc by the recorder?](#)

[\[2-27\] What's a TOC? How does it differ from a directory?](#)

[\[2-28\] What's an ISO? A CIF? BIN and CUE? .DAT?](#)

[\[2-29\] Why was 74 minutes chosen as the standard length?](#)

[\[2-30\] Why is there a visibly unwritten strip near the CD-R hub?](#)

- [\[2-31\] What is "BURN-Proof"? "JustLink"? "Waste-Proof"?](#)
- [\[2-32\] Can playing CD-Rs in a DVD player hurt the discs?](#)
- [\[2-33\] Who \\*really\\* made this CD-R blank?](#)
- [\[2-34\] Can I make copies of DTS-encoded CDs?](#)
- [\[2-35\] Why 44.1KHz? Why not 48KHz?](#)
- [\[2-36\] What format are .CDA files in?](#)
- [\[2-37\] What are DD-R and DD-RW?](#)
- [\[2-38\] What's an ATIP?](#)
- [\[2-39\] What are "ML" discs and devices?](#)
- [\[2-40\] What's CD-MRW? Mount Rainier? EasyWrite?](#)
- [\[2-41\] What's Audio Master Quality \(AMQ\) recording?](#)
- [\[2-42\] Can I draw pictures on a disc with the recording laser?](#)
- [\[2-43\] What are the gory details about how are 1s and 0s encoded?](#)
- [\[2-43-1\] How does the laser read or write a disc?](#)
- [\[2-43-2\] How do pits and lands turn into 1s and 0s? What's EFM?](#)
- [\[2-43-3\] What's a frame? CIRC encoding? How does ECC work?](#)
- [\[2-43-4\] What's in a sector?](#)
- [\[2-43-5\] What's in a subcode channel?](#)
- [\[2-43-6\] I want even more details](#)
- [\[2-44\] Digital is better than analog, right?](#)
- [\[2-44-1\] What is "digital" and "digitization", anyway?](#)
- [\[2-44-2\] How does this relate to CD-DA?](#)
- [\[2-45\] What's a CDR-ROM? CD-PROM?](#)
- [\[2-46\] What's HD-BURN? GigaRec?](#)
- [\[2-47\] What are C2 errors? What do they say about disc quality?](#)
- [\[2-48\] What are CD+R and CD+RW?](#)
- [\[2-49\] What's HighMAT?](#)
- [\[2-50\] What's VariRec?](#)
- [\[2-51\] Will my CDs work on players in other countries?](#)
- [\[2-52\] Do CD-Rs have deeper pits? Are "shallow burns" bad?](#)
- [\[2-53\] What's a stacking ring?](#)

### **[3] How Do I...**

- [\[3-1\] How do I copy a CD-ROM?](#)
- [\[3-1-1\] Why can't I just do a block copy like a floppy?](#)
- [\[3-2\] How do I extract tracks from \("rip"\) or copy an audio CD?](#)
- [\[3-2-1\] How do I remove the voice from a CD track, leaving just music?](#)
- [\[3-2-2\] How do I encode a CD track to MP3?](#)
- [\[3-3\] How do I get rid of hisses and clicks on audio CDs?](#)
- [\[3-4\] How do I copy game console discs \(e.g. Playstation, Dreamcast\)](#)
- [\[3-5\] How do I get long filenames onto a disc?](#)

[\[3-5-1\] ISO-9660](#)

[\[3-5-2\] Rock Ridge](#)

[\[3-5-3\] HFS/HFS+ and Macintosh extensions to ISO-9660](#)

[\[3-5-4\] Joliet](#)

[\[3-5-5\] Romeo](#)

[\[3-5-6\] ISO/IEC 13346 and ISO/IEC 13490](#)

[\[3-5-7\] ISO-9660:1999](#)

[\[3-6\] How do I use a CD-i disc on a PC?](#)

[\[3-7\] How can I extract disc and track titles from an audio CD?](#)

[\[3-8\] How do I write more than 80 minutes of audio or 700MB of data?](#)

[\[3-8-1\] How well do 80-minute CD-R blanks work?](#)

[\[3-8-2\] How well do 90-minute and 99-minute CD-R blanks work?](#)

[\[3-8-3\] How can I exceed the stated disc capacity \("overburning"\)?](#)

[\[3-9\] How do I put photographs onto CD-ROM?](#)

[\[3-9-1\] How do I create a PhotoCD?](#)

[\[3-9-2\] How can I set up a photo album on CD-ROM?](#)

[\[3-9-3\] How can I show digital photos on my DVD player?](#)

[\[3-10\] How do I make a CD that will work on a PC or a Mac?](#)

[\[3-11\] How do I access different sessions on a multi-session CD?](#)

[\[3-12\] How do I transfer my records or cassettes to a CD?](#)

[\[3-12-1\] ...with a stand-alone audio CD recorder?](#)

[\[3-12-2\] ...with a CD recorder attached to my computer?](#)

[\[3-12-3\] How can I clean up the audio before recording?](#)

[\[3-13\] How do I transfer an audio DAT tape to CD?](#)

[\[3-14\] How do I put audio and data on the same CD?](#)

[\[3-15\] How do I make a bootable CD-ROM?](#)

[\[3-16\] How do I convert home movies into video on CD?](#)

[\[3-16-1\] How do I create a VideoCD from AVI or MPEG files?](#)

[\[3-16-2\] How do I create an SVCD?](#)

[\[3-16-3\] How do I create an AVCD?](#)

[\[3-17\] How can I burn several copies of the same disc simultaneously?](#)

[\[3-18\] Can I make copies of copies?](#)

[\[3-19\] How can I compress or encrypt data on a CD-ROM?](#)

[\[3-20\] Can I do backups onto CD-R?](#)

[\[3-21\] How do I automatically launch something? Change the CD icon?](#)

[\[3-21-1\] How does Windows "autorun" work?](#)

[\[3-21-2\] How do I launch a document \(like a web page\)?](#)

[\[3-21-3\] What autorun software is available?](#)

[\[3-22\] How can I be sure the data was written correctly?](#)

[\[3-23\] How do I create, copy, or play Audio Karaoke/CD+G discs?](#)

[\[3-24\] How do I copy a CD-ROM with 3GB of data on it? A huge VideoCD?](#)

- [\[3-25\] How do I get my CD-R pressed into a real CD?](#)
- [\[3-26\] How do I make a CD without that two-second gap between tracks?](#)
- [\[3-27\] How can I record RealAudio \(.ra\), MIDI, WMA, and MP3 on a CD?](#)
- [\[3-28\] How do I add CD-Text information?](#)
- [\[3-29\] Can I distribute a web site on a CD-ROM?](#)
- [\[3-30\] How do I clean my CD recorder?](#)
- [\[3-31\] Is it better to record at slower speeds?](#)
- [\[3-32\] Where do I get drivers for my CD recorder?](#)
- [\[3-33\] Can I copy discs without breaking the law?](#)
- [\[3-33-1\] ...in the United States of America?](#)
- [\[3-33-2\] ...in Canada?](#)
- [\[3-34\] Can CD-Rs recorded at 2x be read faster than 2x?](#)
- [\[3-35\] How do I make my CD-ROM work on the Mac, WinNT, and UNIX?](#)
- [\[3-36\] How do I put "hidden tracks" and negative indices on audio CDs?](#)
- [\[3-37\] Do I need to worry about viruses?](#)
- [\[3-38\] How do I cover up a bad audio track on a CD-R?](#)
- [\[3-39\] How do I duplicate this hard-to-copy game?](#)
- [\[3-40\] Should I erase or format a disc? How?](#)
- [\[3-41\] How do I equalize the volume for tracks from different sources?](#)
- [\[3-42\] How do I make a bit-for-bit copy of a disc?](#)
- [\[3-43\] How do I put punctuation or lower case in CD-ROM volume labels?](#)
- [\[3-44\] How do I extract audio tracks from an "enhanced" CD on the Mac?](#)
- [\[3-45\] How do I disable DirectCD for Windows?](#)
- [\[3-46\] How do I specify the order of files \(e.g. sorting\) on ISO-9660?](#)
- [\[3-47\] How do I put a password on a CD-ROM?](#)
- [\[3-48\] Can I record an audio CD a few tracks at a time?](#)
- [\[3-49\] How do I copy DVDs onto CD-R?](#)
- [\[3-49-1\] I heard about software that copies DVDs with a CD recorder!](#)
- [\[3-50\] How do I copy Mac, UNIX, or "hybrid" CD-ROMs from Windows?](#)
- [\[3-51\] How do I copy something in "RAW" mode? What's DAO-96?](#)
- [\[3-52\] How do I do cross-fades between audio tracks?](#)
- [\[3-53\] How do I create a CD with my favorite songs on it?](#)
- [\[3-54\] How do I record directly onto CD from a microphone?](#)
- [\[3-55\] Is it okay to record a CD from MP3?](#)
- [\[3-56\] How can I test a disc image before recording?](#)
- [\[3-57\] How do I clear the "read-only" flag under Windows?](#)
- [\[3-58\] How do I share a CD recorder across a network?](#)
- [\[3-59\] How do I write a large file across multiple discs?](#)
- [\[3-60\] What's the safest, most reliable way to write data to CD-R?](#)

## [\[4\] Problems](#)

- [\[4-1\] What does "buffer underrun" mean?](#)
- [\[4-1-1\] What's the deal with Windows Auto-Insert Notification \(AIN\)?](#)
- [\[4-1-2\] What's all this about Win9x VCACHE settings?](#)
- [\[4-2\] I can't get long Win95 filenames to work right](#)
- [\[4-3\] I can't read the multisession CD I just made](#)
- [\[4-4\] Write process keeps failing N minutes in](#)
- [\[4-5\] Why did my CD-R eject and re-load the disc between operations?](#)
- [\[4-6\] My CD-ROM drive doesn't like \\*any\\* CD-R discs](#)
- [\[4-7\] How do I avoid having a ";1" on my ISO-9660 discs?](#)
- [\[4-8\] I keep getting SCSI timeout errors](#)
- [\[4-9\] I'm having trouble writing a complete disc](#)
- [\[4-10\] What's the CDD2000 Write Append Error / spring problem?](#)
- [\[4-11\] Getting errors reading the first \(data\) track on mixed-mode CD](#)
- [\[4-12\] My recorder ejects blank discs immediately](#)
- [\[4-13\] I'm getting complaints about power calibration](#)
- [\[4-14\] My Adaptec 2940 pauses after finding my recorder](#)
- [\[4-15\] I can't see all the files on the CD-R](#)
- [\[4-16\] My multi-session disc only has data from the last session](#)
- [\[4-17\] I'm getting SCSI errors](#)
- [\[4-18\] Why doesn't the copy of an audio CD sound the same?](#)
- [\[4-18-1\] Why doesn't the audio data on the copy match the original?](#)
- [\[4-18-2\] The audio data matches exactly, why do they sound different?](#)
- [\[4-19\] Digital audio extraction of a track is shifted slightly](#)
- [\[4-20\] I can't play extracted audio files by double-clicking in Win95](#)
- [\[4-21\] I can't read an ISO-finalized packet-written disc](#)
- [\[4-22\] I'm finding corrupted files on the CD-ROMs I write](#)
- [\[4-23\] Having trouble playing an audio CD in a home or car player](#)
- [\[4-24\] Having trouble using a CD-ROM on a different machine](#)
- [\[4-25\] I can't copy a VideoCD](#)
- [\[4-26\] The test write succeeds, but the actual write fails](#)
- [\[4-27\] I can no longer erase a particular CD-RW disc](#)
- [\[4-28\] Having trouble formatting discs with DirectCD](#)
- [\[4-29\] I can't write CD-Rs after installing Windows 98](#)
- [\[4-30\] I can't use the copy of a CD-ROM after installing Windows 98](#)
- [\[4-31\] The disc I was writing with DirectCD is now unreadable](#)
- [\[4-32\] I'm getting a message about 100 form transitions](#)
- [\[4-33\] My system hangs when I insert a blank disc](#)
- [\[4-34\] My CD-R discs don't work in my DVD player](#)
- [\[4-35\] I need help recovering data from a CD-ROM](#)
- [\[4-36\] What does "not convertible to CD quality" mean?](#)
- [\[4-37\] I inserted a CD-ROM but Windows thinks it's an audio CD](#)

- [\[4-38\] I get read errors when trying to copy a game](#)
- [\[4-39\] Restarting or shutting Windows down after recording causes hang](#)
- [\[4-40\] Why do CD-Rs play poorly when anti-skip protection is enabled?](#)
- [\[4-41\] I'm having trouble recording under Windows 2000 or WinXP](#)
- [\[4-42\] I formatted a CD-RW and only have about 530MB free](#)
- [\[4-43\] My CD recording software keeps crashing](#)
- [\[4-44\] Do I need to update my ASPI layer?](#)
- [\[4-45\] The write process completes, but the disc is still blank](#)
- [\[4-46\] My CD-RW drive doesn't work with my CD-RW blanks](#)
- [\[4-47\] Audio discs have crackling sounds on the last few tracks](#)
- [\[4-48\] Files in deep directories can be seen but not opened](#)
- [\[4-49\] My CD-ROM drive stopped working after uninstalling software](#)
- [\[4-50\] Audio CDs recorded from MP3s play back fast and high-pitched](#)
- [\[4-51\] Windows says access denied, can't create or replace file](#)
- [\[4-52\] I can't see any files on a CD-R or CD-RW from MS-DOS](#)
- [\[4-53\] My OS doesn't support ISO-13346 "UDF"](#)

## **[\[5\] Hardware](#)**

- [\[5-1\] Which CD recorder should I buy?](#)
  - [\[5-1-1\] Yamaha](#)
  - [\[5-1-2\] Sony](#)
  - [\[5-1-3\] Smart & Friendly](#)
  - [\[5-1-4\] Philips](#)
  - [\[5-1-5\] Hewlett-Packard \(HP\)](#)
  - [\[5-1-6\] Plasmon](#)
  - [\[5-1-7\] Kodak](#)
  - [\[5-1-8\] JVC](#)
  - [\[5-1-9\] Pinnacle](#)
  - [\[5-1-10\] Ricoh](#)
  - [\[5-1-11\] Pioneer](#)
  - [\[5-1-12\] Olympus](#)
  - [\[5-1-13\] Optima](#)
  - [\[5-1-14\] Mitsumi](#)
  - [\[5-1-15\] DynaTek Automation Systems](#)
  - [\[5-1-16\] Microboards of America](#)
  - [\[5-1-17\] Micro Design International](#)
  - [\[5-1-18\] MicroNet Technology](#)
  - [\[5-1-19\] Procom Technology](#)
  - [\[5-1-20\] Grundig](#)
  - [\[5-1-21\] Plextor](#)
  - [\[5-1-22\] Panasonic \(Matsushita\)](#)

[\[5-1-23\] Teac](#)

[\[5-1-24\] Wearnes](#)

[\[5-1-25\] Turtle Beach](#)

[\[5-1-26\] Creative Labs](#)

[\[5-1-27\] Taiyo Yuden](#)

[\[5-1-28\] Memorex](#)

[\[5-1-29\] Hi-Val](#)

[\[5-1-30\] Dysan](#)

[\[5-1-31\] Traxdata](#)

[\[5-1-32\] BenQ \(nee Acer\)](#)

[\[5-1-33\] Waitec](#)

[\[5-1-34\] BTC](#)

[\[5-1-35\] Caravelle \(Sanyo\)](#)

[\[5-1-36\] Micro Solutions](#)

[\[5-1-37\] Pacific Digital](#)

[\[5-1-38\] Iomega](#)

[\[5-1-39\] Goldstar \(LG Electronics\)](#)

[\[5-1-40\] AOpen](#)

[\[5-1-41\] Toshiba](#)

[\[5-1-42\] TDK](#)

[\[5-1-43\] Lite-On](#)

[\[5-1-44\] CenDyne](#)

[\[5-1-45\] VST \(SmartDisk\)](#)

[\[5-1-46\] ASUS](#)

[\[5-1-47\] Samsung](#)

[\[5-1-48\] APS / LaCie](#)

[\[5-2\] How long do CD recorders last?](#)

[\[5-3\] What kind of PC is recommended?](#)

[\[5-4\] What kind of Mac is recommended?](#)

[\[5-5\] Which standard CD-ROM drives work well with CD-R?](#)

[\[5-6\] What kind of HD should I use with CD-R? Must it be AV-rated?](#)

[\[5-7\] What SCSI adapter should I use with a CD recorder?](#)

[\[5-7-1\] Adaptec - 1510/1522A/1540/1542CF](#)

[\[5-7-2\] Adaptec - 2840/2910/2920/2930/2940](#)

[\[5-7-3\] ASUS - SC-200/SC-875](#)

[\[5-7-4\] Tekram - DC-390U/DC-390F](#)

[\[5-7-5\] Adaptec - 1350/1460/1480](#)

[\[5-8\] Can I use a CD recorder as a general-purpose reader?](#)

[\[5-9\] To caddy or not to caddy?](#)

[\[5-10\] Can I burn CDs from a Jaz drive? Tape drive?](#)

[\[5-11\] What is "Running OPC"?](#)

- [\[5-12\] What's the story with stand-alone audio CD recorders?](#)
- [\[5-13\] What's firmware? How and why should I upgrade my recorder?](#)
- [\[5-14\] How well do parallel-port, USB, and 1394 recorders work?](#)
- [\[5-15\] How should I configure my system for an ATAPI CD recorder?](#)
- [\[5-15-1\] Should I have DMA enabled for an ATAPI recorder in Windows?](#)
- [\[5-16\] How important is CD-RW?](#)
- [\[5-17\] What is an "MMC Compliant" recorder?](#)
- [\[5-18\] What do I need to record on a UNIX \(Linux, Solaris, etc\) system?](#)
- [\[5-19\] What do I need for recording CDs from a laptop?](#)
- [\[5-20\] I need to make \\*lots\\* of copies](#)
- [\[5-21\] How do I connect two drives to one sound card in a PC?](#)
- [\[5-22\] How fast is 1x? What are CAV, CLV, PCAV, and ZCLV?](#)
- [\[5-23\] Will playing CD-Rs damage my CD player?](#)
- [\[5-24\] Can I "overclock" my CD recorder?](#)
- [\[5-25\] I need some help installing the drive](#)
- [\[5-26\] How much power does a CD recorder use?](#)
- [\[5-27\] Will the laser in my drive wear out?](#)

## **[6] Software**

- [\[6-1\] Which software should I use?](#)
- [\[6-1-1\] Adaptec - Easy-CD, Easy-CD Pro, and Easy-CD Pro MM \("ECD"\)](#)
- [\[6-1-2\] Adaptec - CD-Creator \("CDC"\)](#)
- [\[6-1-3\] Gear Software - GEAR Pro](#)
- [\[6-1-4\] Roxio - Toast](#)
- [\[6-1-5\] CeQuadrat - WinOnCD](#)
- [\[6-1-6\] Young Minds, Inc. - CD Studio+](#)
- [\[6-1-7\] Golden Hawk Technology \(Jeff Arnold\) - CDRWIN](#)
- [\[6-1-8\] Optical Media International - QuickTOPiX CD](#)
- [\[6-1-9\] Creative Digital Research - CDR Publisher](#)
- [\[6-1-10\] mkisofs](#)
- [\[6-1-11\] Asimware Innovations - MasterISO](#)
- [\[6-1-12\] Newtech Infosystems, Inc. \(NTI\) - CD-Maker](#)
- [\[6-1-13\] Cirrus Technology/Unite - CDMaker](#)
- [\[6-1-14\] Hohner Midia - Red Roaster](#)
- [\[6-1-15\] Dataware Technologies - CD Author](#)
- [\[6-1-16\] CreamWare - Triple DAT](#)
- [\[6-1-17\] MicroTech - MasterMaker](#)
- [\[6-1-18\] Angela Schmidt & Patrick Ohly - MakeCD](#)
- [\[6-1-19\] Liquid Audio Inc. - Liquid Player](#)
- [\[6-1-20\] Jörg Schilling - cdrecord](#)
- [\[6-1-21\] Prassi Software - CD Rep and CD Right](#)

- [\[6-1-22\] Zittware - CDMaster32](#)
- [\[6-1-23\] Dieter Baron and Armin Obersteiner - CD Tools](#)
- [\[6-1-24\] PoINT - CDwrite](#)
- [\[6-1-25\] PoINT - CDaudio Plus](#)
- [\[6-1-26\] Roxio - Easy CD Creator Deluxe \("ECDC"\)](#)
- [\[6-1-27\] Padus - DiscJuggler](#)
- [\[6-1-28\] Ahead Software - Nero](#)
- [\[6-1-29\] CharisMac Engineering - Discribe](#)
- [\[6-1-30\] István Dósa - DFY\\$VMSCD](#)
- [\[6-1-31\] RSJ Software - RSJ CD Writer](#)
- [\[6-1-32\] James Pearson - mkhybrid](#)
- [\[6-1-33\] JVC - Personal Archiver Plus](#)
- [\[6-1-34\] Roxio - Jam](#)
- [\[6-1-35\] Pinnacle Systems - InstantCD/DVD \(was VOB\)](#)
- [\[6-1-36\] Sony - CD Architect](#)
- [\[6-1-37\] Eberhard Heuser-Hofmann - CDWRITE](#)
- [\[6-1-38\] CeQuadrat - JustAudio!](#)
- [\[6-1-39\] Digidesign - MasterList CD](#)
- [\[6-1-40\] Thomas Niederreiter - X-CD-Roast](#)
- [\[6-1-41\] Jesper Pedersen - BurnIT](#)
- [\[6-1-42\] Jens Fangmeier - Feurio!](#)
- [\[6-1-43\] Iomega - HotBurn](#)
- [\[6-1-44\] DARTECH, Inc - DART CD-Recorder](#)
- [\[6-1-45\] Interactive Information R&D - CDEveryWhere](#)
- [\[6-1-46\] DnS Development - BurnIt](#)
- [\[6-1-47\] Andreas Müller - CDRDAO](#)
- [\[6-1-48\] Tracer Technologies - \(various\)](#)
- [\[6-1-49\] SlySoft - CloneCD](#)
- [\[6-1-50\] IgD - FireBurner](#)
- [\[6-1-51\] Jodian Systems & Software - CDWRITE](#)
- [\[6-1-52\] Erik Deppe - CD+G Creator](#)
- [\[6-1-53\] Micro-Magic - CD Composer](#)
- [\[6-1-54\] Earjam, Inc. - Earjam IMP](#)
- [\[6-1-55\] Emagic - Waveburner](#)
- [\[6-1-56\] Zy2000 - MP3 CD Maker](#)
- [\[6-1-57\] Integral Research - Speedy-CD](#)
- [\[6-1-58\] Desernet Broadband Media - Net-Burner and MP3-Burner](#)
- [\[6-1-59\] Stomp, Inc. - Click 'N Burn](#)
- [\[6-1-60\] Steinberg Media Technologies - Clean! plus](#)
- [\[6-1-61\] Enreach - I-Author for VCD/SVCD](#)
- [\[6-1-62\] VSO Software - Blindread/Blindwrite](#)

- [\[6-1-63\] Microsoft - Windows XP](#)
- [\[6-1-64\] An Chen Computers - CD Mate](#)
- [\[6-1-65\] E-Soft - Alcohol](#)
- [\[6-1-66\] Stomp Inc. - RecordNow MAX](#)
- [\[6-1-67\] James Mieczkowski - Cheetah CD Burner](#)
- [\[6-1-68\] Blaze Audio - RipEditBurn](#)
- [\[6-1-69\] Acoustica, Inc. - MP3 CD Burner](#)
- [\[6-1-70\] MagicISO, Inc. - MagicISO](#)
- [\[6-1-71\] Simone Tasselli - Burn4Free](#)
- [\[6-1-72\] Sonic Solutions - Record Now!](#)
- [\[6-1-73\] Freeridecoding - BurnAgain](#)
- [\[6-2\] What other useful software is there?](#)
- [\[6-2-1\] Optical Media International - Disc-to-Disk](#)
- [\[6-2-2\] Gilles Vollant - WinImage](#)
- [\[6-2-3\] Asimware Innovations - AsimCDFS](#)
- [\[6-2-4\] Steven Grimm - WorkMan](#)
- [\[6-2-5\] Cyberdyne Software - CD Worx](#)
- [\[6-2-6\] Arrowkey - CD-R Diagnostic](#)
- [\[6-2-7\] DC Software Design - CDRCue Cuesheet Editor](#)
- [\[6-2-8\] Astarte - CD-Copy](#)
- [\[6-2-9\] Frank Wolf - CDR Media Code Identifier](#)
- [\[6-2-10\] Logiciels & Services Duhem - MacImage](#)
- [\[6-2-11\] Erik Deppe - CD Speed 2000](#)
- [\[6-2-12\] Andre Wiethoff - Exact Audio Copy \(EAC\)](#)
- [\[6-2-13\] Earle F. Philhower, III - cdrLabel](#)
- [\[6-2-14\] Adobe - Audition \(formerly Cool Edit\)](#)
- [\[6-2-15\] Elwin Oost - Burn to the Brim](#)
- [\[6-2-16\] Mike Looijmans - CDWave](#)
- [\[6-2-17\] ECI - DriveEasy](#)
- [\[6-2-18\] Jackie Franck - Audiograbber](#)
- [\[6-2-19\] High Criteria - Total Recorder](#)
- [\[6-2-20\] Smart Projects - IsoBuster](#)
- [\[6-2-21\] GoldWave Inc. - GoldWave](#)
- [\[6-2-22\] Naltech - CD Data Rescue](#)
- [\[6-2-23\] Jufsoft - BadCopy Pro](#)
- [\[6-2-24\] CDRoller Soft Co. - CDRoller](#)
- [\[6-2-25\] FlexiMusic - Wave Editor](#)
- [\[6-2-26\] Nic Wilson - DVD Info Pro](#)
- [\[6-2-27\] Audacity](#)
- [\[6-3\] What is packet writing \(a/k/a DLA - Drive Letter Access\)?](#)
- [\[6-3-1\] What's UDF?](#)

- [\[6-3-2\] Do I want to do packet writing?](#)
- [\[6-4\] What packet writing software should I use?](#)
  - [\[6-4-1\] Roxio - Drag-to-Disc \(a/k/a DirectCD\)](#)
  - [\[6-4-2\] CeQuadrat - PacketCD](#)
  - [\[6-4-3\] SmartStorage - SmartCD for Recording](#)
  - [\[6-4-4\] Gutenberg Systems - FloppyCD](#)
  - [\[6-4-5\] Pinnacle Systems - InstantWrite \(was VOB\)](#)
  - [\[6-4-6\] Prassi - abCD](#)
  - [\[6-4-7\] Ahead - InCD](#)
  - [\[6-4-8\] Oak Technologies - SimpliCD ReWrite](#)
  - [\[6-4-9\] NewTech Infosystems, Inc. \(NTI\) - File CD](#)
  - [\[6-4-10\] Veritas - DLA \(Drive Letter Access\)](#)
  - [\[6-4-11\] BHA - B's CLiP](#)
- [\[6-5\] Can I intermix different packet-writing programs?](#)
- [\[6-6\] I want to write my own CD recording software](#)
  - [\[6-6-1\] PoINT - CDarchive SDK](#)
  - [\[6-6-2\] Golden Hawk Technology \(Jeff Arnold\)](#)
  - [\[6-6-3\] Gear Software - GEAR.wrks](#)
  - [\[6-6-4\] VOB - CD-Wizard SDK](#)
  - [\[6-6-5\] Dialog Medien - ACDwrite.OCX](#)
  - [\[6-6-6\] ECI - The Engine](#)
  - [\[6-6-7\] NUGROOVZ - CDWriterXP](#)
  - [\[6-6-8\] Ashampoo - DiscForge Plug & Burn](#)
  - [\[6-6-9\] NuMedia Soft - CDWriterPro](#)
  - [\[6-6-10\] Sonic Solutions - AuthorScript](#)
- [\[6-7\] What software is available for doing backups?](#)
  - [\[6-7-1\] Adaptec - Easy-CD Backup](#)
  - [\[6-7-2\] D.J. Murdoch - DOSLFNBK](#)
  - [\[6-7-3\] Dantz - Retrospect](#)
  - [\[6-7-4\] Veritas - Backup Exec](#)
  - [\[6-7-5\] Symantec - Norton Ghost](#)
  - [\[6-7-6\] PowerQuest - Drive Image Special Edition for CD-R](#)
  - [\[6-7-7\] Centered Systems - Second Copy](#)
  - [\[6-7-8\] FileWare - FileSync](#)
  - [\[6-7-9\] Novastor - NovaDISK](#)
  - [\[6-7-10\] Roxio - Take Two](#)
  - [\[6-7-11\] NTI - Backup NOW!](#)
  - [\[6-7-12\] CeQuadrat - BackMeUp LT](#)
  - [\[6-7-13\] Duncan Amplification - disk2disk](#)
  - [\[6-7-14\] Pinnacle Systems - InstantBackup \(was VOB\)](#)
  - [\[6-7-15\] Microsoft - Backup](#)

[\[6-7-15\] Portlock Software - Storage Manager](#)

[\[6-7-16\] Willow Creek Software - Backup To CD-RW](#)

[\[6-7-17\] TeraByte Unlimited - Image for Windows](#)

[\[6-8\] How do I get customer support for bundled recording software?](#)

## **[\[7\] Media](#)**

[\[7-1\] What kinds of media are there?](#)

[\[7-2\] Does the media matter?](#)

[\[7-3\] Who manufactures CD-R media?](#)

[\[7-4\] Which kind of media should I use?](#)

[\[7-4-1\] What's the best brand of media?](#)

[\[7-5\] How long do CD-Rs and CD-RWs last?](#)

[\[7-6\] How much data can they hold? 650MB? 680MB?](#)

[\[7-7\] Is it okay to write on or stick a label on a disc?](#)

[\[7-8\] How do CD-Rs behave when microwaved?](#)

[\[7-9\] What can I do with CD-R discs that failed during writing?](#)

[\[7-10\] Where can I find jewel cases and CD sleeves?](#)

[\[7-11\] What's "unbranded" CD-R media?](#)

[\[7-12\] How do I repair a scratched CD?](#)

[\[7-13\] What's this about a Canadian CD-R tax?](#)

[\[7-14\] Can I get 80mm \(3-inch "cd single"\) CD-Rs?](#)

[\[7-15\] Where can I find CD-ROM business cards and "shaped" CDs?](#)

[\[7-16\] Can you tell pressed CDs and silver CD-Rs apart?](#)

[\[7-17\] What's the difference between "data" and "music" blanks?](#)

[\[7-18\] How do I convert data CD-Rs into "consumer audio" blanks?](#)

[\[7-19\] Is translucent media bad?](#)

[\[7-20\] How do I destroy CD-R media beyond all hope of recovery?](#)

[\[7-21\] Can I recycle old CDs, CD-Rs, and CD-RWs?](#)

[\[7-22\] Is there really a fungus that eats CDs?](#)

[\[7-23\] How do I clean CD-R and CD-RW discs?](#)

[\[7-24\] Are "black" discs different from other discs?](#)

[\[7-25\] My disc just shattered in the CD drive!](#)

[\[7-26\] How do I tell which side on a silver/silver disc is up?](#)

[\[7-27\] How should I handle and store CDs?](#)

[\[7-28\] What causes the rainbow effect when looking at the data side?](#)

[\[7-29\] Can I print directly on a CD-R?](#)

## **[\[8\] Net Resources and Vendor Lists](#)**

[\[8-1\] Information resources](#)

[\[8-2\] Magazines and other publications](#)

[\[8-3\] Net.vendors](#)

[\[8-3-1\] Consumer software, hardware, and media](#)

[\[8-3-2\] Net.vendors \(duplication services and hardware\)](#)

[\[8-4\] News sources & mailing lists](#)

## [\[9\] Contributors](#)

---

The last-modified date of each section is shown below the Subject line. The date format used is YYYY/MM/DD. The date stamps were added on 1998/04/06, so you won't find any older than that.

This version of the FAQ is generated automatically by *faq2html*, an application developed specifically for converting the plain ASCII version of the CD-Recordable FAQ to HTML. The program is available [in source form](#).

You are visitor  to this page since June 1st, 1998.

---

[Top of page](#)

FAQ Copyright © 2005 by [Andy McFadden](#). All Rights Reserved.

## Service



### DVD-R / DVD-R Testkits

Das folgende Problem kennen Sie bestimmt auch. Sie haben viel Geld für eine Packung DVD Rohlinge bezahlt, aber keiner will sich so richtig brennen lassen oder wird nicht abgespielt.

Damit Sie in Zukunft Zeit, Geld und Ärger sparen können, haben wir uns etwas für Sie ausgedacht. In Zusammenarbeit mit vielen namhaften Herstellern können wir Ihnen ein **exklusives Test-Kit** anbieten.

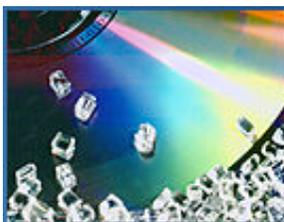
Sie sparen sich die Suche nach den Rohlingen der verschiedenen Hersteller, die es oft auch nicht einzeln zu kaufen gibt. So können Sie zeitgleich und in Ruhe die verschiedenen Rohlinge auf Kompatibilität mit Ihren Geräten testen.

In jedem **Kit** finden Sie 10 verschiedene DVD Rohlinge der wichtigsten Hersteller. Es gibt je ein **DVD-R** und ein **DVD+R Kit**. Sie können natürlich auch beide bestellen. Zusätzlich erhalten Sie eine CD mit einer ausführlichen Anleitung mit allen wichtigen Informationen zum optimalen Einsatz des **Kits**. Eine voll funktionsfähige 30-Tage Testversion von Nero ist ebenfalls enthalten.

Das **Testkit** erhalten Sie nicht im Handel, sondern ausschließlich online direkt bei uns.

**Bestellen Sie hier Ihr Kit.**

## Information



[Optical Storage Directory](#)  
[Rohstoffe & Komponenten](#)  
[Polycarbonat](#)

Sind Sie auf der Suche nach Anbietern und Herstellern von **Polycarbonat** zur Produktion von CDs oder DVDs?

Diese und mehr als 100 weitere Rubriken zum Thema optische Speichermedien finden Sie in unserem **Optical Storage Directory**.

## Interactive Business Card (IBC)



Sie möchten unseren Besuchern mehr Informationen über Ihr Unternehmen und Ihre Dienstleistungen zur Verfügung stellen? Dann sollten Sie ganz einfach eine **Interactive Business Card** mit Ihren Daten und Ihrem Firmenlogo füllen. **Weitere Informationen erhalten Sie hier.**

## News



**28.06.2005** Steigende Ölpreise zwingen optische Medienhersteller weltweit zu Preiserhöhungen

[\[weitere News im Überblick\]](#)

## Marktübersicht

1782 CD-Rohlinge - Die neuesten:



704 DVD-Rohlinge - Die neuesten:



## Service



Sie finden unseren bewährten **Service** zum Thema Überbrennen. Umfassende Informationen zum Thema 90/99 Minuten CD-Rohlinge und Überbrennen. Alles über Medien, Brenner, Einstellungen und mögliche Probleme

Texte und Gestaltung, Copyright © 1998-2005 **InstantInfo** - Alle Rechte vorbehalten. Alle Angaben ohne Gewähr.

Sämtliche aufgeführte Produktbezeichnungen, Logos und Signets sind Warenzeichen der jeweiligen Hersteller.

Anfragen und Anregungen bitte an: [info@instantinfo.de](mailto:info@instantinfo.de)

# Index of /mirrors/faq/storage/

Name	Last modified	Size	Description
 <a href="#">_Parent Directory</a>			
 <a href="#">_comp.arch.storage_FAQ+</a>	03-Aug-04 23:50	78K	
 <a href="#">_comp.arch.storage_FAQ+</a>	03-Aug-04 23:50	105K	

# Index of /mirrors/faq/jpeg/

Name	Last modified	Size	Description
 <a href="#">Parent Directory</a>			
 <a href="#">old/</a>	19-May-98 17:44	1K	
 <a href="#">part1</a>	03-Aug-04 23:50	57K	
 <a href="#">part2</a>	03-Aug-04 23:50	40K	

# Index of /mirrors/faq/mpeg/

Name	Last modified	Size	Description
 <a href="#">Parent Directory</a>			
 <a href="#">MPEG.WWW.faq</a>	03-Aug-04 23:50	1K	
 <a href="#">old/</a>	19-May-98 17:44	1K	
 <a href="#">comp.compression.faq.+</a>	03-Aug-04 23:50	84K	
 <a href="#">jpeg.faq.Z</a>	03-Aug-04 23:50	27K	
 <a href="#">mbone.faq.Z</a>	03-Aug-04 23:50	15K	
 <a href="#">mpeg.faq.30.Z</a>	03-Aug-04 23:50	89K	
 <a href="#">mpeg.faq.31.Z</a>	03-Aug-04 23:50	110K	
 <a href="#">mpeg.faq.32.Z</a>	03-Aug-04 23:50	109K	
 <a href="#">mpeg2.faq.1.Z</a>	03-Aug-04 23:50	5K	
 <a href="#">mpeg2.faq.2.Z</a>	03-Aug-04 23:50	28K	



# Compact Disc Terminology

[home](#)

[blog](#)

[FAQs](#)

[industry](#)

[technology](#)

[applications](#)

[terminology](#)

[bibliography](#)

[history](#)

[humor](#)

[about us](#)

[top of page](#)

## Compact Disc Terminology

**Note:** Other online glossaries of terms used in the compact disc industry can be found at:

- [Octave Systems's glossary](#)
- [Sanyo-Verbatim's CD glossary](#)
- [JVC America's glossary](#)

**A Complete  
Digital Media  
e-Reference  
Library!**

Thousands of  
bestselling  
tech books

**Safari**

Better than  
e-books

**Try it FREE**

- **BLER**

Block Error Rate. This is the "raw" digital error rate before any error correction.

- **BLERmax**

The maximum number of [BLERs](#) allowed on a disc. According to the industry standard, a CD-ROM is allowed a BLER of up to 220 before it is considered a "bad" disc.

- **CD**

Compact Disc, a digital medium formed of a 12cm [polycarbonate substrate](#), a [reflective metalized layer](#), and a protective lacquer coating. The physical format of CDs is described by the ISO9660 industry standard. [CD-Recordable](#) discs also have an organic dye data layer between the substrate and the metal reflective layer.

- **CD-R**

Compact Disc-Recordable. This term is used to describe the technology of recordable CD as well as the equipment, software and [media](#) used to make recordable discs.

- **Cross-talk**

This is a measure of the amount of interference coming from neighboring pit tracks on a CD. As track pitch is tightened (when tracks are packed closer together to put more data on a disc), cross-talk increases. A maximum value of 50% is allowed by Red Book specifications.

- **cyanine**

One type of [organic dye](#) used to form the data layer in [CD-R](#) discs. Cyanine was the first material used for these discs, but presently a

metal-stabilized cyanine compound is generally used instead of "raw" cyanine. An alternative material is [phthalocyanine](#).

- o **data layer**

In [CD-R](#), the [organic dye](#) sandwiched between the [polycarbonate substrate](#) and the [metalized reflective layer](#) of the [media](#). [CD-Recordable](#) discs do not have any data on them until they are recorded. Instead the recording laser selectively melts "[pits](#)" into the dye layer -- but rather than burning holes in the dye, it simply melts it slightly, causing it to become non-translucent so the reading laser beam is refracted rather than reflected back to the reader's sensors. In pressed CDs, the data layer is part of the polycarbonate substrate, and is pressed into the top side of it by a "[stamper](#)" during the [injection moulding process](#).

- o **injection moulding**

A manufacturing method where molten material is forced into a mold, usually under high pressure, and then cooled so the material takes on the shape of the mirror image of the mold.

- o **jitter**

*Definition temporarily unavailable. Please check back later*

- o **lacquer spincoat**

Acrylic lacquer is spincoated in a thin layer on top of the [metal reflective layer](#) of a [CD](#) to protect it from abrasion and corrosion. Usually a decorative label is also applied on top of the

lacquer, but this is not a standard requirement.

- o **mastering**

Mastering is the process of creating a stamper or set of stampers to be used in the [injection moulding](#) stage of manufacturing compact discs. During this process a digital signal from a computer is used to guide a laser beam which etches a pattern of "[pits and lands](#)" (in the case of [CDs](#)) or a continuous groove (for [CD-Rs](#)) onto a highly polished glass disc coated with photoresist. This "glass master" is then cured (developed) with ultraviolet light and rinsed off, and a metal (nickel or silver) mold is electroformed on top of it. This mold is removed and then electroplated with a nickle alloy to create one or more [stampers](#) to be used in the [injection moulding](#) machine to press the data into the [polycarbonate substrate](#) of CDs, or the guiding groove into the substrate of CD-Rs.

- o **media or "blanks"**

[CD-Recordable](#) media are the discs used to record digital information using a special recorder and premastering software with a computer.

These discs are made of a [polycarbonate substrate](#), a layer of [organic dye](#), a metalized [reflective layer](#), and a protective [lacquer](#) coating.

Some discs also have an additional protective coating over the metalized layer, and some discs have a printable surface silkscreened on them.

- o **Orange Book**

The Orange Book is the specification for [CD-Recordable](#).

---

- o **organic dye**

The [data layer](#) of [CD-R discs](#) is made from either [cyanine](#) or [phthalocyanine](#) dye which is melted during the recording process. Where the dye is melted, it becomes opaque or refractive, scattering the reading laser beam so it is not reflected back into the reader's sensors. The difference between reflected and non-reflected light is interpreted by the player as a binary signal.

- o **phthalocyanine**

An organic dye used to form the [data layer](#) in some [CD-Recordable](#) discs. [Mitsui Toatsu Corporation](#) holds the patent on this dye, but has licensed its formula to some other manufacturers.

- o **pits & lands**

In a "pressed" or mass-replicated CD, the bumps and grooves that represent the binary data on a disc's [substrate](#) are pressed into it during manufacture. [CD-R](#) discs do not have true pits and lands, but the unmelted, clear areas and melted, opaque places in the [dye layer](#) fulfill the same function as pits and lands on a pressed disc.

- o **reflective layer**

The metal layer on top of the dye that reflects the laser beam back to the reading assembly. This is usually 24K gold in [CD-Recordable](#) discs, but Mitsubishi has recently introduced a silver disc as well.

- o **stamper**

The data-bearing removable "die" used during the [injection moulding](#) of a [CD](#) to imprint [pits and lands](#) into the [polycarbonate substrate](#) of the disc. In manufacturing [CD-R](#) media, instead of pits and lands, a continuous spiral is pressed into the substrate as a guide to the recorder's laser. The stamper is part of a "disc family" created in the [mastering](#) process.

- o **substrate**

The optical-quality, [injection moulded](#) optical-quality clear polycarbonate plastic "bottom" of a [CD](#) or [CD-R](#). For [CD-Rs](#), this layer does not contain "[pits and lands](#)" but has a single spiral groove that guides the recorder's laser.

Thursday, 16-Dec-2004 00:20:28 EST



this  
Web site

Thursday, 16-Dec-2004 00:20:28 EST

Sponsored Links

<a href="#">YOUR link could be here! Contact us for details.</a>	<a href="#">CD labels - Laser Labels</a>	<a href="#">Wholesale CD-R Media</a>
--	--	--------------------------------------

*Important Notice!* If you read this Web site in a language other than English, please be aware that it has been translated without the author's permission or review, and may not say exactly what was intended. Also, some links and any fill-in forms may not function properly in the translated pages. If you encounter problems such as this, please navigate to our native homepage at <http://www.cd-info.com>. Thank you for your understanding.



**Entertainment**

**Music**  
**Online Casinos**  
**Free Online Games**  
**Lotteries**  
**Dating**

**Shopping**

**Flowers**  
**Furniture**  
**Car Search**  
**Jewelry**  
**Hobbies**

**Internet**

**MP3**  
**Chat**  
**Web Hosting**  
**Domain Registration**  
**Advertising**

**Travel**

**Rental Cars**  
**Vacations**  
**Airline Tickets**  
**Maps**  
**Hotels**

**Finance**

**Mortgages**  
**Credit Cards**  
**Real Estate**  
**Investing**  
**Stocks**

**Health & Fitness**

**Nutrition**  
**Weight Loss**  
**Women's Health**  
**Prescription Drugs**  
**Health Food**



## Library

 [Printer-friendly version](#)

Product Navigation

 [CD DVD Duplicators](#)

 [DVD-R Media](#)

 [DVD+R Media](#)

 [CD-R RW Media](#)

 [DVD CD Recorders](#)

 [Software](#)

 [Accessories](#)

 [Deals](#)

 [Printers CD DVD](#)

 [Digital Photo Archiving](#)

 [Cartridges Ink Thermal](#)

 [DVD Players](#)

 [CD DVD Duplication](#)

### o Manuals

- o [Copy Master Manual](#)
- o [Primera CD DVD Bravo Quick Start Guide PDF \(82KB\)](#)
- o [CopyWriter Live Manual \(PDF 485KBs\)](#)
- o [ASUS CDRW 5232AS Manual](#)
- o [Pioneer A08XL Manual](#)

### o Newsletters

- o [Newsletters](#)

### o Articles

- o [General DVD-R Information](#)
- o [Ink Jet Printers v. Thermal Printers](#)
- o [Recording compact disc Digital Audio \(How to record a red book CD-R\)](#)
- o [How to create a SVCD with Nero](#)

### o Reviews

- o [Primera Ink-Jet Printer](#)

### o Other

- o [FAQs Sections: CD Software CD Duplication CD Recorders DVD](#)
- o [Copy Master FAQs](#)
- o [Official Internet DVD FAQ for the rec.video.dvd Usenet newsgroups \*\*DVD FAQ \\*500 KBs\*\*](#)
- o [DVD Compatibility Chart](#)
- o [DVD - General Purpose v. Authoring Media](#)
- o [How to identify CD media disc manufacturer with Nero](#)

**Happy Thanksgiving!**

Online ordering open.  
Phones re-open Monday.

	20@ \$33.95
	10@ \$34.90
	5@ \$35.99
	3@ \$39.65



**Note to authors: We are accepting articles for publication in our library. Please submit your article to [paul@octave.com](mailto:paul@octave.com) for review. Octave makes the final decision on what is published.**

More to come!

[Home](#)



- [Shipping Weights](#)
- [Resources](#)
- [CD Duplicators & DVD Copiers](#)
- [DVD Duplicators](#) by StorDigital
- 

Prices in U.S. Dollars

Order online or call **1-800-626-8539** Open M-F 8-5 Pacific Time

Octave Systems, Inc.

California Location  
[504A Vandell Way](#)  
[Campbell, CA 95008, USA](#)  
Phone: 408-866-8424  
Fax: 408-866-4252  
[info@octave.com](mailto:info@octave.com)

Washington Location  
[916 N. Wright Blvd.](#)  
[Liberty Lake, WA 99019](#)  
Toll-Free: 800-440-4142  
Phone: 509-922-5718  
Fax: 509-922-5749

[Privacy Policy](#)

©2005

## Glossary

- [A-Time](#)
- [Authoring](#)
- [CD-I](#)
- [CD-ROM/XA](#)
- [ECC/EDC](#)
- [HFS](#)
- [Hybrid Disc](#)
- [Indexing](#)
- [ISO 9660](#)
- [Media Conversion](#)
- [Mixed Mode Disc](#)
- [Mode 1](#)
- [Mode 2](#)
- [PQ Information](#)
- [Premastering](#)
- [SMPTE Time](#)
- [Tagging](#)
- [Transfer Rate](#)
- [Yellow Book](#)

### **A-Time:**

Absolute Time. Elapsed time, referenced to the program start (00:00:00), on a CD or R-DAT. A-Time is measured in minutes, seconds, and frames.

### **Authoring:**

Creation of a database for a CD-ROM. The end product of authoring is usually a search and retrieval type document with the addition of a user interface. Specific authoring functions include tagging and indexing.

### **CD-I:**

Compact Disc Interactive. An interactive multimedia system which connects to a television. The CD-I standard is known as the Green Book.

### **CD-ROM/XA:**

CD-ROM Extended Architecture. A standard which allows interleaving of compressed audio and video data for synchronization purposes.

### **ECC/EDC:**

Error Correction Code/Error Detection Code. Codes specified in the color book standards and imbedded in CD data which facilitate the reconstruction of data if read errors occur.

### **HFS:**

Hierarchical File System, used by the Macintosh platform. HFS formatted CD-ROMs have the same file structure as an Apple hard disk.

### **Hybrid Disc:**

A CD-ROM which can function on either the PC or the Macintosh platform. The disc contains separate ISO 9660 and HFS partitions.

### **Indexing:**

Creation of a data index to speed up search and retrieval.

### **ISO 9660:**

An international standard defining the file and directory structures for CD-ROM. An ISO 9660 formatted CD-ROM will function on any computer platform containing the appropriate driver software. Most

common with PC compatible systems

**Media Conversion:**

The process of converting data from one type of media to another for premastering and mastering. Premastering software typically requires input data on hard disk. 8mm tape and compact disc are preferred as input media for the mastering process.

**Mixed Mode Disc:**

A CD-ROM that contains both a computer data track (#1) and audio tracks (#2-99)

**Mode 1:**

Most common CD-ROM data format. Contains three layers of error correction for computer data.

**Mode 2:**

CD-ROM data format with two layers of data correction for audio and compressed video. Utilized in CD-ROM/XA.

**PQ Information:**

Information on the disc (or tape) that determines track start points, control bits, timing information, etc.

**Premastering:**

The process of formatting data into the exact image that will appear on a CD-ROM, including file structure (i.e. ISO 9660) and file locations. A premastered image is ready to be mastered and replicated.

**SMPTE Time:**

Time code adopted for use with the 3/4" U-matic tape used in CD Production. Originally used for video, this was devised by the Society of Motion Picture and Television Engineers.

**Tagging:**

Placing hidden markers in text to indicate where to insert specific images.

**Transfer Rate:**

The amount of data which is transferred from the CD-ROM to the computer. The CD-ROM transfer rate is limited by the speed at which the disc rotates in the drive. The conventional CD-ROM transfer rate is approximately 150 kilobytes/sec, referred to as 1x. Therefore, a quadruple speed (4x) CD-ROM drive can transfer data at a rate of 600 KB/sec.

**Yellow Book:**

International standard which defines the physical properties of a CD-ROM disc.

- [How CD data structures are formed](#)
- [CD Readback System](#)
- [How "real" computer data is generated from a CD-ROM](#)
- [CD Manufacturing](#)
  - [Step 1 - Pre-Production](#)
  - [Step 2 - Mastering](#)
  - [Step 3 - Replication and Fullfillment](#)

 [Glossary](#)

 [CD Overview](#)

 [Top](#)

---

[Products](#) · [Services](#) · [Specifications](#) · [News](#) · [Order](#) · [Site Map](#) · [Home](#)

---

Optical Disc Solutions  
1767 Sheridan Street  
Richmond, Indiana 47374

BKlaine@odiscs.com  
Phone: 800.704.7648  
Fax: 765.935.0174

© 2005, Optical Disc Solutions, Inc., All Rights Reserved

## DVD Glossary

- [A-Time](#)
- [AC-3](#)
- [Artifacts](#)
- [Aspect ratio](#)
- [Authoring](#)
- [Bit rate](#)
- [CBR](#)
- [Chapter stop](#)
- [Compress](#)
- [CSS](#)
- [DDP](#)
- [DLT](#)
- [DVD-5](#)
- [DVD-9](#)
- [DVD-10](#)
- [DVD-18](#)
- [Decompress](#)
- [ECC/EDC](#)
- [Indexing](#)
- [LBR](#)
- [Macrovision](#)
- [Media Conversion](#)
- [MPEG1, MPEG2](#)
- [NTSC](#)
- [PAL](#)
- [Pan & Scan](#)
- [Pit Art](#)
- [PQ Information](#)
- [Premastering](#)
- [Regional Coding](#)
- [Subpicture information](#)
- [Substrate](#)
- [Tagging](#)
- [Telecine process](#)
- [VBR](#)

### **A-Time:**

Absolute Time. Elapsed time, referenced to the program start (00:00:00), on a DVD. A-Time is measured in minutes, seconds, and frames.

### **AC-3:**

A six-channel digital audio format; its official name is Dolby Digital Surround Sound System.

### **Artifacts:**

Interference or other unwanted "noise" in video such as flickering or changes in color.

### **Aspect ratio:**

The dimensions of the image. A "typical television" ratio is 4 X 3 (also known as 1.33 X 1). Many motion pictures, and some new TVs, have a 16 X 9 format (also known as 1.78 X 1).

### **Authoring:**

Creation, combining and setup of a various files for a DVD video, DVD-ROM, or audio disc. This includes audio, video, graphics and text files. The end result is a DLT tape with DVD image files and DDP descriptor.

### **Bit rate:**

The rate at which bits of data are encoded on the disc. A typical bit rate is 3.5 Megabits per second.

### **CBR:**

Constant bit rate compression. This indicates that in the encoding

process the bit rate does not change, despite the simplicity or complexity of the image being encoded. (See [VBR](#)).

**Chapter stop:**

Programming that allows a viewer to jump immediately to a particular part.

**Compress:**

To reduce or compact audio or video information so that it can more easily be stored or transmitted.

**CSS:**

Content scrambling system, a type of Digital copy protection sanctioned by the DVD Forum.

**DDP:**

Disc Description Protocol is a small file(s) which describe how to master a data image file for optical disc(DVD or CD). This is ANSI industry standard developed by Doug Carson and Associates. This information is used in the mastering process by the Laser Beam Recorders.

**DLT:**

Digital Linear Tape, a high storage capacity (10-20 Gbytes) tape used as the input medium to master DVD. Media designated "type III" or "type IV" tapes are used for DVD.

**DVD-5:**

DVD format in which 4.7 Gigabytes of data are stored on one side of the disc in one layer.

**DVD-9:**

DVD format in which 8.5 Gigabytes of data are stored on one side of the disc in two layers.

**DVD-10:**

DVD format in which 9.4 Gigabytes of data are stored on two sides of the disc in one layer each.

**DVD-18:**

DVD format in which 17.0 Gigabytes of data are stored on two sides of the disc in two layers each.

**Decompress:**

To change audio and video information from a compacted form to its original state.

**ECC/EDC:**

Error Correction Code/Error Detection Code. Codes specified in the color book standards and imbedded in DVD data which facilitate the reconstruction of data if read errors occur.

**Indexing:**

Creation of a data index to speed up search and retrieval.

**LBR:**

Laser beam recorder. It creates the DVD master disc.

**Macrovision:**

An analog protection scheme developed by Macrovision for the protection of analog copying. It is widely used in VHS and has now been applied to DVD.

**Media Conversion:**

The process of converting data from one type of media to another for premastering and mastering. Premastering software typically requires

input data on hard disk. 8mm tape and compact disc are preferred as input media for the mastering process.

**MPEG1, MPEG2:**

Standards for compressing video. MPEG stands for Moving Pictures Expert Group.

**NTSC:**

National Television Systems Committee, which devised the standards for TV broadcasting in the United States, Canada, Japan, and a few other countries. (See [PAL](#))

**PAL:**

Phase Alteration Line. A standard for TV broadcasting in Europe, Australia and New Zealand, and other countries. (See [NTSC](#))

**Pan & Scan:**

A method of transferring movies with an aspect ratio of 16 X 9 to film, tape, or disc to be shown on a conventional TV with a 4 X 3 aspect ratio. Only part of the full image is selected for each scene. Pan & Scan is the opposite of "letterbox" or "widescreen."

**Pit Art:**

A type of DVD labeling in which the pits are cut in a design to resemble writing or another image.

**PQ Information:**

Information on the disc (or tape) that determines track start points, control bits, timing information, etc.

**Premastering:**

The process of formatting data into the exact image that will appear on a DVD, including file structure and file locations. A premastered image is ready to be mastered and replicated.

**Regional Coding:**

Copy-protection coding built into a DVD disc that allow it to be played, or prevent it from being played, in one or more regions of the world. For this purpose, the world is divided into six regions.

**Subpicture information:**

Captions, subtitles, or other text that can be displayed or hidden.

**Substrate:**

A DVD half-disc. Two substrates, each 0.6 mm thick, are bonded together to form a 1.2 mm thick DVD disc.

**Tagging:**

Placing hidden markers in text to indicate where to insert specific images.

**Telecine process:**

The process of turning film into video.

**VBR:**

Variable bit rate compression. This indicates that in the encoding process the bit rate changes, depending upon the simplicity or complexity of the image being encoded. (See [CBR](#)).

 [Back](#)

 [DVD Authoring](#)

- [Mastering](#)
- [Replication](#)
- [Printing and Packaging](#)
- [Services](#)
- [DVD Overview](#)

▲[Top](#)

---

[Products](#) · [Services](#) · [Specifications](#) · [News](#) · [Order](#) · [Site Map](#) · [Home](#)

---

Optical Disc Solutions  
1767 Sheridan Street  
Richmond, Indiana 47374

BKlaine@odiscs.com  
Phone: 800.704.7648  
Fax: 765.935.0174

© 2005, Optical Disc Solutions, Inc., All Rights Reserved

[Author Index](#)

Enter  
keywords you wish to find  
information about or  
words describing a concept



Search Conservation OnLine document library  
except [AIC](#), [Albumen](#) (each of which has its own search engine) and [California Preservation Clearinghouse](#).

[Mailing list archives](#) are excluded from this index. Each mailing list has its own separate search facility, which you'll find when on the main page for that list.

If you don't find what you want with this index, try a Google search, below.

Search CoOL with Google



Entire Conservation OnLine document library

---

[\[Search all CoOL documents\]](#)

[\[Feedback\]](#)

This page last changed: January 30, 2005



# CoOL Feedback

Your comments and suggestions for ways to improve Conservation OnLine will be most appreciated. Please understand that I probably won't be able to respond to every message personally but be assured that I will read it and consider it very carefully.



This form is **only** for feedback about CoOL

**Please do NOT use it to ask for conservation advice nor for any matters not directly related to the operation of this server**

**NO** reference questions, please!

**Name**

**Email**

**Subject**

**Comments:**

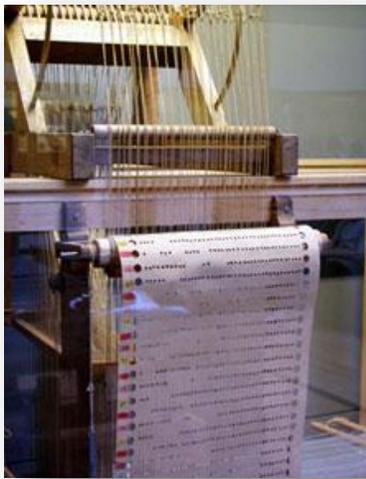
Remember to hit a **Carriage Return** when you reach the right margin (70 chars); there is no word wrap and anything you type past the margin will get truncated!

## Introduction to Data Storage Media

By Craig Ball

Text © 2011

Mankind has been storing data for thousands of years, on media as diverse as stone, bone, clay, wood, metal, glass, skin, papyrus, paper, plastic and film. In fact, people were storing data in binary formats long before the emergence of modern digital computers. Records from 9<sup>th</sup> century Persia describe an organ playing interchangeable cylinders. Eighteenth century textile manufacturers employed perforated rolls of paper to control looms, and Swiss and German music box makers used metal drums or platters to store tunes. At the dawn of the Jazz Age, no self-respecting American family of means lacked a player piano capable (more or less) of reproducing the works of the world's greatest popular and classical composers.

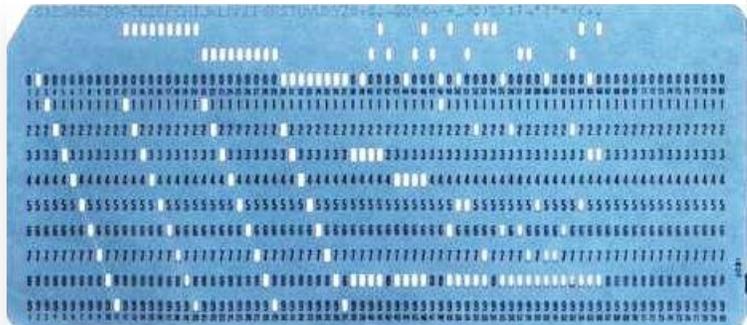


Whether you store data as a perforation or a pin, you're storing binary data. That is, there are two data states: hole or no hole, pin or no pin, zeroes or ones.

### Punched Cards

In the 1930's, demand for *electronic* data storage led to the development of fast, practical and cost-effective binary storage media. The first of these were punched cards, initially made in a variety of sizes and formats, but ultimately standardized by IBM as the 80 column, 12 row (7.375" by 3.25") format that dominated computing well into the 1970's.

IBM 5081 80 column card



[From 1975-79, the author spent many a midnight in the basement of a computer center at Rice University typing program instructions on these unforgiving punch cards].

The 1950's saw the emergence of magnetic storage as the dominant medium for electronic data storage, and it remains so today. Although optical and solid state storage are expected to ultimately eclipse magnetic media for local storage, magnetic storage will continue to dominate network and cloud storage well into the 2020s, if not beyond.



## Tape

The earliest popular form of magnetic data storage was magnetic tape. Spinning reels of tape were a clichéd visual metaphor for computing in movies and television shows from the 1950's through 1970's. Though the miles of tape on those reels now reside in cartridges and cassettes, tape remains an enduring medium for backing up and archiving electronically stored information. The LTO-5 format introduced in 2010 natively holds 1.5 terabytes of uncompressed data and delivers a transfer rate of 140 megabytes per second. Since most data stored on backup tape is compressed, the actual volume of ESI on tape may be 2-3 times greater than the native capacity of the

tape.

Magnetic tape was the earliest data storage medium for personal computers, including the pioneering Radio Shack TRS-80 and the very first IBM personal computer, the model XT.

While tape isn't as fast or capacious as hard drives, it's proven to be more durable and less costly for long term storage; that is, so long as the data is being *stored*, not *restored*.

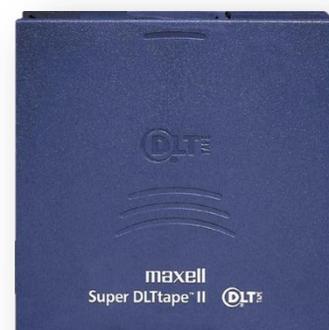
### LTO-5 Ultrium Tape



### Sony AIT-3 Tape



### SDLT-II Tape



## Chronology of Magnetic Tape Formats for Data Storage (Wikipedia)

1951 – UNISERVO	1986 - SLR
1952 - IBM 7 track	1987 - Data8
1958 - TX-2 Tape System	<b>1989 - DDS/DAT</b>
1962 – LINCtape	1992 - Ampex DST
1963 – DECtape	1994 - Mammoth
1964 - 9 Track	1995 - IBM 3590
1964 – MagCard Selectric typewriter	1995 - Redwood SD-3
1966 - 8-Track Tape	<b>1995 - Travan</b>
1972 - QIC	<b>1996 - AIT</b>
1975 - KC Standard, Compact Cassette	1997 - IBM 3570 MP
1976 - DC100	1998 - T9840
1977 - Commodore Datasette	1999 – VXA
1979 – DECtapell	2000 - T9940
1979 - Exatron Stringy Floppy	<b>2000 - LTO Ultrium</b>
1983 - ZX Microdrive	<b>2003 - SAIT</b>
1984 - Rotronics Wafadrive	2006 - T10000
1984 - IBM 3480	2007 - IBM 3592
<b>1984 - DLT</b>	2008 - IBM TS1130

For further information, see Ball, [Technology Primer: Backups in Civil Discovery](http://www.craigball.com/Ball_Technology%20Primer-Backups%20in%20E-Discovery.pdf) at [http://www.craigball.com/Ball\\_Technology%20Primer-Backups%20in%20E-Discovery.pdf](http://www.craigball.com/Ball_Technology%20Primer-Backups%20in%20E-Discovery.pdf)

## Floppy Disks

It's rare to encounter a floppy disk today, but floppy disks played a central role in software distribution and data storage for personal computing for almost thirty years. Today, the only place a computer user is likely to see a floppy disk is as the menu icon for storage on the menu bar of Microsoft Office applications. All floppy disks have a spinning, flexible plastic disk coated with a magnetic oxide (e.g., rust). The disk is essentially the same composition as magnetic tape in disk form. Disks are **formatted** (either by the user or pre-formatted by the manufacturer) so as to divide the disk into various concentric rings of data called **tracks**, with tracks further subdivided into tiny arcs called **sectors**. Formatting enables systems to locate data on physical storage media much as roads and lots enable us to locate homes in a neighborhood.

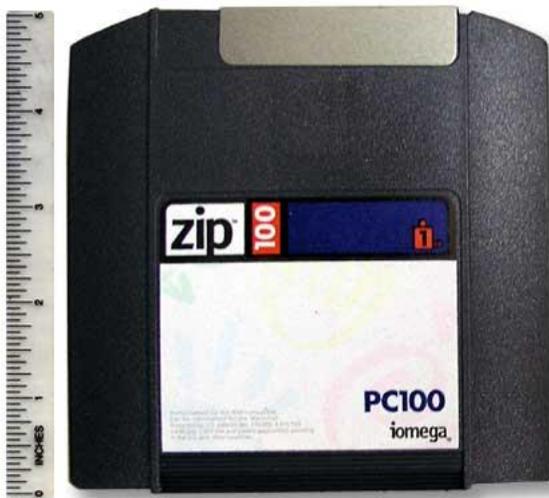
Though many competing floppy disk sizes and formats have been introduced since 1971, only five formats are likely to be encountered in e-discovery. These are the 8", 5.25", 3.5 standard, 3.5 high density and Zip formats. Of these, the 3.5HD format 1.44 megabyte capacity floppy is by far the most prevalent legacy floppy disk format.

The Zip Disk was one of several proprietary “super floppy” products that enjoyed brief success before the high capacity and low cost of recordable optical media (CD-R and DVD-R) and flash drives rendered them obsolete.



8", 5.25" and 3.5" Floppy Disks

### Zip Disk



### 8" Floppy Disk in Use



### Optical Media

The most common forms of optical media for data storage are the CD, DVD and Blu-ray disks in read only, recordable or rewritable formats. Each typically exists as a 4.75" plastic disk with a metalized reflective coating and/or dye layer that can be distorted by a focused laser beam to induce pits and lands in the media. These pits and lands, in turn, interrupt a laser reflected off the surface of the disk to generate the ones and zeroes of digital data storage. The practical difference between the three prevailing forms of optical media are their native data storage capacities and the availability of drives to read them.

A **CD** (for **Compact Disk**) or **CD-ROM** (for **CD Read Only Media**) is read only and not recordable by the end user. It's typically fabricated in factory to carry music or software. A **CD-R** is recordable by the end user, but once a recording session is closed, it cannot be altered in normal use. A **CD-RW** is a re-recordable format that can be erased and written to multiple times. The native data storage capacity of a standard-size CD is about 700 megabytes.



A **DVD** (for **Digital Versatile Disk**) also comes in read only, recordable (**DVD±R**) and rewritable (**DVD±RW**) iterations. The most common form of the disk has a native data storage capacity of approximately 4.7 gigabytes. So, one DVD holds the same amount of data as six and one-half CDs.

By employing the narrower wavelength of a blue laser to read and write disks, a dual layer **Blu-ray** disk can hold up to about 50 gigabytes of data, equalling the capacity of about ten and one-half DVDs. Like their predecessors, Blu-ray disks are available in recordable (BD-R) and rewritable (BD-RE) formats.

### Hard Disk Drives

The hard disk drive has been around for more than fifty years, but it was not until the 1980's that the physical size and cost of hard drives fell sufficiently for their use to be commonplace. Though most attention has been paid to the amazing leaps in microprocessor technology described by **Moore's Law** (named for Intel co-founder, Gordon Moore, and projecting that the number of transistors on a microprocessor doubles every two years), the strides made in hard drive capacity and cost are every bit as breathtaking.

Introduced in 1956, the IBM 350 Disk Storage Unit (pictured with child to illustrate size) was the first commercial hard drive. It was 60 inches long, 68 inches high and 29 inches deep (so it could fit through a door). It held 50 magnetic disks of 50,000 sectors, each storing 100 alphanumeric characters. That is, it held

IBM 350 Disk Storage Unit



4.4 megabytes, or enough for about two cellphone snapshots today. It weighed a ton (literally), and users paid \$130.00 per month to *rent* each megabyte of storage.

Today, that same \$130.00 *buys* a 3 terabyte hard drive from Walmart that stores *3 million times* more information, weighs less than three pounds and hides behind a paperback book.

Over time, hard drives have taken various shapes and sizes (or “form factors” as the standard dimensions of key system components are called in geek speak).

The photo at right depicts six hard drives with covers removed: 8”, 5.25”, 3.5”, and 2.5”, 1.8”, and 1” disks. Of these, three form factors are still in common use: 3.5” (desktop drive), 2.5” (laptop drive) and 1.8” (iPod and microsystem drive).

Hard drives connect to computers by various mechanisms called “interfaces” that describe both how devices “talk” to one another, as well as the physical plugs and cabling required. The five most common hard drive interfaces in use today are:

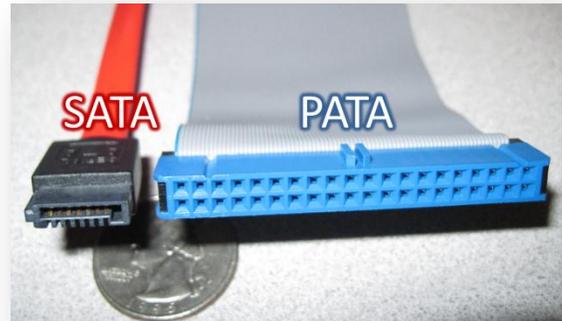
1. **PATA** for **Parallel Advanced Technology Attachment** (sometimes called EIDE for **Extended Integrated Drive Electronics**):
2. **SATA** for **Serial Advanced Technology Attachment**
3. **SCSI** for **Small Computer System Interface**
4. **SAS** for **Serial Attached SCSI**
5. **FC** for **Fibre Channel**

Though once dominant in personal computers, PATA drives are rarely found in machines manufactured after 2006. Today, virtually all laptop and desktop computers employ SATA drives for local storage. SCSI, SAS and FC drives tend to be seen exclusively in servers and other applications demanding high performance and reliability.

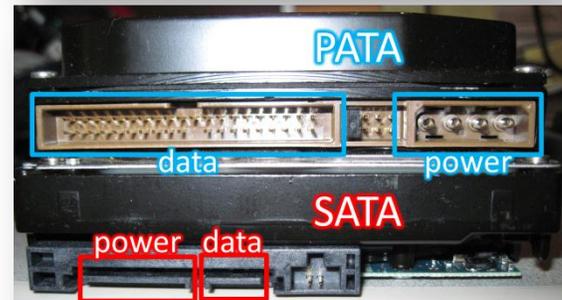


From the user's perspective, PATA, SATA, SCSI, SAS and FC drives are indistinguishable; however, from the point of view of the technician tasked to connect to and image the contents of the drive, the difference implicates different tools and connectors.

The five drive interfaces divide into two employing parallel data paths (PATA and SCSI) and three employing serial data paths (SATA, SAS and FC). Parallel ATA interfaces route data over multiple simultaneous channels necessitating 40 wires where serial ATA interfaces route data through a single, high-speed data channel requiring only 7 wires. Accordingly, SATA cabling and connectors are smaller than their PATA counterparts (see photos, right).



Fibre Channel employs optical fiber (the spelling difference is intentional) and light waves to carry data at impressive speeds. The premium hardware required by FC dictates that it will be found in enterprise computing environments, typically in conjunction with a high capacity/high demand storage device called a **SAN** (for Storage Attached Network) or a **NAS** (for Network Attached Storage).



It's easy to become confused between hard drive interfaces and external data transfer interfaces like USB or FireWire seen on external hard drives. The drive within the external hard drive housing will employ one of the interfaces described above (except FC); however, to facilitate external connection to a computer, a device called a **bridge** will convert data written to and from the hard drive to a form that can traverse a USB or FireWire connection. In some compact, low-cost external drives, manufacturers dispense with the external bridge board altogether and build the USB interface right on the hard drive's circuit board.

## RAIDs

Whether local to a user or in the Cloud, hard drives account for nearly all the electronically stored information attendant to e-discovery. In network server and Cloud applications, hard drives rarely work alone. That is, hard drives are ganged together to achieve greater capacity, speed and reliability in so-called **Redundant Arrays of Independent Disks** or **RAIDs**.

In the SAN pictured here, the 16 hard drives housed in trays may be accessed as **Just a Bunch of Disks** or **JBOD**, but it's far more likely they are working together as a RAID.



RAIDs serve two ends: redundancy and performance. The redundancy aspect is obvious—two drives holding identical data safeguard against data loss due to mechanical failure of either drive—but how do multiple drives improve **performance**? The answer lies in splitting the data across more than one drive using a technique called **striping**.

A RAID improves performance by dividing data across more than one physical drive. The swath of data deposited on one drive in an array before moving to the next drive is called the "stripe." If you imagine the drives lined up alongside one another, you can see why moving back-and-forth across the drives to store data might seem like painting a stripe across the drives. By striping data, each drive can deliver its share of the data simultaneously, increasing the amount of information handed off to the computer's microprocessor.

But, when you stripe data across drives, information is lost if any drive in the stripe fails. You gain performance, but surrender security.

This type of RAID configuration is called a **RAID 0**. It wrings maximum performance from a storage system, but it's risky.

If RAID 0 is for gamblers, **RAID 1** is for the risk averse. A RAID 1 configuration duplicates everything from one drive to an identical twin, so that a failure of one drive won't lead to data loss. RAID 1 doesn't improve performance, and it requires twice the hardware to store the same information.

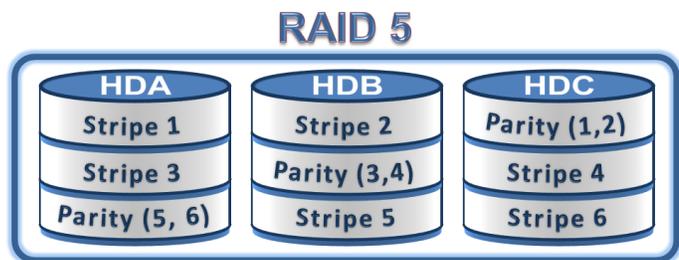
Other RAID configurations strive to integrate the *performance* of RAID 0 and the *protection* of RAID 1.

Thus, a "RAID 0+1" mirrors two striped drives, but demands four hard drives delivering only half their total storage capacity. It's safe and fast, but not cost-efficient. The better solution grows out of a concept called **parity**, key to a variety of RAID configurations. Of those other configurations, the one most often seen is called **RAID 5**.

To understand parity, consider the simple equation  $5 + 2 = 7$ . If you didn't know one of the three values in this equation, you could easily solve for the missing value, i.e., presented with " $5 + \_ = 7$ ," you can reliably calculate the missing value is 2. In this example, "7" is the **parity value** or checksum for "5" and "2."

The same process is used in RAID configurations to gain increased performance by striping data across multiple drives while using parity values to permit the calculation of any missing values lost to drive failure. In a three drive array, any one of the drives can fail, and we can use the remaining two to recreate the third (just as we solved for 2 in the equation above).

In this illustration of **RAID 5**, data is striped across three hard drives, HDA, HDB and HDC. HDC holds the parity values for data stripe 1 on HDA and stripe 2 on HDB. It's shown as "Parity (1, 2)." The parity values for the other stripes are distributed on the other drives. Again, any one of the three drives can fail and all of the data is recoverable. This configuration is called RAID 5, and though it requires a minimum of three drives, it can be expanded to dozens or hundreds of drives.



### Flash Drives, Memory Cards and Solid State Drives

Computer memory storage devices have no moving parts, and the data resides entirely within the solid materials which compose the memory chips, hence the term, "solid state." Historically, rewritable memory was volatile (in the sense that contents disappeared when power was withdrawn) and expensive. But beginning around 1995, a type of non-volatile memory called NAND flash became sufficiently affordable to be used for removable storage in emerging applications like digital photography. Further leaps in the capacity and dips in the cost of NAND flash led to the near-eradication of film for photography and the extinction of the floppy disk, which was replaced by simple, inexpensive and reusable USB storage devices called, variously, flash drives, thumb drives, pen drives and memory sticks or keys.





Solid state drives in standard hard drive form factors are increasingly displacing platter-based hard drives. Solid state drives are significantly faster, lighter and more energy efficient than mechanical drives, but they currently cost anywhere from 20-30 times more per gigabyte than their spinning counterparts. All signs point to the eventual obsolescence of mechanical drives by solid state drives, and some products (notably tablets like the iPad and ultra-lightweight laptops like the MacBook Air) have eliminated hard drives altogether in favor of solid state storage.

Currently, solid state drives assume the size and shape of mechanical drives to facilitate compatibility with existing devices. However, the size and shape of mechanical hard drives was driven by the size and operation of the platter they contain. Because solid state storage devices have no moving parts, they can assume virtually any shape. It's likely, then, that slavish adherence to 2.5" and 3.5" rectangular form factors will diminish in favor of shapes and sizes uniquely suited to the devices that employ them.



### Binary and Decimal Prefixes

As we evolved with ten fingers and toes, our counting systems gravitated to base ten; that is, to the decimal system. But digital storage systems are built around base two, the binary system. Accordingly, what we call a kilobyte isn't really 1,000 bytes ( $10^3$ ). It's actually the binary value  $2^{10}$  or 1,024 bytes. We mostly ignore the difference, until we buy a 100GB drive and wonder why the computer sees just 93.1GB. The gap between binary and decimal capacities grows with larger drive capacities. Efforts have been made to address the discrepancy through the use of new prefixes to denote binary multiples. For example, the counterpart of a decimal kilobyte ( $10^3$ ) would be a binary **kibibyte** ( $2^{10}$ ). A decimal megabyte ( $10^4$ ) corresponds to a binary **mebibyte** ( $2^{20}$ ), a gigabyte ( $10^5$ ) to a **gibibyte** ( $2^{30}$ ), a terabyte ( $10^6$ ) to a **tebibyte** ( $2^{40}$ ) and a petabyte ( $10^7$ ) to a **pebibyte** ( $2^{50}$ ). But if you've never heard of a mebibyte or gibibyte, you can see just how well these initiatives have caught on.

## 6.4 EXAMPLE FILE SYSTEMS

In the following sections we will discuss several example file systems, ranging from quite simple to highly sophisticated. Since modern UNIX file systems and Windows 2000's native file system are covered in the chapter on UNIX (Chap. 10) and the chapter on Windows 2000 (Chap. 11) we will not cover those systems here. We will, however, examine their predecessors below.

### 6.4.1 CD-ROM File Systems

As our first example of a file system, let us consider the file systems used on CD-ROMs. These systems are particularly simple because they were designed for write-once media. Among other things, for example, they have no provision for keeping track of free blocks because on a CD-ROM files cannot be freed or added after the disk has been manufactured. Below we will take a look at the main CD-ROM file system type and two extensions to it.

#### The ISO 9660 File System

The most common standard for CD-ROM file systems was adopted as an International Standard in 1988 under the name **ISO 9660**. Virtually every CD-ROM currently on the market is compatible with this standard, sometimes with the extensions to be discussed below. One of the goals of this standard was to make every CD-ROM readable on every computer, independent of the byte ordering used and independent of the operating system used. As a consequence, some limitations were placed on the file system to make it possible for the weakest operating systems then in use (such as MS-DOS) to read it.

CD-ROMs do not have concentric cylinders the way magnetic disks do. Instead there is a single continuous spiral containing the bits in a linear sequence (although seeks across the spiral are possible). The bits along the spiral are divided into logical blocks (also called logical sectors) of 2352 bytes. Some of these are for preambles, error correction, and other overhead. The payload portion of each logical block is 2048 bytes. When used for music, CDs have leadins, leadouts, and intertrack gaps, but these are not used for data CD-ROMs. Often the position of a block along the spiral is quoted in minutes and seconds. It can be converted to a linear block number using the conversion factor of 1 sec = 75 blocks.

ISO 9660 supports CD-ROM sets with as many as  $2^{16} - 1$  CDs in the set. The individual CD-ROMs may also be partitioned into logical volumes (partitions). However, below we will concentrate on ISO 9660 for a single unpartitioned CD-ROM.

Every CD-ROM begins with 16 blocks whose function is not defined by the ISO 9660 standard. A CD-ROM manufacturer could use this area for providing a

bootstrap program to allow the computer to be booted from the CD-ROM, or for some other purpose. Next comes one block containing the **primary volume descriptor**, which contains some general information about the CD-ROM. Among this information are the system identifier (32 bytes), volume identifier (32 bytes), publisher identifier (128 bytes), and data preparer identifier (128 bytes). The manufacturer can fill in these fields in any desired way, except that only upper case letters, digits, and a very small number of punctuation marks may be used to ensure cross-platform compatibility.

The primary volume descriptor also contains the names of three files, which may contain the abstract, copyright notice, and bibliographic information, respectively. In addition, certain key numbers are also present, including the logical block size (normally 2048, but 4096, 8192, and larger powers of two are allowed in certain cases), the number of blocks on the CD-ROM, and the creation and expiration dates of the CD-ROM. Finally, the primary volume descriptor also contains a directory entry for the root directory, telling where to find it on the CD-ROM (i.e., which block it starts at). From this directory, the rest of the file system can be located.

In addition to the primary volume descriptor, a CD-ROM may contain a supplementary volume descriptor. It contains similar information to the primary, but that will not concern us here.

The root directory, and all other directories for that matter, consists of a variable number of entries, the last of which contains a bit marking it as the final one. The directory entries themselves are also variable length. Each directory entry consists of 10 to 12 fields, some of which are in ASCII and others of which are numerical fields in binary. The binary fields are encoded twice, once in little-endian format (used on example). Thus a 16-bit number uses 4 bytes and a 32-bit number uses 8 bytes. The use of this redundant coding was necessary to avoid hurting anyone's feelings when the standard was developed. If the standard had dictated little endian, then people from companies with big-endian products would have felt like second-class citizens and would not have accepted the standard. The emotional content of a CD-ROM can thus be quantified and measured exactly in kilobytes/hour of wasted space.

The format of an ISO 9660 directory entry is illustrated in Fig. 6-1. Since directory entries have variable lengths, the first field is a byte telling how long the entry is. This byte is defined to have the high-order bit on the left to avoid any ambiguity.

Directory entries may optionally have an extended attributes. If this feature is used for a directory entry, the second byte tells how long the extended attributes are.

Next comes the starting block of the file itself. Files are stored as contiguous runs of blocks, so a file's location is completely specified by the starting block and the size, which is contained in the next field.

The date and time that the CD-ROM was recorded is stored in the next field,

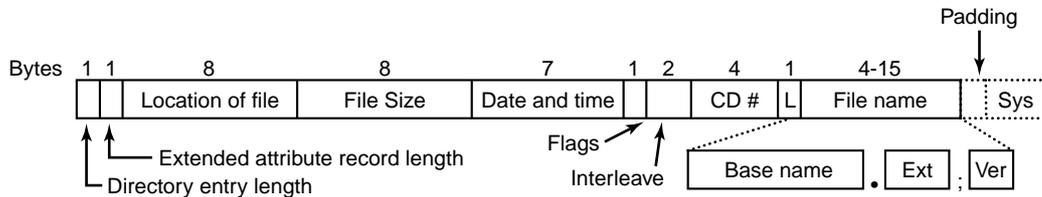


Figure 6-1. The ISO 9660 directory entry.

with separate bytes for the year, month, day, hour, minute, second, and time zone. Years begin to count at 1900, which means that CD-ROMs will suffer from a Y2156 problem because the year following 2155 will be 1900. This problem could have been delayed by defining the origin of time to be 1988 (the year the standard was adopted). Had that been done, the problem would have been postponed until 2244. Every 88 extra years helps.

The *Flags* field contains a few miscellaneous bits, including one to hide the entry in listings (a feature copied from MS-DOS), one to distinguish an entry that is a file from an entry that is a directory, one to enable the use of the extended attributes, and one to mark the last entry in a directory. A few other bits are also present in this field but they will not concern us here. The next field deals with interleaving pieces of files in a way that is not used in the simplest version of ISO 9660, so we will not consider it further.

The next field tells which CD-ROM the file is located on. It is permitted that a directory entry on one CD-ROM refers to a file located on another CD-ROM in the set. In this way it is possible to build a master directory on the first CD-ROM that lists all the files on all the CD-ROMs in the complete set.

The field marked *L* in Fig. 6-1 gives the size of the file name in bytes. It is followed by the file name itself. A file name consists of a base name, a dot, an extension, a semicolon, and a binary version number (1 or 2 bytes). The base name and extension may use upper case letters, the digits 0-9, and the underscore character. All other characters are forbidden to make sure that every computer can handle every file name. The base name can be up to eight characters; the extension can be up to three characters. These choices were dictated by the need to be MS-DOS compatible. A file name may be present in a directory multiple times, as long as each one has a different version number.

The last two fields are not always present. The *Padding* field is used to force every directory entry to be an even number of bytes, to align the numeric fields of subsequent entries on 2-byte boundaries. If padding is needed, a 0 byte is used. Finally, we have the *System use* field. Its function and size are undefined, except that it must be an even number of bytes. Different systems use it in different ways. The Macintosh keeps Finder flags here, for example.

Entries within a directory are listed in alphabetical order except for the first

two entries. The first entry is for the directory itself. The second one is for its parent. In this respect, these entries are similar to the UNIX `.` and `..` directory entries. The files themselves need not be in directory order.

There is no explicit limit to the number of entries in a directory. However, there is a limit to the depth of nesting. The maximum depth of directory nesting is eight.

ISO 9660 defines what are called three levels. Level 1 is the most restrictive and specifies that file names are limited to 8 + 3 characters as we have described, and also requires all files to be contiguous as we have described. Furthermore, it specifies that directory names be limited to eight characters with no extensions. Use of this level maximizes the chances that a CD-ROM can be read on every computer.

Level 2 relaxes the length restriction. It allows files and directories to have names of up to 31 characters, but still from the same set of characters.

Level 3 uses the same name limits as level 2, but partially relaxes the assumption that files have to be contiguous. With this level, a file may consist of several sections, each of which is a contiguous run of blocks. The same run may occur multiple times in a file and may also occur in two or more files. If large chunks of data are repeated in several files, level 3 provides some space optimization by not requiring the data to be present multiple times.

### Rock Ridge Extensions

As we have seen, ISO 9660 is highly restrictive in several ways. Shortly after it came out, people in the UNIX community began working on an extension to make it possible to represent UNIX file systems on a CD-ROM. These extensions were named Rock Ridge, after a town in the Gene Wilder movie *Blazing Saddles*, probably because one of the committee members liked the film.

The extensions use the *System use* field in order to make Rock Ridge CD-ROMs readable on any computer. All the other fields retain their normal ISO 9660 meaning. Any system not aware of the Rock Ridge extensions just ignores them and sees a normal CD-ROM.

The extensions are divided up into the following fields:

1. PX - POSIX attributes.
2. PN - Major and minor device numbers.
3. SL - Symbolic link.
4. NM - Alternative name.
5. CL - Child location.
6. PL - Parent location.
7. RE - Relocation.

#### 8. TF - Time stamps.

The *PX* field contains the standard UNIX *rwrxrwxrwx* permission bits for the owner, group, and others. It also contains the other bits contained in the mode word, such as the SETUID and SETGID bits, and so on.

To allow raw devices to be represented on a CD-ROM, the *PN* field is present. It contains the major and minor device numbers associated with the file. In this way, the contents of the */dev* directory can be written to a CD-ROM and later reconstructed correctly on the target system.

The *SL* field is for symbolic links. It allows a file on one file system to refer to a file on a different file system.

Probably the most important field is *NM*. It allows a second name to be associated with the file. This name is not subject to the character set or length restrictions of ISO 9660, making it possible to express arbitrary UNIX file names on a CD-ROM.

The next three fields are used together to get around the ISO 9660 limit of directories that may only be nested eight deep. Using them it is possible to specify that a directory is to be relocated, and to tell where it goes in the hierarchy. It is effectively a way to work around the artificial depth limit.

Finally, the *TF* field contains the three timestamps included in each UNIX i-node, namely the time the file was created, the time it was last modified, and the time it was last accessed. Together, these extensions make it possible to copy a UNIX file system to a CD-ROM and then restore it correctly to a different system.

### Joliet Extensions

The UNIX community was not the only group that wanted a way to extend ISO 9660. Microsoft also found it too restrictive (although it was Microsoft's own MS-DOS that caused most of the restrictions in the first place). Therefore Microsoft invented some extensions that were called **Joliet**. They were designed to allow Windows file systems to be copied to CD-ROM and then restored, in precisely the same way that Rock Ridge was designed for UNIX. Virtually all programs that run under Windows and use CD-ROMs support Joliet, including programs that burn CD-recordables. Usually, these programs offer a choice between the various ISO 9660 levels and Joliet.

The major extensions provided by Joliet are

1. Long file names.
2. Unicode character set.
3. Directory nesting deeper than eight levels.
4. Directory names with extensions

The first extension allows file names up to 64 characters. The second extension enables the use of the Unicode character set for file names. This extension is important for software intended for use in countries that do not use the Latin alphabet, such as Japan, Israel, and Greece. Since Unicode characters are two bytes, the maximum file name in Joliet occupies 128 bytes.

Like Rock Ridge, the limitation on directory nesting is removed by Joliet. Directories can be nested as deeply as needed. Finally, directory names can have extensions. It is not clear why this extension was included, since Windows directories virtually never use extensions, but maybe some day they will.

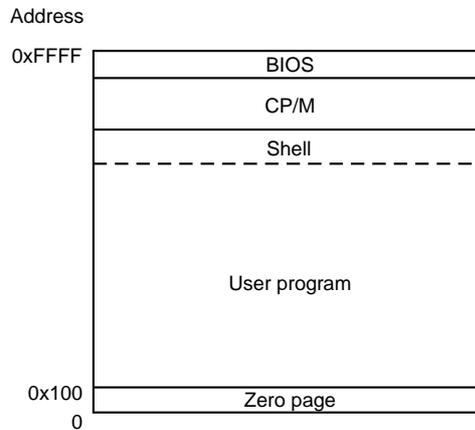
### 6.4.2 The CP/M File System

The first personal computers (then called microcomputers) came out in the early 1980s. A popular early type used the 8-bit Intel 8080 CPU and had 4 KB of RAM and a single 8-inch floppy disk with a capacity of 180 KB. Later versions used the slightly fancier (but still 8-bit) Zilog Z80 CPU, had up to 64 KB of RAM, and had a whopping 720-KB floppy disk as the mass storage device. Despite the slow speed and small amount of RAM, nearly all of these machines ran a surprisingly powerful disk-based operating system, called **CP/M (Control Program for Microcomputers)** (Golden and Pechura, 1986). This system dominated its era as much as MS-DOS and later Windows dominated the IBM PC world. Two decades later, it has vanished without a trace (except for a small group of diehard fans), which gives reason to think that systems that now dominate the world may be essentially unknown when current babies become college students (Windows what?).

It is worth taking a look at CP/M for several reasons. First, it was a historically very important system and was the direct ancestor of MS-DOS. Second, current and future operating system designers who think that a computer needs 32 MB just to boot the operating system could probably learn a lot about simplicity from a system that ran quite well in 16 KB of RAM. Third, in the coming decades, embedded systems are going to be extremely widespread. Due to cost, space, weight, and power constraints, the operating systems used, for example, in watches, cameras, radios, and cellular telephones, are of necessity going to be lean and mean, not unlike CP/M. Of course, these systems do not have 8-inch floppy disks, but they may well have electronic disks using flash memory, and building a CP/M-like file system on such a device is straightforward.

The layout of CP/M in memory is shown in Fig. 6-2. At the top of main memory (in RAM) is the BIOS, which contains a basic library of 17 I/O calls used by CP/M (in this section we will describe CP/M 2.2, which was the standard version when CP/M was at the height of its popularity). These calls read and write the keyboard, screen, and floppy disk.

Just below the BIOS is the operating system proper. The size of the operating system in CP/M 2.2 is 3584 bytes. Amazing but true: a complete operating system



**Figure 6-2.** Memory layout of CP/M.

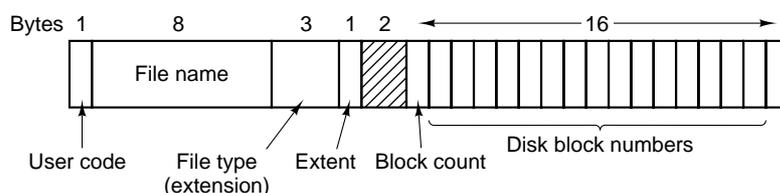
in under 4 KB. Below the operating system comes the shell (command line processor), which chews up another 2 KB. The rest of memory is for user programs, except for the bottom 256 bytes, which are reserved for the hardware interrupt vectors, a few variables, and a buffer for the current command line so user programs can get at it.

The reason for splitting the BIOS from CP/M itself (even though both are in RAM) is portability. CP/M interacts with the hardware only by making BIOS calls. To port CP/M to a new machine, all that is necessary is to port the BIOS there. Once that has been done, CP/M itself can be installed without modification.

A CP/M system has only one directory, which contains fixed-size (32-byte) entries. The directory size, although fixed for a given implementation, may be different in other implementations of CP/M. All files in the system are listed in this directory. After CP/M boots, it reads in the directory and computes a bitmap containing the free disk blocks by seeing which blocks are not in any file. This bitmap, which is only 23 bytes for a 180-KB disk, is kept in memory during execution. At system shutdown time it is discarded, that is, not written back to the disk. This approach eliminates the need for a disk consistency checker (like *fsck*) and saves 1 block on the disk (percentually equivalent to saving 90 MB on a modern 16-GB disk).

When the user types a command, the shell first copies the command to a buffer in the bottom 256 bytes of memory. Then it looks up the program to be executed and reads it into memory at address 256 (above the interrupt vectors), and jumps to it. The program then begins running. It discovers its arguments by looking in the command line buffer. The program is allowed to overwrite the shell if it needs the memory. When the program finishes, it makes a system call to CP/M telling it to reload the shell (if it was overwritten) and execute it. In a nutshell, that is pretty much the whole CP/M story.

In addition to loading programs, CP/M provides 38 system calls, mostly file services, for user programs. The most important of these are reading and writing files. Before a file can be read, it must be opened. When CP/M gets an open system call, it has to read in and search the one and only directory. The directory is not kept in memory all the time to save precious RAM. When CP/M finds the entry, it immediately has the disk block numbers, since they are stored right in the directory entry, as are all the attributes. The format of a directory entry is given in Fig. 6-3.



**Figure 6-3.** The CP/M directory entry format.

The fields in Fig. 6-3 have the following meanings. The *User code* field keeps track of which user owns the file. Although only one person can be logged into a CP/M at any given moment, the system supports multiple users who take turns using the system. While searching for a file name, only those entries belonging to the currently logged-in user are checked. In effect, each user has a virtual directory without the overhead of managing multiple directories.

The next two fields give the name and extension of the file. The base name is up to eight characters; an optional extension of up to three characters may be present. Only upper case letters, digits, and a small number of special characters are allowed in file names. This 8 + 3 naming using upper case only was later taken over by MS-DOS.

The *Block count* field tells how many bytes this file entry contains, measured in units of 128 bytes (because I/O is actually done in 128-byte physical sectors). The last 1-KB block may not be full, so the system has no way to determine the exact size of a file. It is up to the user to put in some END-OF-FILE marker if desired. The final 16 fields contain the disk block numbers themselves. Each block is 1 KB, so that maximum file size is 16 KB. Note that physical I/O is done in units of 128-byte sectors and sizes are kept track of in sectors, but file blocks are allocated in units of 1 KB (8 sectors at a time) to avoid making the directory entry too large.

However, the CP/M designer realized that some files, even on a 180-KB floppy disk, might exceed 16 KB, so an escape hatch was built around the 16-KB limit. A file that is between 16 KB and 32 KB uses not one directory entry, but two. The first entry holds the first 16 blocks; the second entry holds the next 16 blocks. Beyond 32 KB, a third directory entry is used, and so on. The *Extent* field keeps track of the order of the directory entries so the system knows which

16 KB comes first, which comes second, and so on.

After an open call, the addresses of all the disk blocks are known, making read straightforward. The write call is also simple. It just requires allocating a free block from the bitmap in memory and then writing the block. Consecutive blocks on a file are not placed in consecutive blocks on the disk because the 8080 cannot process an interrupt and start reading the next block on time. Instead, interleaving is used to allow several blocks to be read on a single rotation.

CP/M is clearly not the last word in advanced file systems, but it is simple, fast, and can be implemented by a competent programmer in less than a week. For many embedded applications, it may be all that is needed.

### 6.4.3 The MS-DOS File System

To a first approximation, MS-DOS is a bigger and better version of CP/M. It runs only on Intel platforms, does not support multiprogramming, and runs only in the PC's real mode (which was originally the only mode). The shell has more features and there are more system calls, but the basic function of the operating system is still loading programs, handling the keyboard and screen, and managing the file system. It is the latter functionality that interests us here.

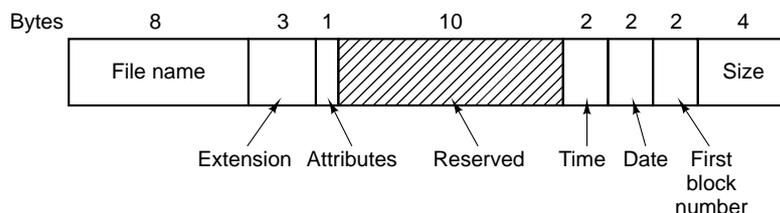
The MS-DOS file system was patterned closely on the CP/M file system, including the use of 8 + 3 (upper case) character file names. The first version (MS-DOS 1.0) was even limited to a single directory, like CP/M. However, starting with MS-DOS 2.0, the file system functionality was greatly expanded. The biggest improvement was the inclusion of a hierarchical file system in which directories could be nested to an arbitrary depth. This meant that the root directory (which still had a fixed maximum size) could contain subdirectories, and these could contain further subdirectories ad infinitum. Links in the style of UNIX were not permitted, so the file system formed a tree starting at the root directory.

It is common for different application programs to start out by creating a subdirectory in the root directory and put all their files there (or in subdirectories thereof), so that different applications do not conflict. Since directories are themselves just stored as files, there are no limits on the number of directories or files that may be created. Unlike CP/M, however, there is no concept of different users in MS-DOS. Consequently, the logged in user has access to all files.

To read a file, an MS-DOS program must first make an open system call to get a handle for it. The open system call specifies a path, which may be either absolute or relative to the current working directory. The path is looked up component by component until the final directory is located and read into memory. It is then searched for the file to be opened.

Although MS-DOS directories are variable sized, like CP/M, they use a fixed-size 32-byte directory entry. The format of an MS-DOS directory entry is shown in Fig. 6-4. It contains the file name, attributes, creation date and time, starting block, and exact file size. File names shorter than 8 + 3 characters are left

justified and padded with spaces on the right, in each field separately. The *Attributes* field is new and contains bits to indicate that a file is read-only, needs to be archived, is hidden, or is a system file. Read-only files cannot be written. This is to protect them from accidental damage. The archived bit has no actual operating system function (i.e., MS-DOS does not examine or set it). The intention is to allow user-level archive programs to clear it upon archiving a file and to have other programs set it when modifying a file. In this way, a backup program can just examine this attribute bit on every file to see which files to back up. The hidden bit can be set to prevent a file from appearing in directory listings. Its main use is to avoid confusing novice users with files they might not understand. Finally, the system bit also hides files. In addition, system files cannot accidentally be deleted using the *del* command. The main components of MS-DOS have this bit set.



**Figure 6-4.** The MS-DOS directory entry.

The directory entry also contains the date and time the file was created or last modified. The time is accurate only to  $\pm 2$  sec because it is stored in a 2-byte field, which can store only 65,536 unique values (a day contains 86,400 unique seconds). The time field is subdivided into seconds (5 bits), minutes (6 bits), and hours (5 bits). The date counts in days using three subfields: day (5 bits), month (4 bits), and year–1980 (7 bits). With a 7-bit number for the year and time beginning in 1980, the highest expressible year is 2107. Thus MS-DOS has a built-in Y2108 problem. To avoid catastrophe, MS-DOS users should begin with Y2108 compliance as early as possible. If MS-DOS had used the combined date and time fields as a 32-bit seconds counter, it could have represented every second exactly and delayed the catastrophe until 2116.

Unlike CP/M, which does not store the exact file size, MS-DOS does. Since a 32-bit number is used for the file size, in theory files can be as large as 4 GB. However, other limits (described below) restrict the maximum file size to 2 GB or less. A surprising large part of the entry (10 bytes) is unused.

Another way in which MS-DOS differs from CP/M is that it does not store a file's disk addresses in its directory entry, probably because the designers realized that large hard disks (by then common on minicomputers) would some day reach the MS-DOS world. Instead, MS-DOS keeps track of file blocks via a file allocation table in main memory. The directory entry contains the number of the first

file block. This number is used as an index into a 64K entry FAT in main memory. By following the chain, all the blocks can be found. The operation of the FAT is illustrated in Fig. 6-0.

The FAT file system comes in three versions for MS-DOS: FAT-12, FAT-16, and FAT-32, depending on how many bits a disk address contains. Actually, FAT-32 is something of a misnomer since only the low-order 28 bits of the disk addresses are used. It should have been called FAT-28, but powers of two sound so much neater.

For all FATs, the disk block can be set to some multiple of 512 bytes (possibly different for each partition), with the set of allowed block sizes (called **cluster sizes** by Microsoft) being different for each variant. The first version of MS-DOS used FAT-12 with 512-byte blocks, giving a maximum partition size of  $2^{12} \times 512$  bytes (actually only  $4086 \times 512$  bytes because 10 of the disk addresses were used as special markers, such as end of file, bad block, etc. With these parameters, the maximum disk partition size was about 2 MB and the size of the FAT table in memory was 4096 entries of 2 bytes each. Using a 12-bit table entry would have been too slow.

This system worked well for floppy disks, but when hard disks came out, it became a problem. Microsoft solved the problem by allowing additional block sizes of 1 KB, 2 KB, and 4 KB. This change preserved the structure and size of the FAT-12 table, but allowed disk partitions of up to 16 MB.

Since MS-DOS supported four disk partitions per disk drive, the new FAT-12 file system worked up to 64-MB disks. Beyond that, something had to give. What happened was the introduction of FAT-16, with 16-bit disk pointers. Additionally, block sizes of 8 KB, 16 KB, and 32 KB were permitted. (32,768 is the largest power of two that can be represented in 16 bits.) The FAT-16 table now occupied 128 KB of main memory all the time, but with the larger memories by then available, it was widely used and rapidly replaced the FAT-12 file system. The largest disk partition that can be supported by FAT-16 is 2 GB (64K entries of 32 KB each) and the largest disk 8 GB, namely four partitions of 2 GB each.

For business letters, this limit is not a problem, but for storing digital video using the DV standard, a 2-GB file holds just over 9 minutes of video. As a consequence of the fact that a PC disk can support only four partitions, the largest video that can be stored on a disk is about 38 minutes, no matter how large the disk is. This limit also means that the largest video that can be edited on line is less than 19 minutes, since both input and output files are needed.

Starting with the second release of Windows 95, the FAT-32 file system, with its 28-bit disk addresses, was introduced and the version of MS-DOS underlying Windows 95 was adapted to support FAT-32. In this system, partitions could theoretically be  $2^{28} \times 2^{15}$  bytes, but they are actually limited to 2 TB (2048 GB) because internally the system keeps track of partition sizes in 512-byte sectors using a 32-bit number and  $2^9 \times 2^{32}$  is 2 TB. The maximum partition size for various block sizes and all three FAT types is shown in Fig. 6-5.

Block size	FAT-12	FAT-16	FAT-32
0.5 KB	2 MB		
1 KB	4 MB		
2 KB	8 MB	128 MB	
4 KB	16 MB	256 MB	1 TB
8 KB		512 MB	2 TB
16 KB		1024 MB	2 TB
32 KB		2048 MB	2 TB

**Figure 6-5.** Maximum partition size for different block sizes. The empty boxes represent forbidden combinations.

In addition to supporting larger disks, the FAT-32 file system has two other advantages over FAT-16. First, an 8-GB disk using FAT-32 can be a single partition. Using FAT-16 it has to be four partitions, which appears to the Windows user as the C:, D:, E:, and F: logical disk drives. It is up to the user to decide which file to place on which drive and keep track of what is where.

The other advantage of FAT-32 over FAT-16 is that for a given size disk partition, a smaller block size can be used. For example, for a 2-GB disk partition, FAT-16 must use 32-KB blocks, otherwise with only 64K available disk addresses, it cannot cover the whole partition. In contrast, FAT-32 can use, for example, 4-KB blocks for a 2-GB disk partition. The advantage of the smaller block size is that most files are much shorter than 32 KB. If the block size is 32 KB, a file of 10 bytes ties up 32 KB of disk space. If the average file is, say, 8 KB, then with a 32-KB block,  $\frac{3}{4}$  of the disk will be wasted, not a terribly efficient way to use the disk. With an 8-KB file and a 4-KB block, there is no disk wastage, but the price paid is more RAM eaten up by the FAT. With a 4-KB block and a 2-GB disk partition, there are 512K blocks, so the FAT must have 512K entries in memory (occupying 2 MB of RAM).

MS-DOS uses the FAT to keep track of free disk blocks. Any block that is not currently allocated is marked with a special code. When MS-DOS needs a new disk block, it searches the FAT for an entry containing this code. Thus no bitmap or free list is required.

#### 6.4.4 The Windows 98 File System

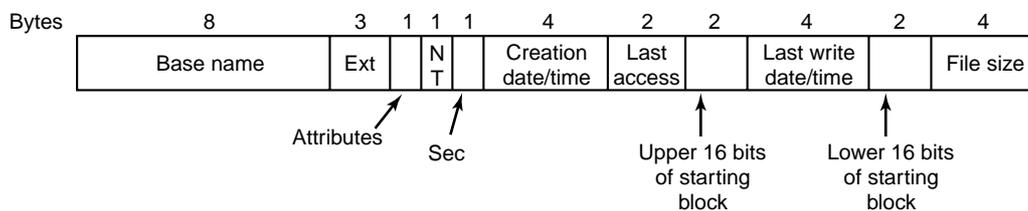
The original release of Windows 95 used the MS-DOS file system, including the use of 8 + 3 character file names and the FAT-12 and FAT-16 file systems. Starting with the second release of Windows 95, file names longer than 8 + 3 characters were permitted. In addition, FAT-32 was introduced, primarily to allow larger disk partitions larger than 2 GB and disks larger than 8 GB, which were then available. Both the long file names and FAT-32 were used in Windows

98 in the same form as in the second release of Windows 95. Below we will describe these features of the Windows 98 file system, which have been carried forward into Windows Me as well.

Since long file names are more exciting for users than the FAT structure, let us look at them first. One way to introduce long file names would have been to just invent a new directory structure. The problem with this approach is that if Microsoft had done this, people who were still in the process of converting from Windows 3 to Windows 95 or Windows 98 could not have accessed their files from both systems. A political decision was made within Microsoft that files created using Windows 98 must be accessible from Windows 3 as well (for dual-boot machines). This constraint forced a design for handling long file names that was backward compatible with the old MS-DOS 8 + 3 naming system. Since such backward compatibility constraints are not unusual in the computer industry, it is worth looking in detail at how Microsoft accomplished this goal.

The effect of this decision to be backward compatible meant that the Windows 98 directory structure had to be compatible with the MS-DOS directory structure. As we saw, this structure is just a list of 32-byte entries as shown in Fig. 6-4. This format came directly from CP/M (which was written for the 8080), which goes to show how long (obsolete) structures can live in the computer world.

However, it was possible to now allocate the 10 unused bytes in the entries of Fig. 6-4, and that was done, as shown in Fig. 6-6. This change has nothing to do with long names, but it is used in Windows 98, so it is worth understanding.



**Figure 6-6.** The extended MS-DOS directory entry used in Windows 98.

The changes consist of the addition of five new fields where the 10 unused bytes used to be. The *NT* field is mostly there for some compatibility with Windows NT in terms of displaying file names in the correct case (in MS-DOS, all file names are upper case). The *Sec* field solves the problem that it is not possible to store the time of day in a 16-bit field. It provides additional bits so that the new *Creation time* field is accurate to 10 msec. Another new field is *Last access*, which stores the date (but not time) of the last access to the file. Finally, going to the FAT-32 file system means that block numbers are now 32 bits, so an additional 16-bit field is needed to store the upper 16 bits of the starting block number.

Now we come to the heart of the Windows 98 file system: how long file names are represented in a way that is backward compatible with MS-DOS. The

solution chosen was to assign two names to each file: a (potentially) long file name (in Unicode, for compatibility with Windows NT), and an 8 + 3 name for compatibility with MS-DOS. Files can be accessed by either name. When a file is created whose name does not obey the MS-DOS naming rules (8 + 3 length, no Unicode, limited character set, no spaces, etc.), Windows 98 invents an MS-DOS name for it according to a certain algorithm. The basic idea is to take the first six characters of the name, convert them to upper case, if need be, and append ~1 to form the base name. If this name already exists, then the suffix ~2 is used, and so on. In addition, spaces and extra periods are deleted and certain special characters are converted to underscores. As an example, a file named *The time has come the walrus said* is assigned the MS-DOS name *THETIM~1*. If a subsequent file is created with the name *The time has come the rabbit said*, it is assigned the MS-DOS name *THETIM~2*, and so on.

Every file has an MS-DOS file name stored using the directory format of Fig. 6-6. If a file also has a long name, that name is stored in one or more directory entries directly preceding the MS-DOS file name. Each long-name entry holds up to 13 (Unicode) characters. The entries are stored in reverse order, with the start of the file name just ahead of the MS-DOS entry and subsequent pieces before it. The format of each long-name entry is given in Fig. 6-7.

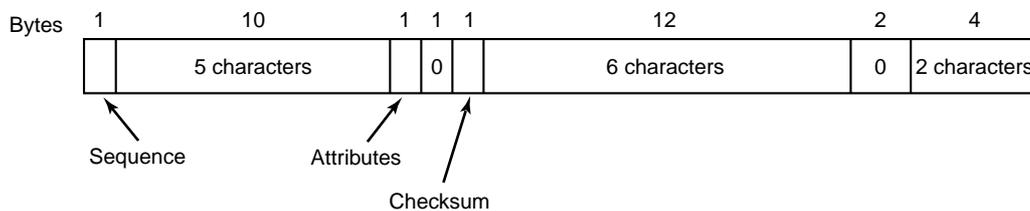


Figure 6-7. An entry for (part of) a long file name in Windows 98.

An obvious question is: “How does Windows 98 know whether a directory entry contains an MS-DOS file name or a (piece of a) long file name?” The answer lies in the *Attributes* field. For a long-name entry, this field has the value 0x0F, which represents an otherwise impossible combination of attributes. Old MS-DOS programs that read directories will just ignore it as invalid. Little do they know. The pieces of the name are sequenced using the first byte of the entry. The last part of the long name (the first entry in the sequence) is marked by adding 64 to the sequence number. Since only 6 bits are used for the sequence number, the theoretical maximum for file names is 63 × 13 or 819 characters. In fact they are limited to 260 characters for historical reasons.

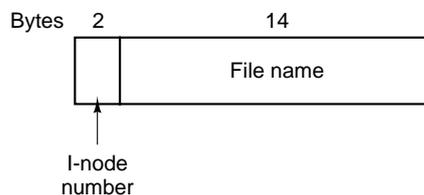
Each long-name entry contains a *Checksum* field to avoid the following problem. First, a Windows 98 program creates a file with a long name. Second, the computer is rebooted to run MS-DOS or Windows 3. Third, an old program there then removes the MS-DOS file name from the directory but does not remove the



versions in Chap. 10.

The file system is in the form of a tree starting at the root directory, with the addition of links, forming a directed acyclic graph. File names are up to 14 characters and can contain any ASCII characters except / (because that is the separator between components in a path) and NUL (because that is used to pad out names shorter than 14 characters). NUL has the numerical value of 0.

A UNIX directory entry contains one entry for each file in that directory. Each entry is extremely simple because UNIX uses the i-node scheme illustrated in Fig. 6-0. A directory entry contains only two fields: the file name (14 bytes) and the number of the i-node for that file (2 bytes), as shown in Fig. 6-9. These parameters limit the number of files per file system to 64K.

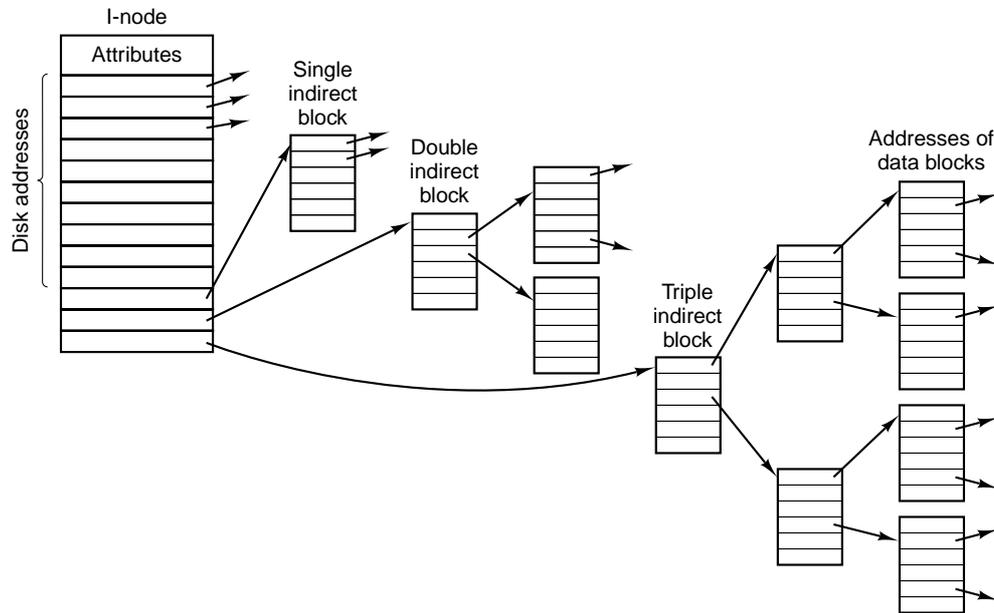


**Figure 6-9.** A UNIX V7 directory entry.

Like the i-node of Fig. 6-0, the UNIX i-nodes contains some attributes. The attributes contain the file size, three times (creation, last access, and last modification), owner, group, protection information, and a count of the number of directory entries that point to the i-node. The latter field is needed due to links. Whenever a new link is made to an i-node, the count in the i-node is increased. When a link is removed, the count is decremented. When it gets to 0, the i-node is reclaimed and the disk blocks are put back in the free list.

Keeping track of disk blocks is done using a generalization of Fig. 6-0 in order to handle very large files. The first 10 disk addresses are stored in the i-node itself, so for small files, all the necessary information is right in the i-node, which is fetched from disk to main memory when the file is opened. For somewhat larger files, one of the addresses in the i-node is the address of a disk block called a **single indirect block**. This block contains additional disk addresses. If this still is not enough, another address in the i-node, called a **double indirect block**, contains the address of a block that contains a list of single indirect blocks. Each of these single indirect blocks points to a few hundred data blocks. If even this is not enough, a **triple indirect block** can also be used. The complete picture is given in Fig. 6-10.

When a file is opened, the file system must take the file name supplied and locate its disk blocks. Let us consider how the path name `/usr/ast/mbox` is looked up. We will use UNIX as an example, but the algorithm is basically the same for all hierarchical directory systems. First the file system locates the root directory. In UNIX its i-node is located at a fixed place on the disk. From this i-node, it

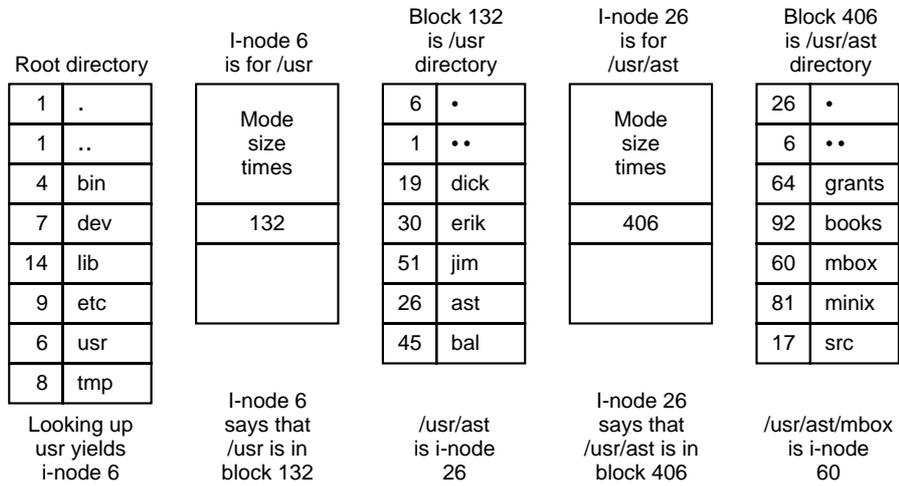


**Figure 6-10.** A UNIX i-node.

locates the root directory which can be anywhere on the disk, but say block 1 in this case.

Then it reads the root directory and looks up the first component of the path, *usr*, in the root directory to find the i-node number of the file */usr*. Locating an i-node from its number is straightforward, since each one has a fixed location on the disk. From this i-node, the system locates the directory for */usr* and looks up the next component, *ast*, in it. When it has found the entry for *ast*, it has the i-node for the directory */usr/ast*. From this i-node it can find the directory itself and look up *mbox*. The i-node for this file is then read into memory and kept there until the file is closed. The lookup process is illustrated in Fig. 6-11.

Relative path names are looked up the same way as absolute ones, only starting from the working directory instead of starting from the root directory. Every directory has entries for *.* and *..* which are put there when the directory is created. The entry *.* has the i-node number for the current directory, and the entry for *..* has the i-node number for the parent directory. Thus, a procedure looking up *../dick/prog.c* simply looks up *..* in the working directory, finds the i-node number for the parent directory, and searches that directory for *dick*. No special mechanism is needed to handle these names. As far as the directory system is concerned, they are just ordinary ASCII strings, just the same as any other names.



**Figure 6-11.** The steps in looking up */usr/ast/mbox*.

# File Systems

A file system is a clearly-defined method that the computer's operating system uses to store, catalog, and retrieve files.

# Module 11: File-System Interface

- File Concept
- Access :Methods
- Directory Structure
- Protection
- Consistency Semantics

# Files & File Systems

- File
  - data, in some format
- File System
  - Set of **named** files, maybe organized (directories)
  - Information on files (metadata)

# Files & File Systems



alan



ambients



assigning\_types\_to\_processes.ps



asynchrony96.ps.gz



attelm.ps



beos.rapport.ps



BibDeskDocs



bigraphs



bis-proof.ps.gz



BRICS-NS-05-3.pdf



caml

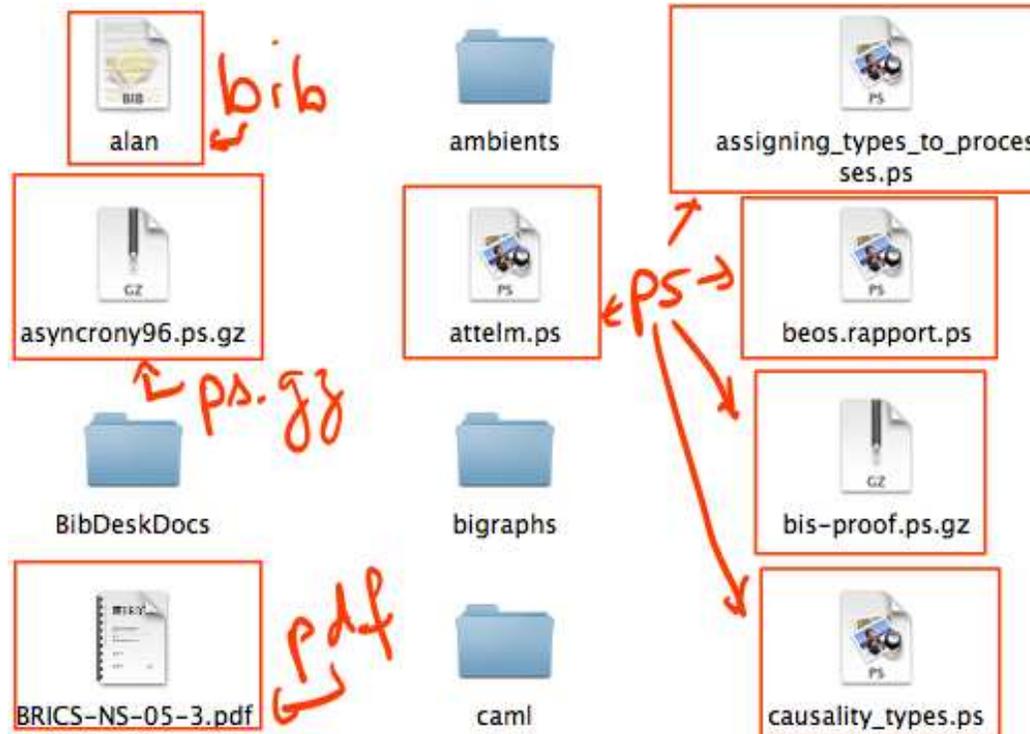


causality\_types.ps

# File Concept

- Contiguous logical address space
- Types:
  - Data
    - \* numeric
    - \* character
    - \* binary
  - Program

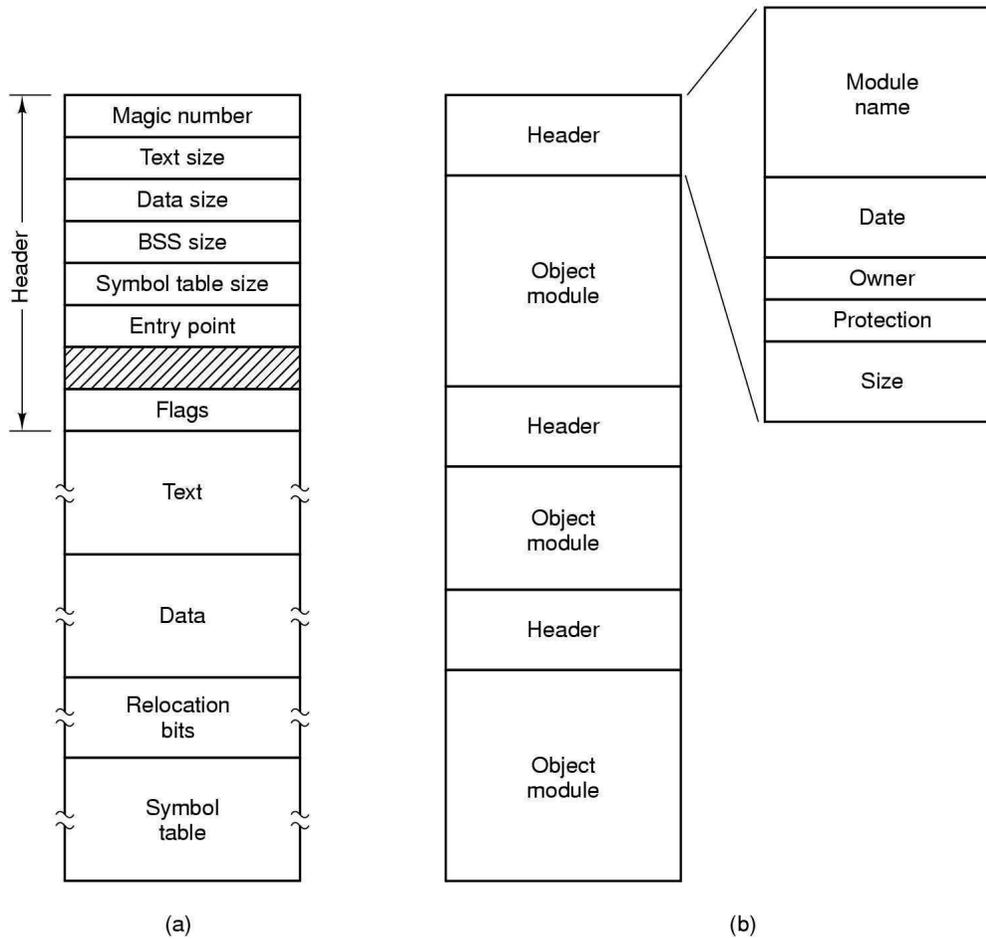
# File Types



## File Types – name, extension

File Type	Usual extension	Function
Executable	exe, com, bin or none	ready-to-run machine-language program
Object	obj, o	compiled, machine language, not linked
Source code	c, p, pas, 177, asm, a	source code in various languages
Batch	bat, sh	commands to the command interpreter
Text	txt, doc	textual data documents
Word processor	wp, tex, rrf, etc.	various word-processor formats
Library	lib, a	libraries of routines
Print or view	ps, dvi, gif, pdf	ASCII or binary file
Archive	arc, zip, tar, gz	related files grouped into one file, sometimes compressed.

# File Types

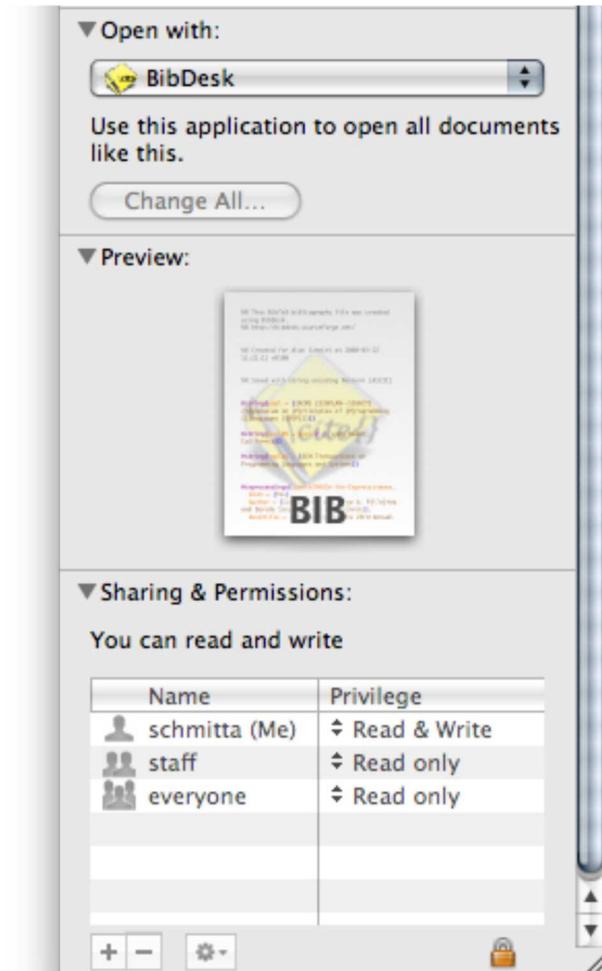
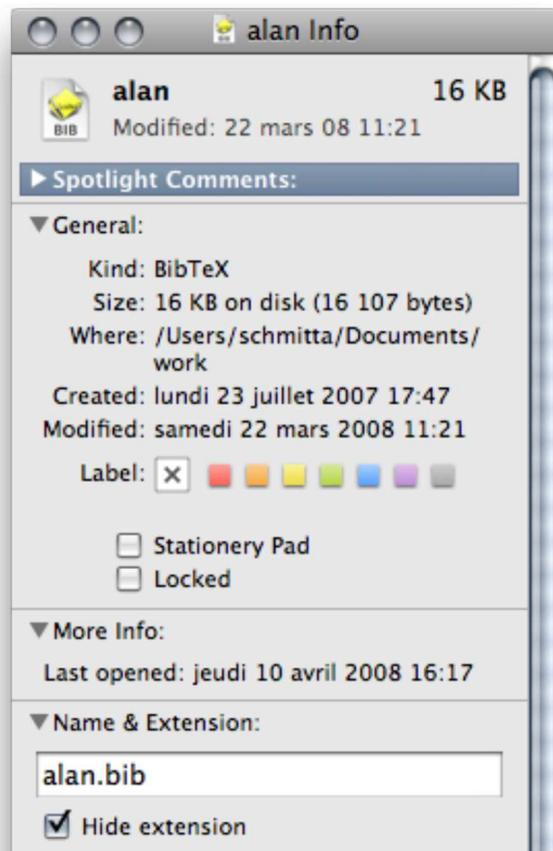


(a) An executable file (b) An archive

# File Structure

- None - sequence of words, bytes
- Simple record structure
  - Lines
  - Fixed length
  - Variable length
- Complex Structures
  - Formatted document
  - Relocatable load file
- Can simulate last two with first method by inserting appropriate control characters.
- Who decides:
  - Operating system
  - Program

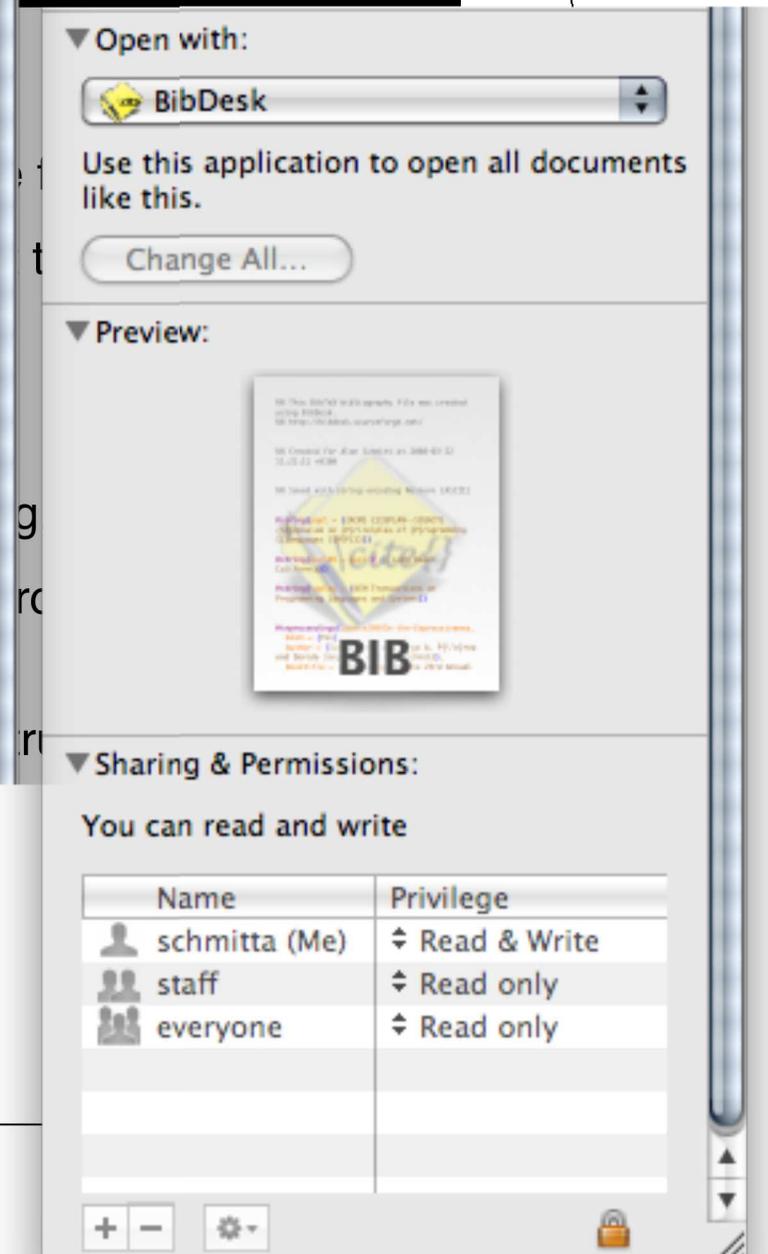
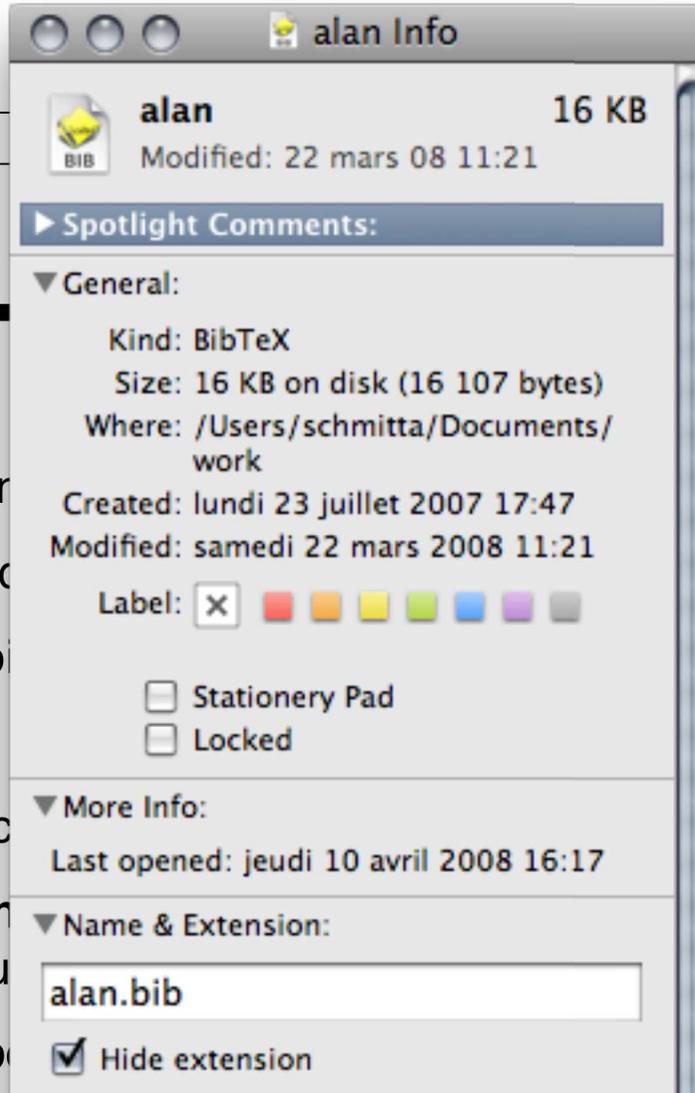
# File Metadata



# File Attributes

- **Name** – only information kept in human-readable form.
- **Type** – needed for systems that support different types.
- **Location** – pointer to file location on device.
- **Size** – current file size.
- **Protection** – controls who can do reading, writing, executing.
- **Time, date, and user identification** – data for protection, security, and usage monitoring.
- Information about files are kept in the directory structure, which is maintained on the disk.

- **Name** – only in
- **Type** – needed
- **Location** – po
- **Size** – current
- **Protection** – c
- **Time, date, an**  
security, and u
- Information ab  
maintained on the disk.



# File Operations

- create
- write
- read
- reposition within file – file seek
- delete
- truncate
- $\text{open}(F_i)$  – search the directory structure on disk for entry  $F_i$ , and move the content of entry to memory.
- $\text{close}(F_i)$  – move the content of entry  $F_i$  in memory to directory structure on disk.

# Access Methods

- Sequential Access

*read next*

*write next*

*reset*

*no read after last write  
(rewrite)*

- Direct Access

*read n*

*write n*

*position to n*

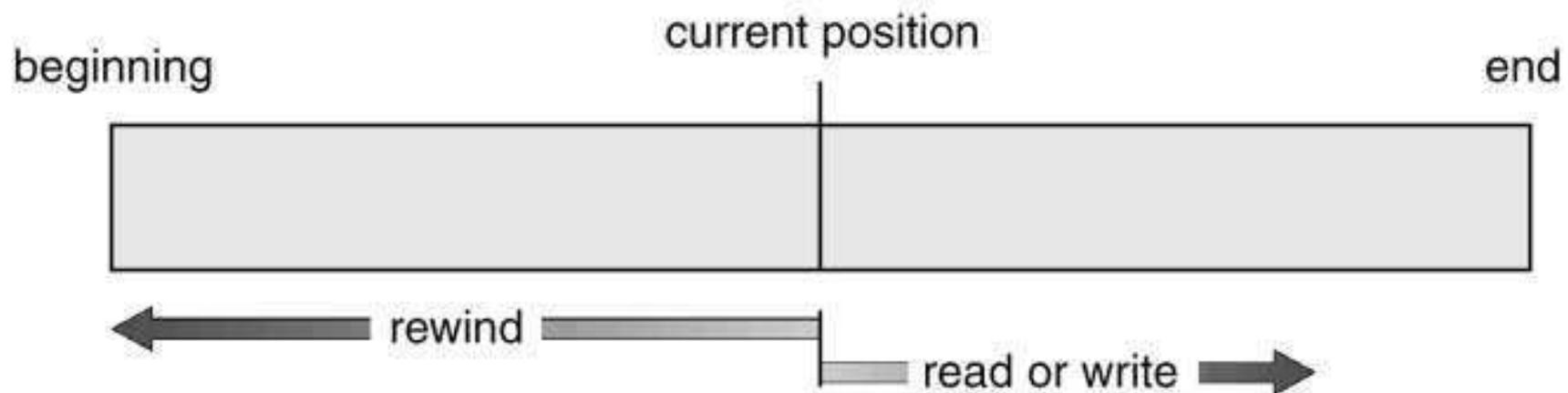
*read next*

*write next*

*rewrite n*

*n* = relative block number

# Sequential-access File



# Directories



alan



ambients



assigning\_types\_to\_processes.ps



asynchrony96.ps.gz



attelm.ps



beos.rapport.ps



BibDeskDocs



bigraphs



bis-proof.ps.gz



BRICS-NS-05-3.pdf



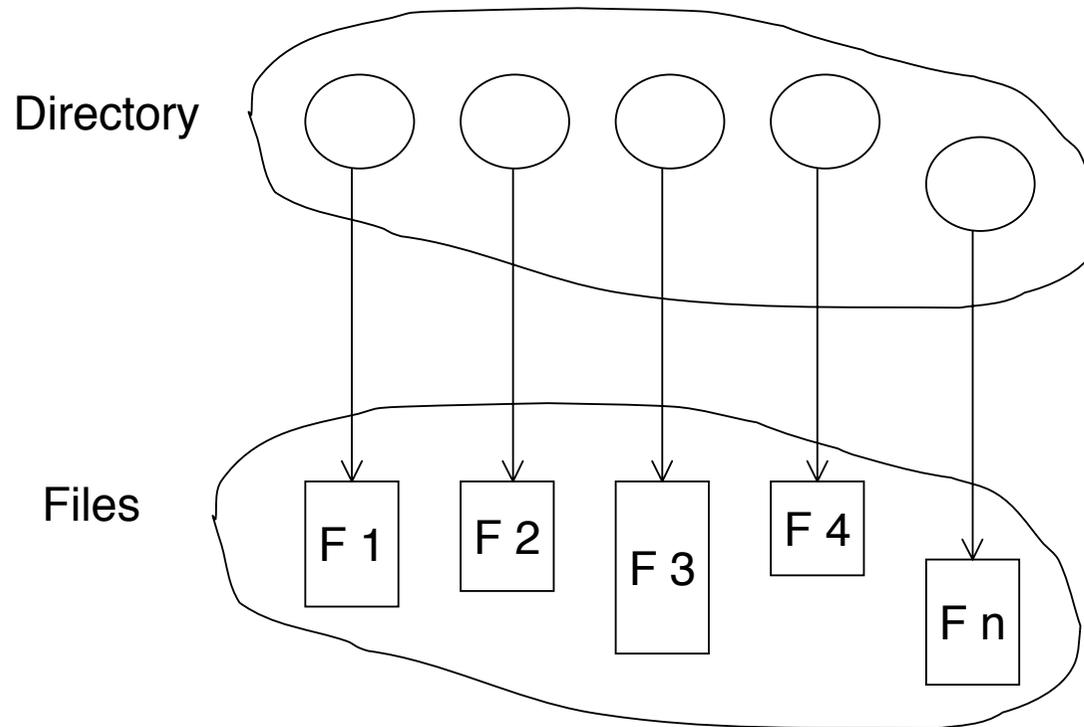
caml



causality\_types.ps

# Directory Structure

- A collection of nodes containing information about all files.



- Both the directory structure and the files reside on disk.

# Information in a Device Directory

- Name
- Type
- Address
- Current length
- Maximum length
- Date last accessed (for archival)
- Date last updated (for dump)
- Owner ID (who pays)
- Protection information (discuss later)

# Operations Performed on Directory

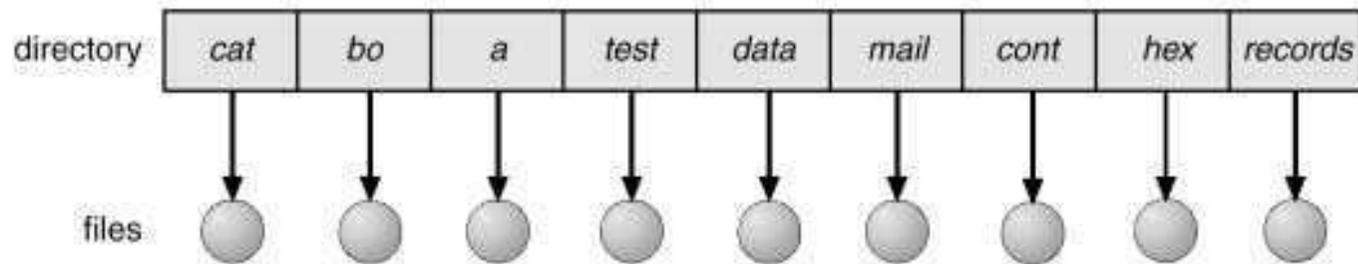
- Search for a file
- Create a file
- Delete a file
- List a directory
- Rename a file
- Traverse the file system

## Organize the Directory (Logically) to Obtain

- Efficiency – locating a file quickly.
- Naming – convenient to users.
  - Two users can have same name for different files.
  - The same file can have several different names.
- Grouping – logical grouping of files by properties, (e.g., all Pascal programs, all games, ...)

# Single-Level Directory

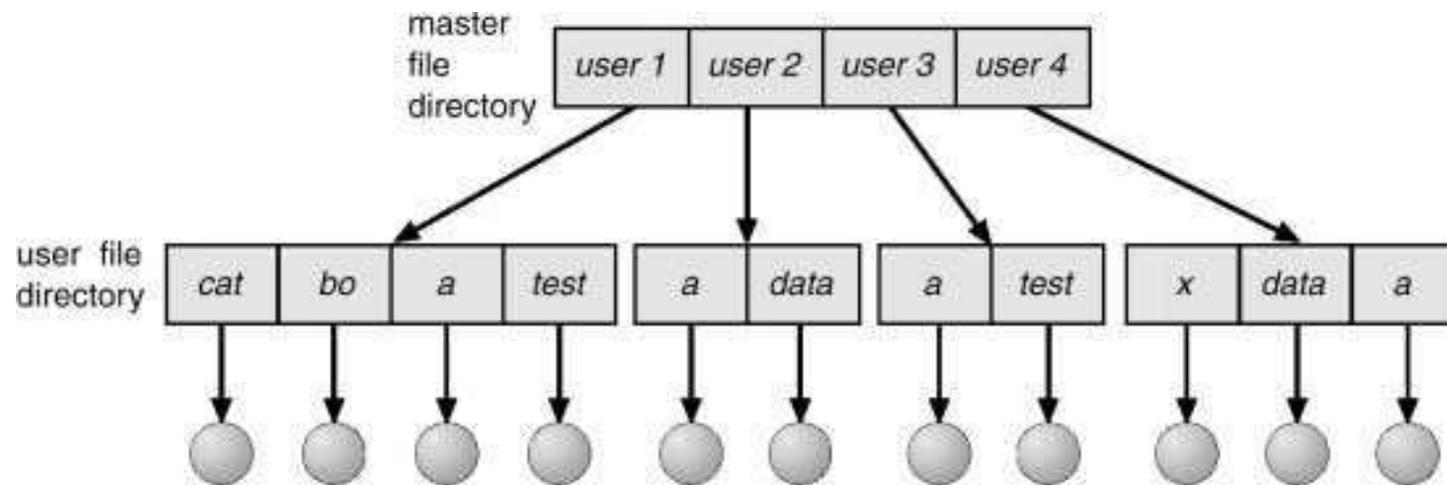
- A single directory for all users.



- Naming problem
- Grouping problem

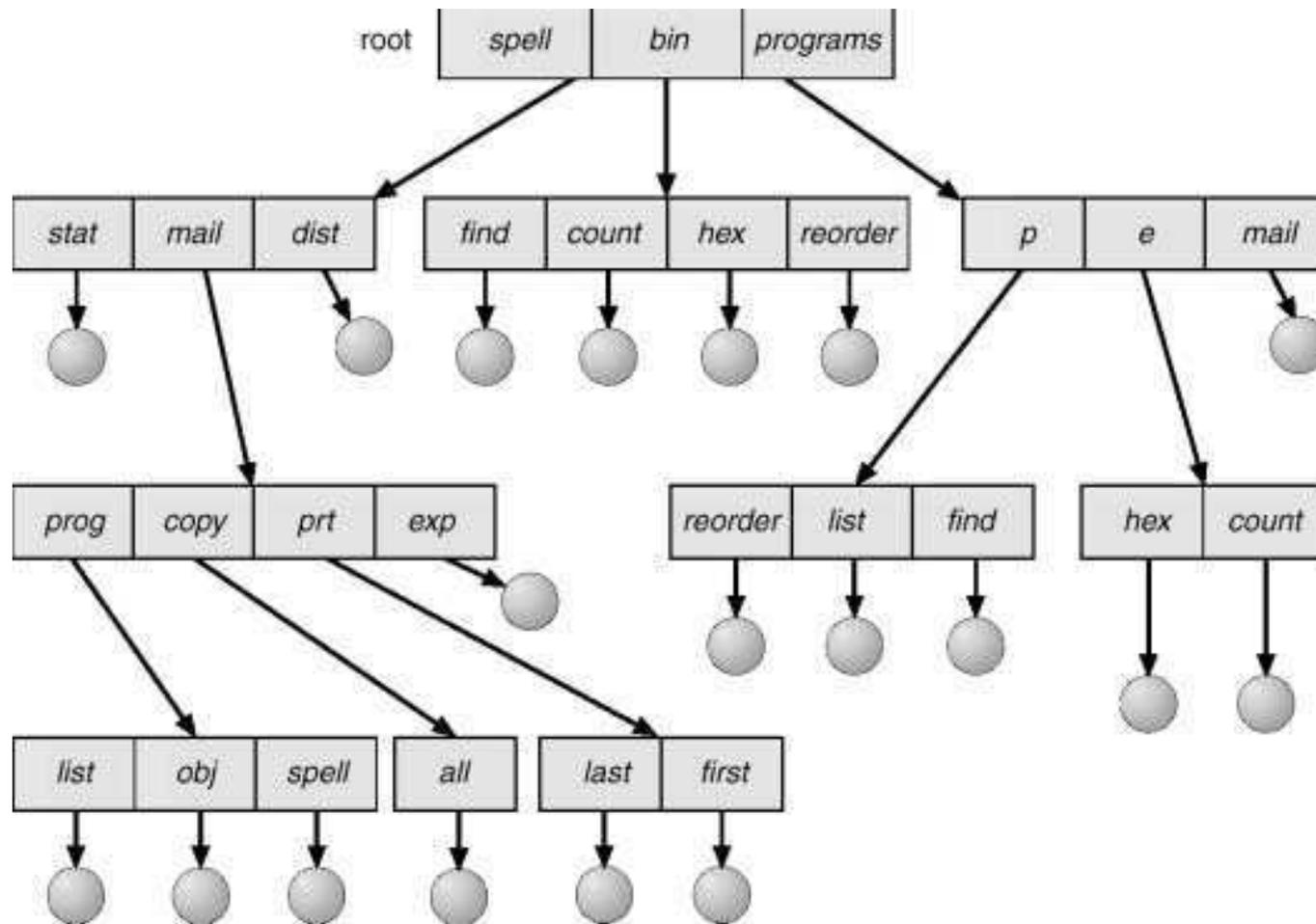
# Two-Level Directory

- Separate directory for each user.



- Path name
- Can have the same file name for different user
- Efficient searching
- No grouping capability

# Tree-Structured Directories



## Tree-Structured Directories (Cont.)

- Efficient searching
- Grouping Capability
- Current directory (working directory)
  - **cd** /spell/mail/prog
  - **type** list

## Tree-Structured Directories (Cont.)

- Absolute or relative path name
- Creating a new file is done in current directory.
- Delete a file

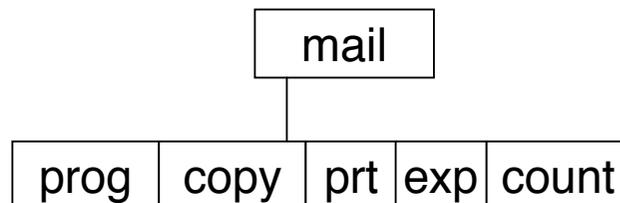
**rm** <file-name>

- Creating a new subdirectory is done in current directory.

**mkdir** <dir-name>

Example: if in current directory /spell/mail

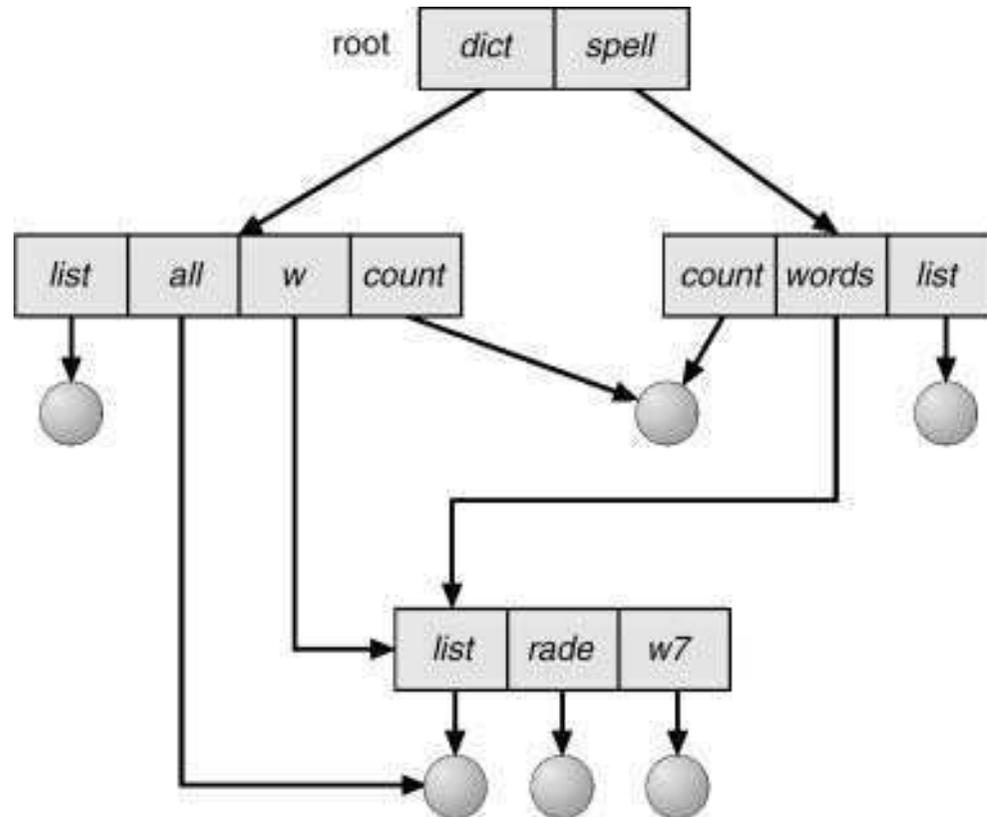
**mkdir** count



- Deleting “mail” ⇒ deleting the entire subtree rooted by “mail”.

# Acyclic-Graph Directories

- Have shared subdirectories and files.



## Acyclic-Graph Directories (Cont.)

- Two different names (aliasing)
- If *dict* deletes *list*  $\Rightarrow$  dangling pointer.

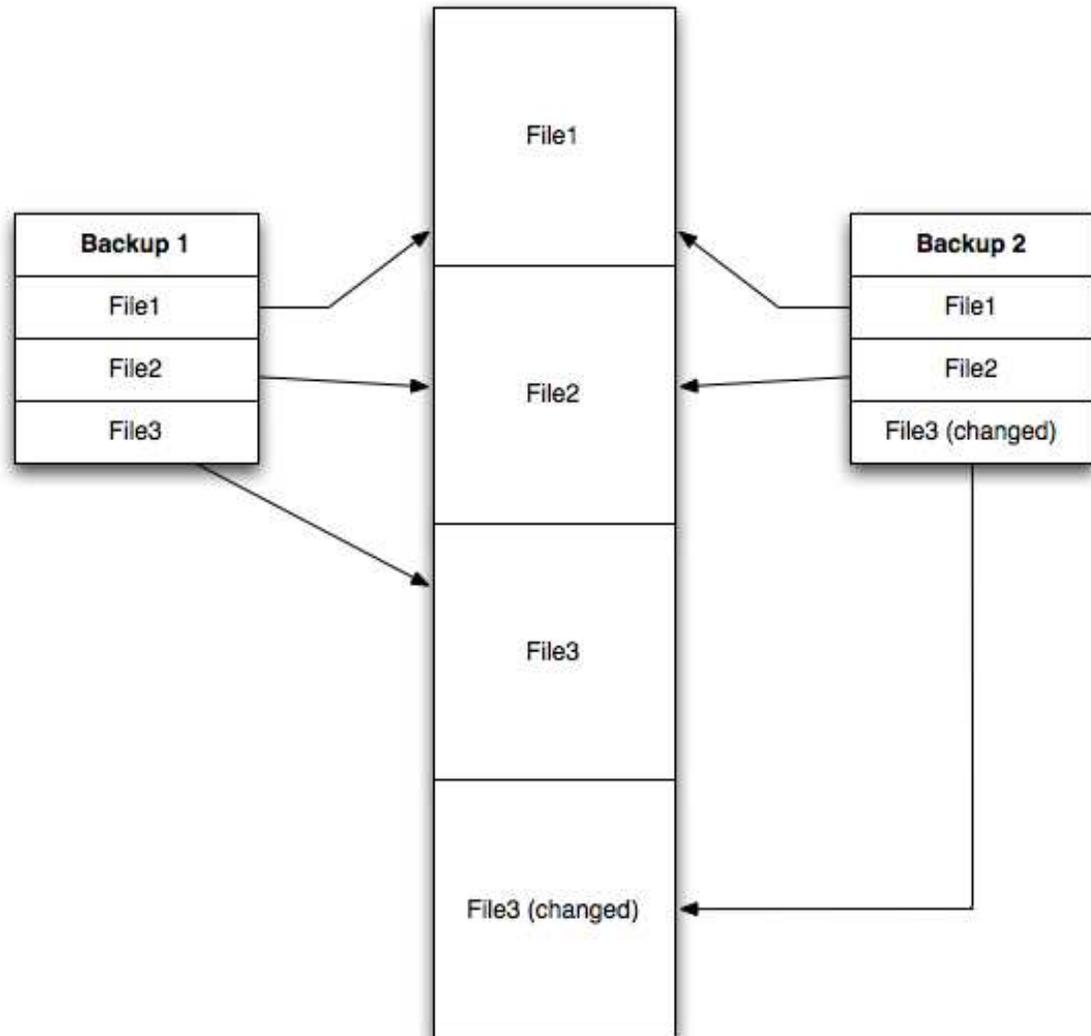
Solutions:

- Backpointers, so we can delete all pointers.  
Variable size records a problem.
- Backpointers using a daisy chain organization.
- Entry-hold-count solution.

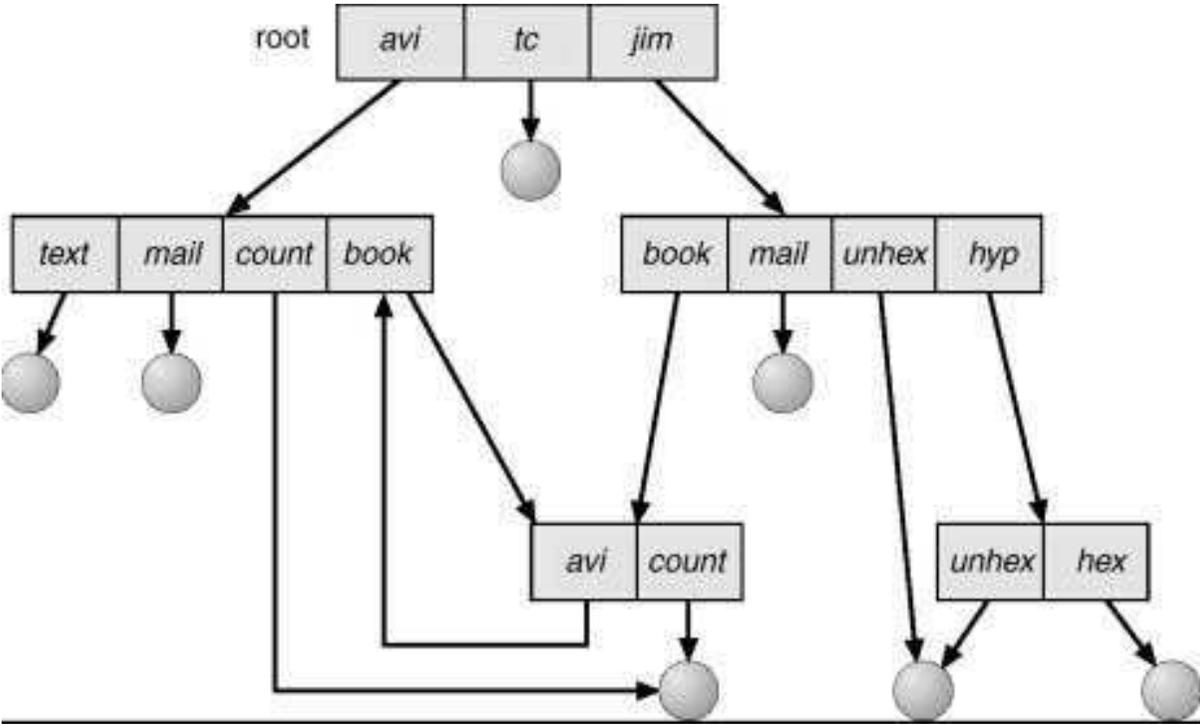
# Using hard links: Time Machine



# Using hard links: Time Machine



# General Graph Directory



## General Graph Directory (Cont.)

- How do we guarantee no cycles?
  - Allow only links to file not subdirectories.
  - Garbage collection.
  - Every time a new link is added use a cycle detection algorithm to determine whether it is OK.

# Protection

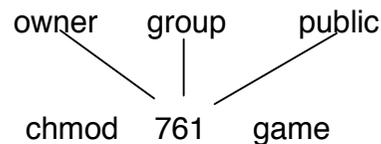
- File owner/creator should be able to control:
  - what can be done
  - by whom
- Types of access
  - Read
  - Write
  - Execute
  - Append
  - Delete
  - List

# Access Lists and Groups

- Mode of access: read, write, execute
- Three classes of users

			RWX
a) owner access	7	⇒	1 1 1
			RWX
b) groups access	6	⇒	1 1 0
			RWX
c) public access	1	⇒	0 0 1

- Ask manager to create a group (unique name), say *G*, and add some users to the group.
- For a particular file (say *game*) or subdirectory, define an appropriate access.



- Attach a group to a file

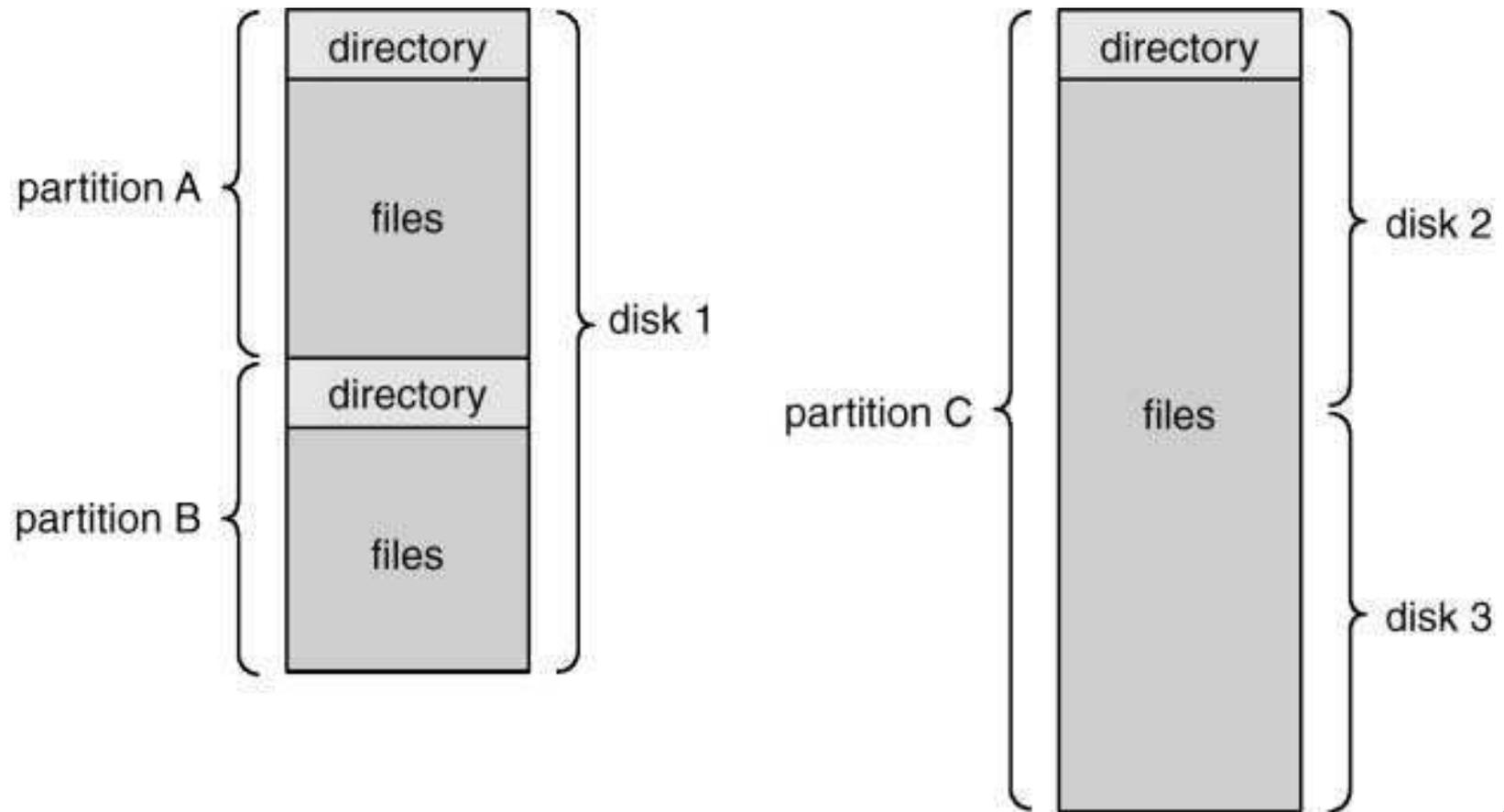
**chgrp** *G* *game*

- File-System Structure
- Allocation Methods
- Free-Space Management
- Directory Implementation
- Efficiency and Performance
- Recovery

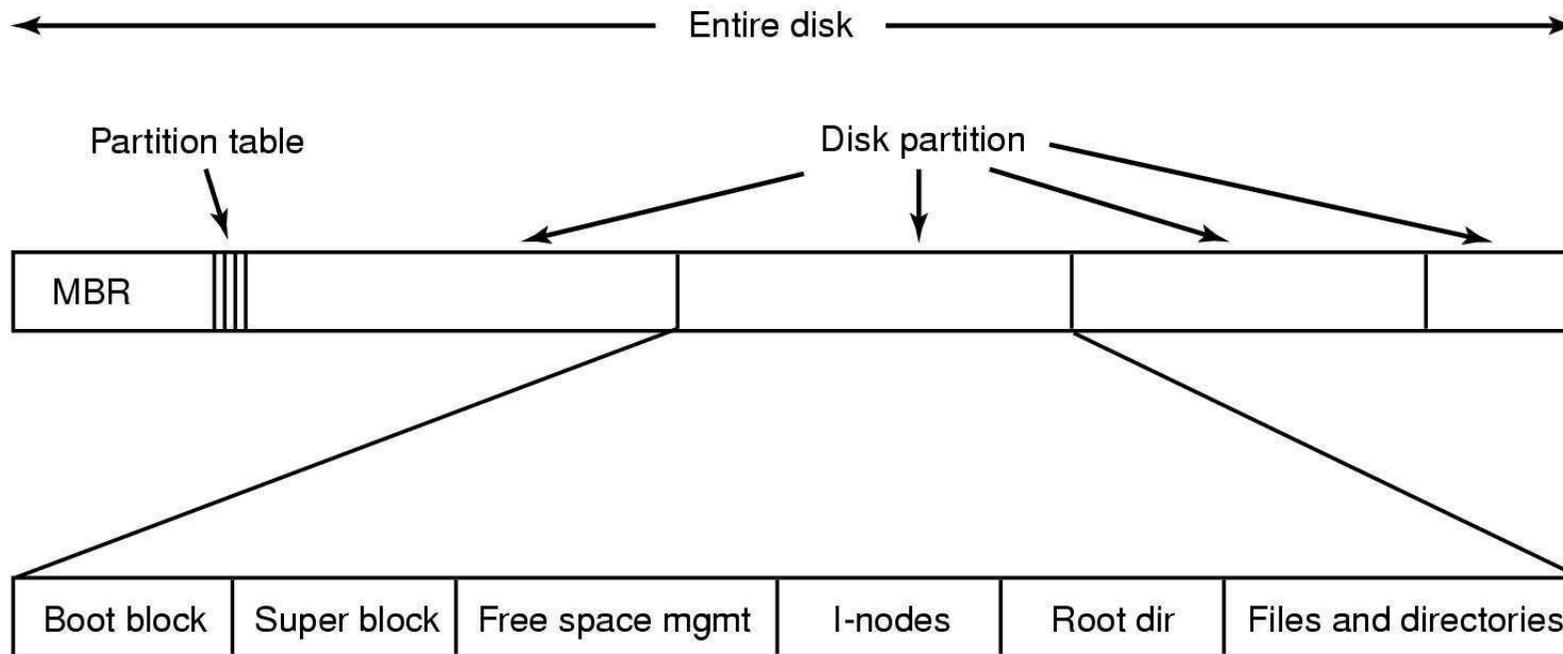
# File-System Structure

- File structure
  - Logical storage unit
  - Collection of related information
- File system resides on secondary storage (disks).
- File system organized into layers.
- *File control block* – storage structure consisting of information about a file.

# Typical File-System Organization



# File System Implementation



A possible file system layout

**Allocation**

# Putting Bytes on Disk

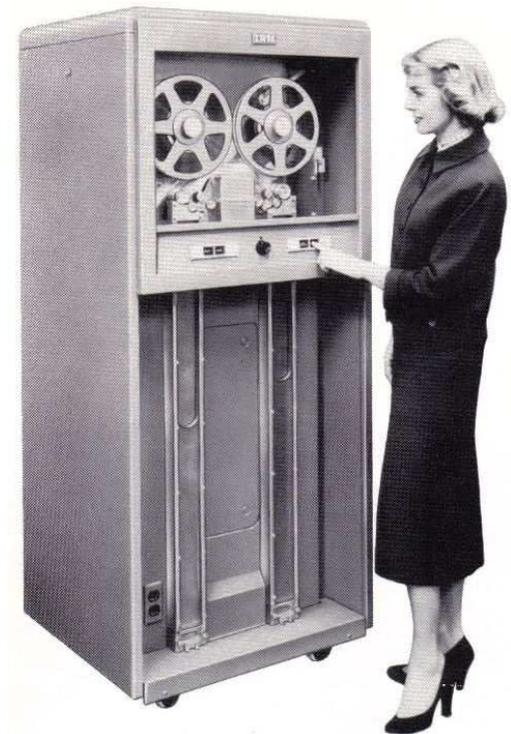
- File *viewed* as a contiguous sequence of bytes
- Allocation is actually *storing* the bytes

# Fragmentation Types

- Data: file not contiguous
- External: unusable empty space between files
- Internal: allocated but unused space  
→ file smaller than block

# Random Access

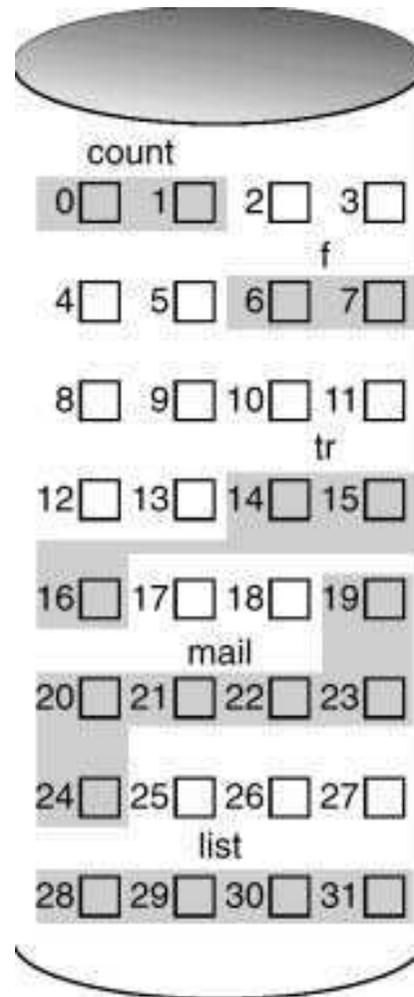
- Access time *independent* of the current block
- Also called *Direct Access*
- RAM: Random Access Memory
- Tape: no direct access



# Contiguous Allocation

- Each file occupies a set of contiguous blocks on the disk.
- Simple – only starting location (block #) and length (number of blocks) are required.
- Random access.
- Wasteful of space (dynamic storage-allocation problem).
- Files cannot grow.

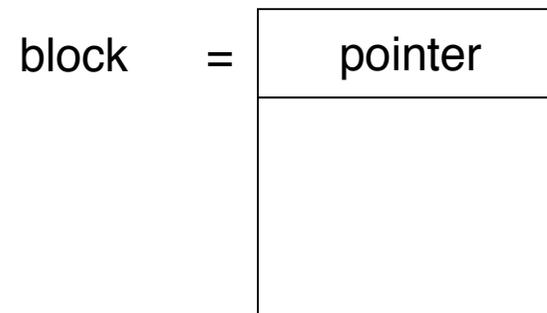
# Contiguous Allocation of Disk Space



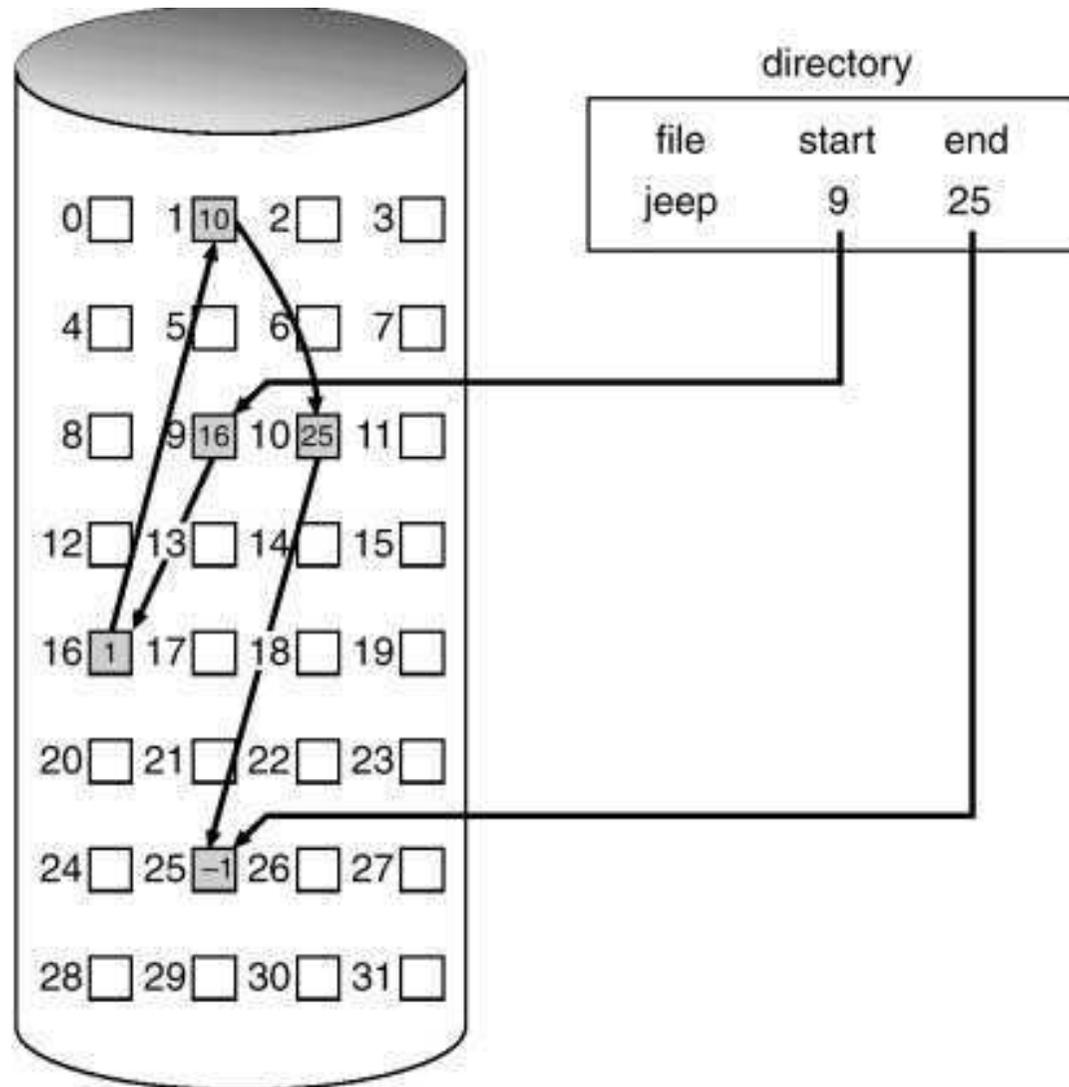
file	start	length
count	0	2
tr	14	3
mail	19	6
list	28	4
f	6	2

# Linked Allocation

- Each file is a linked list of disk blocks: blocks may be scattered anywhere on the disk.



- Allocate as needed, link together; e.g., file starts at block 9



## Linked Allocation (Cont.)

- Simple – need only starting address
- Free-space management system – no waste of space
- No random access
- Clusters of blocks
  - for better performance (disk head moving)
  - to have fewer pointers
- *File-allocation table (FAT)*: disk-space allocation used by MS-DOS and OS/2.

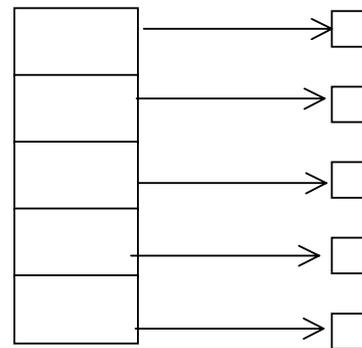
The table is a list of entries that maps each cluster number to:

  - the cluster number of the next entry, or
  - an indication this is the last cluster (end of file), or
  - a special entry to mark bad clusters, or
  - a 0 to mark the cluster is unused

(some cluster may be reserved and are marked in the FAT)

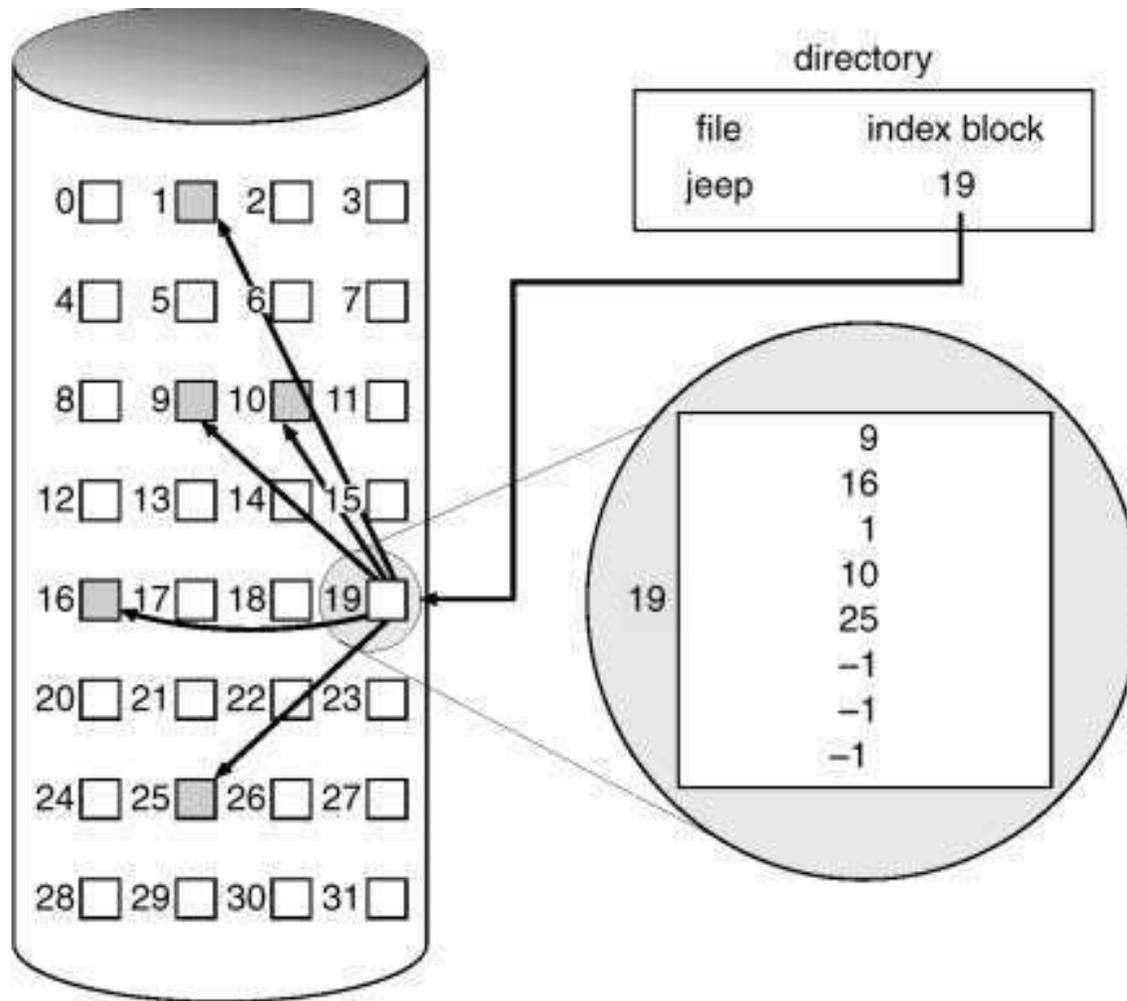
# Indexed Allocation

- Brings all pointers together into the *index block*.
- Logical view.



index table

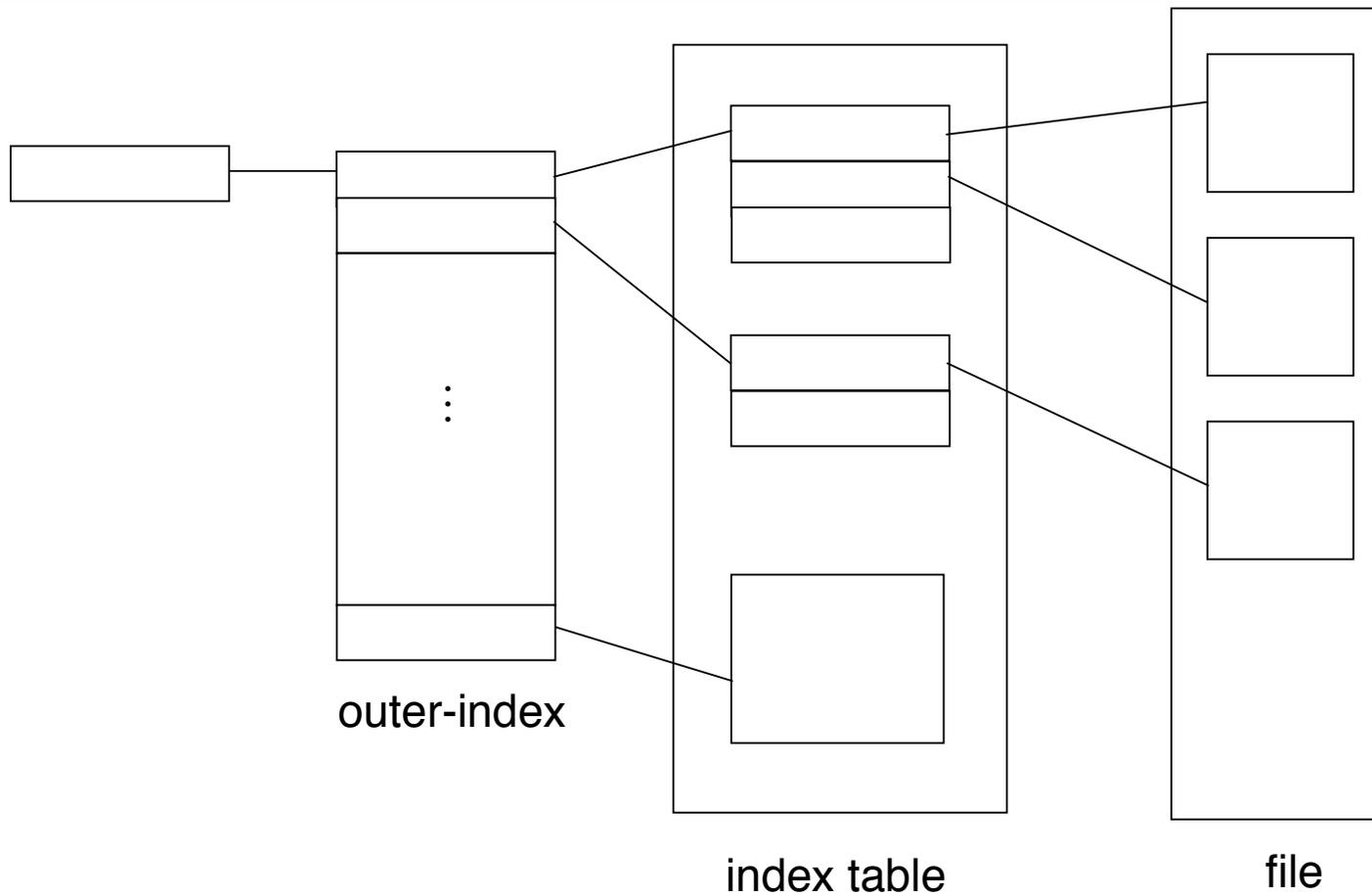
# Example of Indexed Allocation



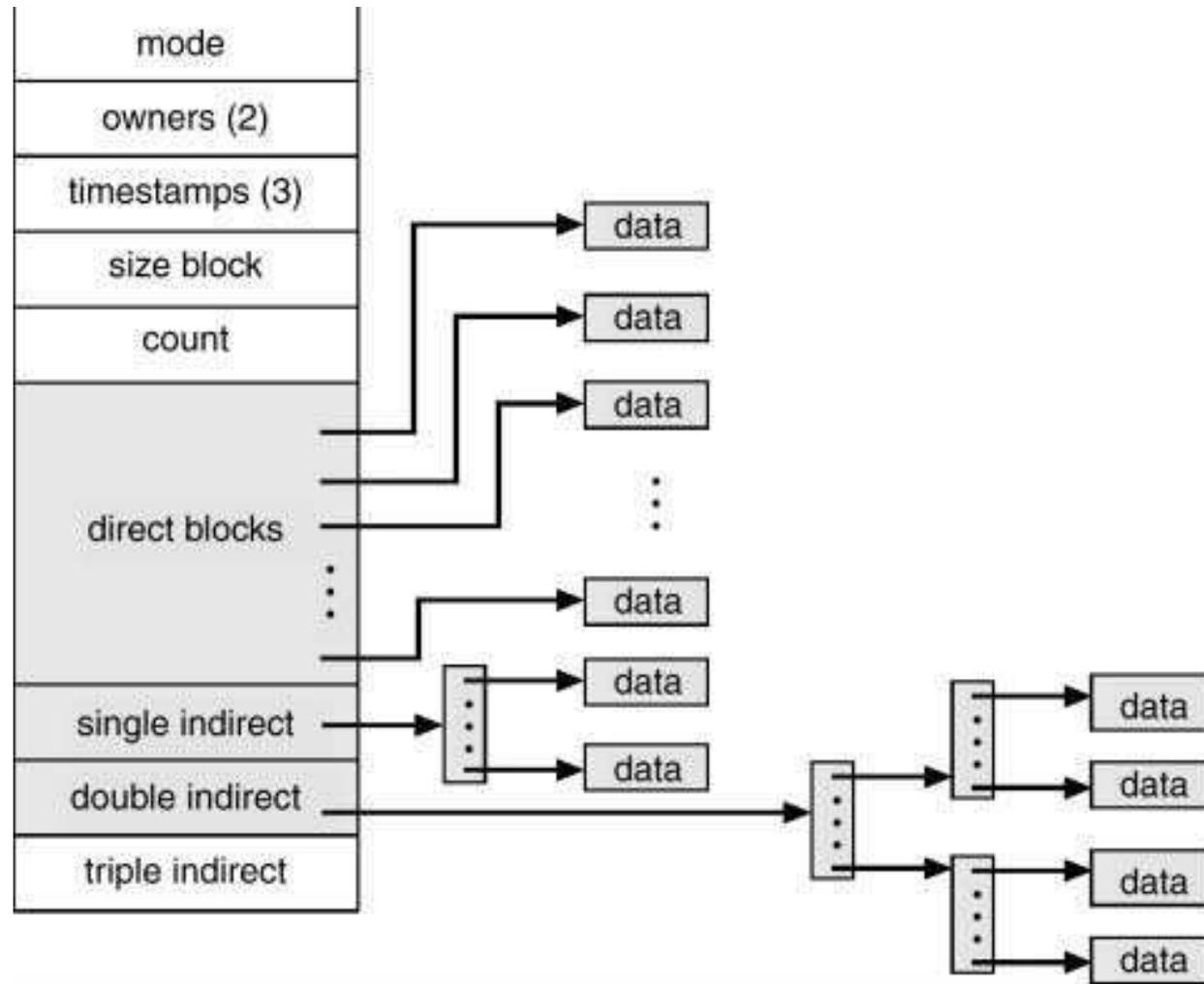
## Indexed Allocation (Cont.)

- Need index table
- Random access
- Random access without external fragmentation, but have overhead of index block.

# Indexed Allocation – Mapping (Cont.)



# Combined Scheme: UNIX (4K bytes per block)



# Comparing Allocation

	Random Access	No Data Frag	No External Frag	Space Waste
Contiguous	✓	✓	✗	0
Linked	✗	✗	✓	# clusters
Indexed	✓	✗	✓	> # clusters



## Free-Space Management (Cont.)

- Bit map requires extra space. Example:

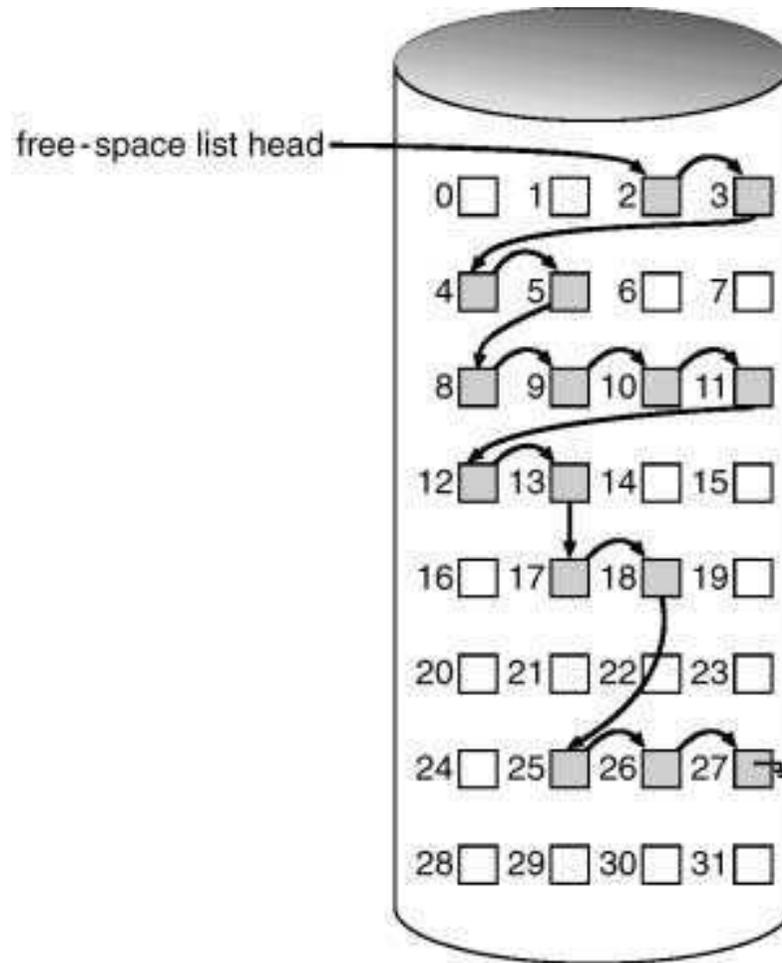
block size =  $2^{12}$  bytes (4K bytes)

disk size =  $2^{30}$  bytes (1 gigabyte)

$n = 2^{30}/2^{12} = 2^{18}$  bits (or 32K bytes)

- Easy to get contiguous files
- Linked list (free list)
  - Cannot get contiguous space easily
  - No waste of space
- Grouping
- Counting

# Linked Free-Space List on Disk





DOE HPC Best Practices Workshop

# File Systems and Archives

SEPTEMBER 26-27, 2011 • SAN FRANCISCO, CA



U.S. DEPARTMENT OF  
**ENERGY**

Office of Science

# The Fifth Workshop on HPC Best Practices: File Systems and Archives

Held September 26–27, 2011, San Francisco

## Workshop Report December 2011

Compiled by Jason Hick, John Hules, and Andrew Uselton  
Lawrence Berkeley National Laboratory

### **Workshop Steering Committee**

John Bent (LANL); Jeff Broughton (LBNL/NERSC); Shane Canon (LBNL/NERSC); Susan Coghlan (ANL); David Cowley (PNNL); Mark Gary (LLNL); Gary Grider (LANL); Kevin Harms (ANL/ALCF); Paul Henning, DOE Co-Host (NNSA); Jason Hick, Workshop Chair (LBNL/NERSC); Ruth Klundt (SNL); Steve Monk (SNL); John Noe (SNL); Lucy Nowell (ASCR); Sarp Oral (ORNL); James Rogers (ORNL); Yukiko Sekine, DOE Co-Host (ASCR); Jerry Shoopman (LLNL); Aaron Torres (LANL); Andrew Uselton, Assistant Chair (LBNL/NERSC); Lee Ward (SNL); Dick Watson (LLNL).

### **Workshop Group Chairs**

Shane Canon (LBNL); Susan Coghlan (ANL); David Cowley (PNNL); Mark Gary (LLNL); John Noe (SNL); Sarp Oral (ORNL); Jim Rogers (ORNL); Jerry Shoopman (LLNL).

---

Ernest Orlando Lawrence Berkeley National Laboratory  
1 Cyclotron Road, Berkeley, CA 94720-8148

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

## **DISCLAIMER**

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

# Contents

Executive Summary .....	3
Introduction.....	5
Workshop Goals.....	5
Workshop Format .....	5
Workshop Breakout Topics .....	5
The Business of Storage Systems .....	5
The Administration of Storage Systems .....	6
The Reliability and Availability of Storage Systems.....	7
The Usability of Storage Systems.....	7
Workshop Findings.....	9
Best Practices .....	9
Gaps and Recommendations.....	13
Appendix A: Position Papers .....	16
Appendix B: Workshop Agenda.....	17
Appendix C: Workshop Attendees .....	21



# Executive Summary

In 2006, the United States Department of Energy (DOE) National Nuclear Security Administration (NNSA) and Office of Science (SC) identified the need for large supercomputer centers to learn from each other on what worked well and what didn't in the areas of acquisition, installation, integration, testing and operation. Previous High Performance Computing (HPC) Best Practice Workshops focused on System Integration in 2007 (<http://outreach.scidac.gov/pibp/>), Risk Management in 2008 (<http://rmtap.llnl.gov/>), Software Lifecycles in 2009 (<http://outreach.scidac.gov/swbp/>), and Power Management in 2010 (<http://outreach.scidac.gov/pmbp/>).

The workshop on HPC Best Practices on File Systems and Archives was the fifth in the series. The workshop gathered technical and management experts for operations of HPC file systems and archives. Attendees identified and discussed best practices in use at their facilities, and documented findings for the DOE and HPC community in this report. The home page for this workshop is <http://outreach.scidac.gov/fsbp/>.

Prior to the workshop, board members identified four critical functionally components to the operation of storage systems: administration, business, reliability, and usability of storage. These components of storage became the breakout session topics. During registration, attendees designated their primary interests for breakout sessions and submitted a position paper to aid in having discussions around proposals or ideas in the position papers provided. Participants submitted 27 position papers, included in Appendix A.

This year the workshop had 64 attendees from 29 different sites from around the world. Participants identified 16 best practices in use:

1. For critical data, multiple copies should be made on multiple species of hardware in geographically separated locations.
2. Build a multidisciplinary data center team.
3. Having dedicated maintenance personnel, vendor or internal staff, is important to increasing system availability.
4. Establish and maintain hot-spare, testbed, and pre-production environments.
5. Establish systematic and ongoing procedures to measure and monitor system behavior and performance.
6. Establish authoritative sources of information.
7. Manage users' consumption of space.
8. Storage systems should function independently from compute systems.
9. Include middleware benchmarks and performance criteria in system procurement contracts.
10. Deploy and support tools to enable users to more easily achieve the full capabilities of file systems and archives.
11. Training, onsite or online, and meeting with the users directly improves the utilization and performance of the storage system.
12. Actively manage relationships with the storage industry and storage community.
13. Monitor and exploit emerging storage technologies.
14. Continue to re-evaluate the real characteristics of your I/O workload.
15. Use quantitative data to demonstrate the importance of storage to achieving scientific results.
16. Retain original logs generated by systems and software.

Many participants do not implement all the best practices, which means they are bringing several new ideas for improvement back to their sites. In addition, attendees identified 15 gaps that require further attention. Gaps likely addressed by evolutionary solutions include:

1. Incomplete attention to end-to-end data integrity.
2. Storage system software is not resilient enough.
3. Redundant Array of Independent Tape (RAIT) does not exist today.
4. Provide monitoring and diagnostic tools that allow users to understand file system configuration and performance characteristics and troubleshoot poor I/O performance.
5. Communication between storage systems users and storage system experts needs improvement.
6. Need tools and mechanisms to capture provenance and provide life-cycle management for data.
7. Ensure storage systems are prepared to support data-intensive computing and new workflows.
8. Measure, and track trends in storage system usage to the same degree we measure and track flops and cycles on computational systems.
9. Tools to provide scalable performance in interacting with files in the storage system.

Gaps requiring revolutionary approaches include:

1. Need quality of service for users in a shared storage environment.
2. Need standard metadata for users to specify data importance and retention.
3. The diagnostic information currently available in today's storage systems is woefully inadequate.
4. Metadata performance in storage systems already limits usability and won't meet the needs of exascale.
5. POSIX isn't expressive enough for the breadth of application I/O patterns.
6. There exists a storage technology gap for exascale.

# Introduction

The U.S. Department of Energy (DOE) has identified the design, implementation, and usability of file systems and archives as key issues for current and future high performance computing (HPC) systems. This workshop was organized to address current best practices for the procurement, operation, and usability of file systems and archives. Furthermore, the workshop addressed whether system challenges can be met by evolving current practices.

## Workshop Goals

The workshop organizers presented the following goals:

- Foster a shared understanding of file system issues in the context of HPC centers.
- Identify top challenges and open issues.
- Share best practices and lessons learned.
- Establish communication paths for managerial and technical staff at multiple sites to continue discussion on these topics.
- Discuss roles and benefits of HPC stakeholders.
- Present findings to DOE and other stakeholders.

## Workshop Format

The organizers requested a short position paper from each attendee to identify best practices in the area of file systems and archives. Each position paper addressed a topic or topics related to best practices for the session they were attending. The session organizers identified the questions below as key topics for each session and suggested that each paper deal with one or more of them. Participants were asked to frame their discussion to identify what they think are best practices in use at their site related to the session topics, and to frame questions within the discussion to help elicit best practices from the other participants. Position papers are collected in Appendix A.

The workshop began with presentations by representatives of the DOE sponsors: Thuc Hoang of the National Nuclear Security Administration (NNSA) and Yukiko Sekine of the Office of Science (SC). The workshop focused primarily on breakout sessions in which participants presented and evaluated selected topics from their position papers. Each day ended with a member of each breakout session presenting the day's results to the entire group. The agenda is presented in Appendix B.

## Workshop Breakout Topics

In their position papers and breakout discussions, participants were asked to consider the topics listed below.

### **The Business of Storage Systems**

For each layer of the storage hierarchy — file systems, archives, others — address these topics (and add to them):

1. HPC facilities across DOE deploy and operate systems at unprecedented scale requiring advanced file system technologies. Achieving the requisite level of performance and scalability from file systems remains a significant challenge.

- A. What are your practices used to plan for future system deployments and system evolution over time?
  - B. How do you establish requirements such as bandwidth, capacity, metadata operations/sec, mean time to interrupt (MTTI), etc. for these systems?
2. It is not uncommon for archival storage deployments to have life spans approaching multiple decades. Growth of archival data at a number of HPC sites is exponential.
    - A. How do you effectively plan for exponential growth rates and archives that will need to serve multiple generations of machines throughout their life within a fixed budget profile?
    - B. Are exponential growth rates sustainable? How do you mitigate if not?
  3. There are relatively few alternatives in parallel file system and archival storage system software that meet the requirements of major HPC facilities. Development of this software varies from proprietary closed source to collaborative open source solutions. Each model has benefits and drawbacks in terms of total cost of ownership, ability to evolve the software to meet specific requirements, and long-term viability (risk) of the system.
    - A. What model do you leverage for your file system or archival system software needs?
    - B. What benefits/drawbacks do you see with these models?
  4. Storage system and tape archive technologies vary from high-end custom hardware developed specifically for the HPC environment to commodity storage platforms with extremely broad market saturation.
    - A. Where do you leverage custom versus commodity storage hardware within your operational environment?
    - B. Do you see opportunities to incorporate more commodity storage technologies within your environment in the future?
    - C. What are the barriers to adopting commodity storage technologies and how can they be overcome in the future?

## The Administration of Storage Systems

For each layer of the storage hierarchy — file systems, archives, others — address these topics (and add to them):

1. Change control and configuration management.
  - A. What specific configuration management tools, methods, or practices does your center use to validate hardware/software changes and releases to minimize production performance degradation or system downtime?
  - B. Can you provide an example of how testing a change/release on a pre-production system provided unique insight into a configuration problem before users detected it?
2. Ongoing system administration.
  - A. What specific file system and/or archive metrics does your center measure and monitor on a regular basis, and how have those findings directed which types of operational tasks your center has automated to minimize frequency and impact of production incidents and optimize system performance and end-user experience?
  - B. Can you provide an example where self-monitoring and detecting production incidents led to an investigation of root cause to reduce mean time to resolution (MTTR) of future file system or archive outages?

3. Technology refresh.
  - A. What unique approach has your center taken to balance the end-user requirement to increase system availability while providing system architects the opportunity and access to the environment to satisfy the ongoing need to refresh underlying file system and archive components?
  - B. How does your center expand capacity of either a file system or archive resource while minimizing user impact?
4. Security management.
  - A. What strategy is your center taking with, for example, operating system (OS) patching or vulnerability scanning, to satisfy the ongoing and rigorous demands of computer security professionals?
  - B. How does your strategy balance the growing need to provide a high degree of collaboration for distributed user communities with access to multiple levels of data sensitivity?
  - C. Assuming your center provides a multi-zoned security architecture with various access control levels and technologies, how would you demonstrate to computer security that it is providing adequate protection and controls?

## The Reliability and Availability of Storage Systems

For each layer of the storage hierarchy — file systems, archives, others — address these topics (and add to them):

1. Resilient architectures and fault tolerance.
  - A. What specific system architectural or configuration practices, decisions or changes (hardware and software) has your center made that have demonstrably improved availability, reliability or performance of your file system or archival system?
  - B. Where in the environment should redundant hardware be bought/deployed?
2. Hardware and software maintenance.
  - A. What is your philosophy or practice for executing system maintenance down times?
  - B. How do those practices contribute to improved system availability and reliability outside of planned outages?
3. Data integrity.
  - A. What strategies do you use to ensure data integrity and what parts of the end-to-end compute/store/visualize/archive cycles does it cover?
  - B. Do you employ end-to-end checksums within the end-to-end cycle and if so where?
4. Off-hours support and availability.
  - A. What mechanisms do you have in place to ensure reliable file system and archive operation during off-hours and, in the event of a facility event, such as power loss, chilled water loss, fire alarm, etc.?
  - B. What mechanisms are in place to quiesce storage and to protect it?

## The Usability of Storage Systems

For each layer of the storage hierarchy — file systems, archives, others — address these topics (and add to them):

1. What are your major usability issues?
2. What applications/tools have you developed or obtained elsewhere and deployed that have made your storage system more effective and useful for end users?
  - A. What tools or methods are available to users for I/O related problem diagnosis? Describe experiences where use of diagnostics resulted in improved outcomes. Are the available diagnostic capabilities sufficiently robust? Scalable?
  - B. Please help us categorize the applications/tools in use.
  - C. Which tools are used by end users, and which primarily by system administrators?
3. Discuss recent challenges in providing I/O service to your user community, and what practices/strategies were used to meet them.
  - A. What trends in user requirements resulted in the need to address the challenges?
  - B. How well are current solutions meeting the demands, or where are they falling short?
  - C. Where might experience at other sites be helpful to your challenges?
  - D. Which of your practices outlined above would you suggest for a best practices list?
4. Large data movement: With respect to internal file and storage systems, do sites dedicate specific resources to data movement internally?
  - A. What (if any) direction is given to users for moving data around internally?
  - B. What tools are available for helping users improve data transfer performance?
5. At what organizational level are users managing data organization, per user, per code, per project, or some larger unit? Are there common approaches or best practices that have been identified which are being leveraged to aid these efforts?
6. How does your site manage health monitoring of I/O services, and how is pertinent information transmitted to users? What feedback do users have on the content and timeliness of the information?
7. What are your user training and documentation practices?

# Workshop Findings

Workshop participants identified the best practices provided in this section of the report as applicable to file systems and archival storage systems operations at high performance computing facilities. The practices summarize the experiences of 29 different sites across many different funding and parent agencies, and represent decades of experience in operations.

In this section, each best practice is accompanied by details to help better understand the meaning of the practice, and implementation examples where practical. Most practices identified have an aspect of usability, reliability, administration and business in operating storage; that is to say that many of the practices were intertwined among breakout sessions at the workshop. For example, reliability investment decisions aren't just about compliance or policies in a particular storage system; they affect the ability to do science and, therefore, the usability of the storage system.

## Best Practices

- 1. For critical data, multiple copies should be made on multiple species of hardware in geographically separated locations.**

There is no substitute for multiple copies on dissimilar hardware and software and media formats. Systems that compromise on either hardware, software, or media distinction are more vulnerable to data loss than systems that enable hardware, software, and media diversity (i.e. a single firmware bug can corrupt data stored geographically apart, but on identical hardware/software). Workshop attendees also recommended that sites should have multiple file systems to provide available storage during downtimes and distribute particular workloads to supporting system configurations (e.g., scratch and home, backup and archive). Participants pointed out that backups and disaster recovery plans are classic methods still relevant today of providing the best data protection for critical data and systems retaining critical data.

- 2. Build a multidisciplinary data center team.**

Robust facilities operations are a key component of high performing HPC data systems. Sites agreed that regardless of the specific method, teaming facility personnel with HPC personnel to design, plan, and maintain equipment in support of data systems requirements increased reliability of the data systems. Regularly scheduled meetings involving facility representatives and personnel from every aspect of HPC center operations are critical to maximizing communication and overall HPC center efficiency. Specific facility recommendations for data systems arose at the workshop: having multiple power feeds, priority for use of universal power supplies (UPS), ensuring that power supplies meet high standards (i.e., CBEMA and SEMIF47), and having automation for facility operations (i.e., emergency power-off, startup and shutdown).

- 3. Having dedicated maintenance personnel, vendor or internal staff, is important to increasing system availability.**

By dedicating personnel to specialize on storage hardware and software, the facility will have increased involvement in the issues, needs, and solutions around operating its storage systems. All workshop participants had dedicated storage staff or vendors for their storage systems. By having dedicated personnel, they are able to have the detailed knowledge and experience on hand to prevent many system failures and respond quickly to return the system to operation. Further discussion also pointed out that having development knowledge (e.g., source code, or developers working on storage software) enhanced the reliability and availability in HPC storage systems. Workshop participants pointed to the success of both Lustre and HPSS at the labs involved in

their development as examples of this. Non-development sites often rely on development sites for achieving high availability storage systems.

**4. Establish and maintain hot-spare, testbed, and pre-production environments.**

Data systems in HPC environments are held to high standards of integrity since data is important to the validity of scientific findings. Workshop participants discussed that increased reliability and availability of file systems and archives can be achieved by using hot-spare/burn-in and pre-production systems where software and hardware is tested prior to being placed into production. Attendees identified that the benefits are twofold: to validate procedures (executing both the deployment steps and failback steps) and to validate the readiness of hardware or software in the specific site's environment. Testbed data systems are key to gaining knowledge and expertise of new technologies, and to evaluating future storage hardware and software in isolation from the production environment. Participants identified the importance of not using the pre-production environment for evaluation of hardware or software that is not intended to move into production rapidly, so as not to undermine the stability of the pre-production environment and its ability to closely represent the production environment.

**5. Establish systematic and ongoing procedures to measure and monitor system behavior and performance.**

The complexity of HPC file systems and archival resources presents a challenge to design, deploy, and maintain storage systems that meet specific performance targets. It is important to understand the anticipated workload and use that understanding to design performance benchmarks that reflect that workload. Performance benchmarks and workload simulators should be used systematically from the time new hardware is first under evaluation, through the deployment of a new resource, and regularly throughout the life of the system. Benchmark tests establish a baseline for expectations of the system, and regular testing will reveal most performance degradations due to software and hardware issues, resource contention, or a change in the workload. Share benchmark results with other sites and compare performance with other comparable systems. In addition to performance benchmarks, testing should also include data integrity checks. Monitor usage statistics over time in order to anticipate the growth of demand for resources.

**6. Establish authoritative sources of information.**

In order to answer questions accurately and consistently, there should be a shared and widely known location for information. User documentation, training, and how-to documents should be available and kept up to date. There needs to be a definitive source code control repository for open source infrastructure and benchmark software. Use configuration and change management tools to assist system staff with the administration of the storage resource. Support staff need a central location for procedures and scripts for service disruptions and problem resolution. Make benchmark results, system metrics, and reports on milestones available.

**7. Manage users' consumption of space.**

This was a hot topic in the workshop, where attendees discussed different aspects of managing usage of the storage system. A variety of solutions came forth, involving both active (allocations and quotas) and passive (accounting and auditing) techniques. Most participants used allocations or quotas to manage consumption of storage resources with good success. Participants unanimously agreed that having transparency of capacity to users was important to success in managing a limited resource in high demand. Those that use allocations and quotas noted that they are important to managing data effectively and keeping the system usable. A popular proposal put forth is to consider allocation renewals as a way to revalidate the need for data

retention. Attendees agreed that automation of storage policies is desired (e.g., information lifecycle management, or ILM, capabilities).

**8. Storage systems should function independently from compute systems.**

Sites that design their file systems, even local scratch, to be independent of their compute system hardware and software improve the availability and reliability of the storage system. Without independence, issues with either the storage or the compute system usually affect each other. There is an added benefit to users in having the systems decoupled: the compute system's data may still be available to users during compute system downtime. For instance, data analysis or visualization could proceed if the file system were available to platforms capable of such work. It was further pointed out that the file and archive systems should be able to be standalone as well. Integrated file and archive systems that are so tightly coupled as to prevent usage of either the file or archive system in the event the other is unavailable are not recommended. Workshop participants identified that this practice is achievable with either dedicated file systems (e.g., local scratch file systems) and center-wide or global file systems. Though workshop participants agreed on this practice, they noted that facilities should also ensure they provide mechanisms to pre-stage or move data between the distinct systems for greater usability.

**9. Include middleware benchmarks and performance criteria in system procurement contracts.**

Users have an expectation of predictable performance. The importance of storage to HPC systems needs to be stressed from the beginning of the procurement process. A fraction of the overall budget, in the range of 10% to 20%, should be dedicated to file system and storage resources in keeping with the goal of achieving target performance. Performance targets should include middleware libraries, and DOE sites should fund performance enhancements to those libraries. The desired result is that users see "portable performance" between sites when using middleware libraries that have been transparently optimized for the local storage architecture.

**10. Deploy and support tools to enable users to more easily achieve the full capabilities of file systems and archives.**

Often users express frustration with achieving expected or reported performance of file systems and archives. Workshop participants shared a number of different tools that aid a user in exploiting the performance of the file system or archive. Specifically, the workshop identified the following tools as particularly useful:

- Parallel Storage Interface (PSI), HTAR, and spdcp for parallelizing metadata and/or data operations
- GLEAN, ADIOS, and Parallel Log-structured File System (PLFS) for restructuring I/O to better match underlying file system configuration
- Darshan, Integrated Performance Monitoring (IPM), and Lustre Monitoring Tool (LMT) for providing monitoring and diagnostic information to troubleshooting poor I/O performance
- Hopper, EMSL, NEWT, and GlobusOnline for providing web interfaces to improve access and operations on data.

**11. Training, onsite or online, and meeting with the users directly improves the utilization and performance of the storage system.**

Training and educating users on storage system operation (e.g., its capabilities and limitations, architecture, and configuration) and recommended usage (e.g., special options or flags to consider or to avoid) benefits the utilization, administration, and user satisfaction with the storage system. Most workshop participants had web pages and various other forms of usage and system documentation; however, some sites had online videos providing details about using their storage

systems, and others conducted onsite training or worked one-on-one with their users. All agreed that the more engaged a site is in providing user training on storage systems, the more satisfied and positive the users of storage tended to be. Several of the breakout sessions proposed that transparency of storage system operations and capabilities is key to managing user expectations and important to operating a successful storage system. Training and documentation are the most successful ways of providing transparent storage system operations.

**12. Actively manage relationships with the storage industry and storage community.**

Attendees agreed that sites should avoid objective metrics and penalties in acquiring and maintaining storage systems, instead working towards partnerships between customer and industry. Some examples of collaborations in storage are HPSS and Lustre (e.g. OpenSFS). In situations without strong collaboration, it was noted that user groups and working groups are an effective way for the storage community and stakeholders to work towards storage system improvement to mutual benefit between the participants. Participants also recommended encouraging diversity in HPC storage solutions to prevent risk from product extinctions. By developing a relationship with the storage industry, sites better understand upcoming technologies and products to improve file or archive system operation. Many sites advocated supporting open source technologies to mitigate risk. All participants recommended stronger communication and collaboration amongst fellow HPC storage administrators as important to improving file and archive system operations.

**13. Monitor and exploit emerging storage technologies.**

To the extent possible, HPC systems benefit when they can exploit commodity products and services. As new technologies emerge, DOE centers need to examine them and determine if and how those technologies fit in the HPC environment. Cloud services, for example, may provide cost benefits in high volume data processing, but may also require scientists to rethink their application architectures. An early and careful evaluation of emerging technologies can mitigate the risk of making a choice with a poor long-term outcome. As the cost and performance of flash and spinning media evolve, the underlying architecture may need to change; for example, the storage hierarchy may replace tape with disk.

**14. Continue to re-evaluate the real characteristics of your I/O workload.**

While data intensive computing is creating I/O workloads that exceed Moore's law, there is a growing gap between computation and storage capabilities. Monitoring tools like Darshan and the Lustre Monitoring Tool (LMT) can help capture changes in user requirements and the I/O workload on currently deployed systems. As the science being conducted evolves, this will lead to changes in future requirements. Those changes must inform the procurement and planning of systems. The underlying science is driving an evolving set of storage requirements both for bandwidth and volume. It is becoming important to track storage metrics they way the HPC community tracks flops and cycles.

**15. Use quantitative data to demonstrate the importance of storage to achieving scientific results.**

Workshop attendees felt there was an element of advocacy in establishing clear and detailed storage system requirements. An "I/O blueprint" can quantify both the costs and benefits of required storage. Since an under-resourced storage hierarchy will impact the amount of science that can be accomplished, that negative impact represents an opportunity cost. A comparison between that opportunity cost and the storage system expense can allow the HPC system to be designed to have storage that balances it.

## **16. Retain original logs generated by systems and software.**

Workshop participants noted that pristine logs are critical to security forensics, problem diagnosis, and event correlation.

While identifying the current best practices for operation of HPC storage systems, we spent time also discussing changes in user needs based on scaling our systems to the exascale level, which implies exabyte file systems and archives. The current gaps and some future recommendations to prepare file and archive systems for the exascale era of storage are provided in the next section.

## **Gaps and Recommendations**

The information that follows is divided into two sections. First are gaps that exist in current operations of HPC storage systems for which solutions do not exist, but which the community or industry is progressing towards. Second are gaps that are not being addressed operationally and that demand revolutionary approaches to solve.

Gaps likely addressed by evolutionary solutions:

### **1. Incomplete attention to end-to-end data integrity.**

While no HPC solution exists today providing end-to-end (client to system and back to client) data integrity, the community contributes and supports the T10PI standard. The standard is currently being adopted for components used in the HPC environment.

### **2. Storage system software is not resilient enough.**

An otherwise normal hardware error can cause a storage system outage due to current software limitations. Today this impacts the choice of hardware and the amount of facilities infrastructure required to sustain reliable storage systems.

### **3. Redundant Array of Independent Tape (RAIT) does not exist today.**

The only data protection that exists today for tape is multiple copies. However, the HPSS collaboration is working on an implementation of RAIT for use within HPSS.

### **4. Provide monitoring and diagnostic tools that allow users to understand file system configuration and performance characteristics and troubleshoot poor I/O performance.**

Today there exist several tools for monitoring file system performance: Darshan, IPM, and LMT. However, these do not have broad deployment across HPC sites and need further support for code maintenance and feature improvement to be more widely accepted.

### **5. Communication between storage systems users and storage system experts needs improvement.**

There are few storage experts at HPC sites, and it is rare to identify individuals on user projects who are responsible for focusing on storage or I/O. Further, there are rarely consultants at HPC centers who focus on or have the skill set to manage storage or I/O issues. As storage and I/O increasingly become focal issues in HPC projects, this will improve. Sites are working on identifying efficient and effective user education mechanisms (e.g., online videos, onsite workshops).

### **6. Need tools and mechanisms to capture provenance and provide lifecycle management for data.**

Workshop participants identified a growing focus on provenance and lifecycle management as new policies that mandate them are being developed under the America COMPETES Act, which invests in science and technology to improve the United States' competitiveness.

**7. Ensure storage systems are prepared to support data-intensive computing and new workflows.**

Storage systems are designed for computational system needs today. However, new instruments such as genomic sequencers and next generation light sources have tremendous bandwidth and capacity requirements alone that will strain existing systems. As well, the push towards data-intensive computing will result in different system architectures than exist at most HPC facilities today.

**8. Measure and track trends in storage system usage to the same degree we measure and track flops and cycles on computational systems.**

This will improve understanding of the capabilities and limitations of the system in use, and will help with improving quality of service for users in a shared storage environment.

**9. Tools to provide scalable performance in interacting with files in the storage system.**

This is focused on tools users require to interact with files (cp, tar, gzip, grep, etc.). Tools help, but there is room for improvement beyond tools with existing systems.

Gaps requiring revolutionary approaches:

**1. Need quality of service for users in a shared storage environment.**

Nearly all sites represented at the workshop have shared or centralized storage environments. Attendees recognized the need to provide a higher level of performance to users that need it. In current shared environments, all users experience the aggregate performance of the shared storage system. Unlike compute resource managers, storage managers have no method of scheduling workloads or even understanding how to prevent competing I/O workloads from affecting each other.

**2. Need standard metadata for users to specify data importance and retention.**

One common example is that it is impossible today to specify the lifetime of a file as an attribute that stays with the file through whatever number of storage systems the file moves through. This is required for management of the data.

**3. The diagnostic information currently available in today's storage systems is woefully inadequate.**

With any storage system in HPC use today, it is still difficult and time consuming to figure out what user is affected by or causing a particular problem. This is primarily because HPC systems in use today are all distributed and have large and complex architecture. For example, the scale of systems today makes seemingly simple operations, such as finding a user (bad actor) who is willfully or unknowingly impacting center performance, impossible in real time.

**4. Metadata performance in storage systems already limits usability and won't meet the needs of exascale.**

There are two main problems with increasing metadata performance in storage systems today: software design for distribution of metadata operations, and reliable hardware limitations for accelerating metadata operations. Storage system software requires design to enable parallel and distributed metadata operations to enable the best performance. Metadata is normally a relatively

small amount of data in even very large storage systems today. Finding a reliable but increasingly fast small storage device that is affordable is challenging. In addition, improving the usefulness of metadata (e.g. indexes, extended attributes) to users of storage systems means having more small fast reliable storage devices.

**5. POSIX isn't expressive enough for the breadth of application I/O patterns.**

POSIX compliance limits the ability to achieve maximum performance in storage systems. To achieve improved performance, most storage systems relax POSIX requirements (e.g. lazy updates to file timestamps). A new approach is needed to enable a broader range of application I/O without burdening the application with the complexity of keeping their data consistent.

**6. There exists a storage technology gap for exascale.**

Storage technology is still improving at a slower rate than compute or networking technology. New forms of storage, namely solid state, will help but not solve the problem. Probe memory is one of the most promising in terms of performance characteristics that could significantly boost storage system capability, but it isn't expected until at least 2015, and it's only being worked by a select group of storage vendors.

## Appendix A: Position Papers

The papers included in this section were the position papers requested of and submitted by workshop attendees. They represent a collection of ideas, architectures, and implementations surrounding the practice of running production storage systems. The papers facilitated discussion and identification of Best Practices for operation of file and archival storage systems.

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper**

<b>First Author Name</b> Affiliation e-mail address	<b>Second Author Name</b> Affiliation e-mail address
---	--

**Philippe DENIEL**  
**CEA/DAM**  
*philippe.deniel@cea.fr*

## **ABSTRACT / Summary**

CEA/DAM manages two compute centers : TERA100 (first SC in Europe) which is dedicated to classified applications and TGCC which is an open compute center for institutional collaboration (see <http://www-hpc.cea.fr/en/> for details). The management of produced data lead CEA's teams to deal with several specific issues, making them develop their own solutions and tools. This paper is focusing on fFiles's Lifetime, and metadata management.

## **INTRODUCTION**

CEA/DAM has been involved in HPC for many years. Because the compute has widely increased, the amount of produced data as drastically increased as well, making it necessary to have dedicated systems and dedicated teams to handle the architecture in charge of storing the data. This situation leads to several challenges : keeping data available to end users is of course one of them, but not the only one. With a huge amount of data comes a huge amount of metadata records. Consideration has to be taken to manage them. Last, the data kept areis not all of same value. When some files are criticals, others are not, but managing this aspect may be painful to the user who has thousands of files to delal with and sort. Tools have then to be made available to users to help them deal with information life cycle.

### **Quotas and retentions**

Ian Fleming said “diamonds are forever”, but for such files are not forever. The main issue there comes from the users. They produce lots of data (a daily production of 30 to 100 TB a day is a very common situation at CEA/DAM), but they often don't care about what the data become. This leads to a perpetually growing storage system where less than 1% of the content is accessed. Finally a big amount of files will never be read and are even totally useless once the run of the code is over (checkpoint/restart files for example). But the truth is this : if not forced, a user will never delete his files. Two main reasons for this:

- Lack of time
- Afraid fear of accidentally deleteing useful data

I suggest two solutions to handle this. The first is an old-fashioned Unix paradigm : quotas. The second is more sophisticated and is based on extended attributes to implement files's retentions.

Quotas usually works on a “space used” and a “used inodes” basis. File's size is not that critical (modern FS and storage system are huge today), but consideration on inodes are more interesting because they depict well the numbers of metadata records owned by a user. This is interesting in today's situation where the metadata footprint becomes the filesystem's limitation. Quotas are simple to set, manage and query (quotactl function in the libC, RQUOTAv2 protocol to be used jointly with NFS), but it has its inconvenient limitations. One of them is the distributed nature of the filesystem used in the HPC world. In a massively distributed product where data are spread across multiple data servers with parallel pattern, it becomes hard to efficiently keep a centralized place to keep user's information on quotas. Anyway, I suggest that when available, quotas are to be used because they are a simple way of setting limits to the users, making them aware of the amount of files and data that they own.

Files rRetentions is a another promising another way. The idea is to associate a specific metadata record to every file and directory. This is done by using extended attributes (aka xattr), which makes the assumption that the underlying storage system's namespace handles such a feature. This xattr will contain an information on the object's lifetime. This can be something like “this file will stop being of interest after a given date” or “this file can be considered useless isf not read/written during a defined period”. The key there is to have this metadata for every file (with users input). Specific tools will then audit the file system, produce a list of files to be deleted based on “retention policies”. The user will be warned (mail...) when some of their his files are candidates for deletion. Finally files are purged. This approach can lead to a virtuous circle : when producing data, users will take the habit to set the parameters to tell how long they'll require need the files, giving to the administrator input on their file's lifetime. This is good for the sysadm that who will save space on his storage system, and this is good for the user can who can schedule the deletion of his files, avoiding the painful task of cleaning his directories when quota limit is reached.

## **Metadata management**

Past challenges tofor filesystems wasere size : would the available resources be large enough to store everything I want to put in the system ? Then come performance consideration, and the idea that the users hate to wait to access their data. Right now, these aspects are addressed by modern filesystems (for example Lustre which is widely used at CEA) that are based on a distributed design relying on multiples data servers.

But many files means many metadata records and this can quickly be problematic, especially in a HPC environment. People who have once seen a single directory with hundreds of thousands of files in it know what I am speaking about. Beyond the technical consideration (big directories are an “edge” situation), the manageability of such exotic objects is a real problem : a single “ls -l” in it may last for hours.

Frequent filesystem audits (like those from CEA's RobinHood product (<http://robinhood.sf.net>)) helps in this : it becomes easy to identify “nasty” patterns in users directories and takes corrective actions. For example, the admin could decide to pack a big directory into a single tar file. Providing users with tools withusing “best practices enforcement” is also a way we follow. Copying Data copydata to the storage system goes is to delegated to a utilitytool that can decide to pack the data automatically.

Metadata volume is definitely an aspect to be seriously considered. Data volume issues have been solved

by striping the data. It may not be so easy to stripe metadata because they carry internal dependencies (a file belongs to a directory and can exist with several names if hard links are available) which may limit the algorithms. I actually believe that the main challenge for the filesystems on exascale compute center will be metadata management. Starting into considering this issue today, by setting limits to users to prevent them for to creating “file systems's monsters” and by teaching them the good practices is definitely something to be done today.

## **Conclusions**

The Exascale systems are coming tomorrow. Beyond the compute power's revolution, there is an incredible technical gap for the storage system. Data management will not be the greatest challenge, but metadata management will. The systems we will have at this time will store data that are produced today or have been generated in the past years. If we are not careful today, we will come to an excruciating situation in the future. And for sure, tomorrow's issues can be smoothed today by setting metadata's useage limits (quotas, retentions) and by providing users with tools to reduce metadata production.

**I/O Performance Measuring – White box test and Black box test-**  
U.S. Department of Energy Best Practices Workshop on File Systems & Archives  
San Francisco, CA September 26-27, 2011 Position Paper

**FUJITA, Naoyuki**

JAXA: Japan Aerospace Exploration Agency  
fujita@chofu.jaxa.jp

**SOMEYA, Kazuhiro**

JAXA: Japan Aerospace Exploration Agency  
someya.kazuhiro@jaxa.jp

**ABSTRACT**

The complexity of the component of file system/storage system (Thereafter, called the system.) is given to one of the reasons that the I/O performance measurement doesn't generalize. Here, let's think about the scene that discusses the I/O performance. Two cases are greatly thought. One is characteristic grasp of the system, and another is comparison between systems. We insist on using the white box test and the black box test properly in this paper. It is necessary to understand a detailed characteristic of the system by the white box test to guide an appropriate I/O operation to the system user and for the I/O tuning. On the other hand, when other systems and one system are compared, you should start from black box test comparison for a constructive discussion because of each system has each design and architecture.

**INTRODUCTION**

CPU benchmarking is widely discussed and some major benchmark suites<sup>1,2</sup> exist. However, I/O benchmarking is not more general than that of CPU. Therefore, generally speaking, I/O performance measuring and discussion are difficult. In this paper, we insist on using the white box test and the black box test properly.

As you know, white box test is a test done under the design and architecture of the system is understood. On the other hand, black box test is a test done without requiring them. In case of this time, design and architecture is a component, and the connection relationship of the system such as file systems and the storage devices.

**WHITE BOX TEST EXAMPLE**

This chapter shows examples of the white box test strategy and result. One result is a system that was operating in 2000(Thereafter, called 2000 System), another one is a system that has been operated since 2010(Therefore called 2010 System).

**White box test strategy**

As a number of nodes increases in HPC system, the system design and architecture becomes complex and changes it's characteristic. We propose layered benchmark as white box test, and show some results. Layered means device level(Measuring Point 1), local file system level(Measuring Point 2), network file system level(Measuring Point 3), and FORTRAN level(Measuring Point 4). Fig. 1 shows measuring points on recent HPC System.

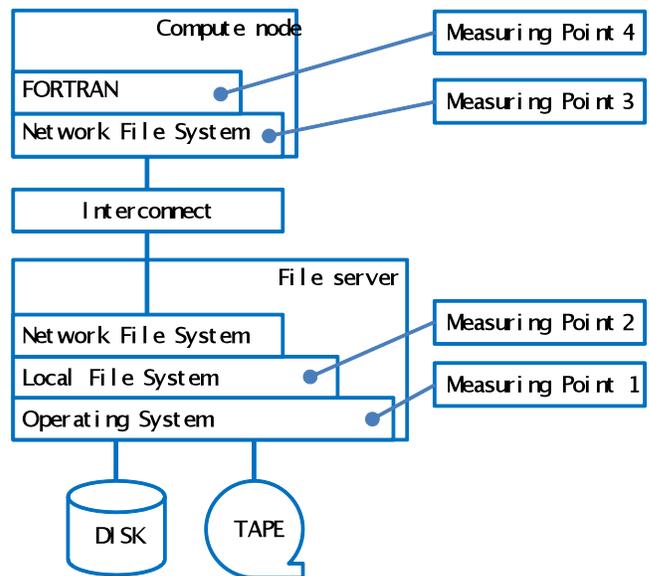


Fig. 1 HPC System I/O measuring points

## White box test results

Fig. 2 shows 2000 and 2010 system configuration chart. Each system has Compute nodes, Interconnect, which are X-bar switch and IB switch, and File server(s). So there is no change in a basic composition. As a partial change point, 2010 System has clustered file servers and storage devices are attached via FC-SAN switches. Fig. 3 shows the result of the layered benchmark of these two systems. There are some

bottleneck points in a system. To analyze a bottleneck, we aim at file system cache, interconnect bandwidth, and DAS/SAN bandwidth and its connection relationship design.

Device level benchmark showed similar result between two designs except disk write performance. But the characteristic of network file system level was very different. In this case, it depends on file system cache effect.

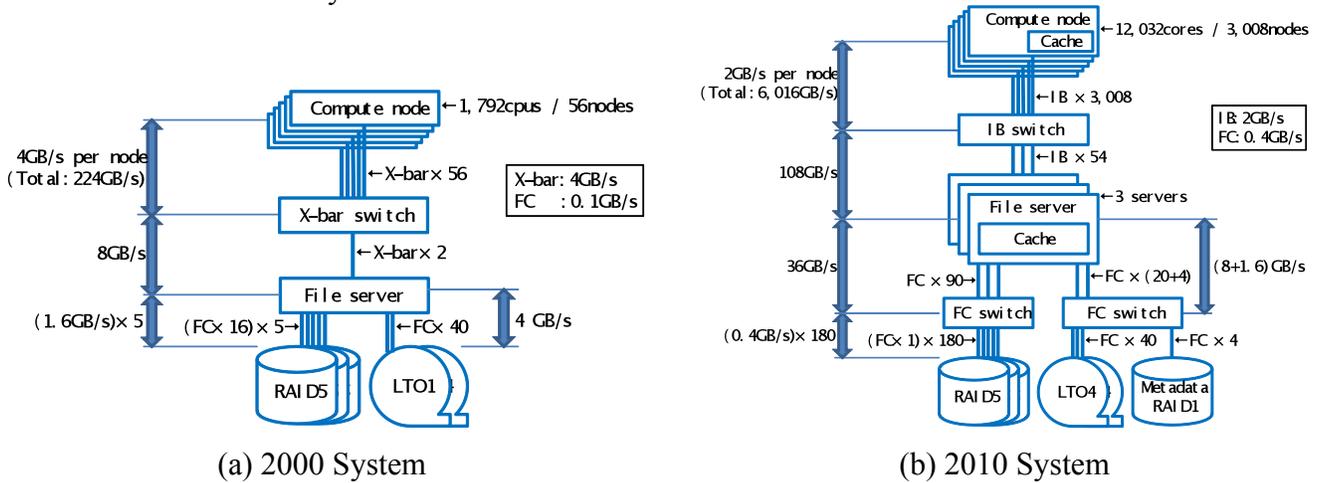


Fig. 2 File system and storage system configuration

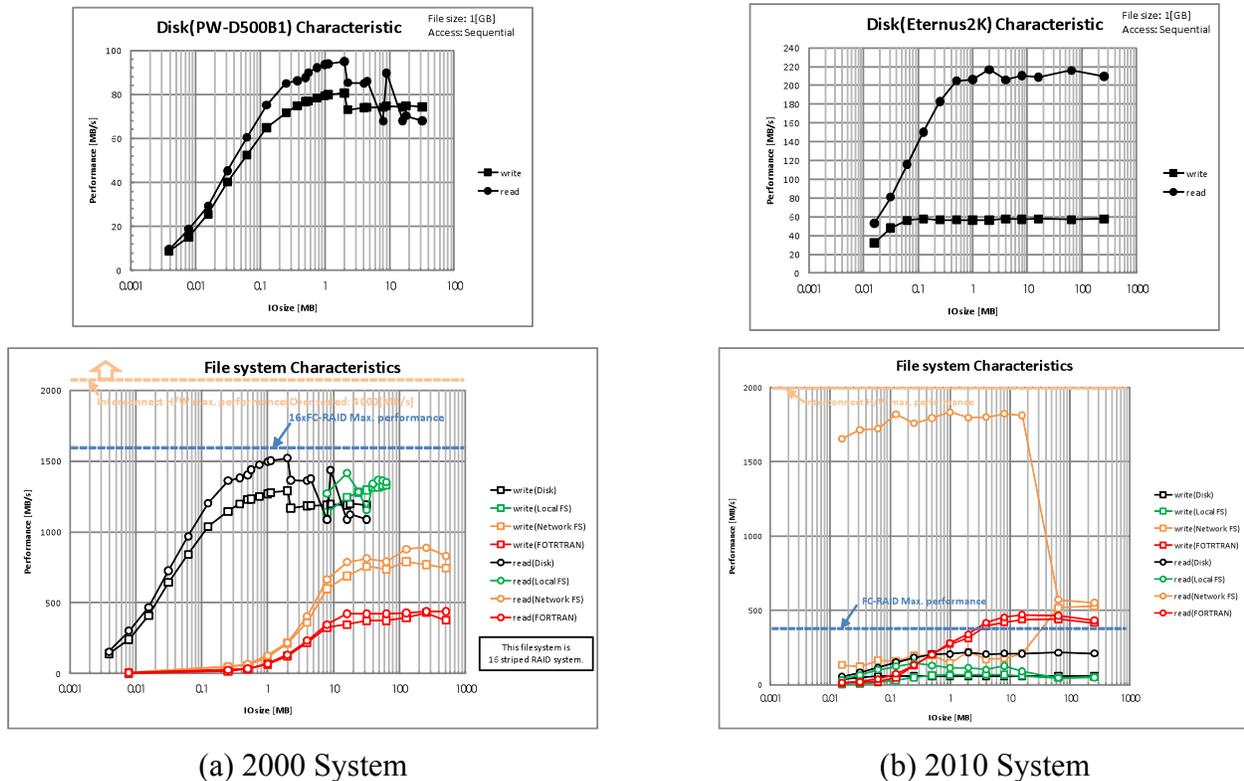


Fig. 3 Layered benchmark results

## BLACK BOX TEST EXAMPLE

This chapter shows examples of the black box test strategy and result.

### Black box test strategy

As we said, each system has each design and architecture, so the comparison of simple I/O performance is not significant. But when we discuss about I/O performance, especially compare with several systems, first of all, it should take a general view of a rough performance. A common tool to measure the file system performance that is appropriate for the measurement of large-scale storage doesn't exist, and the performance measurement tool is made individually in each system and the performance is evaluated individually. In addition, as a peculiar operation to the file system will be needed, it is difficult to compare it with the performance measurement result in another file system. Then, we model the measurement tool and the measurement item, and propose the method of simply diagnosing the characteristic of the large-scale storage system based on the result of a measurement that uses the tool<sup>3</sup>.

### Objective

It aims at the thing that the following two points can be measured generally in a short time.

#### (1) Checkup of installed system

Whether the performance at which it aimed when the system administrator installed the system has gone out is examined.

#### (2) Routine physical examination under operation

Grasp of performance in aspect of user. The operation performance is measured.

### Diagnostic model

#### (1) Checkup of installed system

-Maximum I/O bandwidth performance

Read performance immediately after Write. It is assumed that data are in the client cash.

-Minimum I/O bandwidth performance

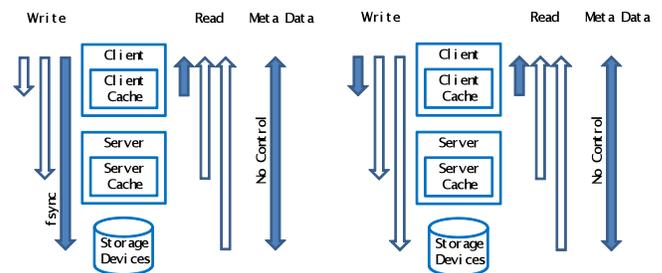
Fsync is assumed after Write and the multiplication cash assumes all things forwarded to the real storage device.

-Meta data access performance

The presence of cash is not considered (Because the cache management cannot be controlled in the black box test).

#### (2) Routine physical examination under operation

This diagnosis tool is regularly made to work while really operating it, and the state grasp is enabled. In this case, it is assumed to gather the maximum performance (cash hit performance) from the viewpoint of the user aspect. An enough prior confirmation by the system administrator is necessary to make the measurement tool work regularly. Moreover, customizing the measurement tool (measurement downsizing etc.) might be needed. The measurement model is shown in Fig. 4.



(1) Checkup of installation (2) Routine examination

Fig. 4 Measurement model

### Measurement tool and item

Using IOR.

(a) Large-scale data transfer (Throughput performance: Constant amount of file for each process, large I/O length)

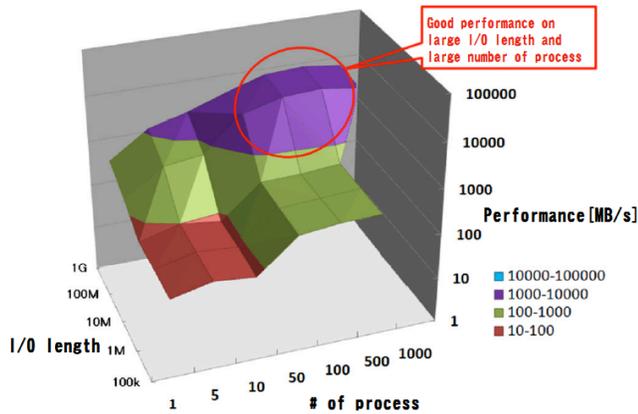
(b) Constant volume of data (Throughput performance: Small file size, small I/O length)

Using mdtest.

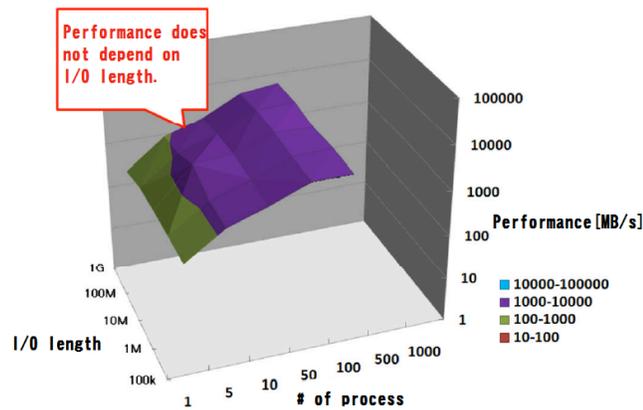
(3) Meta data access (response performance)

## Black box test results

Fig. 5 shows the result example of large-scale data transfer on System A and B.



(a) System A



(b) System B

Fig. 5 Large-scale data transfer results

Both white box test and black box test should be used when we manage file system and storage system.

## REFERENCES

1. Standard Performance Evaluation Corporation  
<http://www.spec.org/benchmarks.html>
2. TOP500 supercomputer Sites,  
<http://www.top500.org/>  
Scientific System Society, "Large-scale storage system working group report," 2011,(in Japanese)

## CONCLUSIONS

"What should be measured?"

Each layer benchmark should be done when we want to know the characteristics of the system. The Layers are device level, local file system level, network file system level, and FORTRAN level. This kind of measurement will be done as a white box test.

Modeled measurement item and tool should be used, when we want to compare several systems. The result is a starting point of the discussion. This kind of measurement will be called black box test.

## LA-UR-11-11388

Approved for public release; distribution is unlimited.

Title: U.S. Department of Energy Best Practices Workshop on File Systems & Archives Position Paper

Author(s): Torres, Aaron  
Scott, Cody

Intended for: U.S. Department of Energy Best Practices Workshop on File Systems & Archives, 2011-09-26/2011-09-27 (San Francisco, California, United States)



**Disclaimer:**

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By acceptance of this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper**

**Aaron Torres**

Los Alamos National Laboratory, HPC-3  
agtorre@lanl.gov

**Cody Scott**

Los Alamos National Laboratory, HPC-3  
cscott@lanl.gov

**ABSTRACT / SUMMARY**

**As HPC archival storage needs continue to grow, we have started to look at strategies to incorporate cheaper, denser, and faster disk as a larger part of the archival storage hierarchy. The archive in Los Alamos National Laboratory's Turquoise open collaboration network has always used a generous amount of both fast and slow disk in addition to tape. Lessons learned during Road Runner Open Science pointed to the need for large amounts of cheaper, slower disk for storage of small to medium sized files and faster disk in order to store and quickly retrieve file metadata. New advances in the last year may signal another transition that altogether eliminates the need for migrating small and medium files to tape. Improvements in disk speed, particularly solid state devices (SSDs), also allow us to operate on billions of files in a reasonable span of time even as archives continue to grow.**

**INTRODUCTION**

The Open Science simulations run on the Road Runner supercomputer at Los Alamos National Laboratory (LANL) in 2009 provided the opportunity to test an archive based on commercial off the shelf (COTS) components. For this archive, we chose the General Parallel File System (GPFS) and Tivoli Storage Manager (TSM) due to robust metadata features, fast data

movement, flexible storage pool hierarchy and migration, and support for a multitude of disk and tape options [1].

This archive joins a long history of archival storage at LANL, including the Central File System (CFS) and High Performance Storage System (HPSS). Thanks to administrative diligence, we have or can recreate records about usage patterns of these archives. One similarity we keep seeing in large archives, with the COTS archive being no exception, is that we primarily store numerous small to medium sized files rather than storing large to huge files. As of May, 2011, HPSS at LANL houses nearly 163 million files with total size of 19.6 PB with an average file size of 131.5MB [2]. NERSC publishes similar statistics with an archive housing over 118 million files and 12 PB for an average size of 109MB [3].

**ROAD RUNNER LESSONS LEARNED**

When designing for archival storage, one often considers the extreme case for file size. In HPC this generally means designing for enormous files on the order of terabytes for current supercomputer sizes. In practice, however, we see a tremendous amount of small to medium files, especially with users performing n-to-n writes or using the Parallel Log-structured File System (PLFS) to effectively convert n-to-1 writes to n-to-n [4]. In the case of Roadrunner, 20 million 8-16 MB files were archived in one weekend [5].

For the COTS archive, this proved to be the largest pain point since the Hierarchical Storage

Manager (HSM) feature of TSM does not currently support aggregating smaller files together when moving them to tape, resulting in poor performance. Users could aggregate their own files using the “tar” command, but they cannot be relied on to do this for all cases. Another option would be to put file aggregation into an archive copy tool such as how the LANL-developed Parallel Storage Interface (PSI) does with the Gleicher developed HTAR [6]. However, doing so breaks POSIX compliance because no other standard file system tool can read or write files aggregated in this way. One of the design goals of the COTS archive was to leverage as much standard software as possible.

For the COTS archive, moving small files to tape without a transparent file aggregation technique did not make sense. So, small files are kept on RAID 6 disk arrays and backed up to tape. RAID provides recovery from minor amounts of single disk failure, and the tape backup provides disaster recovery. Moreover, TSM's backup function does support aggregating small files before sending them to tape.

The COTS archive has 122 TB of fast fiber channel disk to act as a landing area for new files and 273 TB of SATA disk for files under 8 MB to be moved to. Finally, it has 3 PB of tape for files over 8 MB and for the backups of the SATA disk pools. Currently, the archive houses over 107

million files with a total size of 2.1 PB and an average size of 21.22 MB according to our latest statistics as of August, 2011. As shown in Figure 1, 97 million files are less than or equal to 8 MB. This indicates that the general case for our archive is large amounts of smaller files.

## RECENT ADVANCEMENTS

The recent explosion of “cloud” backup providers like Mozy, Backblaze, and others lead to questions about how we store large amounts data and if we are doing it in the most cost effective way. For a cloud-based backup service, density and uptime are the two primary driving forces because users continue to back up ever larger amounts of data as they put more of their life on the computer in terms of photos, videos, etc. and data may be backed up or restored at any time. These are also motivating factors for HPC archives. On September 1<sup>st</sup>, 2009, Backblaze posted an entry to their company blog describing their Backblaze Pod capable of storing 67 TB of data in a 4U enclosure using 47 one terabyte drives for \$7,867, or 11.4¢ per gigabyte [7]. On July 20<sup>th</sup>, 2011, they posted an updated entry now indicating that they can store 135TB in 4U using 47 three terabyte drives for \$7,384 or 5.3¢ per gigabyte [8]. Also, Backblaze notes they have deployed 16 PB of disk in the last 3 years [8]. In terms of raw storage, that is within striking distance of the size of LANL’s largest HPSS archive at nearly 20 PB.

On the other end of the spectrum, eBay recently replaced 100 TB of SAS disk with SSD [9]. They did this to speed up virtualization and reduce the size of their disk farm. They had a 50% reduction in standard storage rack space and a 78% drop in power consumption by moving to SSD. Although it is impossible for an HPC archive to take this approach, it is possible to replace portions of the total system for tremendous benefits.

An example of using SSD in a storage hierarchy is IBM's recent efforts at speeding up GPFS using SSD [10]. By storing GPFS metadata on SSD, IBM saw a 37 times speed improvement for metadata operations and was able to scan 10 billion files in 43 minutes. For comparison, it

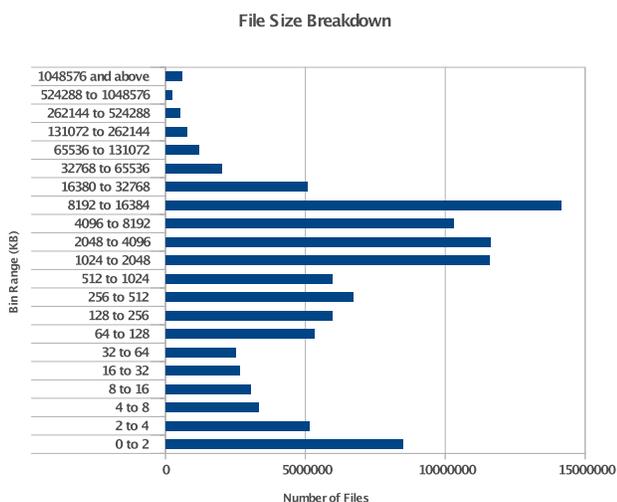


Figure 1. File Size Breakdown of COTS Archive.

takes roughly 20 minutes to scan the 120 million files on the LANL COTS archive using eight 15,000 RPM fiber channel disks in four RAID 1 stripes.

### **CHEAP, DENSE DISK**

The growth of density in hard disks shows little sign of slowing down. In the 2 years between Backblaze posting blog entries about their Pod system, the cost of the Pod actually went down even though the raw capacity of the 4U box doubled. Hard disks in the 4 TB range are on the horizon for desktops and servers in the next year [11], and even laptops are moving to 1 TB disks [12]. The Hitachi 3 TB drives used by Backblaze can be purchased for \$120-130 from a variety of retailers [13]. For comparison, an LTO5 tape that holds 1.5 TB of uncompressed data costs approximately \$60 [14]. For the same capacity, the disk costs as much as the tape.

One interesting move by companies like Backblaze is that they use consumer level hard drives instead of “enterprise ready” drives. Such drives are substantially cheaper; with the enterprise version of the Hitachi 3TB drives costing over \$320-350 per drive as of August 23, 2011 [15]. Backblaze also takes advantage of the manufacturer's 3 year warranty to get a replacement disk if one fails rather than an expensive maintenance contract. HPC archives might be able to leverage the same kind of disk drive by taking into account the disk failure protection afforded by RAID 6 and by having a tape backup of whatever is stored on such disks.

Unlike Backblaze and their Pod, we do not want to be in the custom hardware business. So, we looked for existing commercial hardware that could get the same density of disk. We found the SuperMicro SuperChassis 847E26-RJBOD1 [16]. It is a storage chassis that can support 45 disk drives in 4U. It does not have a built-in motherboard like the Backblaze Pod to manage the disk, but the COTS archive already has machines in its GPFS cluster that can easily take a RAID card with an external SAS connector to plug into this storage expansion chassis. Filling the chassis with 3 TB consumer level disk drives

and including the RAID card costs approximately \$12,000, or 8¢ per gigabyte.

The idea behind these enormous disk pools is not to completely replace tape, but to adjust the size of file that gets moved to tape. It is entirely feasible today to change the threshold used in the COTS archive from 8 MB to 1 GB with the current price of disk. With this change, we can move all files 1 GB and larger to tape. This file size is also much closer to the size of file necessary to get a tape drive up to peak streaming speed based on internal testing done at LANL. In addition, backing up any file that will stay resident on a disk greatly reduces the fear of failed tapes when storing enormous quantities of small files.

One argument against disk in archive is that archives are usually “write once and read never,” so it does not make sense to “waste” power and cooling on spinning disk for data that may never get read. For the COTS archive, the ability of GPFS to move data to different types of storage (ie, fast disk, slow disk, and tape using TSM) based on arbitrary criteria like dates could be leveraged to move rarely or never read data to tape. The tape performance hit is acceptable because the data is essentially “cold”. Similarly, a large disk pool can be used to stage data for reading if a user knows he or she will be pulling some set of data from the archive.

### **FAST DISK CACHE**

As HPC archives ingest ever more data because of exascale supercomputers, the metadata will probably become more and more important. At some point in the not to distant future, users will want to search the archive on metadata instead of being forced to create complex directory hierarchies to find the files they are interested in. An example query could be “find all the checkpoint files that were copied to the archive within the last 3 days.” Thus, it is also important to quickly search an archive’s metadata, whether it is in a file system like GPFS or a database like HPSS using DB2. Here is where faster disk systems like SSD can be used to great effect in HPC archives. As mentioned previously, IBM’s

testing of storing GPFS metadata on SSD and being able to scan billions of files in less than an hour shows how such fast disk can be very useful.

Another pilot program at LANL is testing metadata performance on SSD using the GPFS COTS archive as a basis for the number of files and types of files stored. Having the ability to quickly scan the metadata of the entire archive provides many benefits, particularly to future research projects and in data management including ongoing work to index and quickly search archive metadata.

In addition, as SSD storage becomes cheaper and denser, it may eventually be possible to replace our fast disk cache, currently consisting of fiber channel disk, with a large pool of SSD similar to how eBay replaced their SAS disk environment. With our current data requirements this is still cost ineffective, but it is worth examining and testing now for the future.

## CONCLUSIONS

There are many advantages to having a large, easy to manage pool of disk. When raw speed is not a requirement of this disk, there are solutions available to procure, maintain, and deploy a tremendous amount of disk cost effectively that compares very favorably to the cost of tape. Taking advantage of faster disk like SSD for metadata and disk cache will also benefit future HPC archives. The LANL COTS archive is in a unique position to test and potentially deploy some of these newer solutions in-place with limited negative effect to users.

## REFERENCES

1. Hsing-bung Chen, Grider Gary, Scott Cody, Turley Milton, Torres Aaron, Sanchez Kathy, Bremer John. *Integration Experiences and Performance Studies of A COTS Parallel Archive System*. IEEE Cluster Conference, 2010
2. *Accounting Data*. <http://hpss-info.lanl.gov/AcctProcess.php>
3. *Storage Trends and Summaries*. <http://www.nersc.gov/users/data-and-networking/hpss/storage-statistics/storage-trends/>
4. Bent John, Gibson Garth, Grider Gary, McClelland Ben, Nowocznski Paul, Nunez James, Polte Milo, Wignate Meghan. *PLFS: A Checkpoint Filesystem for Parallel Applications*. Super Computing, 2009
5. Scott, Cody. *COTS Archive Lessons Learned & Fast Data Pipe Projects*. JOWOG-34
6. Gleicher, Michael. *HTAR – Introduction*. <http://www.mgleicher.us/GEL/htar/>
7. Nunfire, Time. *Petabytes on a Budget: How to build cheap cloud storage*. <http://blog.backblaze.com/2009/09/01/petabytes-on-a-budget-how-to-build-cheap-cloud-storage/>
8. Nufire, Tim. *Petabyes on a Budget v2.0: Revealing More Secrets*. <http://blog.backblaze.com/2011/07/20/petabytes-on-a-budget-v2-0revealing-more-secrets/>.
9. Mearian, Lucas. *Ebay attacks server virtualization with 100TB of SSD storage*. [http://www.computerworld.com/s/article/9218811/EBay\\_attacks\\_server\\_virtualization\\_with\\_100TB\\_of\\_SSD\\_storage](http://www.computerworld.com/s/article/9218811/EBay_attacks_server_virtualization_with_100TB_of_SSD_storage).
10. Feldman, Michael. *IBM Demos Record-Breaking Parallel File System Performance*. [http://www.hpcwire.com/hpcwire/2011-07-22/ibm\\_demos\\_record-breaking\\_parallel\\_file\\_system\\_performance.html](http://www.hpcwire.com/hpcwire/2011-07-22/ibm_demos_record-breaking_parallel_file_system_performance.html)
11. Shilov, Anton. *Samsung Shows Off Prototype of 4TB Hard Disk Drive*. [http://www.xbitlabs.com/news/storage/display/20110308081634\\_Samsung\\_Shows\\_Off\\_Prototype\\_of\\_4TB\\_Hard\\_Disk\\_Drive.html](http://www.xbitlabs.com/news/storage/display/20110308081634_Samsung_Shows_Off_Prototype_of_4TB_Hard_Disk_Drive.html)
12. Altavilla, Dave. *A Terabyte For NotebooksL WD Scorpio Blue 1TB Drive*. <http://hothardware.com/Reviews/1TB-WD-Scorpio-Blue-25-HD-QuickTake/>
13. *HITACHI Deskstar 0S03230 3TB 5400 RPM 32MB Cache SATA 6.0Gb/s 3.5" Internal Hard Drive -Bare Drive*. <http://www.newegg.com/Product/Product.aspx?Item=N82E16822145493>
14. *IBM – LTO Ultrium 5 – 1.5 TB / 3 TB – storage media*. <http://www.amazon.com/IBM-LTO-Ultrium-storage-media/dp/B003HKLHZC>
15. *HITACHI Ultrastar 7k3000 HUA723030AKLA640 (OF12456) 3 TB 7200 RPM 64MB Cache SATA 6.0Gb/s 3.5"*

*Internal Hard Drive -Bare Drive.*

<http://www.newegg.com/Product/Product.aspx?Item=N82E16822145477>

16. *SuperChassis 417E16-RJBOD1.*

<http://www.supermicro.com/products/chassis/4U/417/S417E16-RJBOD1.cfm>

# U.S. Department of Energy Best Practices Workshop on

## File Systems & Archives

San Francisco, CA

September 26-27, 2011

### Position Paper

**Nicholas P. Cardo**

National Energy Research Scientific Computing Center  
Lawrence Berkeley National Laboratory  
cardo@nersc.gov

#### ABSTRACT

Disk quotas are a useful tool for controlling file system space consumption. However, each file system type provides its own mechanism for displaying quota usage. Furthermore, each file system could display the information differently. Unifying how quota information is reported would simplify the user's experience.

Also having quotas span multiple file systems would provide users some flexibility in storage usage.

#### INTRODUCTION

The use, management, and enforcement of disk quotas is often difficult to interpret at the user's level as well as being too rigid of an enforcement mechanism.

#### Identification of the issues

While disk quotas are extremely useful in managing disk space, they are often complicate, hard to understand, counter productive to the user community.

Lets first examine quota-reporting utilities. For IBM's General Parallel File System (GPFS), the command `mmlsquota` is used

While standard Linux utilizes the `quota` command.

```
$ quota Disk quotas for user juser (uid 500):
Filesystem blocks quota limit grace files quota limit grace
/dev/efs0 2 100 200 2 10 20
```

So now there are three different commands each with a different syntax showing different information. Instructing users how to interpret the results can be quite involved. This is especially true when the data is closely examined to exactly what the users really care about. At that level all that matters is what is being consumed and what the limit is. Lustre is quite detailed in its output and provides space consumption information down to the Object Storage Target (OST). This presents a case of information overload as file systems could have 100's of OSTs and each one represents one line of output. But the real question is do users really need to see this.

```
nid00011:-> mmlsquota home1
Block Limits
Filesystem type KB quota limit in_doubt grace | File Limits
files quota limit in_doubt grace Remarks
tlhome1 USR 29491548 41943040 41943040 34032 none | 7680 1000000 1000000 429 none fshost
```

to display quota limits and usage.

For Lustre, quota information is obtained with the command `lfs quota`.

```
nid00011:-> lfs quota -u juser /scratch
Disk quotas for user juser (uid 500):
Filesystem kbytes quota limit grace files quota limit grace
/scratch 2666948 0 0 83 0 0
sc-MDT0000_UUID 120 0 0 83 0 0
sc-OST0000_UUID 4 0 0
...
```

File system quota reporting also is highly dependant on the file system architecture, and provides details unique to that file system. GPFS provides the `mmrepquota` command producing:

Block Limits							File Limits				
Name	type	KB	quota	limit	in_doubt	grace	files	quota	limit	in_doubt	grace
fuser1	USR	17684	41943040	4194304	0	none	72	1000000	1000000	0	none
fuser2	USR	180	41943040	4194304	0	none	32	1000000	1000000	0	none

Lustre provides no such reporting functionality. Standard Linux provides the `repquota` command for reporting operations.

```
# repquota /quota
*** Report for user quotas on device /dev/fs1
Block grace time: 7days; Inode grace time: 7day
Block limits      File limits
User      used soft hard grace used soft hard grace
-----
fuser1    -- 1204 0 0          5 0 0
fuser2    -- 10 100 200      9 10 20
```

The more file systems types that are present on a system, the bigger the problem becomes.

Along the same lines is that the actual underlying quotas are per file system and cannot be aggregated across multiple file systems. Users must be granted quotas on each individual file system and managed by that file systems quota utilities.

### Statement of Position

Quota utilities should be externalized from the products where each vendor is encouraged to contribute to them to support their file system. Furthermore, each vendor should supply a standardized API call to retrieve or manipulate disk quotas. It is recognized that each file system may need to present details not applicable to other file systems. In this case, the utilities should use extended flags to control the operation.

The application of disk quotas needs to be externalized from file system. While the accumulation of accounting data needs to be within each file system, the enforcement of quotas can be externalized. This would allow for a single disk quota to span multiple file systems regardless of file system type. A kernel module could open the quota file, holding the file descriptor open for a system call to access directly from within the file systems.

### SUPPORTING DOCUMENTATION

Many quota operations can be easily externalized. Each of the file systems mentioned already provide an API call that can be used to retrieve or manipulate disk quotas. GPFS provides `gpfs_quotactl()`, Lustre provides `llapi_quotactl()`, and standard Linux provides `quotactl()`. This shows that the underlying interface is already in place, but unique to that file system. Linux already can differentiate between the file system types. The mount table contains the field `mnt_type`, which identifies the underlying file system. So why can't a single form of `quotactl()` which utilizes the `mnt_type` to differentiate the file system types be put in place?

The answer is, it can.

### User Quota Report

The first utility to make use of this capability essentially replaces `mmfsquota`, `lfs_quota`, and `quota`, with a single utility that can display quota information to the users, regardless of file system type.

In this example, the `scratch` and `scratch2` file systems are Lustre, while `project`, `common`, `u1`, and `u2` are GPFS. This utility utilizes `getmntent()` to read the mount table in order to access `mnt_type` which is used to determine the file system type. Then the appropriate `quotactl()` system call is used to access the quota information for the file system. The data is then normalized to a consistent format and presented to the users.

Displaying quota usage for user fuser1:								
FileSystem	Space (GB)				Inode			
	Usage	Quota	InDoubt	Grace	Usage	Quota	InDoubt	Grace
scratch	3	-	-	-	83	-	-	-
scratch2	24	-	-	-	334	-	-	-
project	0	-	0	-	1944	-	0	-
common	0	-	0	-	11	-	0	-
u1	28	40	0	-	7680	1000000	429	-
u2	0	40	0	-	2	1000000	0	-

### File System Quota Report

Quota reporting at the file system level is very useful for determining the top consumers of the resources. The issue of different file systems reporting different information can be easily overcome. However, the lack of the capability to simply loop through all quota entries a major obstacle had to be overcome. The solution used was to loop on all users to get their usage information. The downside is that if a user is removed from the system and is consuming resources, it will never be reported.

In a similar manner as in the user quota reporting utility, `statfs()` is used to get the `f_type` of the file system. This is then used to determine the correct `quotactl()` to use for that file system. The user list is obtained simply by looping on `getpwent()`.

```
Filesystem: /scratch2
Report Type: Space
Report Date: Wed Sep 7 07:14:37 2011
```

Username	Space (GBs)		Inode	
	Usage	Quota	Usage	Quota
fuser1	8262	0	663560	0
fuser2	7824	0	225937	0
fuser3	5593	0	216674	0
fuser4	4548	0	111542	0
fuser5	2171	0	436872	0

The report can be sorted either by space or by inodes. Reported is a simple and easy to read output that is the same regardless of file system type.

### Quotas Spanning File Systems

Enforcing file system quotas external to the file system opens up a flexibility to customize the effects when quotas are reached as well as the opportunity to span file systems. Normally quotas are set up with a soft limit that can be exceeded for some grace period while not exceeding a hard limit. The effect of reaching the hard limit is usually the I/O being aborted with the error `EDQUOT` (quota exceeded). Running a large-scale computation for several days that aborts due to quota limits being reached seemed a bit counter productive, not to mention the loss of valuable computational time. Rather than to terminate the run, a better solution would be to allow it to run to completion while preventing further work from starting. A simple check at job submission and another at job startup can prevent new work from being submitted or started without the loss of computational time.

In addition to the flexibility in how to enforce quota limits, the ability to combine usage information from multiple file systems is enabled allowing for a single quota to span file systems. The

process is to simply retrieve the utilization from the desired file systems, accumulate it, and then evaluate it. For batch jobs, this can be performed in submit filters or prologues. This is in production at job submission time. If users are over their quota, they will receive a message:

```
ERROR: your current combined scratch space usage of 6 GBs exceeds
your quota limit of 4 GBs.

You are currently exceeding your disk quota limits. You will
not be able to submit batch jobs until you reduce your usage
to comply with your quota limits.
```

This change has improved the users experience on the system while keeping resource consumption in check.

Externally to the file system, an infrastructure was needed to support the ability to grant a quota that applies to all users, as well as exceptions. Some projects simply require more storage resources than is desired to grant to all users. Having a default quota is easy as it is a value that applies to all users. The challenge was the ability to override this while tracking those with extended quotas.

Another utility was created to manage a data file used to track quota extensions.

```
> chquota -R
----- Space Quota ----- Inode Quota -----
Username Q GigaBs Expiration Ticket Inodes Expiration Ticket Filesystem
-----
fuser1 U 10240 01/10/2012 110112-000033 5000000 01/10/2012 110112-000033 /scratch
fuser1 U 10240 11/15/2011 110714-000039 - - - - - /scratch
```

Not only are the new limits for space and inodes recorded, but also the expiration dates for the extension as well as the problem tracking ticket. From a single report, a clear understanding of all existing quota extensions can be ascertained. A feature of this utility is the ability to automatically remove expired quotas. Each

not via cron, the command is run to evaluate all quota extensions and remove any that have expired.

Another feature that is targeted to improving the users experience is the ability to inform if a quota extension is about to expire.

```
chquota: your 6 GB space quota on /scr expires on 09/09/11
(110901-000001)
```

The number in the parenthesis is the trouble ticket number tracking the request. This can be placed in login scripts to inform users each time they login to the system.

## CONCLUSIONS

Simplifying disk quotas improves usability, reporting, and the user's experience on the system all while controlling consumption of resources.

File system vendors should be encouraged to align their quota implementations into a single command set of tools that provide a consistent interface, regardless of file system type. Until that happens, centers should adopt a plan to develop such tools as they improve the user's experience. Taking this one step further, all centers should adopt the practice of putting these tools into service creating consistency across centers. Many users perform their calculations at several centers and having a consistent set of tools will enhance their ability to work effectively.

By externalizing disk quota enforcement to job submission, users are forced to keep their resource consumption in check without the risk of losing a run due to quota limits. As a result, the computational resources are much more effective as no time is lost due to calculations being cut short when quota limits are hit.

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper**

**Wayne Hurlbert**  
Lawrence Berkeley National Laboratory  
wehurlbert@lbl.gov

**ABSTRACT / SUMMARY**

**This position paper aims to provide information about techniques used by the Mass Storage Group at the National Energy Research Scientific Computing Center (NERSC) to accomplish technology refresh, system configuration changes, and system maintenance while minimizing impact on users and maximizing system availability and reliability. In particular, it addresses the Center's position that shorter, scheduled outages for archival storage system changes, occurring at familiar times, minimizes the likelihood of unscheduled or extended outages, and so minimizes impact on users.**

**INTRODUCTION**

For the purposes of this discussion, it is taken as given that much of the activity involved in technology refresh, along with configuration changes and other system maintenance, requires systems to be off-line. NERSC's approach to these activities in the archival storage systems, is largely driven by the need to minimize the impact on our users. The notion of minimum impact encompasses the scheduled activity itself, potential fallout from the activity, and the concept of preventive maintenance. This has resulted in a conservative attitude toward system maintenance that favors incremental rather than radical change. In the following I will discuss the motivation and

benefits of this approach, and mention some of the real world steps taken at NERSC to implement the approach.

**Little by Little**

While it can be tempting to "just do it", an incremental approach to technology refresh and other system maintenance activities is usually a viable alternative to the more significant outages often required to accomplish the changes in a single sitting.

Types of projects for which this approach might be helpful include:

- Server upgrades or replacement.
- Significant application, OS, or layered software upgrades.
- Replacement or reconfiguration of disk and tape resources.
- Replacement or reconfiguration of large infrastructure components such as SAN switches.

These projects will often take many hours, sometimes days, to accomplish and run a relatively high risk of unanticipated problems or complications.

An incremental approach indicates that these larger projects be broken up into smaller pieces which can be accomplished in an independent and sequential manner. Naturally, there are projects where this is not possible, for various reasons; our

finding is that the reasons are typically not technical in nature.

### **The Benefits**

There are several benefits provided by this approach:

- Less complexity of the tasks executed during an outage, which means a reduction in the likelihood of human mistakes in planning or execution of the tasks.
- Lower risk of aborted or extended outages due to unexpected or unanticipated complications. For example, because fewer tasks are being undertaken, there is a smaller window for hardware failure if devices or servers are being power cycled. Naturally, a device can fail during either an incremental activity or a major project, but the impact on workflow is likely to be smaller, and the impact on the user is likely to be less significant in terms of total time for the outage.
- Easier back out in the case of the need to abort the maintenance activity due to unexpected or unanticipated events.
- Lower likelihood of human error due to the fatigue and stress which usually occur during significant projects.
- When compared with forklift upgrades, lower risk of subsequent fallout due to as yet undiscovered bugs or defects. This is particularly true, obviously, for newer products.
- Where desirable, allows for completing system-down activities during business hours, because of the shorter outages. Business hours may be required in order to insure access to outside expertise.

### **User Expectations**

The incremental approach to performing system maintenance subscribes to the notion that shorter, more frequently scheduled outages will ensure a more stable system, which will better serve users.

Outages should be scheduled for a standard day and time, even if not at standard intervals e.g. weekly, with the intent that users will come to expect that time period and plan around it. For instance, on one end of the spectrum, users can simply plan to not run during the normal hours, on the normal day for outages. However, NERSC does provide a programmatic, network based mechanism for automated jobs to check system availability.

Further, NERSC has developed an effective protocol for suspending user storage transfers during short outages. Referred to as “sleepers”, user interface tools on the compute machines look for lock files which cause these clients to loop on the system sleep call until the lock file disappears. The result is that many user jobs simply pause until the outage is completed.

In annual user satisfaction surveys at NERSC, the archival storage resources typically receive high scores with regard to system availability and reliability. [1] [2]

### **Preventive Maintenance**

Preventive maintenance, in the sense of avoiding unscheduled outages and the associated user interruptions, can be seen as primarily concerned with restarting, rebooting, and/or power cycling equipment. These activities usually take relatively little time, and fit nicely with shorter, more frequently scheduled outages. Examples include:

- Reboot to validate configuration changes made while the system is live, even if a reboot/power cycle is not strictly required.
- Reboot to flush out pending hardware failures, or to reset hardware that is in a confused state.
- Rebooting or power cycling also helps maintain familiarity with the way systems and devices behave during power-down and power-up.
- Restarting applications, and less importantly these days restarting operating systems, can

help avoid outages due to software defects such as memory leaks.

- Build rather than copy: when locally built software must be installed on multiple servers, building it on each server validates the installation and configuration of layered software (in addition to allowing debug activities on the various servers).

### **Example**

Project: application upgrade on the current production server hardware, which requires OS and/or layered software upgrades.

The NERSC storage group will typically build a new system disk, from the ground up, on a second disk in the production server.

This will usually involve an outage to install the new OS followed by one to several 2-3 hour outages to install, build, configure, and test (as appropriate) layered software and application code. Each of these outages will involve a reboot to the second system disk for the work to be done, followed by a reboot back to the production disk.

This activity is usually spread out over a number of weeks, and is typically interleaved with other

activities that may, or may not, involve preparation for the upgrade.

The upgrade is finalized by rebooting to the new system disk and performing any remaining activities required before going live.

### **CONCLUSIONS**

A conservative approach to system outages for technology refresh, system reconfigurations, and other maintenance can be accomplished through a policy which uses multiple short outages to perform the work incrementally. This promotes greater system stability and minimizes the number of unscheduled outages, resulting in better service to users.

### **REFERENCES**

1. NERSC 2010 High Performance Computing Facility Operational Assessment.
2. NERSC User Surveys.  
<http://www.nersc.gov/news-publications/publications-reports/user-surveys>.

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper**

**Todd M. Heer**

Lawrence Livermore National Laboratory  
theer@llnl.gov

**ABSTRACT / SUMMARY**

**Specific to our center, archival data integrity emanates from dual copy files, intensive preproduction environment analysis, and ongoing HSM verification testing. Availability falls from the judicious application of a redundancy model. Efficiency can be obtained by leveraging the large procurements part and parcel of HPC center operations as well as the thinning out of unnecessary costly equipment.**

**A newly implemented soft quota model constrains growth. Flexibility and communication with users ensure success.**

**INTRODUCTION**

The deployment and administrative tasks of an HPC data archive tax credos of data integrity, service availability, and operational efficiency. Couple that with the specter of prodigious growth, and you have a witches' brew of daunting missions.

**DEPLOYING TO OUR CREDOS**

**I. DATA INTEGRITY**

Arguably, the ultimate responsibility of an archive is to protect the data.

- ***Dual Copy, Dual Technology***

Data integrity is achieved by dual copy of files over a specific size range. It is often sufficient to simply have dual copies of a file, unless a specific underlying technology is the source of the problem (e.g. firmware bug causing corruption on a data pattern). In this case, a differing technology must store the second copy.

We dual copy over two tape drive technologies in order to avoid such scenarios. Currently these technologies are Oracle T10000C and IBM LTO-5.

The recent leap in capacity resulting from the barium-ferrite particle of the T10000C media realizes an average of 7.9TB per cartridge with our customer data profile. We have recently been afforded the opportunity to dual copy all files up to 256MB as a consequence. We offer a special class of service customers can specify to obtain dual copy files on tape regardless of size.

Each technology is further separated in two distinct robotic library complexes (Oracle SL8500) separate by a distance of approximately 1 kilometer.

- ***Offline Testing***

As tape drives are either purchased or replaced due to failure, they are tested for integrity and performance before being placed into production. A suite of tools was created to facilitate this out-

of-band testing. Files of known size and composition are written to and read from test media. Timing is conducted and data is examined by means of a checksum. It is important to understand that a performance threshold exists below which drives should be considered faulty for the environment, even if integrity checks pass.

- ***End to end verification***

The largest stride in the quest for complete data integrity can only be realized by testing the entire stack of software and hardware in use by the archive application. We employ a homegrown utility called DIVT – Data Integrity Verification Tool.

DIVT runs as a client on various center platforms while using various source file systems. It transfers files into the archive. The files land on level 0 disk cache. They are then pulled out of the archive and compared against the original. The files are stored again, except this time the file is pushed down to level 1 tape and purged off of level 0 disk. Again it is retrieved from the archive and compared against the original.

Should any anomaly exist, email notification will be sent.

This push and pull against the disk and tape levels of the HSM is constant. Finding problems is a game of percentages. In the last two years, DIVT has found two major problems. The first was a file stat() bug with Lustre parallel filesystem reporting inconsistent file size, the result of which were corrupted tar archive images. The second problem was a tape drive that was silently truncating files, thereby corrupting them on tape. None of these would've been found had it not been for the utility. The opportunity for silent corruption is rampant.

## **II. AVAILABILITY**

The focus is on “nines of availability”. Simply stated, it means reducing the length of planned outages. Our goal is often said to be “two and a half nines,” or 99.5% annual uptime, which translates into 3.65 hours of outage per month or

1.8 days per calendar year. For this reason, each second of outage is tracked.

- ***Pre-Production***

A “Pre-Production” environment is an absolute necessity to an archive. All new device firmware, device drivers, operating system fixes and version upgrades, and application versions are tested rigorously. It is here where the methods and order of complex integrations take shape. Tuning parameters are also sorted. A substantive subset of the exact hardware used in production should be represented in pre-production.

With such an environment comes the need for discipline. A pre-production system must be fed and cared for in the same way a production system would be, otherwise it quickly achieves a state of neglect, requiring significant resources to restore its usefulness.

We have traditionally run two production environments – unclassified and classified. Each of those has a dedicated pre-production environment. Deployments start in unclassified preproduction. Depending on the nature of the changes, testing can be from a couple weeks to a couple months, after which time it's deemed suitable for production and a planned downtime date is set.

Then the process is started all over for the classified side on its pre-production system. These cycles tend to be much shorter as most software has been battle-hardened in our unclassified environment by this time.

More typically, due to its larger scale, unclassified production will uproot a bug that wasn't caught in preproduction testing. All future deployments are put on hold while problems are researched and remedied.

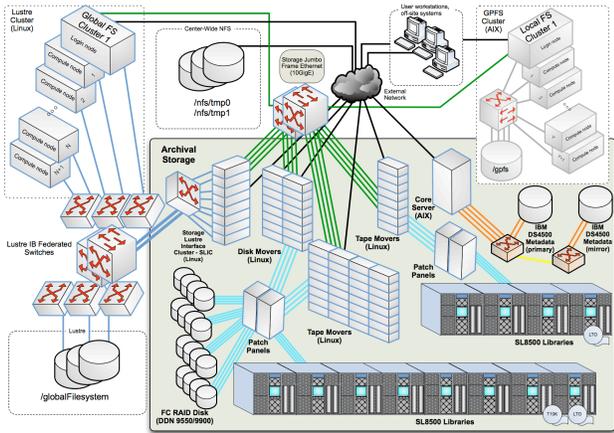
The net result is a well-sorted production rollout that minimizes chances of users finding issues before the deployment staff.

- ***Redundancy***

The current number one reason for loss of service is planned and unplanned electrical outages. Our archive spans four different raised floor

environments. Consequently, we often react to regional raised-floor power work for nearby projects and our own expansion. This is exacerbated by newer electrical safety rules that prohibit electricians from performing “hot work”.

To mitigate, certain hardware has been duplicated in redundant configurations.



Metadata disk for the HPSS Core Server is replicated in two different rooms 1km apart. Either can go down and the application will continue operation. Using operating system disk mirroring on top of RAID controllers across highly available duplicated fiber channel switching technology ensures no one single point of failure between the main core server and its metadata.

Further, robotic software control servers (for Oracle ACSLS) have been made redundant in a cold spare configuration, also located 1km apart.

Identifying single points of failure allows us to concentrate on the biggest bang for our redundancy dollar. The disk and tape mover nodes exist in smaller commodity hardware configurations, in sufficient quantities, so as to allow for individual node failures. Failed nodes are fenced out by our scalable application, HPSS, all while the remaining movers handle the load.

Core server hosts, on the other hand, can be found to have redundant internal drives, fiber HBAs, fans, ethernet cards, power supplies, and ECC memory.

PDUs are specified for twin tailed power sources and are fed from two panels where available.

- **Measured doses of code patching**

Keeping up the nines of availability requires resisting the urge to over-patch the production systems. Security concerns should be thought out and patches tested cohesively in pre-production environments. With few exceptions, the most egregious software security vulnerabilities can be handled by a workaround or an efix which keeps the main archive service available without interruption. Constant patching equals constant downtime.

### III. EFFICIENCY

In many ways, data archives are a study in how to do more with less. Budgets and personnel tend to not grow in step with storage requirements.

- **Trim the fat**

With enough inexpensive data mover hosts, expensive-to-purchase and even more expensive-to-maintain fiber switch technology is not required.

Our data movers are commodity hardware based x86\_64 systems running Linux. All devices are direct attached to the HBA on the host in either FC4 or FC8 native speeds. Fiber trunks running to patch panels handle the interconnects. No electrical is required to these panels.

Should one of these systems crash, there are plenty of remaining nodes to shoulder the load. We mark their associated devices unavailable to the archive application, thus no need exists for a switching architecture to swing devices to online hosts.

- **Piggyback procurements**

Given this commodity hardware data mover design, we are able to leverage the sorts of purchases HPC centers make all the time, namely large cluster and file system disk purchases.

With modest adjustments of node configurations, what was a compute node can be a quite capable and inexpensive I/O data mover machine if tied into the larger procurement process.

- ***Vendor manpower***

Our center has dedicated operations staff well versed in the various hardware types and associated common failure scenarios. Specific vendor gear exists onsite in considerable quantities. Accordingly, we find it possible to negotiate daily onsite vendor CSE/CE support at modest rates. This allows us to have a specialist available for the inevitable unique problems falling outside the scope of an operations staff, as well as for providing a fast track to backend developer support at a moment's notice. This speeds time to resolution and frees our staff to concentrate on the administration of the archive and center at large.

- ***Authoritative sources of information***

An essential component of archive management involves reliably answering questions whose result set changes from frequently to hardly ever. Sources for such questions range from automated scripts to reports written for management. Examples include:

- What milestones were achieved last year?
- What are the firmware versions on the tape drives?
- What fixes make up our previous production code release?

Establishing a single authoritative source abates confusion. The authoritative source often differs for each question, but needs to be identified and communicated to avoid future errors based on incorrect or drifting information gathered from substandard sources (e.g. a file in team member's home directory).

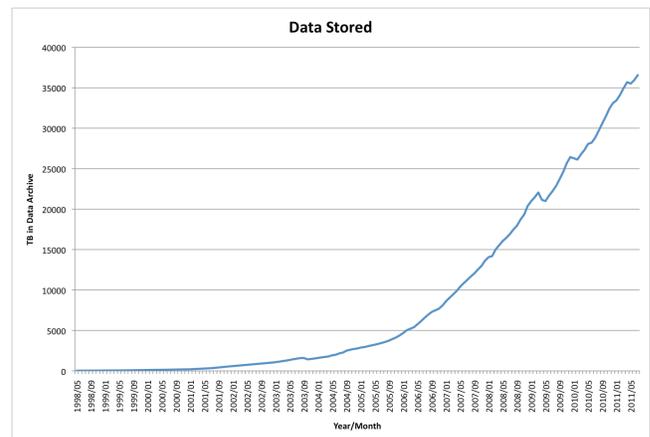
For example, the archive team coalesced on a TeamForge (SourceForge) web utility, which provides a wiki and source control among others. We track project progress here, create How-To's, load key diagram documentation, etc.

Using tape drives as an example, we write utilities that get information in real time by accessing drives over their built in Ethernet connections. Items such as dump status, firmware version, currently mounted cartridge, feet of tape processed, etc. can be gleaned in this fashion.

Our application code and the various local modifications are kept in subversion. We track preproduction and production series. The team members checkout the code, interact with it, and check it back into the central repository. All changes are logged.

## **MANAGING GROWTH**

Fiscal year 2011 marks the first production year of our new Archival Quota system (a.k.a. Aquota). Traditionally, users have been allowed to grow our data archives with few restrictions.



Growth in the last few years suggested that we would need to construct vast new buildings to hold data if this growth curve was to be sustained.

- ***Unique to this quota system***

Two key differences exist comparing Aquota and a traditional disk quota. First is that only annual growth is measured. Data stored the fiscal year prior and before is not considered. Quotas are reset each new fiscal year.

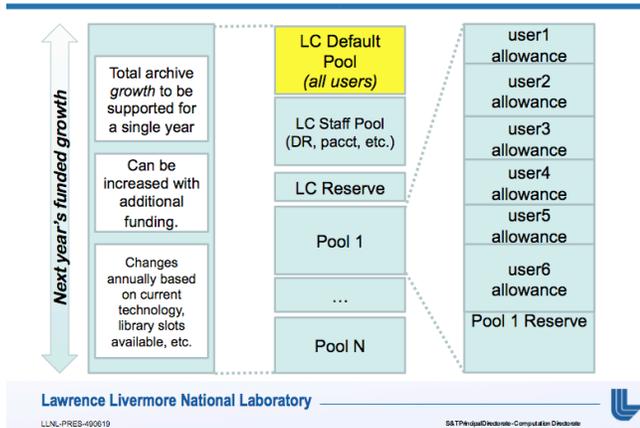
Secondly, it is “soft” enforcement only. Users are still allowed to store after their limits are reached. Users as well as their responsible program managers are contacted when quota is met. It is reported that they have grown beyond their

default allowance and need to seek additional resources.

- **Aquota Model**

Most users live within their yearly budget. The center allocates “pools” of storage to projects. Individuals exceeding their default allowance need to be given space from project pools.

### Soft Annual Quota System Model



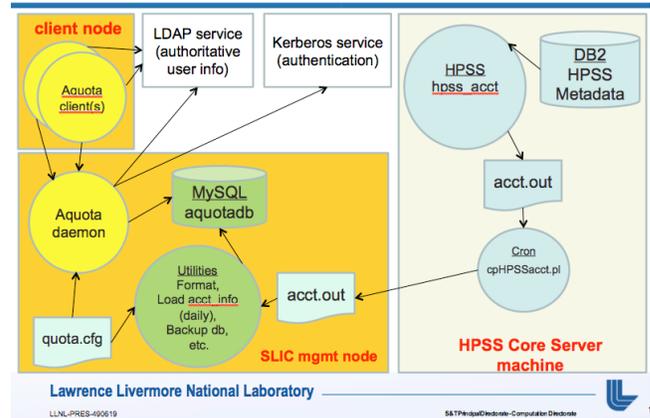
This model of growth control allows the center to predetermine the amount of growth it is willing to sustain for the upcoming fiscal year, rather than attempting to budget based on the previous year's unabated growth.

Once a budget is set, a reasonable set of growth constraints is arrived at based on the amount of media the budget will allow (including potential technology refreshes).

- **Aquota Architecture**

Aquota was built in-house. It is comprised of a server daemon written in C, any number of multiple interactive clients written in C, and a variety of administrative tools written in Perl.

### Aquota Architecture



Nightly exports of HPSS accounting data are imported into a MySQL database. The Aquota daemon handles all client Aquota requests, which can run on a variety of hosts in the center. Users, Pool Managers, and Administrators have increasing levels of authority and interface with the system via the command line client.

- **Impact**

Early evidence for FY11 suggests that overall annual growth will have dropped 14 points from the previous three-year average. The tangible impact is that a tool to facilitate a dialogue has been opened between users, responsible managers, and those of us tasked with offering the archive service. This did not exist in previous years. A common language is now being spoken.

### CONCLUSIONS

Data archives outlive architectures, operating systems, and interconnects. They grow with wild abandon. Bytes churn in a maelstrom of activity as new data arrives and old data is repacked.

Even with a cadre of the latest technological advances and efficient models of deployment, the primary elements of a successful data archive are the people and their willingness to strive to meet the credos of the archive. Key skills in computer science - particularly in languages interpreted and compiled - don't hurt either.

**Challenges in Managing Multi-System Multi-Platform Parallel File Systems**  
**U.S. Department of Energy Best Practices Workshop on**  
**File Systems & Archives**  
**San Francisco, CA**  
**September 26-27, 2011**

**Ryan Braby**

National Institute for Computational Sciences  
rbraby@utk.edu

**Rick Mohr**

National Institute for Computational Sciences  
rmohr@utk.edu

**ABSTRACT / SUMMARY**

The National Institute for Computational Sciences (NICS) is looking to deploy one or more center wide parallel file systems. Doing so should reduce time to solution for many NSF researchers. Researchers who run on multiple systems at NICS will no longer need to move data between parallel file systems and hopefully this will reduce the amount of file system space used for extraneous data replication. However, there are a number of challenges in setting up a multi-system, multi-platform parallel file system. This paper discusses many of the identified challenges for deploying such a file system at NICS and supporting at least the following architectures; Cray XT5, SGI UV, and commodity Linux clusters.

**INTRODUCTION**

The National Institute for Computational Sciences (NICS), a partnership between the University of Tennessee and Oak Ridge National Laboratory, was granted a \$65M award from the NSF in September 2007. A series of Cray HPC systems, named Kraken, were purchased and deployed. Currently, Kraken is a Cray XT5 system with a peak performance of 1.17 PFlops. Lustre is the primary file system for Kraken, and it is built on top of DDN storage directly attached to special I/O service blades in the Cray. These blades act as the MDS and OSS servers for the rest of the system.

In the last year, NICS has deployed a new file system to be shared between Nautilus (a large

SGI UV) and Keeneland (a cluster used for GPGPU development). An evaluation of file system technologies was done, and Lustre was selected for use here. Ideally, the scratch file systems will be shared across all NICS HPC resources. To this end, we have been planning and preparing to upgrade our Infiniband SAN, attach Kraken to this SAN, and migrate Kraken's current Lustre file system to be SAN attached.

While a number of sites have deployed multi-cluster Lustre file systems, unique site requirements prevent the creation of a one-size-fits-all solution. NICS supports a wide variety of platforms (Cray XT, SGI UV, and Linux clusters). Individually, these platforms can present challenges for a site-wide Lustre file system. Combining them further complicates matters.

**CRAY XT**

Cray ships Lustre as part of CLE (Cray Linux Environment), but they are currently shipping an older version (1.6.5) with custom patches. While it is nice to have a vendor supported version, this version is older and lacks features that have been introduced in newer versions. As we move to a center wide file system, there are also concerns about version compatibility between the servers and all the clients.

While it should be possible to put a newer version of Lustre into CLE boot images, there are a number of possible complications with doing so. At this time NICS does not have a file system developer and it is not in our short term plans to

hire one. We could build and install Lustre, but we have minimal resources to test it on. Lacking a file system developer our abilities to fix issues with Lustre in CLE would be limited. The Cray XT systems use a proprietary SeaStar network, which requires it's own Lustre Network Driver (LND) and could complicate LNET routing. Further, Cray support might be hesitant to help on production issues when we are running our own version.

## **SGI UV**

The SGI UV, is a large NUMA architecture with a single system image. Running a single Linux kernel, this architecture tends to get poor IO bandwidth when compared to clustered systems of similar core count. NICS has spent time testing multiple file systems on our 1024 core UV system, and determined that in present day performance Lustre (1.8.6) was the winner (just barely).

Comparing the known road maps for the major parallel file systems, Lustre was the only one that has plans for improving SMP scalability and NUMA performance. In particular, it looks like some improvements in this area have already been added in Lustre 2.1.

Another challenge for parallel file systems on the SGI UV is effectively utilizing multiple network interfaces to our SAN. As a large single system image system, it is important for performance that a file system can drive multiple network interfaces at near line rate. We have had some success scaling Lustre read performance with multi rail infiniband on our UV. This is an area that we hope to see improvements to Lustre for in the future.

## **LINUX CLUSTERS**

Linux clusters with Infiniband interconnects are probably the most common platform for Lustre file systems. As such, including Linux clusters in a multi-cluster Lustre configuration adds some to the complexity. It is another platform to consider and keep track of, but it is also one that you can rely on the community for testing and development.

## **MULTI-SYSTEM CHALLENGES**

Deploying a Lustre file system that spans multiple systems and architectures introduces new challenges apart from the previously mentioned system-specific ones. For example, it may be desirable to run Lustre 2.1 on the SGI UV in order to address some of the SMP scalability issues. However, this would require running Lustre 2.1 on the MDS and OSS servers, which is not compatible with the Lustre 1.6 client on the Cray.

Maintaining compatibility between all of the clients and servers is the first major challenge to a multi-system Lustre deployment. Different platforms may require different patches, and in some cases require different client versions. Knowing which versions are compatible and testing the compatibility is critical to ensuring file system usability.

Some system vendors include a supported version of the Lustre client and publish supported client / server combinations. Merging these requirements from multiple vendors could lead to a situation where the supported versions are not compatible with each other. To reconcile this may require running a version not supported by one or more vendors. One approach to deal with this would be to purchase third party Lustre support.

Managing a multi-system parallel file system makes the file system more of an infrastructure service. Since multiple rely on the availability of the file system, the effects of any disruptions (like maintenance) must be carefully considered. Further, you have to plan upgrades carefully; ensuring that at all points in your upgrade plan you are on compatible versions and not unintentionally running an unsupported combination of server, router, and client versions.

Also, like any infrastructure service, there are possible contention issues. Performance on one system can and will be impacted by access from another system.

## **CONCLUSIONS**

NICS is planning to move to center wide Lustre file systems. There are a number of issues

involved in doing this. While we have identified many of the issues and have ideas of how to deal with them, we do not have the experience and history of implementing these ideas to determine if they are indeed best practices.

## REFERENCES

1. T. Baer, V. Hazlewood, J. Heo, R. Mohr, J. Walsh, *Large Lustre File System Experiences at NICS*. CUG 2009, [http://www.cug.org/5-publications/proceedings\\_attendee\\_lists/CUG09CD/S09\\_Proceedings/pages/authors/11-15Wednesday/12B-Walsh/walsh-paper.pdf](http://www.cug.org/5-publications/proceedings_attendee_lists/CUG09CD/S09_Proceedings/pages/authors/11-15Wednesday/12B-Walsh/walsh-paper.pdf)
2. G. Shipman, D. Dillow, S. Oral, F. Wang, *The Spider Center Wide File System; From Concept to Reality*. CUG 2010, [http://www.nccs.gov/wp-content/uploads/2010/01/shipman\\_paper.pdf](http://www.nccs.gov/wp-content/uploads/2010/01/shipman_paper.pdf)
3. Whamcloud JIRA for Lustre SMP salability: <http://jira.whamcloud.com/browse/LU-56>

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper**

**Kevin Harms**  
LCF/ANL  
harms@alcf.anl.gov

**ABSTRACT / SUMMARY**

**This paper addresses the “Usability of Storage Systems” and the “Administration of Storage Systems” topics. Given the small staff assigned to the LCF Intrepid storage resources we have searched for methods to optimize the use of our storage resources. We present these methods for proactively finding opportunities to tune application I/O and finding degraded hardware that is reducing overall I/O throughput.**

**INTRODUCTION**

The LCF is a relatively new facility and is in the process of developing its storage practices and procedures. One item that has become clear is the need to be proactive about storage usage and administration. We have a limited staff dedicated to the storage system and waiting until an issue turns into a real problem leaves us in an awkward position. In order to address this, we are working on some methods to proactively find issues and start working to solve them before they become worse.

The first method was to install a tool, Darshan, to profile user I/O so that users and staff could have a basic tool to help tune I/O for Intrepid and best utilize the storage resources LCF provides.

The second method is to begin looking at the overall performance of the storage hardware to find and fix marginal hardware without the need to wait for it to degrade to the point of outright failure.

**System Description**

Here is an overview of the core Intrepid storage system. Intrepid has two main storage systems. The home file system is GPFS based and uses 4 DDN9550 SANs that are directly attached via DDR IB to 8 xSeries file servers. The scratch storage has two different file systems running on it, GPFS (intrepid-fs0) and PVFS, (intrepid-fs1) which utilize the same hardware. The scratch area uses 16 DDN9900 SANs, which are directly attached via DDR IB to 128 xSeries file servers. (8 servers per DDN) File system clients are connected over a 10 GB Myrinet fabric.

**THE USABILITY OF STORAGE SYSTEMS**

**Darshan**

Darshan [1] was a tool developed by the MCS department in ANL and deployed on the LCF Intrepid Blue Gene machine. Darshan captures information about each file opened by an application. Rather than trace all operational parameters, however, Darshan captures key characteristics that can be processed and stored in a compact format. Darshan instruments POSIX, MPI-IO, Parallel netCDF, and HDF5 functions in order to collect a variety of information. Examples include access patterns, access sizes, time spent performing I/O operations, operation counters, alignment, and datatype usage. Note that Darshan performs explicit capture of all I/O functions rather than periodic sampling in order to ensure that all data is accounted for.

The data that Darshan collects is recorded in a bounded (approximately 2 MiB maximum) amount of memory on each MPI process. If this memory is exhausted, then Darshan falls back to recording coarser-grained information, but we have yet to observe this corner case in practice. Darshan performs no communication or I/O while the job is executing. This is an important design decision because it ensures that Darshan introduces no additional communication synchronization or I/O delays that would perturb application performance or limit scalability. Darshan delays all communication and I/O activity until the job is shutting down. At that time Darshan performs three steps. First it identifies files that were shared across processes and reduces the data for those files into an aggregate record using scalable MPI collective operations. Each process then compresses the remaining data in parallel using Zlib. The compressed data is written in parallel to a single binary data file.

Darshan was deployed on Intrepid by creating a modified set of mpiXXX compiler wrappers which link in the darshan library code. These modified compiler wrappers are part of the users default path which means many applications link in Darshan with no extra work by the user. These applications put logfiles into a common area and are setup so only the user who produced the logs can read them. Later we change the group permission to a special ‘darshan’ group and then add group read permission. These logs then become accessible by the LCF staff and a few selected MCS research staff.

### User Analysis

The first capability this provides is for users to look at some information about their jobs I/O profile and compare it to common suggestions available via our wiki documentation. If the user feels their I/O performance is not as good as it should be, when contacting the LCF staff, we already have some basic information about the I/O patterns they are using which might give some initial starting suggestions for the user to try for improving I/O performance on Intrepid. This

also addresses a common issue where users are not familiar with how their application does I/O, perhaps because they are using some large application where someone other person or group implemented the IO code. Figure 1 shows an example from the *darshan-job-summary.pl* output.

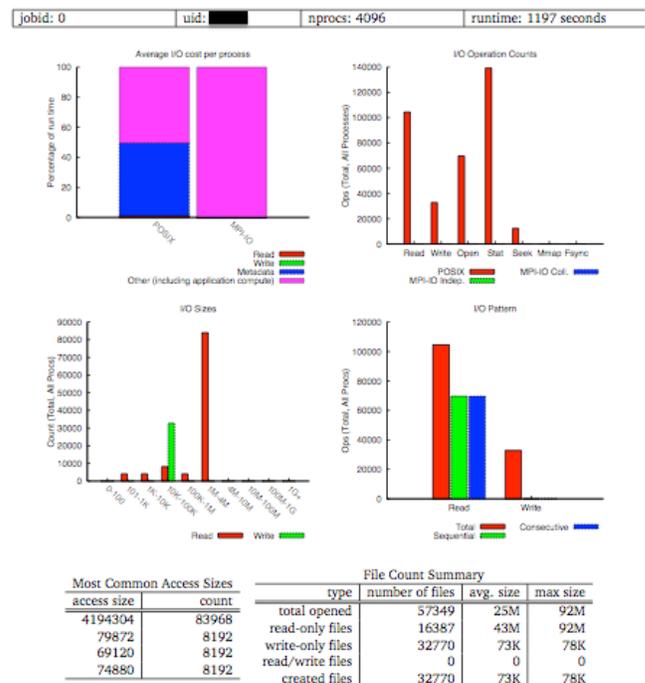
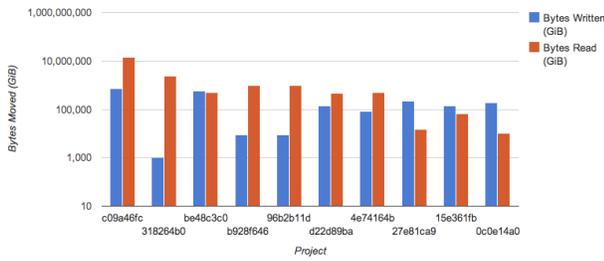


Figure 1 – Darshan Job Summary Example

This summary information can provide a useful starting point for I/O analysis. We are aware of a few applications that have used this output to successfully improve their application I/O for Intrepid.

### Project Analysis

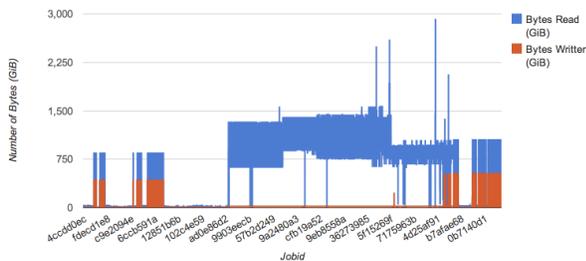
The second capability is to proactively analyze darshan logs to see how users are utilizing the storage system and if they are being effective. We are developing a basic web interface around aggregated log files that can be examined on a per-project basis to find who the major users of the storage system are and how are they using the system. We explored this idea in reference [2]. Figure 2 shows the top 10 projects from 1/1/2011 to 6/30/2011. We can look at these projects individually to see how they are using the I/O system.



Project	Major Project	Bytes Written (GIB)	Bytes Read (GIB)	Bytes Moved (GIB)
c09a46fc	incite	717960.320269	14378691.7527	15096652.073
318264b0	incite	1028.50599501	2415161.5417	2416190.0477
be48c3c0	incite	583163.035626	500354.756459	1083517.79209
b928f646	alcc	8999.55220313	970914.933426	979914.485629
96b2b11d	incite	8912.51224884	940718.474662	949630.986911
d22d89ba	incite	141647.703103	447793.368016	589441.071119
4e74164b	incite	81074.8510083	489839.286861	570914.137869
27e81ca9	incite	222602.486601	14497.4064404	237099.893041
15e361fb	incite	144678.540998	64092.520676	208771.061674
0c0e14a0	incite	181726.374965	10113.3177587	191839.692723

Figure 2 – Top 10 Projects by bytes moved

Once we identify the top I/O users we can examine their I/O usage in more detail. Figure 3 shows an example of aggregate information about a single project. We can look at the percent time spent in I/O and see if we should consider talking with a project about their I/O usage if it looks subpar and thereby improve their utilization of the core-hours they have been granted.



Project	Major Project	Bytes Written (GIB)	Bytes Read (GIB)	Bytes Moved (GIB)
c09a46fc	incite	822788.152176	14693539.2635	15516327.4157

Project	Min (MiB/s)	First Quartile (MiB/s)	Second Quartile (MiB/s)	Max (MiB/s)
c09a46fc	0	42.270659	4827.955307	11199.311494

Project	Job Coverage	Job Coverage Ratio
c09a46fc	1799/3315	0.542684766214

Project	Percent Time in I/O
c09a46fc	0.289274515813

Figure 3 – Darshan Project Information

We had identified a project in 2010 that was a top I/O consumer but had a high percentage of time spent in I/O. We were successful in working with this project to change the method for writing of files, which gave them a 30% improvement in write throughput.

## System Planning

The third capability we get is the ability to look at what I/O patterns users want to use and what they want to do. This information can be used to target how we allocate our resources for next generation systems. Examples from above show users are still obsessed with generating  $O(1000)$ ,  $O(100000)$ ,  $O(1000000)$  files. The file per process model tends to break down at the 8192 node level (or 32768 processes) on Intrepid. For our next generation system, we have planned to split data and metadata and use separate SSD based storage for the metadata in hopes of boosting metadata performance, which would serve as a band-aid for the file-per-process users.

Another point is that we see about 60% of the jobs at large scale go to either shared or partially shared files and fewer use the file-per-process model. Figure 4 shows this distribution. However, in this same time period we saw remarkably few people using high-level libraries such as HDF5 or PnetCDF. This might indicate we need to spend time educating the userbase about these libraries or find out why our users would rather create their own shared file format rather than leverage existing ones.

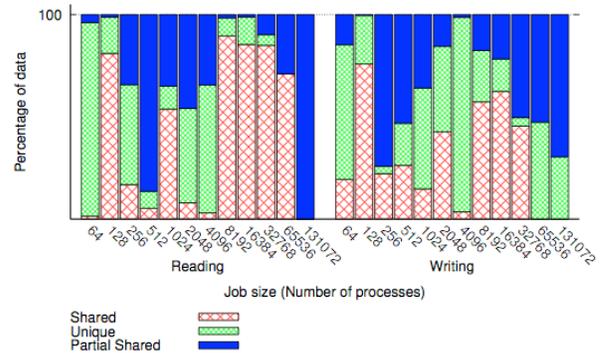


Figure 4 – File usage (1/2010 – 3/2010)

## THE ADMINISTRATION OF STORAGE SYSTEMS

Another aspect of storage system efficiency is ensuring the current hardware is delivering performance up to its useable peak. Anecdotally we have observed a single failing SATA disk can produce a global slowdown of the scratch file

system. During our early test stages when we were tuning the /intrepid-fs1 file system, we would often find a marginal drive would cause a significant slow down in an IOR test case. As an example, we would see something on the order of losing 50% of total throughput. After failing 1 (or more) drives, the system would return to its optimal performance level.

The work we have done in this area is still very preliminary and we have not validated any of our suppositions.

**Log Analysis**

The DDN9900 will report many errors and statistics but it also logs informational events in the system log. These are generally not reflected directly in any of the system statistics. We developed a trivial monitoring tool to check the event logs of each DDN approximately once per day and send an alert if there were a large number of new messages in the log. Here is a short snippet from the monitoring tool, which emails its results.

```

INFO INT_GH 8-29 12:50:31 Recovered:
Unit Attention Disk 9G GTF000PAH51JNF

INFO INT_GH 8-29 12:59:29 Recovered:
Unit Attention Disk 22G GTF002PAHKKXRF

INFO INT_GH 8-29 13:02:43 Recovered:
Subordinate errors detected.

INFO INT_GH 8-29 13:05:18 Recovered:
Unit Attention Disk 2G GTF100PAHW59BF

INFO INT_GH 8-29 12:49:44 Recovered:
Unit Attention Disk 13G GTF002PAHWD21F

INFO INT_GH 8-29 13:00:31 Recovered:
Subordinate errors detected.

New Log Messages: 2650

```

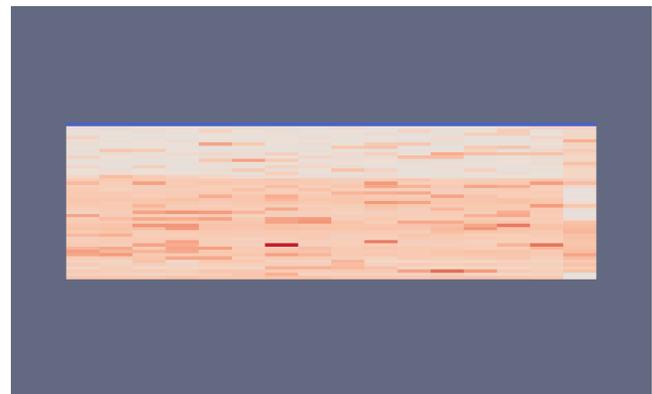
**Example 1 – DDN Log monitoring output**

In Example 1, we see that this DDN had 2600 new log messages and many of those messages are related to problems with disk access on channel ‘G’. In this case, we could have opened a support request with DDN to determine which component was really at fault. In this particular case, disk 7G failed 5 days later. We could have failed disk 7G earlier and presumably not lost any performance during that time period.

**Visualization**

Another method to monitor the storage infrastructure for marginal components is via visualization of I/O metrics. We setup a utility to pull the ‘tierdelay’ metric from all tiers of each of the 32 DDN controllers associated with the scratch file system. We then ran the IOR benchmark with a write workload while we collected samples every 10 seconds. The data was loaded into ParaView and we began looking for patterns.

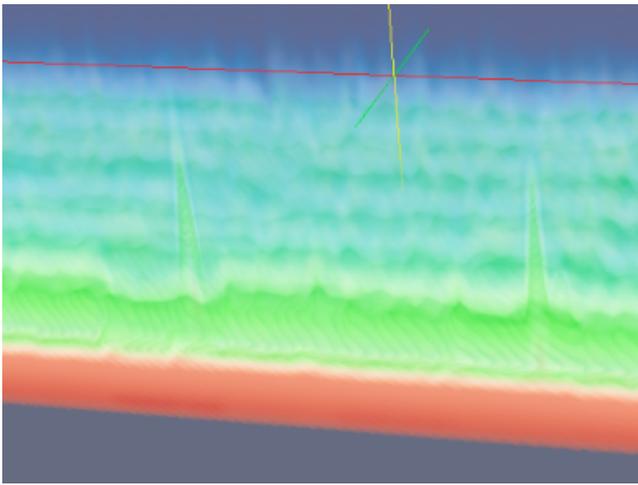
Figure 5 shows a combined visualization of total operation count for each channel/tier combination for all DDNs at the last timestep.



**Figure 5 – Cumulative Operations**

Since the IOR workload was evenly distributing data over all LUNs we should see similar operation counts, but instead we see one tier (dark red) that has significantly more operations than the rest of the tiers. In talking with DDN, the ‘tierdelay’ counter records all operations including internal retries. It would appear that there is some issue on this particular tier resulting in retries being generated.

Figure 6 shows the same metric again but now as a 3D volume.



**Figure 6 – Tier Delay as Volume**

The volume shows a count for the number of operations, which occurred within a defined bucket. For example, 100 operations at 0.2 second delay. The bottom of the volume is the shortest delay and top of the volume is the highest delay. The dark red coloring are higher counts going to blue at the lowest counts. In general the image shows the lowest latency buckets have the highest counts, which is good and the highest latency buckets have the lowest counts, also good. However, you can see a spike on a couple of disks where the higher latency buckets have a much higher total count than most other disks. We don't have conclusive findings that those disk are causing system wide problems, but that is an example of what we hope to find.

## CONCLUSIONS

The Darshan deployment has been successful on the LCF Intrepid system. A few projects have used it to tune I/O characteristics to optimize for Intrepid and seen improvements in throughput. We also identified a project that was significant storage user but also suffered from slow I/O performance. We worked with the members of

this project to update their code with a slightly modified I/O model that used fewer files which resulted in a 30% improvement of their I/O write speed. We plan to continue to enhance our summarization web tools to provide easier access to the darshan data for the LCF staff.

Our progress on identifying faulty hardware prior to failure on the DDN SANs is still very preliminary and we have not validated any of the results. We hope to progress this further by being able to validate performance improvement after a hardware replacement. We would also hope to identify these patterns so that we could create statistical models that would work on the normal I/O load of Intrepid without the need for an invasive diagnostic run.

## ACKNOWLEDGEMENTS

Phil Carns – for all that is Darshan

Neal Conrad – for Darshan web development

Justin Binns – for IO visualizations

## REFERENCES

1. Philip Carns, Robert Latham, Robert Ross, Kamil Iskra, Samuel Lang, and Katherine Riley. 24/7 characterization of petascale I/O workloads. In *Proceedings of 2009 Workshop on Interfaces and Architectures for Scientific Data Storage*, September 2009.
2. Philip Carns, Kevin Harms, William Allcock, Charles Bacon, Samuel Lang, Robert Latham, and Robert Ross. Understanding and improving computational science storage access through continuous characterization. In *Proceedings of 27th IEEE Conference on Mass Storage Systems and Technologies (MSST 2011)*, 2011.

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, Ca  
September 26-27, 2011  
Position Paper**

**Roger Combs  
Navy LSRS Program  
roger.combs1@navy.mil**

**Mike Farias  
Sabre Systems  
mfarias@sabresystems.com**

## **ABSTRACT**

The Navy Littoral Surveillance Radar System (LSRS) Program has demanding streaming aggregate I/O requirements (double-digit GB/sec level). The LSRS Program also has petabyte-level data management issues and accompanying data management policies and procedures that are under constant review.

## **INTRODUCTION**

This document will address current Navy LSRS best-practices within our own High Performance Computing (HPC), capacity environment. Areas of concern will be the following:

- a. Business of storage systems
- b. Administration of storage systems
- c. Reliability of storage systems
- d. Usability of storage systems

**Business of storage systems:** Currently the LSRS Program uses Oracle Storage Archive Manager/Quick File System (SAM/QFS) as the parallel file system and respective Hierarchical Storage Management (HSM) solution to meet our data storage and management needs. Strategically, business viability of SAM/QFS under Oracle, post-Sun Microsystems acquisition, has and continues to be a major concern. As a result of several meetings with Oracle concerning SAM/QFS, ultimately the IBM General Parallel File System (GPFS) and the High Performance Storage System (HPSS) were chosen as the future file system/HSM solution. From both production experience and consensus among some DoE colleagues, a parallel file system is currently regarded as the most challenging and

critical aspect of HPC operations, frequently referred by LSRS as the “backbone.” As fallout of this “backbone” ideology, when faced with an acquisition decision regarding SAM/QFS, only two file systems came into play. Criteria for selection were items such as company viability, development talent, and a deep R&D budget / bench. Ultimately, this list revolved around two solutions, Lustre/HPSS and GPFS/HPSS. Cost was not a criterion for parallel file system selection for CY12 migration.

Historically, cost was a criterion for selection of our SAM/QFS file system and our current migration efforts are serving as a lesson-learned. Moving forward, there has been concern about the viability of the SAM/QFS parallel file system beyond CY11 in terms of development and support. For our “backbone,” there also have been concerns with Lustre and Oracle IP strategy potentially being an issue. Concerns with Lustre stability were also negatively factored into the decision process from reading publications such as the Livermore National Lab (LLNL) I/O “Blueprint” from 2007<sup>1</sup>.

From a business perspective, LSRS best practices dictate that the “backbone” be the most performant solution that can be afforded under the company with the deepest R&D bench. An additional requirement is that the provider of the parallel file system middleware be relevant in the HPC marketplace. Storage acquisition (both cache and tape) are approached from a best-of-breed perspective and not a cost perspective.

**Administration of storage systems:** Currently, storage system administration is handled and led

entirely by private industry personnel. Strategically, LSRS recognizes that this is not best practices, and future administration and management of storage systems will have a government technical lead. Across all HPC functional areas, there will be government division leaders aka department heads (DHs).

Above and beyond organizational layout, monitoring and benchmarking tools could always be better for storage infrastructure in general. Interleaved or Random (IOR) benchmarking is used to get theoretical maximums for I/O on capacity storage. Above and beyond, IOR, solutions from companies such as Virtual Instruments have been explored to potentially better capture Fibre Channel I/O in near-real-time and identify bottlenecks. However, currently Virtual Instruments does not support or project to support Quad Data Rate (QDR) Infiniband, which is orthogonal to our HPC I/O roadmap.

In general, parallel I/O benchmarks seem limited and a bit immature given the projected requirements for data-driven computing currently and in the future<sup>2</sup>. From a tape perspective, minimizing media that is more than a generation behind the current industry products is policy. While tape certainly has value, from our production experience, lifecycle management of tape has proven to be challenging. Subsequently, we are facing the task of ascertaining if obsolete media needs to be discarded or go through a relatively painful conversion process.

**Reliability of storage systems:** Organizationally, file system reliability is believed to be directly related to file system scalability and stability. From this, we borrow from the 2007 LLNL I/O Blueprint<sup>1</sup>, in asserting that in general, file systems are sized to no less than three orders of magnitude below the compute platform(s) they support, i.e., a 10TF system would need no less than 10GB/sec of aggregate I/O bandwidth behind it. Leveraging this approach has significantly increased productivity and nearly eliminated staging. In support of consistent systems reliability and balance, file system and network interconnect acquisition precedes platform acquisition. Systems acquisition

is also approached from a modular perspective in similar fashion to Mark Seager's Peloton and associated Hyperion based initiatives at LLNL.

By definition, we assert that file systems that are not horizontally scalable are intrinsically unstable. QFS currently suffers from the preceding quality with one metadata server per namespace. The current, QFS file system is monolithic; LSRS has established a requirement for no less than two production (primary and secondary) parallel file system namespaces for capacity high-availability. As disk caches for HPC centers enter the petabyte and beyond level, we've found from production that file system scalability capabilities do not necessarily hold up to vendor claims. To provide continuity of daily operations, it is critical that two namespaces are on the floor ready-to-go at any given time. From experience, edge-cases are frequently encountered, taking days or more oftentimes weeks to solve. The preceding service-losses or impacts are compounded when cache-repopulation is considered with file systems at the petabyte level taking weeks to re-populate. With QFS particularly, in terms of monolithic metadata architecture, and the associated production issues that resulted, LSRS realizes the importance of choosing superior architectures and support organizations. LSRS metadata storage is handled from an IOPS-centric point-of-view and RAMSAN technology is used for metadata storage. As a backup, physical solid-state disk is used to complement the RAMSAN. While one monolithic namespace has performance advantages, we plan to leverage two namespaces in the very near future. Post QFS-migration, two GPFS namespaces will be established, prior to QFS-migration a single QFS and GPFS (Data Direct Networks SFA10KE "Gridscaler") namespaces will be established.

Furthermore, to manage job quality of service, Navy LSRS borrowed from the DoD High Performance Computing Modernization Office (HPCMO), and established their Normalized Expansion Factor (NEF) Metric<sup>3</sup>. The details of this metric can be found in an FY2002 whitepaper from HPCMO referenced below, but essentially the metric is a normalized way to measure job quality with no queue-wait time at-all associated with jobs having a NEF of 1.0.

Heuristically, high priority work should not exceed an NEF of 1.7, whereas standard workload should not exceed 2.2. NEF metrics are collected for each individual job and performance data is kept indefinitely.

**Usability of storage systems:** To address usability, LSRS strategically attempts to minimize the number of namespaces deployed to two vice, having multiple in the past. The preceding has obvious usability advantages, but also the performance advantage of having more drive spindles under one namespace. Block-level storage, in general, is abstracted away from analysts using in-house developed mass-storage APIs. In our environment, usability is dominated by performance and concessions are viewed as necessary. Generally, performance, scalability, and stability are the three dominant factors in strategic file system thought. Usability is still a distant fourth-level consideration.

## CONCLUSIONS

The most important aspects of file system and archive best practices are an understanding that the system design-points need to be a function of the application sets, both current and projected, running in production. Heuristics will get you close to balanced, and generally keep architects out of trouble, but to really get outstanding performance requires closer interaction with analysts. At Navy LSRS, the file system is currently regarded as the “backbone” of production operations and subsequently a lot of attention is paid to ensuring that it is sized properly and has an appropriate interconnect and bandwidth.

Tape is effectively sized using the write rate of a typical run of the most write-intensive application in production. While certainly valuable and viable for the long-term, tape has presented Navy LSRS with a number data lifecycle management issues regarding a myriad of end of life tapes and infrastructure (silos). Many of these tapes have questionable value, but due to this uncertainty, they create a lot of work in mining data from useful media and discarding useless media. While valuable, tape certainly presents maintainability

issues if allowed to veer too far from current generations and formats.

Additionally, from a business perspective, much has been learned in terms of interacting with vendors as well as integrators and reading between the lines. From an organizational perspective, Navy LSRS has shifted into an organization that is much more critical of consumed information than in the past. The preceding applies across all functional areas. In other words, asking “is what the vendor is saying useful,” or “is what our integrator is saying practical?” All too often, initially, answers were frequently no. Oftentimes, further investigation led to invaluable insights into real vendor positions vice stated, or performance improvements that were never realized due to inadequate architectural and or operations planning. Particularly with file system and archive materiel, betting on the wrong technology or vendor can be costly, well into the seven-figures and beyond. Subsequently, staying in tune with the HPC community has proven to be a very fruitful investment of both time and energy.

Finally, establishment of file system and I/O roadmaps, i.e., LLNL’s I/O “Blueprints” has helped Navy LSRS tremendously. Moving from ad hoc approaches to file system and archive operations to planned and deliberate signed documentation has forced our organization into making much more informed decisions. Roadmaps, in general are key in supporting a successful HPC program.

## REFERENCES

1. Wiltzius, D. et al. (2007). *LLNL FY07 I/O Blueprint*, 2007. Retrieved from LLNL website: <https://e-reports-ext.llnl.gov/pdf/342161.pdf>
2. Cook, D. et al. (2009). *HPSS in the Extreme Scale Era*, 2009. Retrieved from NERSC website: <http://www.nersc.gov/assets/HPC-Requirements-for-Science/HPSSExtremeScaleFINALpublic.pdf>
3. DoD High Performance Computing Modernization Office (HPCMO) (2002). *HPC System Performance Metrics*, 2002. Retrieved from HPCMO website: [http://www.hpcmo.hpc.mil/Htdocs/HPCMETRIC/fy2002\\_hpcmp\\_metrics.pdf](http://www.hpcmo.hpc.mil/Htdocs/HPCMETRIC/fy2002_hpcmp_metrics.pdf)

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives**

**San Francisco, CA**

**September 26-27, 2011**

**Oak Ridge Leadership Computing Facility Position Paper**

**Sarp Oral, Jason Hill, Kevin Thach, Norbert Podhorszki, Scott Klasky, James Rogers, Galen Shipman**

Oak Ridge Leadership Computing Facility, Oak Ridge National Laboratory  
{oralhs, hilljj, kthach, pnorbert, klasky, jrogers, gshipman}@ornl.gov

**ABSTRACT / SUMMARY**

*This paper discusses the business, administration, reliability, and usability aspects of storage systems at the Oak Ridge Leadership Computing Facility (OLCF). The OLCF has developed key competencies in architecting and administration of large-scale Lustre deployments as well as HPSS archival systems. Additionally as these systems are architected, deployed, and expanded over time reliability and availability factors are a primary driver. This paper focuses on the implementation of the “Spider” parallel Lustre file system as well as the implementation of the HPSS archive at the OLCF.*

**INTRODUCTION**

Oak Ridge National Laboratory’s Leadership Computing Facility (OLCF) continues to deliver the most powerful resources in the U.S. for open science\*. At 2.33 petaflops peak performance, the Cray XT5 Jaguar delivered more than 1.5 billion core hours in calendar year (CY) 2010 to researchers around the world for computational simulations relevant to national and energy security; advancing the frontiers of knowledge in physical sciences and areas of biological, medical, environmental, and computer sciences; and providing world-class research facilities for the nation’s science enterprise.

The OLCF is actively involved in several storage-related pursuits including media refresh, data retention policies, and file system/archive performance. As storage, network, and computing technologies continue to change; the OLCF

is evolving to take advantage of new equipment that is both more capable and more cost-effective. A center-wide file system (Spider) [1] is providing the required high-performance scratch space for all OLCF computing platforms, including Jaguar. At its peak, Spider was serving more than 26,000 clients and providing 240 GB/s aggregate I/O throughput and 10 PB formatted capacity. For archival storage OLCF uses the high-performance tape archive (HPSS). Currently HPSS version 7.3.2 at OLCF is housing more than 20 PB of data with an ingest rate of between 20–40 TB every day. This paper presents the lessons learned from design, deployment, and operations of Spider and HPSS and future plans for storage and archival system deployments at the OLCF.

**THE BUSINESS OF STORAGE SYSTEMS**

Storage requirements for both Spider and HPSS continue to grow at high rates. In September 2010, two new Lustre file systems were added to the existing center-wide file system. These two file systems increased the amount of available disk space from 5 to 10 PB and will help improve overall availability as scheduled maintenance can be performed on each file system individually. The addition of these file systems provided a 300% increase in aggregate metadata performance and a 200% increase in aggregate bandwidth. Additional monitoring improvements for the health and performance of the file systems have also been made.

In August 2010, a software upgrade to version 7.3.2 on the HPSS archive was completed, and staff members began evaluating the next generation of tape hardware. Implementation of this release has resulted in performance improvements in the following areas.

- *Handling small files.* For most systems it is easier and more efficient to transfer and store big files; these modifications made improvements in this area for owners of smaller files. This has been a big gain for the OLCF because of the great number of small files stored by our users.

---

\* The submitted manuscript has been authored by a contractor of the U.S. Government under Contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This research used resources of the Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

- *Tape aggregation.* The system is now able to aggregate hundreds of small files to save time when writing to tape. This has been a tremendous gain for the OLCF.
- *Multiple streams or queues (class-of-service changes).* This has enabled the system to process multiple files concurrently and, hence, much faster, another huge time saver for the OLCF and its users.
- *Dynamic drive configuration.* Configurations for tape and disk devices may now be added and deleted without taking a system down, giving the OLCF tremendously increased flexibility in fielding new equipment, retiring old equipment, and responding to drive failures without affecting user access.

Following this upgrade, in April 2011, twenty STK/Oracle T10KC tape drives were integrated into the HPSS production environment. This additional hardware is proving to be very valuable to the data archive in two distinct ways. The new drives provide both a 2x read/write performance improvement over the previous model hardware and a 5x increase in the amount of data that can be stored on an individual tape cartridge. Along with improved read/write times to/from these new drives, the OLCF now benefits from being able to store 5 TB on each individual tape cartridge—effectively extending the useful life of the existing tape libraries. This has allowed the OLCF to postpone its next library purchase until the first half of FY12.

The OLCF HPSS archive has experienced substantial growth over the past decade (Figure 1). The HPSS archive currently houses more than 20 PB of data, up from 12 PB a year ago. The archive is currently growing at a rate of approximately 1PB every 6 weeks, and that rate has doubled on average every year for the past several years.

Planning around such extreme growth rates, from both a physical resource perspective and an administrative perspective, while operating within a limited budget capacity, presents several challenges. The fact that tape technology and performance traditionally lags behind that of its disk/compute counterparts presents a fiscal challenge in supporting such a large delta in the amount of data taken into the archive each year. We are forced to purchase additional hardware (tape libraries, tape drives, data movers, switches, etc.) each year in order to meet operational and performance requirements. Add in the fact that much, if not the majority of the archived data needs to remain archived for multiple generations of media (a very significant amount of resources are spent in the process of repacking data from older tapes onto newer media), and a tremendous amount of money is spent simply maintaining “status quo” of the archive each year.

The OLCF recognizes that such a model of exponential archive growth is unsustainable over the long-term. We have taken steps to mitigate this problem and slow the growth rate down by introducing quotas on the amount of data users can store in their respective home and/or project

areas within the archive. In addition, we recently made a request to the Top 10 users of the archive to purge any unnecessary data from the archive, and that request to voluntarily remove data yielded well over 1 PB of data that was purged from the archive.

The OLCF has two Sun/STK 9310 automated cartridge systems (ACS) and four Sun/Oracle SL8500 Modular Library Systems. The 9310s have reached the manufacturer end-of-life (EOL) and are being prepared for retirement. Each SL8500 holds up to 10,000 cartridges, and there are plans to add a fifth SL8500 tape library in 2012, bringing the total storage capacity up to 50,000 cartridges. The current SL8500 libraries house a total of 16 T10K-A tape drives (500 GB cartridges, uncompressed), 60 T10K-B tape drives (1 TB cartridges, uncompressed), and 20 T10K-C tape drives (5 TB cartridges, uncompressed). The tape drives can achieve throughput rates of 120–160 MB/s for the T10KA/Bs and up to 240 MB/s for the T10K-Cs.



**Figure 1. OLCF HPSS Archive Growth**

OLCF follows a collaborative open source development model for its scratch space storage system. A multi-national and multi-institutional collaboration, OpenSFS [2] was formed in 2010 by ORNL, LLNL, Cray, and Data Direct Networks. The goals of the OpenSFS organization are to provide a forum for collaboration among entities deploying file systems on leading edge high performance computing (HPC) systems, to communicate future requirements to the Lustre file system developers, and to support a release of the Lustre file system designed to meet these goals. OpenSFS provides a collaborative environment in which requirements can be aggregated, distilled, and prioritized, and development activities can be coordinated and focused to meet the needs of the broader Lustre community. This collaborative open source development model allows the OLCF to have more control and input in high-performance scalable file system development. OpenSFS recently awarded a development contract for future feature development required to meet the requirements of our next-generation systems. OpenSFS has been extremely successful in organizing the Lustre community, providing a

forum for collaborative development, and embarking on development of next-generation features to continue the progression of the Lustre roadmap. OpenSFS is working closely with its European counterpart (EOFS) and has signed a memorandum of understanding to align our respective communities. At LUG 2011 all communities aligned with OpenSFS providing a unified platform from which to carry Lustre well into the future, meeting not only our current petascale requirements but providing an evolutionary path to meeting our Exascale requirements.

For archival storage systems OLCF is participating in a collaborative proprietary closed source model led by IBM. OLCF is currently providing more than 2 FTEs for this collaboration. While this model provides faster development cycles and better maintenance support compared to the open source collaborative model, the cost and business related risks associated with the private company leading the development project are the drawbacks of this model.

OLCF resources are classified as medium-confidentiality and low-availability according to the Federal Information Processing Standards (FIPS). While OLCF recognizes the cost benefit of using commodity storage hardware, the current state of technology does not allow us to deploy such technology in our archival storage systems. However, as these technologies continue to mature, it might be possible to take advantage of commodity storage hardware.

#### **THE ADMINISTRATION OF STORAGE SYSTEMS**

The day-to-day administration of a large parallel file system requires coordination between not only the members of the team working on the file system (both hardware and software), but coordination throughout the computational center as these activities have the potential to cause service outages and impact performance. The OLCF has successfully deployed packages for version control of key administration scripts, as well as centralized configuration management to handle individual node configuration convergence.

The OLCF uses Nagios [3] to monitor the health of the components of the system. Custom checks have been implemented to additionally validate the correctness of the file system – specifically are the devices mounted where they are supposed to be. Additional performance monitoring of the Lustre Network layer (LNET) are done for the Lustre servers and routers in Nagios. Currently this information is not archived for future analysis it is only used for failure detection.

We use the Lustre Monitoring Toolkit [4] developed at Lawrence Livermore National Laboratory (LLNL) to grab periodic Lustre level performance snapshots. Our current dataset is quite small so it is not useful for future predictions, but we have seen interesting trends.

Finally the OLCF has written a tool that can query the Application Programming Interface (API) to the DDN S2A9900 storage controllers. We use it to monitor the performance of the backend controllers. Currently we capture Read and Write Bandwidth and IOPS. This quick glance of the overall system performance can give administrators a fast track to problem diagnosis if say the IOPS are orders of magnitude higher than the Bandwidth. In that case we know to search out an application that is using one of the Lustre OST's served by that DDN 9900.

Being a center-wide file system, Spider is key to simulation, analysis, and even some visualization for the OLCF. Great care is taken to preserve system uptime, and maintenance activities are deferred to at a minimum once per quarter downtime. This outage affects all users of OLCF compute resources, but can help to address performance issues and overall system maintenance tasks that are harder to do real-time. Much of our administrative tasks are coordinated and done live, but with the Jaguar XT5 resource in maintenance period to limit the potential issues for users if something were to go wrong. An example is the DDN controller firmware. We can upgrade one controller out of every couplet, reboot it, and then do the partner controller without causing a file system outage. This can help push the potential quarterly outage to twice per year or even once per year depending on the software releases from DDN.

After a hardware failure caused partial file loss from the Lustre file system, a full root cause analysis led to procedural changes as well as changes in e2fsprogs packages, and spurred development of fast metadata and Lustre object storage targets for determining files that are affected by a large failure of the RAID devices on the backend.

#### **Change Control**

The OLCF has used the configuration management package CFengine [5] for several years. In our implementation of CFengine we have chosen to manage configuration files at a node level (host), a cluster level (groups of hosts related by system task), the operating system level (for each version of the OS), and a generic level that applies to all systems within the center. Additional work has gone in to configure systems that are diskless requiring some workarounds within Cfengine and the rest of the OLCF infrastructure.

In our case we use it to manage the configuration of the Lustre OSS/MDS/MGS servers – we are unable to use it to manage the storage controllers themselves. Additionally we use version control to manage the configuration of the Ethernet switches and routers for simple rollback. Managing the configuration of the Lustre file system is somewhat more difficult, but we wrote scripts and configuration files that describe the file system and can be used to start/stop the file system as well as monitor the health and status of the file systems.

We can additionally use the DDN API tool to query the configuration of the storage controllers and note a deviation from the baseline configuration specified. Work is ongoing to both correctly define the baseline as well as what acceptable deviations and periods of deviation are before notifying administrators. The storage controllers are configured to send their log data to a centralized syslog server that is running the Simple Event Correlator [10] and SEC can alert for matches to pre-configured rules. We also have SEC configured to send all log data captured over a 1 hour period to the administrators for help in solving issues like failed disks or diagnosing performance problems like SCSI commands timing out.

Our current configuration management solution does not perform validation of the configuration or syntax checking for the configuration itself. The next generation configuration management solution (BCFG2) contains input validation and syntax checking on commit.

The OLCF has both development testbeds and a pre-production testbed for verifying both changes as well as system upgrades. We have however not found any bugs at this small scale that have saved problems when the change/upgrade is deployed. Some problems only reveal themselves under sufficient load.

### **Cyber Security**

For the Spider parallel file systems at the OLCF we commit to quarterly OS patching (matching the above mentioned quarterly planned system outages), based on analysis of risk and the location in the network. This is a delicate balance of keeping the system stable/available and satisfying the desires of Cyber Security personnel in keeping systems at the most recent patch levels. The HPSS side has weekly maintenance windows (not always taken), and has the ability to roll out security patches through those windows. Outages of the archive do not affect the production compute clusters where outages for the Spider file systems would take down the compute resources.

One example of how we can demonstrate certain file systems only being available to certain nodes is via the /proc file system on the Lustre OSS and MDS servers. We have a category 3 sensitivity file system and are working to monitor the mounts of that file system via the proc file system on the Lustre servers. If a client that is not authorized to mount the file system is detected an alert is sent to the security team and logs from the non-authorized node are gathered to see who was logged in at the time of the un-authorized access.

The OLCF has three categories of data protection that map to “publicly available information” (Category 1); data that is proprietary, sensitive, or has an export control (Category 2); or data that has additional controls required based on the sensitivity, the content being proprietary, or export control (Category 3). The OLCF has a very small amount of Category 3 data and has a separate file system for Category 3 processing. The OLCF uses Discretionary Access

Controls (the Unix group memberships) for controlling access to data. The OLCF project ID is a logical container for access control; where sets of users are members of projects and have access to the same information. These mappings also carry over to the HPSS archive.

Additionally the OLCF sets secure defaults for permissions on scratch, project, and HPSS directories. The default of project team only for project areas, and user only for user scratch areas, Global home areas, and HPSS “home areas”. These permissions are enforced through our configuration management process (Cfengine), and users can change them by requesting the change via our [help@nccs.gov](mailto:help@nccs.gov) e-mail address.

Managing the Unix group memberships for users closely is a requirement in our environment as these group memberships control access to data that can be considered under export controls or confidential under industrial partnership agreements. Ongoing audits of the memberships of groups is part of the day-to-day accounts processing done by our User Assistance Group and the HPC Operations Infrastructure team. Additional logging infrastructure is being setup in conjunction with the Lustre purging process developed through cooperation with Operations staff at NERSC.

### **Technology refresh**

A key goal of the Spider parallel file system was to decouple the procurement and deployment of storage systems with that of large-scale compute resources at the OLCF. This goal has been realized and we are currently in the process of procuring our next generation file system for OLCF. To ease transition to these new file systems, for a period of 1 – 2 years, the current generation and next generation file systems will be operated in parallel. We have had success in migrating users between file systems, but the process is not without pitfalls and prone to compute users not paying attention to e-mail notifications and then having their jobs terminate abnormally as their application may expect to use a file system that is no longer in operations. Operating the file systems concurrently will allow users to make a gradual transition thereby minimizing the impact to our users.

Based on our enhanced understanding of I/O workloads of scientific applications, garnered from over 12 months of continuous monitoring of our file system environment coupled with a detailed understanding of our applications I/O kernels, we have developed an extensive set of benchmarks to evaluate storage system technologies offered by vendors. Our benchmarks are designed in a way that they mimic the realistic I/O workloads and also allow integrated and traditional block-based storage solution providers to bid on our RFP.

One of the biggest challenges in tape archiving lies in the area of media refreshment. While replacing, updating, or increasing the amount of front-end disk cache or servers responsible for data movement to/from disk and/or tape is a

relatively straightforward and non-intrusive process, the process of media refreshment presents many challenges. In a large-scale tape archive such as that at the OLCF, where 10s of thousands of individual tape cartridges are managed, at any given time there may be thousands of tapes housing multiple PBs of data needing to be retired from service. Unfortunately, the data on those tapes no longer resides on disk cache in most cases, and must be read from the older tapes in order to be written to newer media. Under real life conditions, where resource constraints such as utilization on data mover server(s) and the number of drives available to mount such media are a reality, the process of refreshing older media can literally take years. For example, here at the OLCF we are actively retiring 10,000+ 9840B tapes from service, and based on the performance to date, we expect that process to continue for the next 2.5 to 4 years. The OLCF has recently purchased a small quantity of 9840D drives so we can read the 9840B tapes at a 30% faster rate—in order to bring us closer to the 2.5 year figure. While that process is underway, we are simultaneously retiring several thousand 9940 tapes from service, and that initiative is expected to take approximately one year to complete as well. Media refresh(es) will continue to be a “day-to-day” operation going forward. For purposes of planning and procurement, it is assumed that 5-10% of total HPSS system resources will be utilized for media refresh operations.

### THE RELIABILITY AND AVAILABILITY OF STORAGE SYSTEMS

The OLCF tracks a series of metrics that reflect the performance requirements of DOE and the user community. These metrics assist staff in monitoring system performance, tracking trends, and identifying and correcting problems at scale, all to ensure that OLCF systems meet or exceed DOE and user expectations.

$$SA = \left( \frac{\text{time in period} - \text{time unavailable due to outages in the period}}{\text{time in period} - \text{time unavailable due to scheduled outages in the period}} \right) \times 100$$

**Scheduled Availability (SA)** measures the effect of *unscheduled* downtimes on system availability. For the SA metric, scheduled maintenance, dedicated testing, and other scheduled downtimes are not included in the calculation. The SA metric is to meet or exceed an 85% scheduled availability in the first year after initial installation or a major upgrade, and to meet or exceed a 95% scheduled availability for systems in operation more than 1 year after initial installation or a major upgrade. Reference Table 1.

**Table 1. OLCF Computational Resources Scheduled Availability (SA) Summary 2010–2011**

System	CY 2010		CY 2011 YTD (Jan 1-Jun 30, 2011)		
	Target SA	Achieved SA	Target SA	Achieved SA through June 30,	Projected SA, CY 2011
HPSS	95%	99.6%	95%	99.9%	>95%
Spider	95%	99.8%	95%	98.5%	>95%
Spider2	N/A	N/A	95%	99.9%	>95%
Spider3	N/A	N/A	95%	99.9%	>95%

System	CY 2010		CY 2011 YTD (Jan 1-Jun 30, 2011)		
	Target OA	Achieved OA	Target OA	Achieved OA through June 30, 2011	Projected OA, CY 2011
HPSS	90%	98.6%	90%	98.9%	>90%
Spider	90%	99.0%	90%	96.5%	>90%
Spider2	N/A	N/A	90%	99.1%	~99%
Spider3	N/A	N/A	90%	99.2%	~99%

**Table 2. OLCF Computational Resources Overall Availability (OA) Summary 2010–2011**

System	CY 2010		CY 2011 YTD (Jan 1-Jun 30, 2011)		
	Target OA	Achieved OA	Target OA	Achieved OA through June 30, 2011	Projected OA, CY 2011
HPSS	90%	98.6%	90%	98.9%	>90%
Spider	90%	99.0%	90%	96.5%	>90%
Spider2	N/A	N/A	90%	99.1%	~99%
Spider3	N/A	N/A	90%	99.2%	~99%

$$OA = \left( \frac{\text{time in period} - \text{time unavailable due to outages in the period}}{\text{time in period}} \right) \times 100$$

**Overall Availability (OA)** measures the effect of both *scheduled and unscheduled* downtimes on system availability. The OA metric is to meet or exceed an 80% overall availability in the first year after initial installation or a major upgrade, and to meet or exceed a 90% overall availability for systems in operation more than 1 year after initial installation or a major upgrade. Reference Table 2. As indicated by these numbers, both HPSS and our Spider file systems provide extremely high availability. Overall availability of these systems continues to dramatically exceed our operational requirements. The decrease in overall availability in one of our Spider file systems in 2011 compared to 2010 was due to an increase in the number of dedicated system times taken to evaluate new features and stabilize the next Lustre release. Spider2 and Spider3 remained available during these dedicated system times thereby minimizing impact to users.

Within HPSS, DB2 is used as the storage mechanism for all file/device metadata (ownership, status, location, etc.). DB2 has been proven in the field over many years and is well known for its reliability and availability features.

The front-end disk cache for the HPSS tape archive is comprised of several RAID6 arrays, with individual LUNS “owned” by mover servers responsible for data flow to/from disk. Currently, in our configuration here at the

OLCF, each mover has a single FC or IB path to a target LUN, but we are actively working on modifying that configuration in order to provide multipathing for our disk cache.

HPSS has the ability to store data on multiple levels of tape if so desired. Here at the OLCF, by default, data is written to one level of tape when migrated from the front-end disk cache. Users have the option of specifying a different “Class of Service” in order to have their data written to two levels of tape—providing an extra level of protection in case a media problem is encountered. Due to cost concerns, that is only encouraged and/or recommended for critical data.

While currently not in use at OLCF, HPSS does have High Availability capabilities based on Red Hat Linux cluster services. In this model, HPSS can provide failover redundancy for critical HPSS components—core server, data movers, and gateway nodes.

A feature that will soon be incorporated into HPSS is RAIT—Redundant Array of Independent Tape. RAIT will provide an additional level of redundancy and fault tolerance related to media failures without suffering the full cost penalty associated with the traditional method of having data on more than one level of tape.

#### **Maintenance Activities**

Maintenance activities for the Spider file systems are planned for once per quarter and planning for the “next” maintenance window begins shortly after the “previous” maintenance ends. It starts with a post-mortem analysis of the previous maintenance, and then developing a list of items to perform. At ~2 weeks pre-outage tasks are capped for the upcoming maintenance. A full outage plan is developed including any dependencies that the Lustre team has on other teams inside HPC Operations. This plan is documented on the internal wiki, and is shared through several normally scheduled weekly meetings as well as any outage/maintenance prep meetings. Coordination with the Facilities group is also necessary if one of the reasons for the outage is work being done to the power or cooling infrastructure. This planning process helps us to document upcoming changes/modifications, record their completion date, and also learn from issues that may come up during the maintenance – making the next maintenance hopefully smoother. They also help to enforce overall system knowledge in the administrative team and enforce, through the evaluation of the planned steps, a best practices approach to system administration.

#### **Data Integrity**

For Spider the RAID protections are the only data protections that are in place system wide. Applications can choose to add data protections in their simulation and modeling, but we’ve found that if we enforce anything it hinders performance and may not be what the application needs. End-to-end checksums are currently under evaluation for the next-generation Spider deployment.

End-to-end checksums is a feature recently introduced to HPSS. While not currently in use at OLCF, checksum utilities allow a user to perform a checksum of file content and place the results in a User Defined Attribute for later comparison if/when the file is retrieved [6]. At this time at the OLCF, individual users/departments in some cases perform pre and post retrieval checksums in order to verify data integrity.

#### **24 x 7 Support Model**

In support of a 24x7 operation, we use Nagios to monitor the correct configuration of the file system, and either through SMS messaging or direct phone calls from the 24x7 Computing Operations Center, notify the on-call administrator for the system of any critical event that causes availability to be degraded or lost. In the event of facility events during off hours, the Operations Center will call the HPC Operations Group Leader and then will notify affected teams. In the case of the Lustre team, scripts have been developed to quickly put the DDN controllers into power saving mode, power down the OSSes, MDSes, etc. to lower the heat load in the room. The DDN S2A9900 controllers have a disk sleep mode that parks the heads and spins down the disk.

#### **THE USABILITY OF STORAGE SYSTEMS**

Conventional methods for addressing I/O bottlenecks, such as increasing I/O backend capability by adding more disks with higher speeds, are unlikely to keep up with the performance issues due to the costs associated with storage. The problem is further exacerbated by the inefficiency of I/O performance; some applications are unable to achieve a significant fraction of the peak performance of the storage system. This can be due to a variety of factors the complexity of traditional I/O methods, where the developer has to make a heroic effort to optimize the application I/O. This limit on usability directly impacts the possible performance of the application. The OLCF has implemented a multi-point approach to addressing these challenges.

The ADIOS I/O framework was designed with the aforementioned concerns in mind. The ADIOS I/O framework [9] not only addresses the system I/O issues, but also provides an easy-to-use mechanism for the scientific developers to work from I/O skeleton applications. Through the use of an optional metadata file, which describes the output data and enables automatic generation of the output routines, the burden on the user is substantially reduced. ADIOS componentizes different methods of I/O, allowing the user to easily select the optimal method. In concert with data staging, this work exemplifies a next generation framework for I/O.

As common in many next-generation software projects, the the biggest challenge is often one of technology adoption, that is, getting users to change from current I/O implementations to ADIOS. As the ADIOS ecosystem continues to grow, we believe that ADIOS will gain a wider spread acceptance.

ADIOS and our eSiMon dashboard are used by the combustion, climate, nuclear, astrophysics, and relativity communities. In particular we have created a I/O skeleton generation system, using ADIOS, and have applied this in 10 applications, to make it easy for computing centers to analyze I/O performance from many of the leading LCF applications, with virtually no working knowledge of each application on their systems.

ADIOS has worked well with all current users, and have often shown over a 10X improvement of using other I/O implementations; see Figure 2 for I/O performance of the S3D and PMCL3D simulations on the Jaguar system.

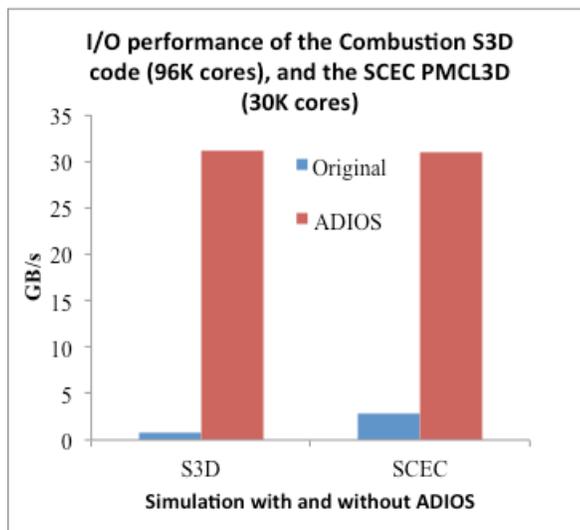


Figure 2. ADIOS performance.

## CONCLUSIONS

Oak Ridge Leadership Computing Facility has developed extensive developed key competencies in architecting and

administration of large-scale Lustre deployments as well as HPSS archival systems. Lessons learned from past Lustre and HPSS deployments and upgrades help us to better adopt to changing technology and user requirements.

## REFERENCES

1. Shipman, G; Dillow, D.; Oral, S.; Wang, F. *The Spider Center Wide Filesystem; From Concept to Reality*. In Proceedings of Cray User's Group 2009.
2. OpenSFS. <http://www.opensfs.org>
3. Nagios. <http://www.nagios.org>.
4. Lustre Monitoring Toolkit (LMT). <https://github.com/chaos/lmt/wiki>.
5. CFEngine. <http://www.cfengine.com>
6. Extreme Scale Storage for a Smarter, Faster Planet <http://www.hpss-collaboration.org/documents/HPSSBrochure.pdf>
7. Schlep, F. *The Guide to Better Storage System Operation*. Indiana University Press, Bloomington, IN, USA, 1973.
8. [Polte2009] Polte, M., Lofstead, J., Bent, J., Gibson, G., Klasky, S.A., Liu, Q., Parashar, M., Podhorszki, N., Schwan, K., Wingate, M. and others. "...And eat it too: High read performance in write-optimized HPC I/O middleware file formats". Proceedings of the 4th Annual Workshop on Petascale Data Storage, pp 21-25, 2009.
9. Lofstead2009] J. Lofstead, F. Zheng, S. Klasky, K. Schwan. Adaptable, Metadata Rich IO Methods for Portable High Performance IO. Parallel & Distributed Processing, IPDPS'09, Rome, Italy, May 2009, DOI=10.1109/IPDPS.2009.5161052.
10. SEC - Simple Event Correlator: <http://simple-evcorr.sourceforge.net/>.

# U.S. Department of Energy Best Practices Workshop on

## File Systems & Archives

San Francisco, CA

September 26-27, 2011

### Position Paper

Dominik Ulmer

CSCS

dulmer@cscs.ch

Stefano C. Gorini

CSCS

gorini@cscs.ch

#### ABSTRACT / SUMMARY

**The Swiss National Supercomputing Center CSCS has introduced a business model which turns data services from a reactively to a pro-actively managed service. A clearly defined center-wide file system hierarchy in conjunction with a set of specialized computers for data analysis allows to optimize storage systems characteristics like bandwidth or latency for different systems and workloads, to plan and manage capacities within the resource allocation process for computing time, and to leverage technical and financial synergies between the different service categories. Storage services are based on Luster and GPFS software with TSM/HSM extensions. A combination of different storage hardware technologies like SATA, SSD, and tape are used for the services depending on the individual requirements.**

#### INTRODUCTION

For many years, data was a side-business to computing for HPC centers. Large-scale storage systems were architected and installed as peripherals of a supercomputing procurement. However, data growth rates exceed performance

growth rates of HPC systems and therefore storage systems become an increasingly more significant part of the investment and operational budget of the computing centers. While the Swiss National Supercomputing Centre CSCS recognizes the importance of data for computational sciences, it does not have the intention to turn from a high-performance computing center to a data storage and management center. It is therefore essential to understand the role of data in the workflow of computational scientists using supercomputers and to accordingly architect the data services offered by the center.

#### SYSTEM BOUNDARIES

The main function of a HPC center is to enable computational scientists to use supercomputers for their research. This involves the preparation and the processing of large data sets, either for preparing input for computing runs or for analyzing data, which may be both, measured data or the result of a computational job. In both cases, one deals with living data for ongoing research projects, i.e. a time span that is well below 10 years. Long-term archiving for documentary purposes is not in the core business of a supercomputing center. A data service at a HPC center in this framework has to address the following topics:

- support of the computational workflow by means of an integrated architecture of computing, storage, and data analysis systems
- a storage hierarchy which is easy to understand by the user and provides a clear basis for the management of technical requirements
- a business model that allows the center to plan investments and operational costs in advance and which is aligned with the business model for providing computational resources.

## THE CSCS STORAGE HIERARCHY

CSCS distinguishes three different levels of storage (see [Error! Reference source not found.](#)):

### A) SCRATCH file system

The purpose of the scratch file system is to provide a storage container for running an individual computational job resp. an individual suite of computational tasks. Data remains only temporarily on the file system and must be copied to a different storage level for permanent storage. The file system has no quotas for user or groups. Old files are automatically deleted in order to maintain capacity. The scratch file system is local to an individual computer and its technical characteristics are specified according to the architecture of the system and the expected workload.

### B) PROJECT file system

The project file system provides a data management and storage space for an individual computational project. CSCS issues a call for project proposals twice a year. Researchers can

	Scratch	Project	Store
Size	Large	Very Large	Extreme size
Quota	No	By group	By consortium
Backup	No	Yes	HSM
Data life time	Wiped regularly by system every few weeks	Duration of project + 6 months	As contractually agreed
Locality	Local	Global	Global
Bandwidth	Very high	High	Good (if file on disk)
Current technology	Lustre	GPFS	GPFS
Allocation mechanism	None	Capacity requested and justified in project proposal	Contract; either matching funds or fully paid by customer

Figure 1: Hierarchy of file systems at CSCS

request computational and storage resources in their proposals, which are evaluated by an external committee with respect to their scientific quality and impact. The size of the storage request and of the compute cycle request must be justified in the proposal and must be coherent to each other. The project receives a storage quota which is shared between all members of the project team. The project file system is globally mounted on all CSCS user facilities and provides enough bandwidth for efficiently transferring large data sets to and from the scratch file systems. It provides extended user functionalities like snapshots. Data is kept on the project file system for the duration of the computational project (up to 3 years) plus 6 months in order to allow the user transferring the final data to a longer-term storage system or to the storage resource of a successor project.

### C) STORE file system

Large research projects are often carried out by consortia, which combine many research groups and projects as identified by the CSCS call for proposal process. Research consortia share data between the individual projects and teams and they manage the data sets over a longer timespan.

CSCS offers the store file system for such consortia. In contrast to the scratch and project file systems, resource on /store is not for free, but requires a financial contribution. Up to a certain limit, academic consortia can get storage space on store on the basis of matching funds. Above the limit and for non-academic consortia, direct investment and operational costs must be fully paid. A consortium must describe its overall research plan and goals, in order to assess the strategic importance of the consortium to science and the HPC center and to define the duration of the contract.

/store is a global file system that can be accessed from all user-accessible computers at CSCS. As it is based on a hierarchical storage management system, which is to a large extent based on tape, bandwidth is lower than to the project file system.

## TECHNICAL IMPLEMENTATION

All three storage levels are built with parallel file system technology in order to ensure performance scalability.

The scratch file systems are currently mainly based on Lustre, which allows for optimal read/write performance. Stability is sufficient and enhanced functionalities are not required because of the shared nature of the file system. Both, LSI and DDN storage controllers have been deployed for different implementations, mainly as direct attached scratch. Because of the meta-data performance bottlenecks in the current Lustre architecture, SSDs have been successfully tested for improving meta-data performance, although the fundamental problem of a non-distributed meta-data store can only be eased but not completely resolved with this approach.

The project file system is characterized by the combination of parallel HPC-type file system features with some enterprise storage requirements. It must be able to handle a large number of files with very good meta-data performance and has to offer functionalities like quota, snapshots, and integration with backup software. CSCS uses very similar storage hardware as on the scratch file systems, driven from separate storage servers that are connected to a high-speed Infiniband network backbone. GPFS has been selected as software technology for this file system because of its RAS features but also because superior meta-data performance compared to Lustre.

For the store file system, raw I/O performance is not as important as for the other two file systems. Technical and financial analysis showed that it is easily implemented with the same GPFS technology as /project combined with the TSM/HSM product of IBM. The TSM solution at CSCS also includes a backup and disaster/recovery functionality which enables us in the case of the total loss of the file system to recover all GPFS metadata within a few hours and all critical files within two days. By sharing licenses, infrastructure, and knowhow, operational costs can be kept low.

CSCS would be interested to change from the proprietary GPFS technology to open-source/public domain software. Lustre in its current state does not seem to be a viable option. If Lustre will be developed further in a coherent fashion, with stable funding and a clear roadmap, it could be envisaged to use in the future Lustre as the fundamental file system technology in combination with pNFS for mounting non-HPC clients.

As described above, we consider data analysis systems as an integral part of a data service. CSCS has decided to offer a portfolio of different computer architectures for data analytics: a standard, fat node cluster; a GPU cluster; a large-shared memory system based on the SGI Altix UV architecture; and a massively-multithreaded Cray XMT2 system. Access to these systems is granted within the allocation process for computing time on the main HPC systems.

## CONCLUSIONS

By defining a coherent business model for data and storage, the Swiss National Supercomputing was able to simultaneously optimize costs and scientific workflow at the center. For long-term sustainability, however, users additionally have to be educated to rethink their storage needs and patterns by means of in-situ data analysis and rewriting the I/O in their codes.

CSCS considers IBM's GPFS technology to currently be the most advanced solution for highly available and powerful *global* parallel file systems. We use GPFS with its characteristics of an enterprise file system only for the global levels of the storage hierarchy. Thus, the number of required client licenses can be drastically reduced by using a small number of I/O forwarding nodes per system. The bandwidth-hungry local scratch file systems, in which every compute node is a client, are built with Lustre. Solid-state memory technologies have developed into a viable alternative or addition to storage hardware solutions, boosting latency and IOPS-sensitive components of the storage system to new performance levels.



**U.S. Department of Energy Best Practices Workshop on**

**File Systems & Archives**

**San Francisco, CA**

**September 26-27, 2011**

**HPC Enhanced User Environment (HEUE) Position Paper**

**Thomas M. Kendall**

U. S. Army Research Laboratory  
DoD High Performance Computing Modernization  
Program

[Thomas.m.kendall4.civ@mail.mil](mailto:Thomas.m.kendall4.civ@mail.mil)

**John Gebhardt**

Lockheed-Martin/U. S. Air Force Research  
Laboratory  
DoD High performance Computing Modernization  
Program

[John.gebhardt@lmco.com](mailto:John.gebhardt@lmco.com)

**Cray J. Henry**

DoD High performance Computing Modernization  
Program

[cray@hpcmo.hpc.mil](mailto:cray@hpcmo.hpc.mil)

**ABSTRACT / SUMMARY**

The DoD High Performance Computing Modernization Program (HPCMP) is now implementing a major change to at all its DoD Supercomputing Resource Centers (DSRC) through the introduction of a center-wide file system (CWFS) and an integrated life-cycle management hierarchical storage manager (ILM HSM).

Following discussions with its top consumers of archival capacity, the HPCMP architected a strategy to enable its customers to reduce archival requirements. The key elements of the HPCMP's strategy are:

- Provide tools to enable customers to associate project specific metadata with files in the archive; enable automated scheduled actions keyed against specific metadata; enable users to control second copy behavior; and enable user specified logical data constructs suitable for building case management features.

- Provide an intermediate level of storage between the HPCMP's traditional two tier scratch and archive architecture. This intermediate storage (i.e. CWFS) will enable customers sufficient time to analyze results, and archive analysis results rather than 3-dimensional restart files. The center-wide file system is sized to allow 30 days of analysis before transfer to archive.
- The introduction of the center-wide file system also creates the opportunity to enhance and upgrade interactive customer support with high performance graphics and large memory to support the analysis efforts.

**INTRODUCTION**

The HPCMP built forecasts of future archival storage capacity needs based upon past history and concluded that the current growth rate was unsustainable. Left unchecked, storage would consume the majority of the HPCMP's budget by the end of the decade. A key finding of the subsequent analysis was that the archive costs remained in an affordable range if the archive

growth rate was constrained to 1.4 times the growth of the previous year's growth. This finding is tied to an assumption that industry doubles tape capacity every 24 months. If the growth rate of the archive exceeds the rate of tape capacity increase, the HPCMP has to fund tape libraries, slots, and potentially licensing for the additional capacity.

After gathering input from the principal investigators of the projects that consumed the vast majority of the program's archival capacity, several recurring themes emerged:

- Existing storage tools were insufficient to manage large datasets and the use of filenames to capture relevant metadata was no longer practical.
- Raw computational outputs were being archived due to insufficient analysis time for data stored in scratch space.
- Performing analysis using batch resources was adding to the problem of insufficient time for analysis.

These observations were further vetted and ultimately formed into requirements for the HPCMP's next generation storage solution. A working group, with representatives from the HPCMO, the DSRCs, and user advisory groups, was formed. The group was chartered to further develop and refine the requirements and to develop the architecture for data flow within the HPCMP.

The architecture that the storage working group arrived at included a combined information lifecycle management and hierarchical storage management layer.

A subsequent effort surveyed the information lifecycle management and high performance storage markets, leading to the creation of an acquisition strategy. A key element of this strategy was the separation of hardware and software requirements and provisioning.

A market survey determined, to no great surprise, that a mature information lifecycle management solution integrated with hierarchical storage management did not exist. The strategy that

emerged was to seek a partnership with industry aimed at fostering the integration of a leading information life cycle management solution with a leading hierarchical storage manager. Our software requirement allowed for an initial capability that could evolve into a fully integrated solution over 10 years.

The combined ILM HSM requirement was called "HPCMP Storage Lifecycle Management." A Request for Proposals was released in March of 2009 and a contract awarded in August of 2009.

With the ILM+HSM addressing the software requirements for improved tools to manage data, the remaining primarily hardware requirements were for the Center Wide File System and the Utility Server. This second component of the acquisition strategy was focused on the required hardware to deliver the new services.

Much of the requirements for the Center Wide File System (CWFS) and utility server derived from the winning ILM+HSM solution. Subsequently, two Requests for Proposals were issued. The RFPs required responses for the six DSRC locations and for a range of file system capacities and performance levels. They also required the inclusion of a 10 gigabit network fabric for connection of the Center Wide File System components, the utility server nodes, and the HPC system login nodes at each DSRC. The storage capacities requirements ranged from 250 TB to 2 PB. The I/O performance requirements ranged from 8.0 to 40.2 GB/s and from 70,000 to 320,000 file open/creates per second.

Awards for the CWFS and utility server were made by Lockheed-Martin in September 2010 and deliveries were completed in December, followed by acceptance and integration. The systems were transitioned into production sequentially center by center between June and August 2011.

In 2011, the HPCMP also took steps to refresh its tape archive hardware. Based on the earlier analysis showing that a doubling of tape capacity every 24 months was a key component of cost containment, the program compared commodity LTO drives with the proprietary Oracle T10000

line. Although the LTO family drives have a lower initial purchase price than the T10000 family drives, the lifecycle costs for LTO were found to be significantly higher. Drivers were the need to replace the media with each generation of LTO drive in order to realize the increase in capacity and the slightly slower capacity growth rate (15x over ten years for LTO and 10x over five years for T10000).

## **CONCLUSIONS**

It is too early to gage the impact on the growth of the HPCMP's archive as a result of the acquisition and deployment of the HPCMP

Enhanced User Environment. The Program's advisory bodies have responded positively to the goals and progress. The initial feedback for the additive analysis capability provided by the utility server and Center Wide File System has been positive. Once the Storage Lifecycle Management solution is fully deployed for production use in October 2011, the full effects of HEUE will be measured and reported.

In terms of the adoption of commodity hardware, the tape industry would need to seek ways to reduce the frequency of complete media replacements while meeting or exceeding the 1.4 compound annual capacity increase target.

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper**

**Stephan Graf**  
Jülich Supercomputing Centre  
st.graf@fz-juelich.de

**Lothar Wollschläger**  
Jülich Supercomputing Centre  
l.wollschlaeger@fz-juelich.de

**ABSTRACT / SUMMARY**

The storage configuration for the supercomputer *JUGENE* in Jülich consists of a GPFS cluster (*JUST*) and two Oracle STK SL8500 tape libraries. In this paper the actual configuration and the next upgrades are described. Furthermore a project for using flash storage as a kind of cache memory for the disk storage is introduced.

**INTRODUCTION**

The Jülich Supercomputing Centre (JSC) operates two supercomputer: The BlueGene/P System *JUGENE* and the x86 based *JuRoPA* system. While the *JUGENE* uses the remote GPFS cluster *JUST*, the *JuRoPA* users works on a local Lustre based storage. There the users can access their files in the GPFS file system via dedicated nodes.

The consideration in this paper for the actual and the future storage configuration/implementation are focused on the *JUGENE* and the *JUST* GPFS cluster.

The users can access three types of file systems:

On \$HOME they should store there code and develop their program.

For the job run they are urged to use the scratch file system \$WORK to get the maximum IO performance.

To archive their results the data should be moved to the \$ARCHIVE file system.

**JUGENE STORAGE PERFORMANCE TODAY**

The *JUGENE* is build up of 72 BlueGene/P Racks with 1 PF peak performance and 144 TiB

main memory. Each rack contains 1024 compute nodes (CN) and 8 IO nodes (576 IO nodes in total), with each one connected via 10GbE to the *JUST* storage. Measurements show that a single IO node gets an IO performance of 450 MB/s reading and 350 MB/s writing. For a whole rack it is 3.6 GB/s reading and 2.8 GB/s writing. The maximal peak IO for the full system is 260 GB/s reading and 200 GB writing. Assuming that 50% of the main memory of one rack (1024 CN) is to be written on file system (e.g. for checkpointing), the required time is 5 minutes for reading and 7 minutes for writing. To write 50% of the main memory of the full system in 15 Minutes requires  $0.5 * 144 \text{ TiB} / 1800\text{s} = 44 \text{ GB/s}$ . The *JUST* cluster based on DS5300 storage devices provides 66 GB/s. But only half of the cluster is used for the fast scratch file system \$WORK. The other half of the clusters hosts the \$HOME and \$ARCHIVE file system. This implicates that the \$WORK can be saturated by  $33 \text{ GB/s} / (8 * 0.35 \text{ GB/s}) = 12$  racks writing to the file system.

On the *JUST* cluster 8 building blocks provides the \$WORK file system, each containing a DS5300 with 36 LUNs per DS5300 having a size of 8 TB (RAID6). This leads in a total capacity for \$WORK of 2.3 PB.

**BLUEGENE/Q INSTALLATION IN 2012**

In 2012 the *JUGENE* will be replaced by a BlueGene/Q system consisting of 6 racks. There are 8 IO nodes per rack, each having a dual 10GbE port with an aggregated bandwidth of 1.5

GB/s. So the maximum throughput of a rack is 12 GB/s and the full system 72 GB/s.

If 50% of the main memory of on rack (1024 CN with 16 GB RAM per node) are to be written on disk it will last (approximately) 12 minutes. So to write 50% of the full system main memory to the storage in 15 minutes, a bandwidth of  $50\% \cdot 384 \text{ TiB} / 1800\text{s} = 115 \text{ GB/s}$  are required. Therefore we will get a storage upgrade for the *JUST* GPFS cluster. We are planning to install 8 DDN SFA12000 and getting an aggregated bandwidth between 100GB/s and 160 GB/s for the scratch file system \$WORK. The performance for \$HOME and \$ARCHIVE will also increase, but this is not concerning us.

## FLASH MEMORY AS SCRATCH FILE SYSTEM

In parallel the JSC will investigate a new storage concept using flash memory cards as a kind of cache between the IO nodes and the ordinary disk storage. It is a European Union funded project, a PRACE (Partnership for Advanced Computing in Europe) prototype for next generation supercomputers.

4 x86 systems each with 2 fusionIO ioDrive Duo 320GB SLC will be set up. The bandwidth of the flash card is 1.5GB/s. The cumulated performance of these 4 nodes should be 12 GB/s, a similar value as 8 BlueGene/Q IO nodes (one rack). Using the GPFS features to setup different kind of storage pools and to implement placement and migration policy rules a concept will be modeled, that new created files will be created on flash, and GPFS will migrate the files to disk in the background automatically.

This concept will be implemented with real BlueGene/Q hardware as soon as it is available.

## ARCHIVE STORAGE EXPANSION

The users should store their results on the \$ARCHIVE file system. There the data will be migrated by Tivoli HSM on tapes (weighted by *size* and *last access time*). For safety two versions (COPYPOOL) will be held on tape. Furthermore every file of the \$HOME and the \$ARCHIVE file system will be backed up. Files on the scratch file

system \$WORK are not backed up and will be deleted after 90 days.

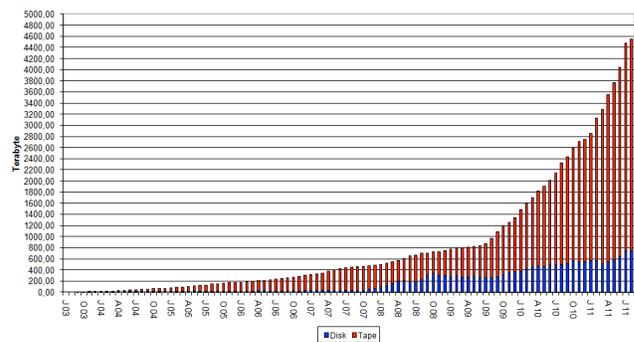


Figure 1: Data growth on the GPFS cluster JUST

The JSC operates two Oracle STK SL8500 libraries with an aggregated capacity of 16.6 TB (T10K-B tape drives). In figure 1 the exponential data growth on our storage cluster can be seen. We expected to run out of space in the third quarter 2011. Because of the ordering and shipping delay of the new hardware it became critical the last month. But now the new hardware has arrived and is going in production. 16 T10K-C tape drives have been added and the new tape generation (which is able to store 5 TB) will replace the old tapes step by step.

This kind of upgrade is the typically way for us to manage the growth of data amount. For the next 6 years we plan to enlarge the capacity of the two libraries to 80 PB just by upgrading to the next tape drive generation T10K-D.

## CONCLUSIONS

On our supercomputer a specific maximal I/O performance is available and for the user it is reasonable to get the maximum performance from the file system. But this is often difficult to achieve. Therefore it is mandatory to train the users and give them the knowledge to speed up their jobs I/O. For this purpose we have developed the *SIONlib* in Jülich. The users can use this library in there code to map very easily local task I/O to one file. By using the *SIONlib* it is possible to get nearly 100% of the performance on the scratch file system \$WORK from the JUGENE. We also use this tool for benchmarking parallel file systems.[1]

The other subject concerns the long time data storing. Till now and for the next years we are able to store all user data in our archive system. The new technologies keep up with the data growth in Jülich. But there are upcoming questions like how long must the data be hold or

what happens when a project ends. These problems must be tackled in mid or long term.

## **REFERENCES**

. [1] <http://www.fz-juelich.de/jsc/sionlib/>

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011**

**Andrew Uselton**  
NERSC/LBL  
acuselton@lbl.gov

**Jason Hick**  
NERSC/LBL  
jhick@lbl.gov

**ABSTRACT / SUMMARY**

This position paper addresses the business of storage systems and practices related to planning for future systems (I-1A) and establishing bandwidth requirements (I-1B), with some discussion also relating to the administration of storage systems and the monitoring of specific metrics (II-2A). The best practice is to balance I/O with compute capability.

We present a quantitative characterization of “HPC and I/O system balance” by examining the relative costs of compute resources and I/O resources on the one hand and the relative impact of compute and I/O activities on the other.

**INTRODUCTION**

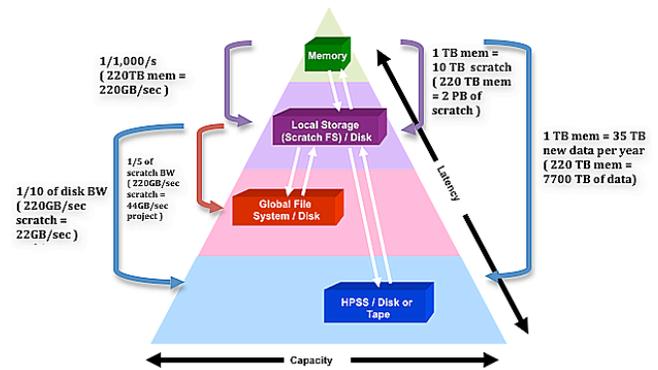
An HPC system with too little I/O infrastructure to support its workload could leave much of the compute resource idle as it waits for I/O operations to complete. The idle compute resource represents an opportunity cost in that it may have no other useful work to do during the wait.

**BACKGROUND**

One study [2] suggests that memory capacity is the key determinant of necessary I/O bandwidth and capacity. Figure 1 presents a traditional guideline for balancing I/O.

The relationship between performance and memory comes from the need to flush the

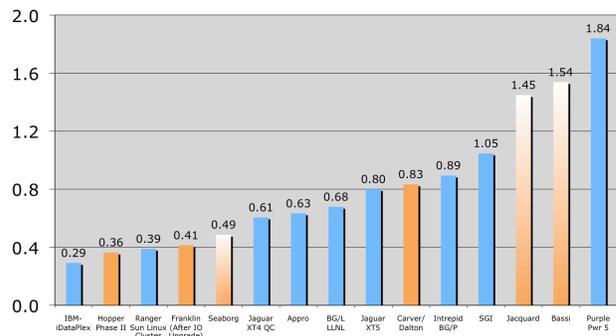
contents of memory to persistent local storage in a combination of reasonable time and cost.



**Figure 1. Conventional HPC I/O Planning Guidelines**

Additional I/O resources provide diminishing returns, so there is a point of balance at which bandwidth is “just enough”, and in this case the heuristic is to move all of memory in about 1000 seconds.

*Estimated Peak IO Bandwidth (GB/Sec) / Total System Memory (TB)*



**Figure 2. Peak bandwidth to system memory**

Figure 2 presents this heuristic as applied to several HPC systems. By that metric, systems with a value over 1.0 have over-provisioned I/O

subsystems relative to system memory capacity. A subjective review of such systems reveals that users are happy with the I/O bandwidth they deliver.

The purpose of this paper is to propose an alternative characterization of balance using a cost-based model in conjunction with the compute and I/O workloads of the HPC system. As a starting point, this discussion abstracts away much of the complexity to arrive at some core ideas.

## SYSTEM BALANCE

As a first simplifying assumption, suppose that the cost of an HPC system is composed entirely of the budget for the compute capability and the budget for the I/O capability. Next, suppose that the work produced by an HPC system is measured as the number of jobs completed weighted by the size of each job in two dimensions: the number of *node-seconds* used in the computation and the number of *node-seconds* used in I/O.

Further, assume that the aggregate compute capability is near linear in the cost of the of compute nodes:

$$C(n) = M_n \times n$$

where  $M_n$  is the marginal cost of nodes. Similarly, assume the aggregate I/O capability (measured as its peak rate) is near linear in the cost of the I/O infrastructure:

$$I(r) = M_r \times r$$

where  $M_r$  is the marginal cost of adding a unit of bandwidth  $r$ .

Now let the utilization  $U$  of the HPC system be given by the fraction of *node-seconds* spent on compute activity, given a particular workload. Our characterization of "system balance", given  $n$  and  $r$ , is given by:

$$B = \left( \frac{U}{1-U} \right) \left( \frac{I(r)}{C(n)} \right)$$

Our claim is that at  $B = 1$ , the system is in balance in that it achieves the maximum amount

of workload per dollar spent. As an example, if you spend 10% of your HPC system budget on I/O infrastructure ( $\frac{I(r)}{C(n)} \cong 0.1$ ), then the nodes should be spending 10% of their time on I/O, and the rest on computation ( $\frac{1-U}{U} \cong 0.1$ ).

This is a relatively intuitive idea given the simplifying assumptions, but it begs the question, "What is  $B$  on my system, given its workload?" Those who design and purchase HPC systems are very familiar with the total cost and the fraction spent on I/O infrastructure. On the other hand, it is not at all clear what the value of  $U$  is. It will certainly be different at different times and for different workloads. We propose that monitoring the jobs and I/O on the system for any given day's activity and for longer intervals will yield the value of  $U$ .

## CHALLENGES

Some system designs and I/O strategies attempt to improve I/O performance by departing from this simple model. For example, a strategy that overlaps computation and I/O will yield a higher utilization. In that case it becomes important to estimate both the expected impact and the extent to which the strategy is implemented in practice. If a strategy can entirely "hide" I/O activity but only affects 10% of the workload, then the simplified model is still close to correct.

The model has plenty of room for improvement. For example, the cost model does not need to as simple as presented. There may be fixed costs and nonlinearities, and the model could incorporate them without difficulty. The model can also include other aspects of HPC system architecture, for example, adding *node-seconds* spent in (node to node) communication. In some cases that communication will compete for bandwidth with the I/O requirements, leading to additional complexities.

## CASE STUDY

The *Carver* IBM Dataplex cluster at NERSC was provisioned with approximately 15% of its budget dedicated to I/O infrastructure. The

system has 30 TB of memory suggesting a target bandwidth to storage of 30 GB/s using the heuristic from the background discussion. *Carver*'s measured bandwidth is about 25 GB/s, so it is designed to be at about 83% of that target. *Carver* runs the Integrated Performance Monitoring (IPM) library [3] with every scheduled job. Each job produces a report at the end of its execution giving the time spent in computation, the time spent in I/O, and the amount of data moved (among other quantities). IPM provides a comprehensive profile of compute and I/O activity for a given interval. From that profile it is possible to directly calculate the utilization. For example, in June 2011  $U \approx 0.94$ . The balance factor for the actual workload is around 2.5. By this measure, the system's balance favors I/O and could handle a heavier load.

### CORRELATING I/O ACTIVITY WITH JOBS

Most HPC systems do not have IPM or other direct measures of the utilization. Without that information we do not know what balance has been achieved in practice after having applied the heuristics from Figure 1. An alternative strategy is under development at NERSC that infers the utilization  $U$  from server-side I/O monitoring with the Lustre Monitoring Tool (LMT) [4]. Server-side data is anonymous with respect to the nodes that generate the I/O. Nevertheless, it is often possible to infer the job from the I/O pattern. When that can be done comprehensively it will yield the utilization as before, and therefore give a quantitative gauge of the balance.

On NERSC's *Franklin* Cray XT4 there are commonly more than one hundred jobs running at a given time, and the I/O workload resulting from that compute workload is potentially composed of I/O from many jobs simultaneously. Often, an application runs many times repeating the same I/O pattern each time. From that collection of jobs (call it a *job class*) we calculate the average I/O behavior for the application, which is an approximation of its expected behavior in isolation from other jobs. The individual calculated behavior of each of the whole suite of

applications provides the initial estimate for the behavior of the system as a whole, and the estimates can be iteratively refined via a generalized linear regression. This is a computationally expensive task but straight forward, in principle.

As an example, we examine the IOR [5] file system benchmark, which runs as a regularly scheduled test of the *Franklin* scratch file systems. 175 such tests were run in July 2011, and the system job log records the start and stop time of each job along with the number of nodes used. The IOR application runs using the same parameters in order to provide a repeatable health-check of the file system. Each job writes 4GB to the file system from each of 64 tasks on 16 nodes. It then reads that data back in. Jobs generally run for 150 to 200 seconds, but can run much longer when the file system is occupied with other I/O. The jobs are submitted to an I/O-oriented scheduling queue, which (voluntarily) serializes I/O intensive applications.

LMT records the bytes written and bytes read for each server every five seconds. That data shows the I/O resulting from the IOR jobs and anything else running at the same time. In order to calculate the average behavior we "warp" (artificially lengthen or shorten) the sequence of LMT observations for each job so that they fit the same length-scale – chosen as the median job run length. A standard linear regression on that data set provides the calculated average behavior.

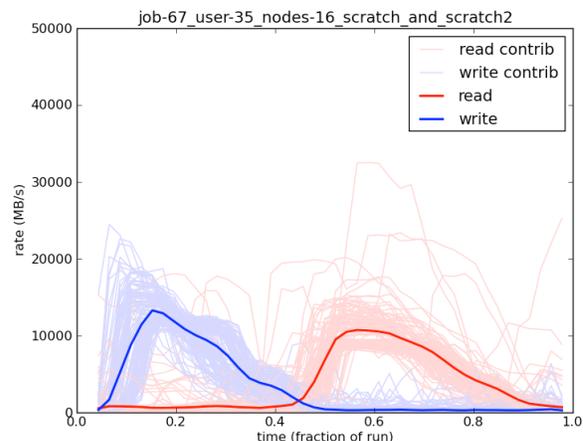


Figure 3.

Figure 3 displays the result of carrying out this analysis. The  $x$ -axis is the artificial time scale – arbitrarily set to 0 to 1 – to which each series of observations is warped. The  $y$ -axis gives the aggregate data rate (blues for writes and red for reads) of the application over the course of the idealized run. The single dark line of each color is the calculated average behavior of the application. Shown in a lighter shade is the collection of 175 separate contributing runs as they appear after being warped. Most of the contributing runs follow the average behavior closely, and demonstrate that the IOR test was running without much interference. A few traces depart wildly from the average and it is those runs that were in contention for I/O resources. Once we calculate the idealized average behavior all of the applications with significant amount of I/O, those idealizations become initial estimates for the coefficients in a big matrix implementing the generalized linear regression. For a file system that does not have a lot of I/O contention the initial estimates will be close to their final values and the computation will converge quickly. In other cases the computation may take significant resources. The end result is a quantified, job-by-job measure for the impact of the application on the I/O system from which we recover the utilization  $U$  and therefore the balance  $B$ .

## CONCLUSIONS

The HPC community has developed a set of heuristics to guide the design of HPC systems so that the I/O capability is matched to the users' needs. In one case where the heuristic was applied in an effort to make a system *I/O-friendly*, a comprehensive characterization of balance using our metric showed that the system deployment was successful. Our measure for balance,  $B = 2.5$ , says that the system is cost effective for an even heavier I/O load than was observed.

The proposed job-log-and-LMT analysis extends the applicability of our metric to cases where direct observation of the utilization  $U$  is impossible or impractical. That characterization can be combined with system cost details to

establish a rigorous evaluation of the balance of the system.

It is always difficult to argue that past performance is guide to future behavior. When planning for a new HPC system the application behavior produced in this analysis must be combined with theoretical considerations for how the new system might behave differently. Nevertheless, the application characterizations and the underlying model that produced them are a valuable starting point to be used during the procurement of new systems.

Once deployed, a new system needs data acquisition systems like IPM, Darchan [6], and LMT in order to evaluate the system balance actually achieved.

## ACKNOWLEDGEMENTS

We would like to thank Dr. Anthony Gamst of UC San Diego for discussion, explanations, and guidance of the generalized linear regression techniques alluded to in this paper.

## REFERENCES

1. NERSC Storage Policies, produced by the Storage Policies Working Group, 2010.
2. J. Shoopman, LLNL internal study on memory capacity to archive data generated 2008-2009.
3. D. Skinner, *Integrated Performance Monitoring: A Portable Profiling Infrastructure for Parallel Applications*. Proc. ISC 2005, International Supercomputing Conference, Heidelberg, Germany, 2005
4. A. Uselton, *Deploying Server-side File System Monitoring at NERSC*, Cray User Group Conference, Atlanta, GA, 2009
5. H. Shan, K. Antypas, J. Shalf, *Characterizing and Predicting the I/O Performance of HPC Applications Using a Parameterized Synthetic Benchmark*. Proc. SuperComputing 2008, Austin, TX, 2008
6. P. Carns, R. Latham, R. Ross, K. Iskara, S. Lang, K. Riley, *24/7 characterization of petascale I/O workloads*. Proc. Of 2009 Workshop on Interfaces and Architectures for Scientific Data Storage, Sep. 2009.



# U.S. Department of Energy Best Practices Workshop on

## File Systems & Archives

San Francisco, CA

September 26-27, 2011

### Position Paper - Business Breakout

**Kim Cupps**

Lawrence Livermore National Laboratory  
cupps2@llnl.gov

#### **ABSTRACT / SUMMARY**

**This “Business of File Systems and Archives” position paper will describe several LLNL best practices that help formulate and optimize the cost/benefit analysis all center’s face in their quest to provide an ongoing, exceptional computing environment within a finite budget.**

#### **INTRODUCTION**

LLNL’s primary computing complex serves approximately 2800 users with access to 1.7 peak PetaFLOPs of compute, 14PB and 300GB/s of parallel file system capacity and bandwidth and 42 PB of archival storage. LLNL seeks to optimize the user experience, providing a long-lived, highly productive environment for our customers, while staying within our budget. While cost/benefit is easy to say, it’s a complicated balance of usability, availability, flexibility of administration, longevity, productivity and many other factors that form the basis of our spending decisions for file systems and archives. There are several practices we use to help us make decisions on what and how much to buy, what improvements must be made and which software to develop ourselves and which to buy off the shelf. First, we place a high value on gathering direct user input via a variety of mechanisms including user meetings, surveys and customer interviews. Another best practice is to

monitor and measure use of resources and plan buys “just in time” (JIT). Strong partnerships with vendors as well as hedges against the trap of becoming beholden to a single vendor or technology for file systems or archive is another best practice. Lastly, planning is crucial to any coherent business strategy. LLNL formalizes the plan for file system and archive resources by producing the “I/O Blueprint”, a procurement and effort planning prioritization document.

#### **Gathering User Feedback**

Making sound business decisions requires a good understanding of the current state of affairs. It’s quite easy to live in a world isolated from those who use the file systems and archives every day, just as it is common for users to work around issues and inconveniences rather than report them. We have the typical trouble ticket system user surveys to help us understand issues; this is feedback we receive on a daily basis. In addition, LLNL holds quarterly user meetings that include a user talk as well as a set of talks on relevant center activities. These meetings include a general feedback session. Most important, LLNL rotates through “Science Team Interviews” so that we meet with teams every two years or so to elicit actionable feedback. A team of center personnel representing management, platform, file system, archive and user services personnel goes out to the customer work area and asks

pointed questions about the compute environment. In general, we find that problem areas spring out of discussion and are not the sorts of issues that people call and report, often they don't even write down the issue in advance of the meeting. For example, the development of HTAR, a multi-threaded file packaging and transfer mechanism from the local file system to the HPSS archive, was the direct result of complaints received from users regarding slow transfers of small files to HPSS. The development of Lorenz, our user dashboard that shows file system usage, NFS quotas and many other customizable fields was also the result of strong user collaboration and input. All of these user feedback mechanisms serve to inform the center about where our customers feel we have the most room for improvement – this is critical to our planning.

### Measuring and Metrics, JIT

Another way we identify areas for improvement is by collecting data on various aspects of the production environment. We gather data on everything from component failure rates, to sizes of files stored in the file system, to file system specific and center-wide uptime percentages for both classified and unclassified file systems and archives.

Unclassified Lustre File System Availability Statistics					
8/1/11 - 9/1/11					
File system	Unplanned		Planned		Uptime %
	Impaired	Down	Impaired	Down	
Center Wide	0.83	3.42	0	0	99.86%
lscratch a	0	1.58	0	0	99.79%
lscratch b	0	0	0	0	100%
lscratch c	0.08	1.75	0	0	99.75%
lscratch d	0.75	0.08	0	0	99.89%

Tracking isn't limited to analysis of failures and availability, it's also crucial to our "just in time" purchase strategy for archive media and tape drives. Cartridges are used up by both a steady stream of new data being written to tape as well as an ongoing repack from soon-to-be-retired media of about 500 cartridges a month. Careful tracking of cartridges insures that tape buys are done on-time, but not so far in advance that the tape is never used. Just-in-time has been shown to

be the most cost efficient way to purchase consumables, and with tape densities doubling (recently quintupling) every year and a half, the value of this best practice is clear.

### Partnerships Coupled with In-House Software Expertise

Strong vendor partnerships are crucial to successful operation of a center. Changing vendors incurs added costs such as retraining staff, forming new relationships and learning new support processes. However, becoming a strictly one vendor operation significantly increases risk. These risks include the company going out of business, dropping support for the product line or unreasonably raising prices. A best practice at LLNL that reduces the inherent risk of the strong vendor partnership, is a staff of in-house software developers for both open source projects (Lustre, SLURM, RHEL) and joint development contracts (HPSS). The software developers provide key value by:

- 1) Solving production problems immediately (increasing system uptime and thereby user productivity);
- 2) Providing a strong voice for DOE HPC requirements
- 3) Providing a mechanism for the center to remain technically competent and engaged in leading edge technology

Other important components of strong vendor partnerships include membership on vendor customer advisory boards, leadership of product user groups and regular attendance at vendor executive level roadmap briefings.

### Advanced Technology and Testbeds

The Hyperion testbed at LLNL includes an 1152 node QDR IB interconnected commodity Linux cluster with two SANs (GE, IB) connected to multiple vendor storage subsystems. The testbed serves multiple purposes. It is a partnership that allows vendor partners to test their software and hardware at scale. It is a platform that allows LLNL to investigate interesting technologies

(NAND Flash, tiered storage, NFS accelerators...) as they become available. WhamCloud performs Lustre testing at scale. Mellanox tests new cards and drivers. DDN and Netapp test new controller technologies and LLNL investigates all of this new technology before it comes to market.

The Lustre testing at scale on Hyperion, and LLNL’s participation with others in the concept of creating “Lustre Centers of Excellence” is an excellent example of how investment in strong vendor partnerships and testbeds can significantly impact the quality of a product.

### Planning

LLNL produces a yearly planning document called the I/O Blueprint. The goal of the Blueprint is to achieve a balanced infrastructure to support the Center’s compute platforms. The Blueprint documents planned purchases in global parallel file systems, NAS, visualization, network and archive areas. It also discusses area specific Center issues and plans for remediation.

The FY05 Blueprint is the document that called for converting from dedicated filesystems to a site-wide global parallel file system. During the era of local file systems, each platform purchase required that dedicated file system hardware be bought for use solely by that platform. Without global file systems, a platform was only able to leverage the speed and capacity that it came with and data needed to be moved or copied to each platform when required. Today, global file systems allow new platforms to leverage existing disk resources and allow existing platforms to take advantage of global resources added over time. As a result we are able to enhance file system resource utilization, eliminate the copying of data, ensure that file system hardware is best-of-breed rather than that available from a particular platform vendor, and focus on center-wide I/O requirements rather than that of individual machines. In short, there is a clear cost/benefit win.

Calculating bandwidth and capacity requirements for archive, file system and networks is not an exact science, and we have used different “rules of thumb” over time to plan purchases. For example, directly copied from the FY08 I/O Blueprint: *“The rule of thumb used in the past for capability platforms was that the file system should provide between 100MB/s and 1GB/s of bandwidth for every TeraFLOP. Dawn file system bandwidth requirements are projected to be 200MB/s per TeraFLIN and Sequoia is projected to be 100MB/s per TeraFLIN leading to 100GB/s and 500GB/s estimates for delivered SWGFS bandwidths for these machines.”*

From our FY11 Blueprint: *“For many years now, bandwidth has been the basis for our file system procurements. We believe that our bandwidth requirement is a function of platform memory. Typically our ratio of file system bandwidth to platform memory (GB/s per TB of memory) has varied from 0.6 to 0.8. Currently that ratio is at 0.5GB/s/TB in the OCF and 0.6GB/s/TB in the SCF. Note that the Sequoia file system is currently 0.4GB/s/TB.”* As they should, requirements definition methodologies have changed as architectures evolve and lessons are learned.

SCF Max Lustre Bandwidths

	Iscratch1	Iscratch2	Iscratch3	Iscratch4	Iscratch5
Coastal	15GB/s	15GB/s	15GB/s	40GB/s	40GB/s
CSLIC	1.25GB/s	1.25GB/s	1.25GB/s	1.25GB/s	1.25GB/s
Eos	10GB/s	10GB/s	10GB/s	10GB/s	10GB/s
Gauss	15GB/s	15GB/s	15GB/s	15GB/s	15GB/s
Graph	15GB/s	20GB/s	15GB/s	15GB/s	15GB/s
Inca	125MB/s	125MB/s	125MB/s	125MB/s	125MB/s
Juno	15GB/s	15GB/s	15GB/s	40GB/s	40GB/s
Minos	15GB/s	15GB/s	15GB/s	30GB/s	30GB/s
Muir	15GB/s	15GB/s	15GB/s	40GB/s	40GB/s
Rhea	15GB/s	15GB/s	15GB/s	20GB/s	20GB/s

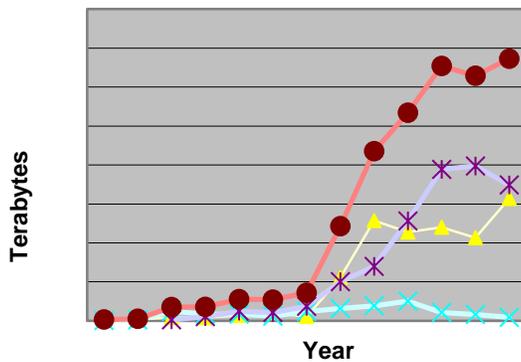
- File System Disk Limited
- Network/Lustre Router Limited
- Node limited (i.e. Single node job scheduling)

The Blueprint is also the document where LLNL outlined an initial plan to address exponential archive growth and associated unmanageable out-year costs. The LLNL Archive Quota implementation was planned as a first mitigation. Archive advisory quotas were implemented in

December of 2010. Initial talks with users were held beginning in August of 2009. Archive growth rates have slowed. We conjecture that this slow down is due to a number of factors, not just the quota implementation. First, simply communicating the cost of storing a particular user's data resulted in that user deleting over 1PB of data in the archive. Raising awareness of costs is a best practice that we have used with very good results. At LLNL, nothing is archived automatically; all transfers are initiated by users. Activities that increase storage to the archive include aggressive global parallel file system purge policies, file system retirements and planned file system down times. A reduction in one causes a reduction in the other. Finally, the biggest factor in archive growth is platform memory capacity. As new large platforms are added, archive growth increases. We expect substantial archive growth with Sequoia and we expect the archive advisory quota implementation to help contain the rate of growth over time. While we expect the Quota implementation to help, it's too early to claim it as a best practice.

forward and providing the best environment possible.

### SCF Writes



### CONCLUSIONS

Providing a balanced infrastructure to optimize user productivity, while minimizing costs, requires attention and focus in a number of areas. The cycle of events includes formalized planning, which is informed by regular collection of data and metrics, user feedback, advanced technology investigations and testbed evaluations. Strong vendor partnerships and in-house software expertise are key enablers to quickly moving

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper**

**David Cowley**  
Pacific Northwest National Laboratory  
david.cowley@pnnl.gov

**ABSTRACT / SUMMARY**

**The EMSL facility, located at Pacific Northwest National Laboratory (PNNL), operates terascale HPC and petascale storage systems to support experimental and computational researchers in molecular sciences. This position paper addresses the Workshop's Business of Storage Systems track and describes EMSL's approach to operating file systems and data archives.**

**INTRODUCTION**

The Environmental Molecular Science Laboratory (EMSL) is a scientific user facility located at PNNL. EMSL houses PNNL's largest concentration of high performance computing systems and data storage systems. While other organizations within the Laboratory are working on obtaining their own significant HPC and data storage resources, the center of mass has not shifted yet. We will be careful in this document to distinguish between PNNL and other sub-organizations within PNNL, including EMSL.

EMSL operates a suite of cutting-edge scientific instruments, capable of generating terabytes of data per week. EMSL has operated HPC systems ranked in the top 20 of the Top500 list since 2003, in addition to a multi-petabyte archive for scientific and computational data. EMSL has been working since 2010 on a scientific data and metadata management system known as MyEMSL.

**GENERAL APPROACH TO STORAGE SYSTEMS**

EMSL HPC systems have had more types of filesystems than at most HPC sites. Each is intended to meet different levels of capacity, performance, and accessibility. So far each of EMSL's HPC systems has been procured with its own filesystems of 3 types:

<b>Filesystem Type</b>	<b>Capacity</b>	<b>Nominal Bandwidth</b>
Global Home	20 TB	1 GByte/sec
Global Scratch	277 TiB	30 GiByte/sec
Node Scratch	350 GiB/node 808 TiB Aggregate	400 MiByte/sec/node 924 GiByte/sec Aggregate

The global home filesystem is available to all nodes in the HPC cluster. its capacity is determined loosely by a "not too big to be backed up" rule of thumb, its performance is determined loosely by a "'cd' and 'ls' commands have to not be slow" rule of thumb.

The global scratch filesystem is a parallel filesystem both larger and higher performance than the home filesystem. It is available to all nodes in the HPC cluster, and its performance and capacity requirements have been derived from a formula based on the theoretical peak

performance of the system. EMSL does not have a requirement to checkpoint whole-system jobs as some other sites do, so this eases some of the requirements on this filesystem.

The node scratch filesystems provide high disk bandwidth per Flop to each node. For these filesystems, performance in terms of write bandwidth and Ops/second are again derived from theoretical peak performance on the compute node. Capacity has been a side effect of the need to provision enough disk spindles to meet the required performance. This may change as magnetic disk and solid-state disk technologies evolve. While providing a scratch filesystem on each compute node does involve considerable cost and added maintenance, the aggregate performance has been more scalable and better in absolute terms than shared parallel filesystems. EMSL has found this to be a differentiating and enabling capability, and will carefully consider it in its upcoming system procurements.

EMSL is planning to move to a "two systems" approach where rather than procuring one large system every 3 to 4 years, we will procure smaller systems every two years and overlap their lifecycles. We will switch to having the home filesystem shared between compute clusters. We expect that each cluster will have its own high performance parallel global scratch filesystem. We will consider critically whether new systems require node scratch filesystems.

### **MANAGING ARCHIVE GROWTH**

EMSL's growth in archive capacity is driven by two factors, the output of scientific instruments and the output of its HPC systems. In effect, the scientific instruments are computers themselves, as is the HPC system, so Moore's law drives data growth rates in both cases. Fortunately magnetic media growth rates (sometimes cited in "Kryder's Law") are on a similar trajectory so storage systems likewise exhibit the behavior of offering twice the capacity for roughly the same cost year over year. This behavior is expected to continue through 2020<sup>[1,2]</sup>.

This allows us to provide space for exponential data growth as long as a relatively consistent

storage budget is available year-to-year. Successive generations of storage have so far had the sheer capacity to swallow up data from earlier generations of technology, provided there is a bridge between the technologies. Ensuring that there is such a bridge between generations is feasible provided there is sufficient planning and investment both in time and dollars to execute it.

Exponential growth rates *are* sustainable with proper planning and funding, but this only provides for storage *space*. By itself, this does not address the problems of managing, understanding, or using the accumulation of data. To that end, EMSL is investing in creating a new scientific data and metadata system known internally as MyEMSL. MyEMSL is addressed in the PNNL position paper for the Usability of Storage Systems track.

### **SOFTWARE FOR FILE SYSTEMS AND ARCHIVES**

EMSL uses the software technologies that best fit its needs and budget, whether open source or proprietary. As much as possible, we wrap proprietary solutions so that they play well in an open-source environment. We were an early adopter of the Lustre filesystem, having used it since the implementation of our MPP2 system in 2003. We have built low-cost filesystems out of commodity hardware up to 1.2 petabytes (the "NWfs" storage system in 2008), and PNNL is building a similar institutional Lustre storage system that will have a 4-petabyte capacity by the end of fiscal year 2011. In 2008, EMSL identified a need to implement a hierarchical storage system, and in 2009 retired NWfs in favor of a new HPSS system.

HPSS provides the right mix of capacity, expandability, and scalable performance for EMSL's needs. The EMSL HPSS system provides archive storage capacity, and we have implemented open source filesystem-like interfaces to it, in addition to the traditional native HPSS interfaces.

EMSL and the rest of PNNL continue to make use of Lustre and will continue to do so until it is clearly dead or orphaned. At this point, PNNL

has enough experience and expertise to not require Lustre support. Even if advanced and long-promised features (e.g. multi-way clustered metadata) are never delivered, Lustre's cost, performance, scalability, and "good enough for us" reliability meet our needs very well.

The difficult to control costs are in additional work scope, i.e. supporting more systems or more users without attendant increases in budgets. Inflation alone causes increased labor costs over time, creating difficulty in operating with flat or declining budgets. Additional work scope compounds this problem if not very carefully managed.

## **HARDWARE FOR FILE SYSTEMS AND ARCHIVES**

Being at the upper-mid range of HPC in terms of system sizes and performance, EMSL is not using and does not expect to use custom hardware in the foreseeable future. We take the best advantage we can of common off the shelf hardware and the economies of scale that come with it.

EMSL does expect to continue to take advantage of commodity storage technologies for the foreseeable future, mostly in conjunction with the Lustre filesystem. Selected high-value storage systems may be constructed of enterprise-grade storage for serviceability features. While we may apply creative engineering approaches to commodity or enterprise-grade building blocks, we do not foresee significant use of custom storage hardware.

I/O capacity and bandwidth requirements for filesystems on EMSL HPC systems are established as a function of peak performance ratings. We have not carefully specified metadata operation or operations/second requirements on our filesystems, though we have re-engineered metadata servers to improve performance when there is a need to do so. MTTI requirements have not been rigorously specified either, though we do specify that common failures (e.g. single disk failure on a node) must not interrupt computation or I/O. During technical review, we assess whether the I/O system is robust enough to

remain serviceable with good maintenance procedures. It has been said, "we don't need five nines, we just need two or three!"

Most of the barriers we see to adoption of commodity storage have to do either with low performance or lack of Reliability, Availability and Serviceability (RAS) features. In our experience, neither of these has presented insurmountable difficulties. The engineering approaches and software tools we apply allow performance to be scaled linearly (or nearly so) by adding more components. The essential RAS features we need are typically available in mid-grade commodity or enterprise hardware. At the other end of the spectrum, many higher end RAS features such as active-active failover/failback prove to cause as much downtime as they are advertised to prevent!

## **SYSTEM EVOLUTION**

EMSL plans a three to four year lifetime for its HPC systems, and has recently decided to switch from operating one large HPC system to two smaller systems with overlapping lifecycles. With this change, we will pull the persistent "home" filesystem out of the cluster and place it where it can be shared between systems and provide continuity between HPC systems as they age out and are replaced.

EMSL procured a new HPSS storage system for archive purposes in 2009, and plans to operate it through at least 2017, with planned lifecycle replacements and technology refreshes for the storage (disk and tape) components.

## CONCLUSIONS

EMSL employs multiple tiers of data storage systems with different capacity and performance characteristics to satisfy various needs. Storage system capacities are planned based upon projected output from the facility's scientific instruments and from HPC system performance. All storage systems have a planned lifecycle with expansions, technology refreshes, and retirement as appropriate. EMSL generally uses commodity or enterprise-class components as building blocks, in concert with a mixture of open source and proprietary software.

## REFERENCES

1. Kryder, H, Kim, C. *After Hard Drives – What Comes Next?*. IEEE Transactions on Magnetics, Vol. 45, No. 10, October 2009  
[http://www.dssc.ece.cmu.edu/research/pdfs/After\\_Hard\\_Drives.pdf](http://www.dssc.ece.cmu.edu/research/pdfs/After_Hard_Drives.pdf).
2. Zyga, L. *What Comes After hard Drives?*  
<http://www.physorg.com/news175505861.html>  
.

# U.S. Department of Energy Best Practices Workshop on

## File Systems & Archives

San Francisco, CA

September 26-27, 2011

### Position Paper: Reliability and Availability

#### John Gebhardt

Lockheed-Martin/U. S. Air Force Research  
Laboratory  
DoD High performance Computing Modernization  
Program  
[john.gebhardt@wpafb.af.mil](mailto:john.gebhardt@wpafb.af.mil)

#### Thomas Kendall

U. S. Army Research Laboratory  
DoD High performance Computing Modernization  
Program  
[thomas.m.kendall4.civ@mail.mil](mailto:thomas.m.kendall4.civ@mail.mil)

#### Cray J. Henry

DoD High performance Computing Modernization  
Program  
[cray@hpcmo.hpc.mil](mailto:cray@hpcmo.hpc.mil)

#### ABSTRACT / SUMMARY

The DoD High Performance Computing Modernization Program (HPCMP) has implemented a multilayered storage approach to cost effectively meet the storage needs of a diverse customer base. Users' can wait in the batch queue indefinitely (but typically start within seventy-two hours) and then can run for up to fourteen days (or longer with special arrangements). To maximize systems availability, several layers of storage and storage use policy are implemented.

#### INTRODUCTION

The HPCMP has layered several storage and file systems within the environment. Each system has specific reliability and availability characteristics and use policy driven by system availability requirements.

This paper discusses the reliability and availability of the following types of storage constructs within the HPCMP:

- HPC scratch space file system

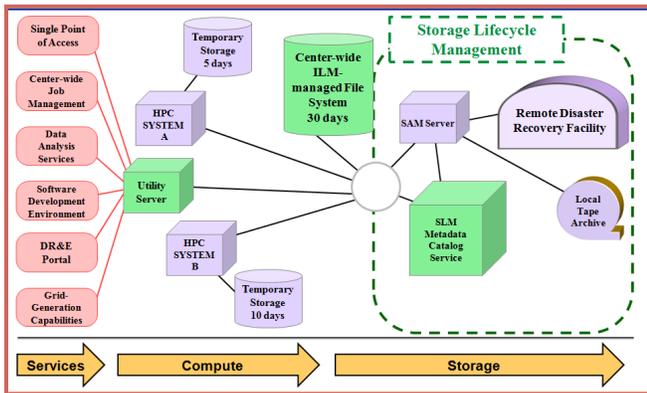
- HPC Home and Applications file system
- Root services file system
- Center Wide File System (CWFS)
- Lifecycle Management System
  - Archive system
  - Tape storage

#### Data Center Facilities

The Department of Defense Supercomputing Resource Centers (DSRC) operates twenty-four hours a day, seven days a week. Each of the DSRCs utilize different combinations of UPS, redundant commercial power feeds and diesel backup power generation capabilities to allow operations to continue through minor power fluctuations and allow for graceful equipment shut for prolonged power outages. In the event of a prolonged power or cooling failure, procedures are activated to shutdown the systems, which in turn quiesces the storage. This approach nearly eliminates unplanned, abrupt outages.

#### . File System Overview

Each DSRC hosts a CWFS which supports the lifecycle management system, the utility server and each HPC system. Each HPC system typically includes a combination of RAID storage devices that are logically decomposed into three file systems -- Scratch space, Home and Applications and Root services. The utility server is similarly configured and the lifecycle management system includes multiple data stores, archival servers and agents described later.



### Center Wide File System

The HPCMP has recently deployed Center Wide File Systems (CWFS) at each DSRC. The purpose of the CWFS is to provide users with a fast central storage capability that can be easily accessed by all major HPC systems and servers within the center. It is intended to serve as the “near-HPC” intermediate storage between scratch file system on each HPC system and the long-term archive. It has been sized to providing a minimum of thirty days of quick intermediate storage. Through CWFS users can move their entire data sets among the scratch file systems and the archive file system. They can perform pre and post processing on their data conveniently avoiding the slower access times associated with archived data. This approach affords users the time necessary to make more thoughtful decisions on what data, for example after a large run, really needs to be archived and what can be deleted.

The CWFS is not backed up. It does contain all the redundancy features of HPC scratch with the

addition of check sums on read from storage to host.

### HPC Scratch Space File Systems

The HPC systems are in high demand. The data sets used to set up runs and the data resulting from runs is very large and very transitory. In order to assure there is sufficient scratch space to stage the next job, HPCMP policy allows for user data to exist on HPC scratch storage for 10 days. Within ten days after data creation, users must move their data to the center wide file system or archive. After ten days the data it is subject to removal to make space available for the future jobs.

Like the CWFS, the HPC scratch space is typically not backed up due primarily to the transient nature of the data and the amount of data.

HPC scratch storage systems are normally procured with the HPC system. The HPC vendors propose the file systems and storage architecture as components within an overall HPC system. The HPCMP request for quotations (RFQ) for HPC systems states that the storage system must be architected to be resilient and robust; highly reliable components are to be utilized.

The HPCMP’s RFQ defines the minimum aggregate data transfer rates between the compute nodes and the disk subsystem are based on specified ratio values for total system memory bandwidth (GB/s) to 1000 times the disk subsystem I/O bandwidth (GB/s). These ratios are 2.07, 1.68, and 1.34 for read, write, and full-duplex respectively. For example, a 200 node system with 50 GB/s of memory bandwidth per node would have total system memory bandwidth of 10000 GB/s. To meet the minimum full-duplex requirement, the system would require an I/O bandwidth of 7.46 GB/s (i.e.  $(200 \times 50 \text{ GB/s}) / (1000 \times 1.34)$ ). The minimum formatted usable disk storage size must be at least 40GB per processor core.

HPC vendors must also commit to monthly interrupt counts and overall systems availability (> 97%). This encourages the vendors to offer

reliable storage systems due to the penalties imposed for not meeting the system availability commitments.

The files systems end up on RAID protected storage that is either RAID 5 or RAID 6. RAID 6 is becoming more common place which is due to the increasing scratch space sizes which are architected with increasingly larger and lower costs SATA disk drives. These large drives take much longer to rebuild leaving the RAID set vulnerable to another drive failure. Vendors architect the storage with redundant paths from the hosts and multiple controllers. Metadata redundancy and availability is expected.

In order to maximize performance HPC scratch storage does not have end-to-end protection mechanisms or check summing.

Downtime for preventative maintenance may be executed at the recommendation of the HPC vendor. Typically, these downtimes are to update the HPC operating environment and not necessarily for the storage systems.

Support for the systems and storage is twenty four hours a day, seven days a week, four hour onsite support for hardware problems.

### **HPC Home and Applications**

Home and application file system on the HPCMP HPC systems are considered more permanent then the HPC scratch file system. These file systems typically are hosted on the same storage as the HPC scratch space and utilize the same file system software.

With only minor exception, home and application file systems protected in the same manner as the HPC file system. Home and application file system are backed up on a daily basis.

### **Root File Systems**

Root drives on major infrastructure servers and key elements in an overall HPC system (login nodes, admin nodes, etc) predominately are architected with multiple disk drives that are protected with RAID 1 or RAID 10. The costs for additional disk drives vs. performance make this a very worthwhile architecture decision.

Compute nodes within an HPC system that are architected with disk drives do not include redundancy. Since the disk drive images on compute nodes are easily reproducible, redundant drives in a RAID configuration are not employed.

### **Archive File Systems**

Arguably, the tape archive systems are one of the HPCMP's most important systems which require the highest level of availability. Prior to integrating the CWFS, and the short time to live for data on HPC scratch the archive had to always be available to the user.

In addition, a user can have their HPC batch jobs in the work load management system queues in some cases up to fourteen days prior to the job running on the HPC system. Jobs that require input data from the archive system would not necessarily want to move their data immediately with the short time to live of the data in HPC scratch.

The DSRCs have architected redundant servers, redundant server component, redundant SAN switches, tape drives, very high speed RAID 5 or RAID 6 disk caches and metadata devices for the archive systems.

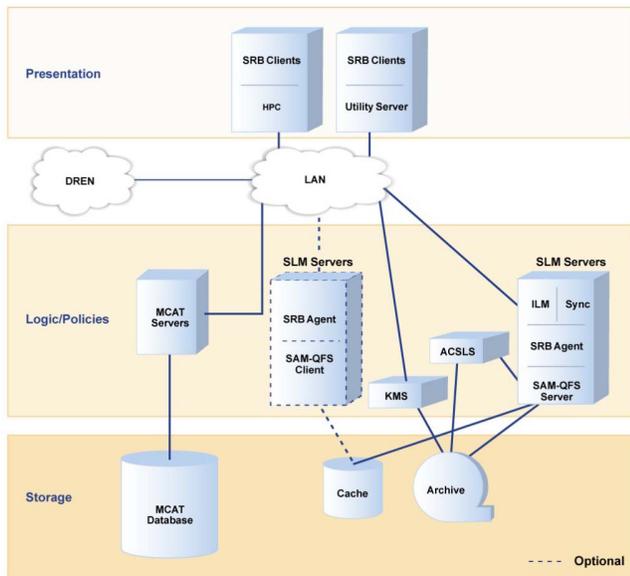
To maintain this high availability, the archive systems have been maintained at twenty four hours a day, seven days a week, with four hour response time.

Implementation of an active-active redundant archive server solution remains a priority.

### **Tape Archives**

The HPCMP currently provides the user by default, two copies of their data on tape. One copy is at the DSRC and the other copy is sent via network to an archive system in another facility and then copied to tape. Archive file system metadata is backed-up daily and stored locally as well as remotely.

### **Storage Lifecycle Management**



The HPCMP is currently implementing storage life cycle management (SLM). SLM tightly couples the current archive systems with an Integrated Lifecycle Management (ILM) management tool. The ILM is an Oracle Real Application Cluster (RAC) environment that will contain the file metadata from the archive as well as user applied metadata such as whether or not to make a disaster recovery copy, when the data can be deleted, what project(s) that data belongs to, etc.

The ILM is architected with multiple Oracle servers via Oracle RAC for redundancy and scalability. Additional reliability features incorporated in the design include redundant server components, a performance disk subsystem for the Oracle databases, utilizing multiple fiber channel paths per server, RAID 5 and RAID 10 volumes, redundant controllers, and redundant network interfaces connected to redundant switches.

## CONCLUSIONS

In order to maximize HPC cycles for researchers, the HPCMP will continue to employ redundancy and other availability measures where practical to maintain availability of the systems, file systems and storage. The HPCMP would like to see the vendor community continue to develop the capabilities for end-to-end data protection (e.g T10DIFF) to ensure bit error rates are extremely low and that bit errors are identified and corrected. As storage space on disk drives continues to grow, new RAID schemes to decrease rebuild times and maintain file system performance are desired and would be implemented.

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper**

**M'hamed Jebbanema**  
Los Alamos National Labs  
mjebb@lanl.gov

### **ABSTRACT / SUMMARY**

This paper will discuss a strategic and automated scripted solution to ensure High Performance Storage System (HPSS) metadata integrity, availability and recoverability in the event of a disaster.

### **INTRODUCTION**

An essential component of High Performance Storage System (HPSS) is the metadata and tools to manage and retrieve the metadata. Metadata can be described as the DNA of the storage system as metadata defines the elements of the transactional data and how they work together. Any loss of metadata proves to be disastrous to data integrity.

In addition to implementation of resilient and fault tolerance hardware and software best practices, we must guarantee the high availability and recoverability of metadata.

Since HPSS uses IBM DB2 as its metadata management system, manual and conventional DB2 standard backups and lack of logs archival monitoring tools dramatically increase administrator workload and probability of data loss.

A Perl language based comprehensive DB2RS (DB2 Recovery Solution) delivers a robust, configurable, and customizable recovery management with minimal efforts.

The Scheduler, Backup, Verifier and Checker(s) services work intrinsically in a holistic approach by using SQLite database as a central repository for all DB2RS activities.

This presentation will cover the design, and integration of DB2RS to ensure metadata integrity and recoverability when disaster occurs.

### **Paper Content**

In order to achieve transaction integrity and zero or little data loss, DB2RS makes automated scheduled local

backups including logs to different media combined with an-offsite hosting similar backups in which to restore from in the event of a disaster.

#### **Goal:**

Develop a metadata backup/recovery solution that is simple, customizable, robust, and easy to maintain.

#### **Objective:**

- Provides recoverable copy of databases.
- Metadata can be recovered to any Point In Time (PIT).
- Guarantees transactional consistency.

#### **Purpose:**

Develop scripts that would leverage DB2 integrated utilities and tools to automate all functions ensuring metadata recoverability.

## **1.Design**

### **1.1 Logging**

Proper DB2 log file configuration and management is an important key to data stewardship and operational availability.

All database changes (inserts, updates, or deletes) are recorded in the DB2 transaction logs. Transaction logs are primarily used for crash recovery and to restore a system after a failure.

DB2 does have the ability to perform dual logging on different volumes as well as different media thereby increasing redundancy for both active logs and archive logs.

We Configure DB2 log sets by implementing MIRRORLOGPATH and dual log archives where each active & archive log set on independent LUN that uses separate physical disks and different type media (TSM) (Figure 1&2).

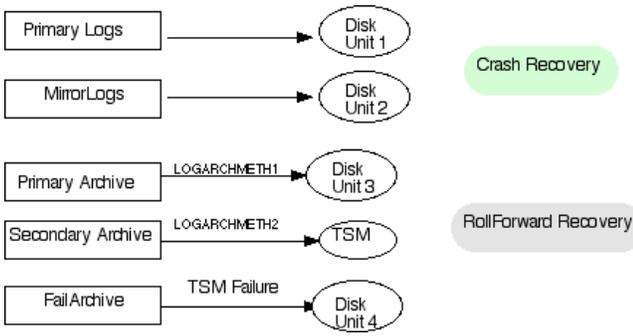


Figure 1. Logging Scheme

Disk Unit 1		Disk Unit 2		Disk Unit 3	
LUN 1	LUN 2	LUN 1	LUN 2	LUN 1	LUN 2
Table Space		Log Files	Log Mirror 1 Archives	Log Files	Log Mirror 2 Archives

- Disaster recovery: Excellent. Short of fire, flood, etc., DB2 can always be recovered.
- Operational availability: Excellent, as good as you can get without going to high availability DB2s. Only interruption would be loss of Disk Unit 1 or simultaneous loss of Disk Unit 2 & 3

Figure 2: Metadata Hardware Fault Tolerance Implementation

### 1.2 Backup service

The scope of this design goes beyond the process of backing up objects to disks or tape but rather encompasses several functions that ensure the validity and integrity of the backups.

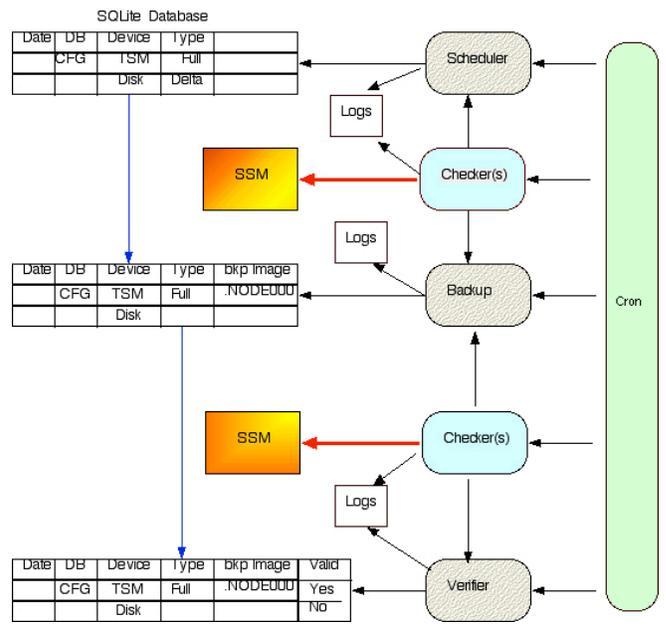


Figure 3: Overview of DB2RS

Perl Language was an evident choice to automate these tasks in order to take advantage of our locally developed Perl modules and for backward compatibility reasons.

Scripts were structured based on type of service or function and SQLite database (Example 1), an open source, self-contained, embeddable, zero-configuration was chosen as a central repository for all services activities.

Four types of services (Figure 3) were automated and can be run either via cron (Example 2) or manually.

- Scheduler : Schedules backup in SQLite database.
- Backup : Checks SQLite for scheduled backups and perform backup service.
- Verifier : Validate the integrity of backups.
- Checker & High Level Checker : perform diagnostics, sanity checks, monitoring and error reporting

Each service has multiple parameters entries in configuration (Example 3) file subject to customization based on disaster recovery requirements.

To prevent invalid data and allow synchronization among all services, applications state and locks were included in the initial design.

All services activities are captured in a centralized log and a man page was integrated into the code for quick reference. (Example 4)

Other useful utilities were coded and added to the mix to ensure completeness and automation of DB2RS.

Example 1: SQLite database service activities

```
20110816000601 CFG TSM FULL 1
20110816041003 Verified
20110817000306 SUBSYS1 TSM INCREMENTAL 1
20110817051002 Verified
20110817000609 CFG TSM FULL 1
20110817045902 Verified
20110818000301 SUBSYS1 DISK FULL 1
20110818045903 Verified
20110818000606 CFG DISK FULL 1
20110818041032 Verified
20110819000301 SUBSYS1 DISK INCREMENTAL 1
20110819045902 Verified
```

Example 2: Services scheduled via cron

```
06 0 * * 1,2,3 schedulme -cron -db cfg -tsm -full -sessions 1
10,59 4,5 * * * bkp -cron > /dev/null 2>&1
02 6,7 * * * bkp -cron > /dev/null 2>&1
0,41 * * * * bkplogtrim -cron > /dev/null 2>&1
09 12 * * 4 checker -cron > /dev/null 2>&1
09 12 * * 1-3,5 checkerhl -cron -disk > /dev/null 2>&1
23 * * * * ensure_archlogs -cron > /dev/null 2>&1
```

Example 3: DB2RS configuration file

[timemachine]

```
# bkp: should we make another bkp if one full bkp already exists
within bkpDiffHrs (integer hrs);
bkpDiffHrs = 20
#bkpv: looks back in SQLite for unverified bkp images earlier
than diffWkBkpv; 24hrs
diffWkBkpv = 86400
#checkerhl & checker: each service must have run successfully in
the last (n) days; 3 day
ChkDysServSuc = 259200
#chekerhl: calculate timestamp before which combo bkps and logs
should exists; 1wk
ChkhlWks = 604800
# checker: all services Service must have run within; 8 hrs
ChkHrsSerbkp = 64800
ChkHrsSerbkpv = 28800
ChkHrsSersched = 28800
#checker: check bkp service progress run..(avoid runaway and
hangs)..service must not be running
# for more than ChkHrsRunbkp; 2 hrs
ChkHrsRunbkp = 7200
# checker: if no bkp within ChkHrsBkpSched after being
scheduled; 20 hrs
```

```
ChkHrsBkpSched = 64800
# checker: stats Failover paths for existence of logs (any files)
within the last ChkHrsFailover; 1 hr
ChkHrsFailover = 3600
# ChkHrsFailover = 0 # test only; make sure checks paths
immediately..no delays
# checker: row created in SQLite for each db combo int he last
ChkWksSqlite; 1wk
#ChkWksSqlite = 604800
ChkWksSqlite = 3600
[constantvars]
DB2DIR = /opt/ibm/db2/path
[db2]
databases = cfg, etc
[db2:cfg]
images = /usr/db2/image/path
archlogs = //path/path/etc..
failarchlogs = /usr/db2/path/path/etc....
imagespct = 60
archlogspct = 60
tsmstartdate = 20110202
```

Example 4: Man page

```
ENSURE_ARCHLOGS(1) User Contributed Perl
Documentation ENSURE_ARCHLOGS(1)
NAME
```

```
ensure_archlogs - Ensure that database has truncated
logs recently
```

SYNOPSIS

```
ensure_archlogs <options>
Within root's or instance owner crontab...
01 * * * * /lan/hpss/path/db2rs/ensure_archlogs -cron
On the command line as root or instance owner...
# ensure_archlogs -force Force a log archive right
now
# ensure_archlogs -force=2h Archive if not performed
in last 2 hours.
# ensure_archlogs Same, but use default intervals from
the dbconfig
# ensure_archlogs -disable=1h Disable scripts in cron
for 1 hour
# ensure_archlogs -enable Re-enable scripts in cron
# ensure_archlogs -help Show the synopsis for this
script
# ensure_archlogs -man Show the man page for this
script
```

DESCRIPTION

This should generally be run via cron. It makes sure that DB2 has archived a log within a certain amount of time for each HPSS database. If it hasn't it tells it to do that with the "db2 archive log" command. This limits the exposure of

Un-archived logs to a certain period of time while also allowing minimum impact on DB2. This should probably not run more frequently than one hour.

### 1.3 Checker(s) and error reporting service

Applies to Logs and backups.

#### 1.3.1 Checker (Example 5)

##### 1.3.1.1 Logs Checks:

Performs the following tasks:

Exclusive analysis utilizing log data mining list of events.

Disk & TSM logging failures detection.

FailArchive path check.

Immediate reporting and notifications to SSM.

##### 1.3.1.2 Backup Checks:

Performs the following tasks:

Sanity Checks

Diagnostic Checks

Immediate reporting and notifications to SSM.

#### Example 5 : Check output

*info: No scheduled backup found at this time!*

*info: checker: No DB2 Logs in Failover paths...good thing*

*info: checker: Found all 4 scheduled combination backups in Sqlite database..schedulme is working fine*

*info: checker: Last successfull bkp run at 20110817055901*

*info: checker: Last successfull bkp Run at*

*20110817071029*

*info: checker: Last successfull schedulme run at*

*20110817000609*

*info: checker: bkp service has run within defined time (Hrs).*

*info: checker: bkp service have run within defined time (Hrs).*

*info: checker: schedulme service have run within defined time(Hrs).*

#### 1.3.2 High-level Checker (Example 6)

Performs the following tasks:

Recent backup for each (database, device) pair completed successfully.

Recent TSM backup for each database must exist and verified

Ensure that TSM copies of all logs since the last verified TSM backup exists.

Immediate reporting and notifications to SSM.

#### Example 6: High Level Checker (Checkerhl) output

*info: Found 56 logs for CFG DISK backup taken at 20110813041003 (2314 - 2369)*

*notice: Everything looks good for CFG DISK backup taken at 20110813041003*

*info: Found 99 logs for SUBSYS1 DISK backup taken at 20110811045903 (3116 - 3214)*

*notice: Everything looks good for SUBSYS1 DISK backup taken at 20110811045903*

*info: Found 96 logs for CFG TSM backup taken at 20110810041003 (2274 - 2369)*

*notice: Everything looks good for CFG TSM backup taken at 20110810041003*

*info: Found 146 logs for SUBSYS1 TSM backup taken at 20110808045903 (3069 - 3214)*

*notice: Everything looks good for SUBSYS1 TSM backup taken at 20110808045903*

*info: Found all 4 backup combinations*

### 1.4 Error Reporting to SSM (HPSS interface):

Exit codes are called to generate sub-class of errors based on custom binary scheme for multiple error reporting.

To simplify error reporting and monitoring, three (3) categories were considered to match HPSS error reporting style.

#### Minor

Backup, verifier, or scheduler service did not run within defined time (Hrs).

Backup have exceeded estimated allowable time to successfully complete backup.

#### Check ASAP (considered critical)

Failure to schedule expected pair (db,device) within the last (days).

TSM backups and associated logs are not found.  
One or more services failed in the last (n) days.

#### MAJOR

Scheduled Backups are behind schedule.

Backup service did not run.

Backup failed (n) times as specified in cron.

Backup could have completed successfully but took longer than expected/estimated.

Failover DB2 log paths contain logs.

Filesystem(s) error.

Log(s) script failure –internal error-.

Log archiving failure: Disk or/and TSM.

## **2. CONCLUSIONS**

DB2RS not only adds value to HPSS native metadata integrity monitoring and reporting tools but also ensures

that our operational staff are monitoring the health and status of the metadata and thus reducing dramatically the risk of loss of data and respectively the time of recoverability.

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper**

**Mark Gary**  
Lawrence Livermore National Laboratory  
mgary@llnl.gov

## **ABSTRACT / SUMMARY**

**This position paper addresses the reliability and availability of storage systems. Specifically it introduces LLNL storage best practices in the areas of resilient architectures and daily operations which contribute to enhanced availability, reliability and computational integrity in LLNL's 24x7 HPC centers.**

## **INTRODUCTION**

Livermore Computing operates multiple 24x7 "lights on" HPC environments and has done so for over 40 years. Availability, reliability and computational integrity are of paramount concern in the Center due to the tremendous investment in, and the importance of, our HPC machines and the data they generate. This position paper outlines, at a very high level, some of the storage system best practices followed by LLNL. The two focus areas covered in this paper are:

- **Resilient Storage Architectures:** Best practices surrounding the storage hardware architectures employed in the LC and their impact on availability and data integrity.
- **Daily Operations:** Storage system best practices surrounding daily operations (from outages and maintenance to training and communications).

Within these areas I very briefly identify best practices as fodder for Workshop discussion.

## **Resilient Storage Architectures**

Allowing storage operations to continue in the face of failure or outage is critical. Among the hardware architecture best practices followed in the LC are:

- ***Scalable Unit Architecture***

Following the lead of our computing platforms we deploy storage hardware using the concept of a Scalable Unit (SU). An SU is the smallest unit of hardware (storage and associated servers) by which you can grow a storage subsystem. Well identified *identical* SUs allow for ease of repair, maintenance, administration, expansion, and sparing. The purchasing power of buying identical hardware in volume is an added benefit.

- ***Leveraging Compute Platform Hardware***

In our file system and archive environments we, whenever possible, leverage and duplicate the server hardware technologies used on our compute platforms. As in the SU area, this helps ease repair, maintenance, administration, expansion and sparing and takes full advantage of the purchasing power and technology investigation efforts made during platform procurements.

- ***Failover Partners***

The LC has nine very large Lustre file systems. The Object Storage Server (OSS) nodes controlling subsets of disk are architected into failover pairs allowing a failed OSS to have its disk taken over by its healthy partner. This

architecture is leveraged constantly and aids not only the case of node failure, but also increases availability during software and firmware deployments and electrical work.

- ***Targeted Use of Uninterruptible Power Supplies (UPS)***

The amount of electrical power required by LC compute and infrastructure hardware makes full coverage with backup power/UPS power economically infeasible. Instead the LC uses its UPS budget in a targeted manner with a concentration on support of metadata services, network infrastructure, and home directory file systems. This strategy focuses on the protection of the most critical data and aids in a rapid return to service of Center operations following a power outage.

- ***Dual Power Sources***

The majority of LC storage infrastructure racks are wired in a redundant manner to be able to survive the planned or unplanned loss of any one electrical subpanel. This is particularly critical in our very dynamic machine room environment during this era of enhanced electrical safety rules.

- ***Archive Dual Copy***

While budgets do not allow us to keep multiple copies of the PetaBytes of simulation data stored in the LC, our HPSS systems do allow a user to direct that dual copies be made of targeted files. The two copies are stored in geographically separated locations cross-Laboratory.

- ***Degraded Mode Archive Operation***

Because of the distributed, multi-level hierarchy architecture of our HPSS archive implementation we are commonly able to provide users with various levels of degraded mode service during outages. In degraded mode not every file is accessible, but typically the most recently written files and those with dual copies can be accessed.

## **Daily Operations**

On a daily basis the LC implements a number of operational best practices surrounding our storage systems including:

- ***“Lights On” Operation***

It is our philosophy, in part driven by the tremendous dollar investment in our computing environment, that the LC provide 24x7x365 customer service and compute availability. Our cross-trained operations staff is always on site monitoring our systems and answering off-hours user questions - around the clock. In the event of an environmental emergency (e.g., loss of cooling) they can immediately react following prescribed/ordered power down procedures. Operations staff are trained in basic file system and archive administration and do hardware repair as well. They have full access to on-call storage system and archive administrators at any hour.

- ***Self-maintenance***

Much of our hardware maintenance is performed by LC employees. This allows us to perform maintenance immediately when a problem occurs, eliminates security escort requirements, and allows us to closely track and learn from system failures. The fact that storage hardware leverages platform hardware procurements means that our personnel need be trained on only a limited number of equipment types.

- ***Hardware/Spare Burn-in***

We have a full hardware spare/RMA center supporting our local maintenance operations. Rather than pulling spare parts off of the shelf, we maintain a burn in environment where we have spare storage hardware under continuous test, exercise, and burn in. When equipment fails, hardware is pulled directly from the burn in environment. Before tape drives are allowed to be placed into service they first undergo a suite of performance tests and integrity tests.

- ***Testbeds***

Livermore Computing has a variety of testbeds in which we test pre-production hardware and

software. These testbeds range from single racks, to the well-known multi-vendor Hyperion test environment where software can be tested at scale with thousands of clients.

- ***Data Integrity Checking***

Continuously in the background, the LC runs a tool called DIVT - the Data Integrity Verification Tool. DIVT checks the data integrity of our archive and our parallel file systems by writing known data patterns to files from different platforms, forcing the data to flow through file systems and down to archival tape, and then checking data integrity upon fetch back to the platform. Over the years DIVT has caught data corruption ranging from on-platform component problems to corrupting drive firmware.

- ***Planned Downtimes***

The LC has a philosophy that planned downtimes happen during the work week from Tuesday through Thursday unless particular circumstances dictate otherwise. While this has an impact on interactive users, Center resources are fully subscribed 24x7 including weekends. Our philosophy allows us to have experts from all disciplines on hand in case of problem in order to improve availability. Fridays are avoided to limit the introduction of problems impacting the weekend. Mondays are avoided to allow users to process the results of their weekend runs.

Software rollouts are planned in such a manner as to minimize impact on the programs supported by the Center. We rollout software to our unclassified systems first which allows us to bring outside experts to bear on problems encountered. Recently we leveraged a Six Sigma

quality project to improve our software rollout process and reduce the length of planned downtimes.

- ***Impacts and File System Meetings***

Every Monday representatives from every facet of the LC (including Facilities) have a formal meeting to manage any outage or operation that has impact on Center customers or has cross-cutting impact among Center discipline areas. This meeting has tremendous value and allows us to combine outages and plan forward in order to maximize the availability of all center resources including storage. A separate meeting which pulls together Operations staff, storage system administrators, and hardware repair personnel occurs weekly. This meeting improves file system specific communication across all involved disciplines and shifts.

## **CONCLUSIONS**

Large HPC environments are extremely complex. They require that particular attention be paid to operational and architectural storage system best practices in order to ensure availability, reliability and computational data integrity.

\* This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-CONF-497278

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper**

**Evan J. Felix**

Pacific Northwest National Laboratory  
evan.felix@pnnl.gov

**ABSTRACT / SUMMARY**

**The EMSL Molecular Science Computing team manages 3 main storage systems to provide scientific data storage to the users of EMSL scientific instruments and computing resources. These storage systems are managed in different ways to protect the availability of data and protect from data loss. This position paper will address the reliability and availability of storage systems track of the Best Practices Workshop**

**The management team has designed and monitors the systems to protect the scientific investment that is stored within the systems. Widely available open-source tools, and home-grown tools are used to monitor and track usage. Communication with users has also been a key element in keeping the systems stable.**

**INTRODUCTION**

The Environmental Molecular Sciences Laboratory(EMSL) at PNNL is a center for scientific research. The building houses many scientific instruments and tools, along with a large computational center. This arrangement of science and computing creates a large amount of scientific data, that we must preserve and store for many years to come. A mix of raw experimental data, processed data, and simulation

data is processed and stored with EMSL's file systems and data archive.

EMSL has three main data storage systems, with specific purposes. A home file system for active user codes and data on the cluster, a high speed global file system attached to the large 167 TF Chinook[3] cluster, and a 6+ Pebibyte archive system for long-term storage. These data storage areas are summarized in Table 1.

Table 1

Type	Size	Type	Speed
Home	20 TiB	Lustre	1GB/s
Global temp	270 TiB	Lustre	30GB/s
Long-term archive	4+ PiB	HPSS	200MB/s single stream

All of these file systems are accessible to users directly on cluster login nodes. Home and temp space is available to all cluster nodes.

**1. CLUSTER FILE SYSTEMS**

The home space and temporary space used for the Chinook cluster are connected directly to the QDR infiniband interconnect. Each system utilizes an active-passive fail-over pair of meta-data servers to manage the file system. The block device for the file system is a RAID1 mirror of two fibre-channel based Virtual RAID5 LUNs. Each system has multiple paths through a switch

to each LUN. These storage systems are HP EVA6000 based arrays.

Lustre OST's are also built using a failover pair, but utilize an active-active strategy to balance the load of 8 large LUNs served by the HP EVA technology. Each of these LUNs use a Virtual RAID5 protection scheme. We have been successful in using active-active, as we put all the heartbeat traffic on a very quiet network, which seems to alleviate the dual node power off issue we have seen in many failover solutions. There are 4 servers for the home file system and 38 servers for the temporary file system.

The infiniband connections for the storage servers are balanced across lower-ranked switches of the federated infiniband network. We have also enabled a QOS strategy with OpenSM, using an EMSL created routing algorithm (Down-Up) that has reduced congestion on the network.

Configuration data, including failure states is gathered from all HP EVA systems and stored in the EMSL MASTER database<sup>[2]</sup>. This database keeps a historical record of all hardware assets, including serial numbers, firmware versions, and status information. A nagios monitoring script can query this database to alert administrators when components fail. Most replacements can be done online, and do not require a file system shutdown. Documentation for all replacement procedures is kept in the system wiki. This database also allows us to look at failure history over the life of the system, and track when components are changed.

The home portion of the file system is backed up on one dedicated node on a daily basis. IBM's Tivoli Storage Manager is used for this purpose. The backup tape system is housed in another building. To perform daily backups a multi-process script was created to keep many streams moving. The temporary space is not backed up.

## **ARCHIVE SYSTEM**

EMSL procured a new HPSS storage system for archive purposes in 2009, and plans to operate it through at least 2017, with planned lifecycle

replacements and technology refreshes for the storage components.

The archive system is an HSM based system using the IBM HPSS software stack. We currently have .5 PiB of disk as the first layer of the stack. It consists of one DDN 9900 couplet, serving data to the HPSS mover nodes, data is stored on DirectRAID™ 6 protected LUNS. As data initially moves into the archive it is stored on this disk cache.

The HPSS system has access to an IBM 3584 tape library, which has a mix of LTO4 and LTO5 tapes. Each data block written to tape is stored twice, to protect against tape loss. This duplication policy was implemented as external backups were no longer feasible. The tape library is in another datacenter on the PNNL campus, which is connected with a 2 10Gbits/s network links for a redundant network system.

One key design point we have used for our archive over the last two iterations has been to never lose data. We allow for more downtime and administrator control to accomplish this, and do not require a specific uptime requirement, but do treat it as a production system which should be online as much as possible.

The EMSL MSC team wrote a FUSE<sup>[1]</sup> based file system for access to the archive, that presents to the users a POSIX-like interface. This system allows us to control more aspects of what a user can do on the system, and increase transfer rates than other tools. It also allows us to 'catch' any file removal actions and move them to a special 'trash' location so that accidental removals are not catastrophic. Since the HPSS unlink commands remove all references to a file in the meta-data store, recovery is very difficult or impossible.

Another management tool we use periodically scans the file system collecting information on users use, and provides to users and administrators information on file counts, and size used by each user. This database also contains historical size of the file system. You can see the size of the file system change in

Figure 1, when we moved from the old archive to the new one as multiple copies were made. The sharp increase in data in the chart is caused by HPSS having more than one copy of each file. By policy it should have two copies on tape, and may have one in the disk cache as well.

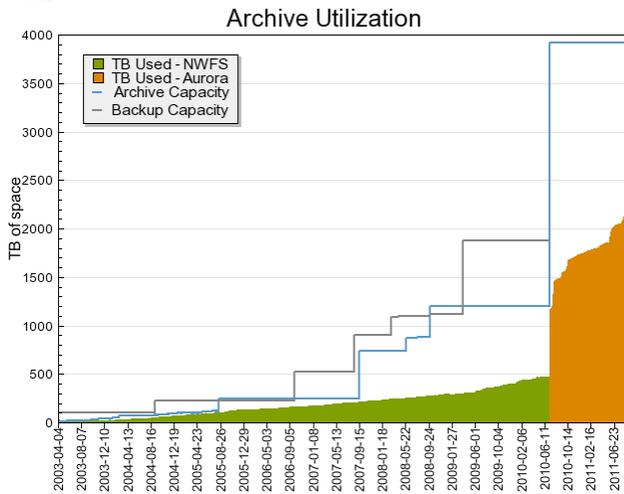


Figure 1

Communication with specific users has been a key aspect of our ability to upgrade and change parts of the archive. Our biggest user by space has assisted in helping us test new changes and been helping in working out any problems before the archive is released back to other users. We also maintain a test system that is built using similar hardware to the production system. This test system allows us to develop and test upgrades before users are on the system.

## MANAGEMENT

Much of our success in managing and protecting users data comes from having experienced administrators and programmers directly responsible for the management of storage resources.

When the archive tools provided by the vendor proved in-adequate for our needs, we wrote an appropriate interface to the archive in a few weeks, and added on features, as we needed them. This change to our strategy did not delay our schedule in deploying the archive. And we found when we attended a Users Group meeting for HPSS that others also had similar issues and were

very interested in know what we had done, and we were able to share our code with them

When Lustre problems have come up, our extensive knowledge of the internal workings of the Lustre source code has been invaluable in saving, and in one catastrophic case saved us from weeks of backup restores. In our new archive backups are no longer used, and so restoring over a pebibyte of data is no longer an issue.

When a vendor specific monitoring system does not integrate with our monitoring infrastructure, we have been able to write wrappers, or use their low-level API to collect data, and detect failures.

We also maintain a MSC wiki that contains the administrative procedures for handling issues, and routine maintenance tasks. We keep an offline copy of this wiki on a USB drive for reference during complete power or network outages.

One member of our team is also on-call at any time. We have found that having every member of the team being a least front-line response, each member learns basic administration of our critical systems, even when they must bring in other experts to solve unexpected issues.

## CONCLUSIONS

The MSC team has found that a deep knowledge of the storage systems that will important scientific data we can design and protect against various failure scenarios and keep the data online and available for our users. Being able to write customized portions of our system allows better integration and management of our resources. This knowledge allows us to be pro-active in finding problems in our system before any data-loss is seen or users experience problems.

## REFERENCES

1. FUSE: Filesystem in user space  
<http://fuse.sourceforge.net/>
2. Simmons, Chris, Evan Felix, and David Brown. *Understanding a Supercomputer: Utilizing Data Visualization*. Tech.

3. Turner AS, KM Regimbal, MA Showalter, WA De Jong, CS Oehmen, ER Vorpapel, EJ Felix, RJ Rousseau, and TP Straatsma. 2009. *"Chinook: EMSL's Powerful New*

*Supercluster."* *SciDAC Review* 13(Summer 2009):60-69.

4. "DirectRAID - DataDirect Networks." *DDN | DataDirect Networks* |. Web. 31 Aug. 2011. <http://www.ddn.com/products/directraid>

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper**

**Venkatram Vishwanath, Mark Hereld and Michael E. Papka  
Argonne National Laboratory  
<venkatv, hereld, papka>@mcs.anl.gov**

## **ABSTRACT**

The performance mismatch between the computing and I/O components of current-generation HPC systems has made I/O a critical bottleneck for scientific applications. It is therefore crucial that software take every advantage available in moving data between compute, analysis, and storage resources as efficiently as networks will allow. Currently available I/O system software mechanisms often fail to perform as well as the hardware infrastructure would allow, suggesting that improved optimization and perhaps adaptive mechanisms deserve increased study.

We describe our experiences with GLEAN – a simulation-time data analysis and I/O acceleration infrastructure for leadership class systems. GLEAN improves the I/O performance, including checkpointing data, by exploiting network topology for data movement, leveraging data semantics of applications, exploiting fine-grained parallelism, incorporating asynchronous data staging, and reducing the synchronization requirements for collective I/O.

## **INTRODUCTION**

While the computational power of supercomputers keeps increasing with every

generation, the I/O systems have not kept pace, resulting in a significant performance bottleneck. The *ExaScale Software Study: Software Challenges in Extreme Scale Systems* explains it this way: "Not all existing applications will scale to terascale, petascale, or on to exascale given current application/architecture characteristics" citing "I/O bandwidth" as one of the issues. On top of this, one often finds that existing I/O system software solutions only achieve a fraction of quoted capabilities.

We have developed an infrastructure called GLEAN [1,2] to accelerate the I/O of applications on leadership systems. We are motivated to help increase the scientific output of leadership facilities. GLEAN provides a mechanism for improved data movement and staging for accelerating I/O, interfacing to running simulations for co-analysis, and/or an interface for in situ analysis via a zero to minimal modification to the existing application code base. GLEAN has scaled to the entire infrastructure of the Argonne Leadership Class Facility (ALCF) comprising of 160K Intrepid IBM Blue Gene/P (BG/P) cores and demonstrated multi-fold improved with DOE INCITE and ESP applications. We discuss some of the lessons learned which could be considered for best practices on file systems and archives.

## OUR POSITION

Based on our experiences with GLEAN, we believe the useful components to improve the I/O performance on leadership class systems include topology-aware data movement, leveraging data semantics, incorporating asynchronous data staging, leveraging fine-grained parallelism, and non-intrusive integration with applications. We briefly elucidate these.

***Topology-aware Data Movement:*** As we move towards systems with heterogeneous and complex network topologies, effective ways to fully exploit their heterogeneity is critical. The IBM BG/P has five different networks with varying throughputs and topologies. The 3D torus interconnects a compute node with its six neighbors at 425 MB/s over each link. In contrast, the tree network is a shared network with a maximum throughput of 850 MB/s to the I/O nodes. The tree network is the only way to get to the I/O nodes in order to perform I/O. BG/Q is expected to have a more complex network topology. Similarly, several other Top-500 supercomputers have complex topologies. As seen in Figure 1, by leveraging the various network topologies, in GLEAN, we achieve up to 300-fold improvement in moving data out from the BG/P system. Another critical aspect is that our data movement mechanism uses reduced synchronization mechanisms wherein only neighboring processes need to co-ordinate their I/O. This is critical as we move towards future systems with millions of cores.

***Fine-grained Parallelism:*** GLEAN's design employs a thread-pool wherein each thread handles multiple connections via a poll-based event multiplexing mechanism. This is critical in future many-core systems with low clock-frequency per core, where multiple threads are needed to drive the 40 Gbps and higher network throughputs per node to saturation.

***Asynchronous data staging*** refers to moving the application's I/O data to dedicated nodes and next writing this out to the filesystem asynchronously

while the application proceeds ahead with its computation. Asynchronous data staging helps satisfy the bursty nature of application I/O common in computational science and blocks the simulation's computation only for the duration of copying data from the compute nodes to the staging nodes. Data staging also significantly reduces the number of clients seen by the parallel filesystem, and thus mitigates the contention including locking overheads for the filesystem. Staging mitigates the variability in I/O performance seen in shared filesystems on leadership systems when accessed concurrently by multiple applications.

***Leveraging Application Data Models:*** I/O system software typically use stream of bytes and files to deal with an application's data. A key design goal in GLEAN is to make application data models a first-class citizen. This enables us to apply various analytics to the simulation data at runtime to reduce the data volume written to storage, transform data on-the-fly to meet the needs of analysis, and enable various I/O optimizations leveraging the application's data models. Toward this effort, we have worked closely with FLASH, an astrophysics application, to capture its adaptive mesh refinement (AMR) data model. We have interfaced with PHASTA, which uses an adaptive unstructured mesh, to make unstructured grids supported in GLEAN, and with S3D, a turbulence simulation, to capture its structured grid model. We have worked with many of the most common HPC simulation data models ranging from AMR grids to unstructured adaptive meshes.

***Non-Intrusive Integration with Applications:*** Application scientists are very interested in I/O solutions wherein they can get the added performance improvements without having to change their simulation code (or with minimal changes). To achieve this, we have mapped Parallel-netCDF and hdf5 APIs, commonly used high-level I/O libraries in simulations, to relevant GLEAN APIs, thus enabling us to non-intrusively interface with simulations using pnetcdf and hdf5.

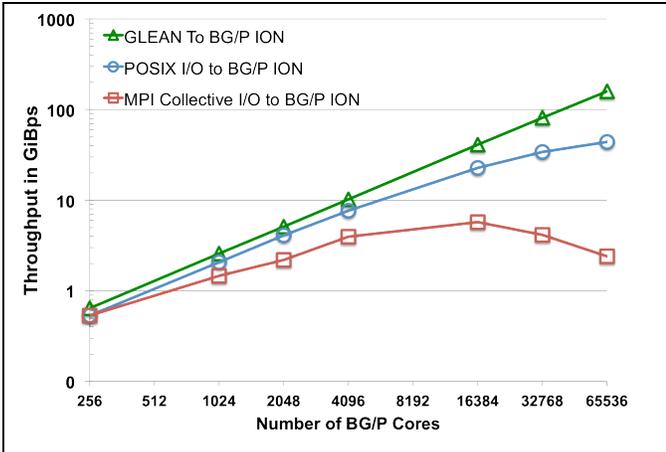


Figure 1: Strong scaling performance of the I/O mechanisms to write 1 GiB data to the BG/P IONs (log-log scale) on ALCF infrastructure

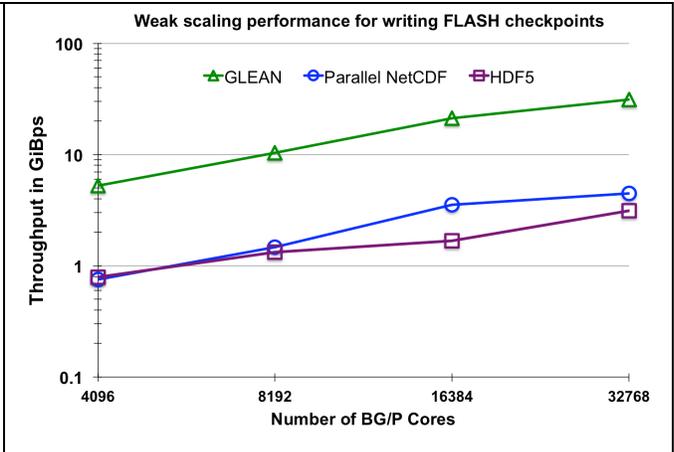


Figure 2: Weak scaling results for writing FLASH checkpoint data

Networking, Storage and Analysis (SC 2011), Seattle, USA, November 2011.

## REFERENCES

- 1.V. Vishwanath, M. Hereld, V. Morozov, and M. E. Papka, "Topology-aware data movement and staging for I/O acceleration on Blue Gene/P supercomputing systems", To appear in IEEE/ACM International Conference for High Performance Computing,
- 2.V. Vishwanath, M. Hereld, and M. E. Papka, "Simulation-time data analysis and I/O acceleration on leadership-class systems using GLEAN", To appear in IEEE Symposium on Large Data Analysis and Visualization (LDAV), Providence, RI, USA, October 2011.

**U.S. Department of Energy Best Practices Workshop on**

**File Systems & Archives**

**San Francisco, CA**

**September 26-27, 2011**

**Position Paper**

**HPC Data Storage Strategy at the UK Atomic Weapons Establishment (AWE)**

**Mark Roberts**

AWE High Performance Computing  
mark.roberts@awe.co.uk

**Paul Tomlinson**

AWE High Performance Computing  
paul.tomlinson@awe.co.uk

**ABSTRACT / SUMMARY**

**AWE has adopted a resilient and globally accessible storage model to support concurrent desktop and compute cluster access. We now provide a global persistent multi-tiered storage which has enhanced usability and reliability. Increasing pressure on budgets has recently focused efforts to reduce and consolidate the directly-attached parallel cluster file system into several global scratch file systems cross mounted on all compute clusters.**

**INTRODUCTION**

Historically, AWE has focused on one or more compute clusters, each with its own local parallel scratch file system and global persistent storage provided by commodity Network Attached Storage (NAS) filers or in-house Network File System (NFS) servers.

Following a major facility issue a few years earlier, which resulted in the sole persistent data store, based on IBMs General Parallel File System (GPFS), being corrupted and had to be restored (very slowly) from backup, it was decided to upgrade to a resilient GPFS cluster. This was designed to provide multi-site resilience, protecting from loss of either site.

This has allowed us to maintain native multicluster GPFS access to the numerous cluster

log-in nodes, visualisation clusters, and secure NFSv4 access direct to the desktop.

**COMMON ENVIRONMENT**

A common name space, providing a consistent view of file systems on desktops, compute and visualization clusters, helps users locate their data and has encouraged well structured work flow with respect to makefiles, shared libraries and code areas.

This, combined with other common environment features, has aided the trend towards more local prototyping, with easier scaling up onto larger platforms.

**DESKTOP ENHANCEMENT**

The Hierarchical Storage Managed (HSM) file systems are now exposed to desktops, running file managers and search utilities, that scan, index and try to determine type by content, all potentially triggering unwanted retrievals, which can block interactive access until the recall from tape completes.

To mitigate this, we modified the KDE3 file manager and GNU utilities (*ls*, *find*, *stat*, *file*) to be aware that a file is migrated based on the naive concept that a migrated file has a non-zero file size with zero data blocks. Users are given visual cues with syntax highlighting or icon overlays, along with extra options for handling such files. Preview operations that would generate multiple recalls are skipped. Obtaining the migration state

via an API or extended attributes that could propagate through software and network protocols would be preferable. However the simple approach has worked well in our environment.

## **FILESYSTEM USAGE TRENDS**

For many years, we recorded per-user GPFS file system usage with a simple POSIX *find* command in conjunction with the HSM *dsmls* command. This was highly inefficient, taking over 24 hours to complete a scan, as well as adding unnecessary CPU and I/O load on the Tivoli Storage Manager (TSM) server.

We now utilize the GPFS Information Lifecycle Management (ILM) interface to scan the file systems and output records containing the extended information such as name, size, access, modify and create time. The resulting data file is processed and imported into a MySQL database. This allows for fine granular analysis at the user and file system level day by day or over a time period.

Currently, with 35 million objects, the new method takes under 20 minutes. We believe that providing such granular information influences users' data storage behavior for the better, and lets us quickly identify unreasonable or unintended usage when problems occur. It also provides long-term trend information to aid future file system capacity planning and procurement.

We have also engaged with Lustre developers to see if future Lustre releases can incorporate a similar capability.

## **DECOUPLED PARALLEL FILESYSTEMS**

Traditionally, as part of a compute cluster procurement, we purchased storage for a dedicated parallel file system to provide the localised fast bandwidth required by codes.

With the potential of cluster lifetimes becoming shorter due to the the ever increasing pace of technology releases, the cost of the disk infrastructure is now becoming a factor. Once our clusters were decommissioned the disk was also

removed and disposed of. With a recent capability cluster procurement the local parallel file system hardware accounted for approximately 15% of the total expenditure. The demand for storage is increasing, so as the total of memory and cores on clusters the associated storage costs may have a larger impact.

We have decided to “decouple” the local parallel file system which allows the storage to be used by multiple clusters. This gives the freedom to either use the cost reduction by increasing the compute size or directing the saving elsewhere. One immediate advantage of no directly attached storage is more rapid initial cluster deployment and, possibly, regular scheduled maintenance without affecting access to the data via other clusters.

The community gain improved useability by not having to transfer the data between the local parallel file systems on the clusters and then to persistent storage.

By extending the concept of resilience to global scratch with a global scratch file system cluster in each facility (three planned) we can now factor in scheduled downtime and upgrades to a chosen global scratch file system cluster more effectively.

## **DATA AND FS AWARE SCHEDULING**

It is our intention that each compute cluster should use by default the global scratch file system in its local facility.

Users may, however, wish to use another global scratch file system in a different building or use the “local” global scratch in conjunction with it. In order to prevent jobs failing due to an unavailable global scratch, we are investigating the concept of storage as a consumable resource in the same manner as a node or cores is used today.

By integrating awareness of global storage into the scheduler/resource manager, jobs may be prevented from failing prematurely. Also when global scratch or persistent storage clusters are

scheduled for maintenance then scheduler system reservations can be placed on the file system preventing jobs from being dispatched, if the job requested that file system as a resource.

approach for efficient global scratch storage.

## **DATA EXPLOSION**

With an increasing amount of scratch, persistent storage and upgraded network links, it becomes relatively easy for the user to copy everything everywhere. This leads to wasted bandwidth, disk storage and tape backups due to duplicated data. File system de-duplication on persistent storage may be possible but ultimately undesirable.

With early MPP clusters it was often quicker to move data from disk or recall from off-line storage than regenerate the data. With the large fast clusters available today it may, in some cases, be quicker and cheaper to save network, disk, and tape resources by regenerating data. This is ultimately a decision that only the user is best placed to make but having to weight up the impact on QA reproducibility and provenance.

Existing compute cluster parallel file systems were designed with the ability to hold four times the amount of system memory from a OS initiated checkpoint. The majority of the mainstream codes are now using restart dumps generated via an in-house I/O library. Code users can then choose whether to perform a full state restart or focus on only saving selected data within the run for analysis. Intelligent software-based restarts greatly reduce the amount and bandwidth required and could allow considerable cost savings, but non-restartable third party codes remain a barrier.

## **CONCLUSION**

Implementation of a fast, secure, exported and resilient global parallel file system for persistent and archive storage has proved invaluable for unifying compute resources at all scales. However the ease of accessibility has created some additional problems and raised user expectations, requiring adaptation of the user interface. We are now exploring a similar

# U.S. Department of Energy Best Practices Workshop on File Systems & Archives:<sup>\*</sup> Usability at Los Alamos National Lab<sup>†</sup>

John Bent  
Los Alamos National Lab  
johnbent@lanl.gov

Gary Grider  
Los Alamos National Lab  
ggrider@lanl.gov

## Abstract

*There yet exist no truly parallel file systems. Those that make the claim fall short when it comes to providing adequate concurrent write performance at large scale. This limitation causes large usability headaches in HPC computing.*

*Users need two major capabilities missing from current parallel file systems. One, they need low latency interactivity. Two, they need high bandwidth for large parallel IO; this capability must be resistant to IO patterns and should not require tuning. There are no existing parallel file systems which provide these features. Frighteningly, exascale renders these features even less attainable from currently available parallel file systems. Fortunately, there is a path forward.*

## 1 Introduction

High-performance computing (HPC) requires a tremendous amount of storage bandwidth. As computational scientists push for ever more computational capability, system designers accommodate them with increasingly powerful supercomputers. The challenge of the last few decades has been that the performance of individual components such as processors and hard drives as remained relatively flat. Thus, building more power supercomputers requires that they be built with increasing numbers of components. Problematically, the mean time to failure (*MTTF* of individual components has over remained relatively flat over time. Thus, the larger the system, the more frequent the failures.

Traditionally, failures have been dealt with by periodically saving computational state onto persistent storage and then recovering from this state following any failure (*checkpoint-restart*). The utilization of systems is then measured using *goodput* which is the percentage of computer time that is spent actually making progress towards

the completion of the job. The goal of system designers is therefore to maximize goodput in the face of random failures using an optimal frequency of checkpointing.

Determining checkpointing frequency should be straight-forward: measure *MTTF*, measure amount of data to be checkpointed, measure available storage bandwidth, compute checkpoint time, and plug it into a simple formula [3]. However, measuring available storage bandwidth is not as straightforward as one would hope. Ideally, parallel file systems could achieve some consistent percentage of the hardware capabilities; for example, a reasonable goal for a parallel file system using disk drives for storage would be to achieve 70% of the aggregate disk bandwidth. If this were the case, then a system designer could simply purchase the necessary amount of storage hardware to gain sufficient performance to minimize checkpoint time and maximize system goodput. However, there exist no currently available parallel file systems that can provide any such performance level consistently.

## 2 Challenges

Unfortunately, although there are some that can, there are many IO patterns that *cannot* achieve any consistent percentage of the storage capability. Instead, these IO patterns achieve a consistently low performance such that their percentage of hardware capability diminishes as more hardware is added! For example, refer to Figures 1a, 1b, and 1c, which show that writing to a shared file, *N-1*, achieves consistently poor performance across the three major parallel file systems whereas the bandwidth of writing to unique files, *N-N*, scales as desired with the size of the job. The flat lines for the *N-1* workloads actually show that there is no amount of storage hardware that can be purchased: regardless of size, the bandwidths remain flat. This is because the hardware is not at fault; the performance flaw is within the parallel file systems which cannot incur massively concurrent writes and maintain performance. The challenge is due

<sup>\*</sup>San Francisco, CA; September 26-27, 2011

<sup>†</sup>LANL Release LA-UR 11-11416

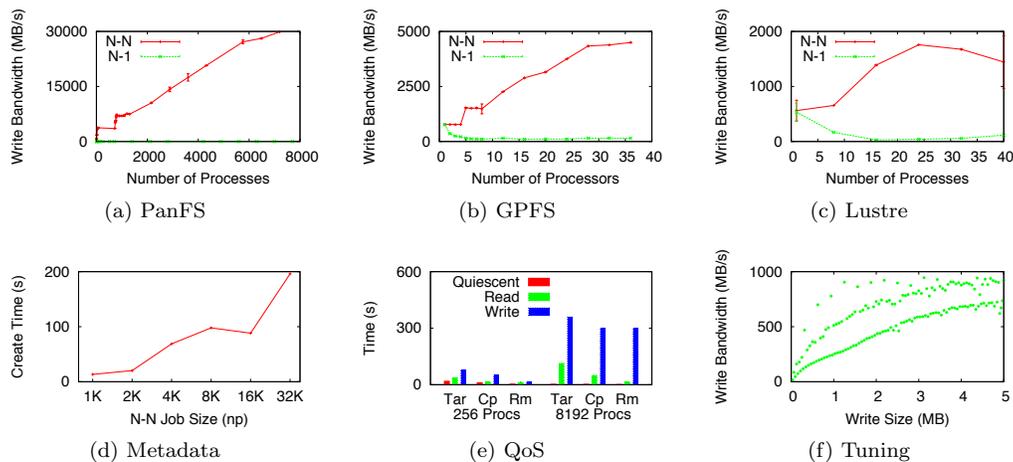


Figure 1: **Usability Challenges.** These graphs address the key usability challenges facing today’s users of HPC storage systems. The top three graphs demonstrate the large discrepancy between achievable bandwidth and scalability using N-N and N-1 checkpoint patterns on three of the major HPC parallel file systems. The bottom left graph shows the challenge of metadata creation at large job size, the bottom middle shows how the notion of interactivity is a cruel joke when small file operations are contending with large jobs performing parallel IO, and finally, the bottom right graph shows the reliance on magic numbers that plagues current parallel file systems.

to maintaining data consistency which typically requires a serialization of writes.

An obvious solution to this problem is for all users to always perform N-N file IO in which every process writes to a unique file. This approach does not come without trade-offs however. One is a performance limitation at scale and the other is a reduction in usability as will be discussed later in Section 3.

The system problem is the massive workload caused by by large numbers of concurrent file creates when each process opens a new file. Essentially this causes the same exact problem on parallel file systems as does writing in an N-1 pattern: concurrent writes perform poorly. In this case, the concurrent writing is done to a shared directory object. These directory writes are handled by a metadata server; no current production HPC parallel file system supports distributed metadata servers. As such, large numbers of directory writes are essentially serialized at a single metadata server thus causing very large slow-downs during the create phase of an N-N workload as is shown in Figure 1d.

### 3 Implications for Usability

This causes large usability headaches for LANL users. All of the large computing projects at LANL are well-aware of, and dismayed by, these limitations. All have incurred large opportunity costs to perform their primary jobs by designing around these limitations or paying large performance penalties. Many create archiving and analysis

challenges for themselves by avoiding writes to shared objects by having each process in large jobs create unique files. Some have become parallel file system experts and preface parallel IO by doing complicated queries of the parallel file system in order to rearrange their own IO patterns to better match the internal data organization of the parallel file system.

#### 3.1 Tuning

Many users have learned that parallel file systems have various *magic numbers* which correspond to IO sizes that achieve higher performance than other IO sizes. Typically these magic numbers correspond with various object sizes in the parallel file system ranging from a disk block to a full RAID stripe. The difference between poorly performing IO sizes and highly performing IO sizes is shown in Figure 1f which was produced using LBNL’s PatternIO benchmark [8]. Also, this graphs seems to merely show that performance increases with IO size, a closer examination shows that there are many small writes that perform better than large writes. In fact, a close examination reveals three distinct curves in this graph: the bottom is IO sizes matching no magic numbers, the middle is for IO sizes in multiples of the object size per storage device, and the upper is for IO sizes in multiples of the full RAID stripe size across multiple storage devices.

The implication of this graphs is that those users who can discover magic numbers and then use those magic numbers can achieve much higher bandwidth than those users who cannot. Unfortunately, both discover-

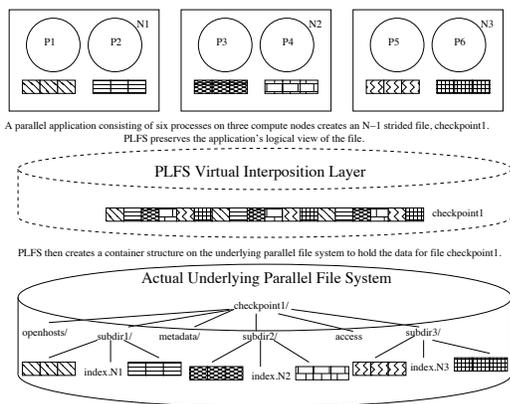


Figure 2: **PLFS Architecture.** This figure shows how PLFS maintains the user's logical view of a file which physically distributes it into many smaller files on an underlying parallel file system.

ing and exploiting magic numbers is difficult and often intractable. Magic numbers differ not only on different parallel file systems (*e.g.* from PanFS to Lustre) but also on different installations of the same file system. Tragically, there is no simple, single mechanism by which to extract magic numbers from a file system.

We have a user at LANL who executes initialization code which first queries *statfs* to determine the file system *f\_type* and then, based on which file system is identified, then executes different code for each of the three main parallel file systems to attempt to discover the magic number for that particular installation. Once discovering this value, the user then reconfigures their own, very complicated, IO library to issue IO requests using the newly discovered magic number. Of course, most users would not prefer to jump through such hoops, and frankly, many users should not be trusted with low-level file system information. Not because they lack intelligence but because they lack education; they are computational scientists who should not be expected to become file system experts in order to extract reasonable performance from a parallel file system.

Of course, even if all users could easily discover magic numbers, they could not all easily apply them. For example, many applications do adaptive mesh refinement in which the pieces of the distributed data structures are not uniformly sized: neither in space nor in time. This means that users looking for magic numbers will need some sort of complicated buffering or aggregation. An additional challenge is that magic numbers are not as easy as merely making the individual IO operations be of a particular size; they must also be correctly aligned with the underlying object sizes. So not only must users attempt to size operations correctly, they must also attempt to align them correctly as well. There are other approaches to

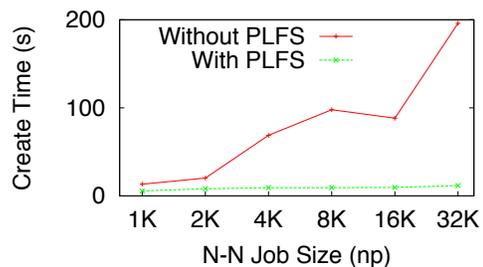


Figure 3: **Addressing Metadata Challenge.** This graph shows how distributed metadata keeps create rates manageable at large scale.

address this problem such as collective buffering [10] in MPI-IO. As we will show later in Section 4.1, collective buffering is beneficial but is not a complete solution.

### 3.2 Quality of Service

Finally, although checkpoint-restart is a dominant activity on the storage systems, obviously it is not the only activity. Computational science produces data which must then be explored and analyzed. As the output data is stored on the same storage system which services checkpoint-restart, data exploration and analysis workloads can contend with checkpoint-restart workloads. As is seen in Figure 1e, the checkpoint-restart workloads can wreak havoc on interactive operations. In this experiment, the latency of small file operations, such as untarring a set of files, copying that same set of files, and then removing the files, was measured during periods of quiescence and then compared to the latency of those same operations when they were contending with large parallel jobs doing a checkpoint write and a restart read. The most painful latency penalties are seen when the operations contend with a 8192 process job doing a checkpoint write.

## 4 Path Forward

There are many emerging technologies, ideas, and potential designs that offer hope that these challenges will be addressed in time for the looming exascale era.

### 4.1 PLFS

Our earlier work in SC09 [2] plays a prominent role in our envisioned exascale IO stack. That work showed how PLFS makes all  $N-1$  workloads achieve the performance of  $N-N$  workloads and also how PLFS removes the need for tuning applications to the underlying system (*i.e.* in PLFS, every number is a magic number!). Those results will not be repeated here but suffice it to say that they

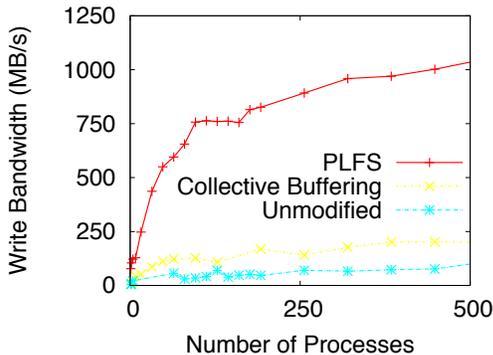


Figure 4: **Collective Buffering.** *This graph shows that collective buffering may not be sufficient for many workloads.*

eliminate the challenges show in Figures 1a, 1b, 1c, and 1f. From a usability perspective, PLFS is an important contribution: in addition to removing the need for IO tuning, PLFS is transparently accessible by *unmodified* applications using either POSIX IO or the MPI IO libraries.

Note that collective buffering [10] is another approach to dealing with the thorny problem of magic numbers. Figure 4 shows that, for one particular workload, collective buffering is an improvement over an unmodified approach to IO but underperforms the bandwidth obtainable using PLFS. In fairness, however, we are not collective buffering experts and perhaps collective buffering could be tuned to achieve much higher bandwidth. Ultimately though, our usability goal is to remove file system and parallel IO tuning from the user’s purview.

Figure 2 shows the architecture of PLFS and how it preserves the user’s logical view of a file while physically striping the data into many smaller files on an underlying parallel file system. This effectively turns all large parallel workloads into N-N workloads. Of course, as we saw in Figure 1d, even N-N workloads suffer at very large scale. Additionally, we know that this performance degradation is due to an overloaded metadata server which will destroy interactive latency as we saw in Figure 1e.

Borrowing ideas from GIGA+ [7], PLFS now addresses these challenges as well. Recent versions of PLFS (since 2.0) can stripe the multiple physical files comprising a logical file over multiple file systems. In the case where each file system is served by a different metadata server, this distributes metadata load very effectively as can be seen in Figure 3 which is the same as Figure 1d but with an added line showing how metadata distribution within PLFS can remove metadata bottlenecks. Note that the workload shown was run using an N-N workload. Although PLFS was originally designed for N-1 workloads, this new functionality will allow PLFS to address metadata concerns for all exascale checkpoint workloads.

## 5 Redesigning the IO Stack

PLFS has proven to be a very effective solution for current IO challenges: it allows all workloads to easily achieve a consistently high percentage of the aggregate hardware capability.

PLFS is not sufficient however to solve the looming exascale IO challenges before us. Recent work [9] shows that the checkpointing challenge is becoming increasingly difficult over time. The checkpoint size in the exascale era is expected to be around 32 PB. To checkpoint this in thirty minutes (a decent rule of thumb) requires 16 TBs of storage bandwidth. Economic modeling shows that current storage designs would require an infeasible 50% of the exascale budget to achieve this performance.

### 5.1 Burst buffer

We must redesign our hardware stack and then develop new software to use it. Spinning media (*i.e.* disk) by itself is not economically viable in the exascale era as it is priced for capacity but we will need to purchase bandwidth. Additionally, the storage interconnect network would be a large expense. Thus far, we have required an external storage system for two main reasons: one, sharing storage across multiple supercomputers improves usability and helps with economies of scale; two, embedding spinning media on compute nodes decreases their MTTF.

Our proposal is to make use of emerging technologies such as solid state devices, *SSD*. This media is priced for bandwidth and for low latency so the economic modeling shows it is viable for our bandwidth requirements. Additionally, the lack of moving parts is amenable to our failure models and allows us to place these devices within the compute system (*i.e.* not on the other side of the storage network). Unfortunately, being priced for bandwidth means these devices cannot provide the storage *capacity* that we require. We still require our existing disk-based parallel file systems for short-term capacity needs (long-term capacity is served by archival tape systems not otherwise discussed here).

We propose adding these devices as a new layer in our existing storage hierarchy between the main memory of our compute nodes and the spinning disks of our parallel file systems; we call this interposition of SSD a *burst buffer* as they will absorb very fast bursts of data and serve as an intermediate buffer to existing HPC parallel file systems. This is not a new idea and is commonly suggested as a solution to the well-known latency gap between memory and disk. Our proposal however is how to specifically incorporate these burst buffers into the existing HPC storage software stack.

## 5.2 E Pluribus Unum

Our envisioned software stack incorporates many existing technologies. The SCR [6] software is a perfect candidate for helping schedule the burst buffer traffic and to enable restart from neighboring burst buffers within the compute nodes. However, we envision merging SCR and PLFS to allow users to benefit from PLFS's capability to handle both N-1 and N-N workloads and to allow use by unmodified application.

We have already add PLFS as a storage layer within the MPI IO library. This library has many important IO optimizations in addition to collective buffering described earlier. One such optimization is available using *MPLFile\_set\_view*. This is an extremely nice feature from a usability perspective. This is clear when we consider what computational scientists are doing: they stripe a multi-dimensional data structure representing some physical space across a set of distributed processors. Dealing with these distributed multi-dimensional data structures is complicated enough without even considering how to serialize them into and out of persistent storage. *MPLFile\_set\_view* lesses this serialization burden; by merely describing their distribution, the user then transfers the specific serialization work to the MPI IO library.

Note that other data formatting libraries such as HDF [1], Parallel netCDF [4], SCR, and ADIOS [5] provide similar functionality and have proven very popular as they remove computer science burdens from computational scientists. These data formatting libraries are the clear path forward to improve usability of HPC storage. However, they will not work in their current form on burst buffer architectures. We envision adding our integrated PLFS-SCR storage management system as a storage layer within these data formatting libraries just as we have done within the MPI IO library. A key advantage of a tight integration between PLFS-SCR and these data formatting libraries is that semantic information about the data can be passed to the storage system thus enabling semantic retrieval.

## 5.3 In situ data analysis

There are two key features of our proposal that enable *in situ* data analysis. The first is that the burst buffer architecture embeds storage much more closely to the compute nodes which drastically reduces access latencies for both sequential and random accesses. The second is that because the data has been stored using data formatting libraries, semantic retrieval of data is possible. This means that we can more easily attempt to co-locate processes within the analysis jobs close to the burst buffers containing the desired data. Finally, even when the data is not available on a local burst buffer, we can take ad-

vantage of the low-latency interconnect network between the compute nodes to transfer data between burst buffers as needed.

## 6 Conclusion

In this proposal, we have described how current usability of HPC storage systems is hampered by two main challenges: poor performance for many large jobs, and occasional intolerably slow interactive latency. We have offered PLFS as a solution for these challenges on today's systems.

Finally, we point out the inability of PLFS to address exascale challenges by itself. We then offer a proposal for integrating PLFS with a burst buffer hardware architecture PLFS and a set of other existing software packages as one path towards a usable and feasible exascale storage solution.

## References

- [1] The HDF Group. <http://www.hdfgroup.org/>.
- [2] J. Bent, G. Gibson, G. Grider, B. McClelland, P. Nowoczynski, J. Nunez, M. Polte, and M. Wingate. Plfs: a checkpoint filesystem for parallel applications. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, SC '09*, pages 21:1–21:12, New York, NY, USA, 2009. ACM.
- [3] J. T. Daly. A higher order estimate of the optimum checkpoint interval for restart dumps. *Future Gener. Comput. Syst.*, 22(3):303–312, 2006.
- [4] J. Li, W. keng Liao, A. Choudhary, R. Ross, R. Thakur, W. Gropp, R. Latham, A. Siegel, B. Gallagher, and M. Zingale. Parallel netcdf: A high-performance scientific i/o interface. *SC Conference*, 0:39, 2003.
- [5] J. F. Lofstead, S. Klasky, K. Schwan, N. Podhorszki, and C. Jin. Flexible io and integration for scientific codes through the adaptable io system (adios). In *CLADE '08: Proceedings of the 6th international workshop on Challenges of large applications in distributed environments*, pages 15–24, New York, NY, USA, 2008. ACM.
- [6] A. Moody, G. Bronevetsky, K. Mohror, and B. R. d. Supinski. Design, modeling, and evaluation of a scalable multi-level checkpointing system. In *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, SC '10*, pages 1–11, Washington, DC, USA, 2010. IEEE Computer Society.
- [7] S. V. Patil, G. A. Gibson, S. Lang, and M. Polte. GIGA+: Scalable Directories for Shared File Systems. In *Petascale Data Storage Workshop at SC07*, Reno, Nevada, Nov. 2007.
- [8] Rajeev Thakur. Parallel I/O Benchmarks. <http://www.mcs.anl.gov/thakur/pio-benchmarks.html>.
- [9] B. Schroeder and G. Gibson. A large scale study of failures in high-performance-computing systems. *IEEE Transactions on Dependable and Secure Computing*, 99(1), 5555.
- [10] R. Thakur and E. Lusk. Data sieving and collective i/o in romio. In *In Proceedings of the Seventh Symposium on the Frontiers of Massively Parallel Computation*, pages 182–189. IEEE Computer Society Press, 1999.

# PSI – A High Performance File Transfer User Interface

Mark A. Roschke

*High Performance Computing Division,  
Los Alamos National Laboratory  
mar@lanl.gov*

C. David Sherrill

*High Performance Computing Division,  
Los Alamos National Laboratory  
dsherril@lanl.gov*

## Abstract

*Transferring and maintaining large datasets requires parallel processing of both data and metadata for timely execution. This paper describes the work in progress to use various processing techniques, including multi-threading of data and metadata operations, distributed processing, aggregation, and conditional processing to achieve increased transfer performance for large datasets, as well as increased rates for metadata queries and updates.*

## 1. Introduction

Ever-increasing computing capabilities result in ever-increasing data sets to be transferred. Such data sets can consist primarily of large files, many small files, or both. Transferring data sets with large files requires an emphasis on parallel file transfer, utilizing as much bandwidth as possible. And it is in this area that the majority of data parallel transfer development has occurred. But, it is no longer rare for a user to generate data sets of 100,000 to one million files. And when data sets reach this size, it is imperative that support be provided for high performance metadata operations, not only in support of file transfer, but also to support browsing and maintaining the data set.

## 2. Overview of PSI

The Parallel Storage Interface (PSI) is a data transfer user interface designed to provide high speed transfer for large data sets, with a special emphasis on utilizing as many resources as possible for a single user request. Developed by the authors, PSI is the main user interface to the High Performance Storage System

(HPSS) at Los Alamos National Laboratory. This paper describes the efforts to provide a full-featured data transfers capability for archival transfer, local transfer, and wide area host-to-host transfer, providing both high-speed data transfer as well as high-speed metadata processing.

PSI uses a parallel workflow model for processing both data and metadata. Work is parallelized and scheduled on available server and client resources automatically, using a priority and resource-based approach. Optimization is performed automatically, including areas such as parallelization, optimized tape transfer, load leveling, etc.

## 3. Unix syntax and Semantics

PSI utilizes UNIX-like syntax and semantics. For example, the following commands are available for data transfer and manipulation of file attributes: **cd**, **chmod**, **chgrp**, **cp**, **du**, **find**, **grep**, **ls**, **mkdir**, **mv**, **rm**, **rmdir**, and **scp**.

## 4. Multi-mode Operation

PSI offers three modes of operation, providing the same syntax, semantics, and look and feel for the three most frequently used data transfer situations, which are 1) local transfer, 2) archival transfer, and 3) host-to-host transfer. The particular interface command determines the context of the specified commands. For example,

<b>sh</b>	<b>cp -R a b</b>	copy files locally
<b>psiloc</b>	<b>cp -R a b</b>	parallel copy locally
<b>psi2ccc</b>	<b>cp -R a b</b>	parallel copy on cluster <b>ccc</b>
<b>psi</b>	<b>cp -R a b</b>	parallel copy in the archive

This approach provides a consistent look and feel, allowing the user to move between the 3 major transfer situations, eliminating the time necessary to learn the command set for each situation.

## 5. Automatic Optimization

The general design approach for PSI is that the user simply specifies the files to be operated on, PSI determines the resources available to the command, and then executes the command, with all optimization being performed automatically, including such features as adjusting all types of thread counts dynamically, optimizing the order of any data transfers to/from tape, assignment of operations across multiple hosts (including load leveling), and splitting large transfers across hosts when appropriate.

To support automatic optimization, all activity within PSI is controlled using a priority-based resource management scheme, limiting the amount of bandwidth and memory that each type of activity can consume. Scheduling of activities such as file transfers are performed via an internal job scheduler, which dispatches activities across available hosts in an optimal order, load leveling all activities as necessary.

## 6. Conditional Transfers

To address occurrences of interrupted transfers as well as that of newly arriving (or updated) data within a data set, PSI can scan both the input tree and output tree, examining file attributes to determine which files need to be transferred. This feature alone can routinely save hours of time that would be spent on re-transferring the entire tree.

## 7. Parallel Archival Tarring Option

When transferring to the archive, the user can select the PSI tar option, which automatically utilizes parallel tar transfers to/from the archive. The parallel tar capability in PSI typically constructs one or more tar files per directory, preserving the original tree structure. Large files are normally transferred un-tarred to the archive. Multiple tar processes are load leveled across available hosts, providing scalable multi-host performance, even for small files.

The archive namespace is extended into the tar files present, utilizing the index file that is stored with each tar file. This namespace extension prevents the archive from becoming a large black box of data. The user can

browse through the original tree, and execute such commands as **ls**, **find**, **grep**, **rm**, and **scp** with references to files within the tar files, and can also utilize globbing (i.e. wild cards) in such references. Conditional transfers are also supported, so that newly arrived files can be placed within new tar files in a directory. In addition, commands such as **scp**, **chmod**, **grep**, **ls**, and **rm** are specially aware of the tar files, and can take advantage of operating on whole tar files when feasible.

## 8. Techniques to Increase Performance

The general approach chosen involves the use of parallel data and metadata processing, automatic optimized file aggregation and de-aggregation, and conditional operations when feasible. Combining these three features provides a variety of performance increases. For example, multi-threading to a degree of 40 threads might increase performance by a factor of 30, while operating on a file aggregate of 1000 files can provide a performance boost of up to 300. Conditional operations can provide a factor of 20 or so. By combining these three features, performance gains of over 1,000 have been observed, as outlined below.

## 9. Multiphase Parallel Work Flow

To facilitate efficient control of the various steps required to execute user requests in a parallel fashion, For example, tasks are organized into phases, e.g. 1) stat source files, 2) stat destination files, 3) transfer files. Work progresses through each phase. Each phase can consist of many threads, each requiring different resources. Achieving high performance in processing metadata requires a reasonably high degree of parallelism; typical thread counts for all three phases is 100 to 200, depending upon the mix of metadata and file transfer operations being performed.

## 10. Areas of Performance Increase

Work at increasing performance has fallen in two general areas – increasing parallelism, and decreasing latency with the latter area receiving the most effort. Increasing parallelism generally falls in the predictable categories of more threads, and more hosts, with some miscellaneous optimization applied to areas such as when to transfer large files across nodes, etc.

Effort to decrease latency has been largely in the area of various types of aggregation, namely 1) data aggregation, 2) control aggregates, 3) metadata query aggregates, and 4) metadata update aggregates.

Metadata query work has involved experiments with striping directory queries across multiple hosts, with an eye toward support of massive directories (directories with greater than 50,000 files).

Since aggregation is largely connected with latency, the benefits from aggregation tend to be shared across the areas of faster scheduling, more scalability, and faster WAN operations.

## 11. Conclusion

Combining the techniques of multi-threaded processing of data and metadata with the concept of small file aggregation can result in significant performance increases. These increases can be further improved by adding techniques such as conditional updates or conditional file transfers. Performance increases above factors of 1000 have been observed. In addition, using user-generated aggregates can result in significant decreases in archival system metadata.

## 12. Performance Results

The following results were obtained on a cluster of 4 client nodes connected to a Panasas file system, For various files sizes and commands.

### Local Mode

Mode	cp 1KB (files/s)	cp 10MB (MB/s)	cp 1GB (MB/s)	conditional transfer (files/s)	chmod (Files/s)	find (Files/s)	rm (Files/s)
sh	32	15	55	-	47	312	247
psi (local)	1,813	398	431	2364	2,090	1,743	2,999

Also, in a recent large scale test on 16 nodes connected to a Panasas file system, 295 TB of data (consisting of 967,000 files) were copied at an average rate of 2.85 GB/sec.

The following results were obtained on a cluster of 4 nodes, from a Panasas file system to Los Alamos HPSS.

### Archive Mode (HPSS)

Mode	cp 1KB (Files/s)	cp 10MB (MB/s)	cp 1GB (MB/s)	cond (files/s)	chmod (Files/s)	find (Files/s)	rm (Files/s)
psi (HPSS)	80	1,071	580	102	295	599	155
psi (HPSS, TAR)	1,139	256	269	2,365	31,188	2,999	15,210

The following results were obtained on a cluster of 4 nodes, from a local Panasas file system to a remote Lustre file system, with a round trip time of 38 ms (Los Alamos, NM to Livermore, CA)

### Host-to-Host Mode

Mode	cp 1KB (Files/s)	cp 10MB (MB/s)	cp 1GB (MB/s)	cond (files/s)	chmod (Files/s)	find (Files/s)	rm (Files/s)
psi (h2h)	1,933	423	480	2,396	2,433	3,012	229

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper**

**Richard Hedges**

Lawrence Livermore National Laboratory  
hedges1@llnl.gov

**ABSTRACT / SUMMARY**

**This position paper discusses issues of usability of the large parallel file systems in the Livermore Computing Center. The primary uses of these file systems are for storage and access of data that is created during the course of a simulation running on an LC system.**

**INTRODUCTION**

The Livermore Computing Center has multiple, globally mounted parallel file systems in each of its computing environments. The single biggest issue of file system usability that we have encountered through the years is to maintain continuous file system responsiveness. Given the back end storage hardware that our file systems are provisioned with, it is easily possible for a particularly I/O intensive application or one with particularly inefficiently coded I/O operations to bring the file system to an apparent halt.

The practice that we will be addressing is one of having an ability to indentify, diagnose, analyze and optimize the I/O quickly and effectively.

**Tools applied**

**LMT:** LMT[1] (Lustre monitoring tool) is run by the Livermore Computing system administration staff to monitor operation of the Lustre file systems. It is generally used to probe and isolate reported problems rather than to identify the problem before or as it develops.

Having an earlier version of LMT accessible to users proved problematic. The particular issue was that some users would “cry wolf” when they saw periods of heavy usage of a file system. These notifications were generally self serving and counter productive, so presently LMT is available for system administrators only.

In daily operations, Lustre system administrators may have running instances of LMT, but would not necessarily be tracking the output, unless a problem (such as a file system being sluggish or unresponsive) had been reported. Due to the architecture of Lustre, LMT leads one down an indirect path in identifying the source of a file system load. Load is observed on storage or metadata servers, next correlated with client activity, and finally (hopefully) identified with a single users job. This detective work can take some time, so it can be a challenge to do all of the tracing while the offending code instance is still active.

**Darshan, strace:** Since a single application program with inefficiently coded I/O operations can have center wide negative impact on parallel file system function and usability, It is critical to be able understand the sequence of I/O operations that a code is generating and to understand the effects of those on file system behavior.

For some time we have been in the business of profiling file system I/O for selected applications to diagnose and resolve performance problems causing center wide impact. Initially profile data was extracted exclusively from strace [2], and

application runs and analyzed essentially by manually reviewing the data.

We generally trace with the options “strace -tt -etrace=file,read,write,close,lseek,iocntl” which provides time stamped system call traces to standard error for the I/O related system calls identified. We can collect the system call traces on a per process basis. It is possible to trace a running process, or to incorporate the tracing in a job run script.

More recently we also use Darshan[3] from Argonne National Laboratory. Darshan is a petascale I/O characterization tool. Darshan is designed to capture an accurate picture of application I/O behavior, including properties such as patterns of access within files, with minimum overhead. Darshan includes scripting to analyze and aggregate the data.

A code can be instrumented with Darshan by utilizing wrapper scripts, or by interposing the libraries using LD\_PRELOAD. Being as lightweight as it makes it suitable for full time deployment, although we at LC do not apply it in that manner.

These methods are available to users, but have primarily been applied by an LC staff member on behalf of a user or application team. Note that these methods are also applied in the case where the performance issues impact the user’s productivity, even if the center-wide impact is minimal.

### **User training and documentation**

Training specific to application I/O performance issues is summarized in two documents maintained on the clusters in /usr/local/docs: (1) Lustre.basics and (2) Lustre.striping. We have also included I/O specific discussions in user oriented system status meetings on a regular basis. Consulting is available, and is offered on a general and on an intervention basis.

Let me interject a personal comment here related to user training, because I am eager to see if others at the workshop have observed similar. Relative to other parts of parts of a complex HPC system (e.g. processor architecture, code parallelization) I find that our user community seems generally more resistant to learning about the I/O architecture and how to use and code to it effectively. I suspect that this is a historically bias that the CPU processing is the valuable resource and the I/O bandwidth and storage capacity are free. We at the center may have reinforced this, if subtly, by our accounting and allocation policies.

### **CONCLUSIONS**

For the key initial step of identifying a code or user who, by their I/O actions are impacting the user, we have a workable tool. LMT provides an path, albeit indirect to associate system load with a particular root cause.

The strace and Darshan offer approaches (some overlapping and some complementary) to analyze the I/O execution of an HPC application. They allow one to identify and localize a problem in a code.

We have an issue on the user training side. Some code teams have taken on the challenge of understanding good application I/O practices. Others have been motivated only when I/O performance was an insurmountable hurdle.

### **REFERENCES**

1. LMT github site  
<https://github.com/chaos/lmt/wiki>
2. strace: standard Linux command to “trace system calls and signals” see “man strace”
3. Darshan: Petascale I/O Characterization Tool  
<http://www.mcs.anl.gov/research/projects/darshan/>

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper  
NCAR's Centralized Data Services Design**

**Pamela Gillman**

NCAR Computational & Information Systems Lab  
pjpg@ucar.edu

**Erich Thanhardt**

NCAR Computational & Information Systems Lab  
erich@ucar.edu

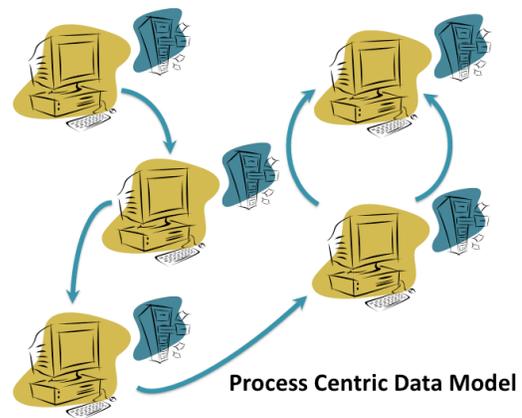
**ABSTRACT / SUMMARY**

**As NCAR designs and builds our next generation computational center, we are exploring ways to evolve scientific data workflows from a process centric model to a more information centric paradigm. By looking at cyberinfrastructure design, resource allocation policies and software methodologies, we can help accelerate scientific discoveries possible from computational resources of this scale. We will explore the challenges we have identified in our data architecture and present some of our current projects moving us towards an information centric solution.**

**INTRODUCTION**

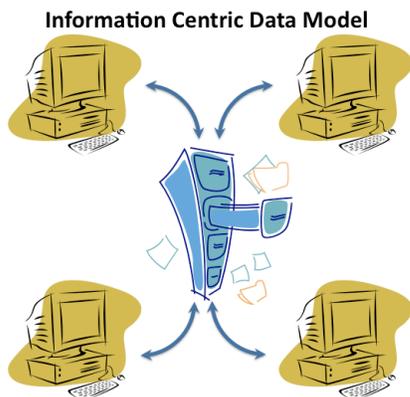
Computational centers have traditionally provided systems architected for a single use such as computation, data analysis or visualization. Each resource has local storage configured for the typical task performed and the center provides a common tape based archival system. This encourages a scientific workflow that typically involves retrieving input data from the archive system, generating new data, including intermediate files necessary for the next run, and finally storing all data back on the archive system. If you wish to post-process, analyze or visualize the data, you need to read the data back

onto a different resource. This process may repeat multiple times as either issues are identified in the data, or discoveries spark a new direction of research.



When surveyed, users identified the movement of data between resources as a significant bottleneck in their workflow. Therefore, armed with this information, and driven by the ever-escalating costs of archival systems and the increases in ability to produce data, we started looking at architectural solutions to evolve workflow. The traditional workflow model is very process centric. Data moves between resources dedicated to a single step in the overall process and the archive essentially becomes a file server. What if we started looking at the data as the center of the workflow? Not only would this decrease the number of data movements in the workflow, but it can potential decrease the amount of data

ultimately targeted for actual archiving. We now refer to this as an information centric model.



### Evolving Scientific Data Workflow

Based on an analysis of current workflows, we identified several challenges in evolving data workflow toward the new paradigm. Many bottlenecks exist in current workflows; it's time consuming to move data between systems; bandwidth to the archive system is insufficient; and available disk storage space is insufficient.

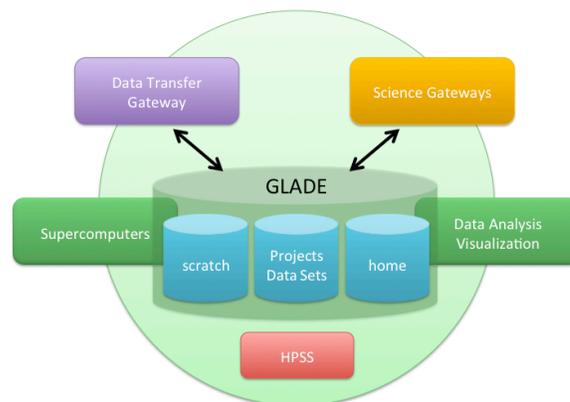
This presents a bit of a 'chicken and egg' problem. The current environment potentially shapes current user behavior. How do we anticipate the behavioral changes that will occur with a significant change in the environment? Storage cost curves are steeper than compute costs so how do we find the right balance between storage and compute investments? Archive costs are on an unsustainable growth curve so how do you better balance usage of disk space versus archive space?

### Globally Accessible Data Environment

The first step we took was to centralize our data analysis and visualization resources around a centralized storage system. This was a lower bandwidth parallel file system solution that provided users of these resources a single namespace for their work. Space was provided not only for short-term 'scratch' usage but also for longer-term project spaces necessary for data analysis and visualization work.

The next step was designing a scalable architecture that encompassed all HPC resources

including access to data collections managed by NCAR. This architecture needed to be flexible enough to support current systems and science gateways, and also able to scale as HPC resources grow with the new data center. And to ensure the shift in workflows, the user needs to be able to interface with this environment in the same way no matter which resource they are working from or what task they are trying to accomplish. High bandwidth connectivity to any resource within this environment and a choice of interfaces support current projects and are flexible enough to support future requirements.



The GLADE data architecture becomes the centerpiece of the new information centric workflow. Data can stay in place through the entire process as GLADE provides not just 'scratch' space for computational runs, but persistent longer-term storage for data processing, analysis and visualization. This persistent storage allows completion of the entire workflow prior to final storage of results either at NCAR or offsite. The addition of high-bandwidth data transfer services with direct access to the GLADE environment provides efficient methods for moving data between NCAR and peer institutions.

### Accounting Systems Enhancements

To allow for better tracking of resource usage, the NCAR accounting system is being redesigned to account for computational resources, data analysis and visualization resources, disk storage usage, data transfer services and archival services. These tools will allow NCAR to fine

tune the balance between resources based on evolving usage patterns due to changes in workflow.

## **Workflow Examples**

### *Case 1: Nested Regional Climate Model (NRCM)*

The project group has common access to ‘scratch’ space and a dedicated longer-term project space. The computational team submits a model run to the supercomputing queue. The model outputs approximately 100 variables per time step along with intermediate data files associated with startup of the next time step to the ‘scratch’ file system. Once the model completes, a post-processing job pulls approximately 20 variables of interest into data analysis files writing these into their project space. The analysis files are now available for analysis by the research team. This data will stay in place as long as necessary to complete the analysis. A final job step writes the final full output files to the HPSS archive. Since the ‘scratch’ file system is purged regularly, intermediate files that are no longer needed are never stored on the archive, and smaller data sets needed for analysis are available to the team right after the computational run completes.

### *Case 2: Research Data Archive (RDA)*

The Research Data Archive provides access to common data sets to the research community. Prior to the implementation of the GLADE architecture this data was only available from the archive system. By allocating space within GLADE for the RDA data, access can now be granted directly from NCAR’s HPC resources. Previous workflows needed to first copy this data from the archive to a ‘scratch’ area before running the computational job. There were costs in time required to access the data, space required to hold the data and the side effect of numerous copies of the same data being on disk at the same time. With direct access now available, jobs use the data from a central location that’s immediately available to all jobs and doesn’t rely on archival access..

## **CONCLUSIONS**

We feel that we have made progress towards a better architecture to meet the diverse needs of our user community. We believe that this architecture is sustainable into the future and will help balance the costs associated with compute/storage/archival. Checks are in place so adjustments can be made as user behaviors change and hopefully data management becomes not just a tedious task, but also something that results in more productive scientific discovery.

# U.S. Department of Energy Best Practices Workshop on

## File Systems & Archives

San Francisco, CA

September 26-27, 2011

### Position Paper

Kevin Glass

PNNL

kevin.glass@pnnl.gov

#### ABSTRACT / SUMMARY

**MyEMSL is a data management system for scientific data produced at EMSL. The data must be made accessible to users either by a simple directory-style search, a metadata search or as part of a workflow. To provide these features the system requires several points of interaction between users and the EMSL archive.**

**This position paper addresses the workshop's Usability of Storage Systems track and summarizes what we have accomplished in the development of MyEMSL and some of the challenges that remain with regard to the interaction between users and our archive.**

#### INTRODUCTION

The MyEMSL data management system was developed to collect and archive data produced by EMSL instruments. The data is made available to the scientists who are allowed access to the data via the web and to analyze that data using EMSL's computational resources, including Chinook, EMSL's supercomputer.

EMSL boasts more than 150 scientific instruments that are capable of producing terabytes of data each week. The data collected

by EMSL instruments comes in a wide variety of formats including images, spectrographs, single value outputs. The data will appear as files ranging from a single file on the order of 100 GB file to thousands of files on the order 1 MB each. Some of the instruments will produce only a single file, which currently is entered manually into a lab notebook.

In addition to getting EMSL users their data, the MyEMSL archive was designed to be a node in a workflow. Raw and processed data is stored in the archive, which is connected to computational resources. For example, the user can transfer data from the archive to a search node running Hadoop. The results of this search will be transferred back to the MyEMSL archive, then passed to Chinook for further processing and ultimately presented to the user through a visualization tool.

As with most of MyEMSL, this technique is in its infancy. We are examining options for critical parts of the infrastructure though most much of it is in place and working.

#### 1. DESIGN OF MyEMSL

The design philosophy for MyEMSL was to let someone else do the work. The system relies on several open source software products such as Apache and SLURM. The goal was to select the optimal product for each type of component by testing the performance and ease of use for each

product. However, budget, time and other factors limited the amount of testing and development available to the team.

Exacerbating this problem was the need to understand the use cases for hundreds of scientists using hundreds of different instruments. Prior to the development of MyEMSL, the development team reviewed information collected from previous attempts to develop data management systems for EMSL. These records included information regarding the amount of data collected by each instrument, the number and sizes of the files produced, the format or formats of the data and whether or not further processing of the data was required before it was released to the end user. The instrument information survey was originally collected on paper. It has since been moved to a web form and repopulated. As new instruments are added to EMSL's instrument suite, the person responsible for the instrument will complete this form.

In addition to collecting fundamental information regarding new equipment, the developers need to collect information regarding the metadata associated with the system. Given the number and specialization of each instrument, collecting the appropriate information from a one-on-one interview is infeasible given our time constraints. To address this problem, we are implementing two features in this system: a metadata form builder and metadata extractors to collect metadata that is automatically collected by the instrument and stored in the raw or processed data files.

## **2. COLLECTING METADATA**

Given the evolving nature of science and scientific inquiry, the required metadata collection set will necessarily evolve. This requires MyEMSL to allow for a flexible metadata storage system. We define metadata to be data used by the end user for searching; a description of where the results of an experiment reside. When data is transferred from an instrument to the archive, the data is divided into metadata and raw or processed data. The raw or processed data is stored in the archive as a set of

files and the metadata is stored in a database. To facilitate searches, MyEMSL generates specific databases based on the field of interest. These act as index caches for common—field based—searches. For example, searches for system biologists may include genetic information and ignore thermodynamic properties gathered from a single experiment. The main point is the scientist must be the person defining what is important for the scientist.

The specifics of metadata storage are still under investigation. We are currently considering two possibilities: a large relational database and a quad store. We are experimenting with these options using three criteria: performance, ease of implementation and ease of use. Of these options, only performance is an objective measure. The others are clearly subjective and will be tested within the constraints of our budget.

The management of raw, processed and metadata lead to two questions: how do we collect provenance data and how do we define data using common or standardized mechanisms? The collection of provenance data begins with an EMSL user submitting a sample for analysis. As this is the only person who can legitimately define the nature of the sample, the user is required to complete a “sample submission” form. Currently, this form collects only the basic data regarding a sample, such as the name of the submitter, the date of submission and so on. As with the collection of experiment metadata, the information required for a given sample type must be defined by the scientists responsible for operating the instrument.

Other sources of provenance data include the configuration of each instrument used to process a sample, the operating environment of the instrument and a description of the workflow used to process a sample. Some of the data is automatically collected from the instrument or detectors near the instrument. Others must be manually entered before data is loaded into the archive. This presents a user-interface problem. The system must minimize requirements for

manual data collection and it must ensure the data is collected. We are still working on this problem.

### **3. UNIFYING DATA STANDARDS**

The question regarding common or standardized data representations is a difficult one to answer. The scientific community agrees on few standardized representation and any level. Thus, there are several “standards” for each field and implementing all of them would be impossible. Our approach to this problem is to adopt a single, recognized standard for storing metadata and to use field-specific representations for the index caches. We are currently working with Nanotab [1] to represent both nanoparticle and some biological data. To generate the index caches, MyEMSL requires translators from Nanotab to other domain standards. We are currently investigating, with input from the scientific community, how to implement this feature.

The central feature of MyEMSL is its ability to move files of varying sizes to the archive and to retrieve those files. Before a scientist performs and experiment, he or she must configure a listener called ScopeSave. ScopeSave will monitor a specified directory and, when the experiment is complete, it will archive and

compress the directory. The archived data files are sent to the archive via an Apache server to a storage queue managed by SLURM. When the archive is ready to store the input data, the data is transferred from the queue to the archive.

### **CONCLUSIONS**

MyEMSL is a data management system currently in operation at EMSL however we are continuing to investigate some features. The system can take data from scientific instruments, load the raw, processed and metadata in a database and makes this data available to users. The main issues under investigation relate to the ease of use of the system by end users. These include features that are directly accessible by the user, for example, data transfer from an instrument to the archive and features that are indirectly accessible, such as the metadata database.

### **REFERENCES**

1. caBIG Nanotechnology Working Group  
<http://sites.google.com/site/cabignanowg/home>

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper**

**Ruth Klundt**  
Sandia National Labs  
rklundt@sandia.gov

**John Noe**  
Sandia National Labs  
jpnoe@sandia.gov

**Stephen Monk**  
Sandia National Labs  
smonk@sandia.gov

**James Schutt**  
Sandia National Labs  
jaschut@sandia.gov

## **ABSTRACT**

**We here present an overview of our current file system strategies, and a brief mention of planning for the future. The focus of the discussion is the link between usability issues and implementation decisions.**

## **INTRODUCTION**

Our primary drivers in the design of file system solutions are reliability and performance. In addition we attempt to provide solutions covering a spectrum of user needs, which also include convenience of use, backup capability, and high availability.

### **Overview of Current File Systems**

User requirements in the storage arena are often difficult or impossible to satisfy simultaneously with a single global solution. Simulation codes generate a large quantity of restart data that must be stored quickly, as a defense against system outages. Most of this data is transitory, so does not need to be backed up. Other types of user data such as application codes and input data must be stored reliably. During periods of maintenance, it is important to users for the continuity of their

work that some portions of the infrastructure remain available.

At present we maintain three basic categories of storage.

- Site-Wide Parallel File System

Our parallel file system is implemented using Lustre [1] running on commodity servers, backed by DDN 9900/9550 raid cabinets. This file system serves ~2PB of fast scratch space to 4 different clusters, via LNET routers. Testing is under way on an upgrade to DDN SFA10K hardware providing ~3PB space for the new TLCC2 installation. Software support for Lustre is provided by Whamcloud [2].

- Intermediate NFS File System

On all clusters, a large storage space is delivered by means of Sun Unified Storage (7410) using ZFS. This is not purged, and not backed up.

- Traditional NAS

Less than 100TB, provided by NetApp hardware backed up to corporate archives. Stable, safe, slow location serving user /home and /projects space commonly across the clusters.

## Usability Impact

The parallel file system satisfies the need for fast storage of large data sets. Although no backups can be done at this size, all possible efforts are made to avoid data loss, by means of hardware RAID configurations and continuity by means of Lustre failover and Multi-path IO. The local Red Sky Lustre implementation, which requires use of software RAID on the Sun equipment, has encountered some difficulties due to increased operational demands and is slated to be shutdown in favor of site file systems.

The intermediate NFS file system provides an alternate location for users to continue work during maintenance periods on the parallel file system. The longevity of the Sun 7410 platform is not clear given the lack of a clear hardware roadmap from Oracle. Although it has proven to be a solid product within this role, we are moving to a solution that is less of a “black box” from the view of the hardware (see below).

The NetApp filers serving /home and /projects have a fairly long history of providing robust reliable service here, although of limited size. New or different solutions have a high bar to meet in order to be considered as replacements for this functionality.

## Future Plans

- GPFS NAS

Some DDN 9550 cabinets are currently being re-purposed for use with IBM’s GPFS file system [3] as an alternate highly available storage space, implemented at minimum cost. Production deployment is imminent.

- Ceph

An effort is in progress to test the robustness, usability, and performance of the Ceph file system [4]. Early results show promise for this open source solution as a potential alternate in the NAS file system space in the near future. In addition, a variety of use cases other

than HPC are being actively explored elsewhere, such as the ability to export as NFS, integration with PNFS [5], and access via user space clients. Interest in Ceph from disparate data storage venues can only improve the robustness of the implementation, and a broad user base provides some confidence that the file system has a productive future ahead.

Some key design elements that make Ceph a high performance file system of interest:

- Workload scalability (lots of servers/clients)
- On-line expansion (easy to add capacity and performance)
- Data replication (fault tolerance without RAID controllers)
- Adaptive meta-data server (scalable)
- Ability to reliably use commodity storage platforms

In conjunction with the Ceph testing effort, a heterogeneous test bed is being expanded and shared as a release test platform for production machines.

## CONCLUSIONS

Challenges in maintaining multiple types of storage might be mitigated in the future, with improvements in current parallel file systems with respect to reliability and availability. Ideally a single global file system solution with pools of storage configured for different use cases would streamline the delivery of the disparate services needed. A single solution capable of providing sufficient bandwidth to parallel platforms, differential backup capabilities, and 24/7 availability to users does not yet exist.

## REFERENCES

1. Lustre <http://www.lustre.org>
2. Whamcloud <http://whamcloud.com>
3. GPFS <http://www-03.ibm.com/systems/software/gpfs/>
4. Ceph File System <http://ceph.newdream.net>
5. PNFS <http://www.pnfs.com>

**U.S. Department of Energy Best Practices Workshop on  
File Systems & Archives  
San Francisco, CA  
September 26-27, 2011  
Position Paper  
[doebpw5@nersc.gov](mailto:doebpw5@nersc.gov).**

**Yutaka Ishikawa**  
University of Tokyo  
[ishikawa@is.s.u-tokyo.ac.jp](mailto:ishikawa@is.s.u-tokyo.ac.jp)

**ABSTRACT**

Two usability issues in storage systems connected with supercomputer are described. One is metadata access and the other one comes from open source codes handling file I/O. In the latter one, because the users do not want to improve such codes, they request us to install faster disks such as SSD. According to our experiment, after improving the code, the performance is twice faster. Though twice faster disk was used, the performance was only 10 to 20 % gain. Another topic is related to a distributed shared file system being designed and deployed in Japan.

**INTRODUCTION**

Information Technology Center at University of Tokyo provides two supercomputer resources, SR11000 and HA8000 cluster, for domestic academic users. SR11000 is six years old machine and will be replaced with this October. HA8000 cluster consists of 952 nodes each of which has two AMD Opteron 8356s (16 cores). Each supercomputer connects with the proprietary parallel file system called HSFS. The total storage size is 1.5 PB.

In addition of computational resource services, we are currently designing and deploying distributed shared storage system, whose total size will be more than 100 PB, accessed by Japanese supercomputers including K computer, as the nation-wide high performance computing infrastructure. This infrastructure is called HPCI (High Performance Computing Infrastructure) supported by Ministry of Education, Culture,

Sports, Science, and Technology. As shown in Figure 1, there are two storage HUBs, West and East. In West HUB, 10 PB storage and 60 PB tape archive will be deployed with K computer. 12 PB storage and 20 PB storage will be deployed in our university. The system is currently under construction and it will be operated from fall in 2012.

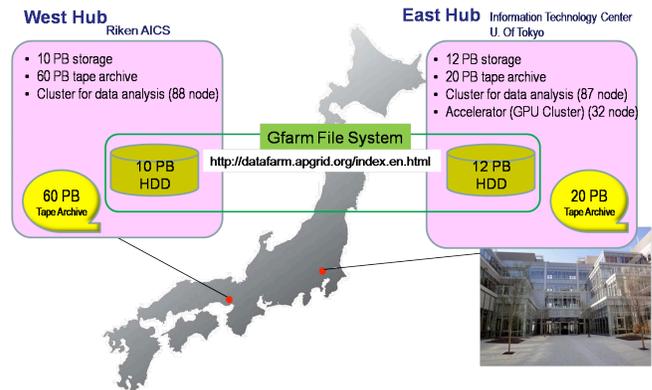


Figure 1 HPCI Storage

**Usability Issues in HA8000 cluster**

As many others pointed out, the interactive users complained slow meta data accesses in the HSFS parallel file system especially “ls -l” command that involves many meta data accesses to obtain all file statuses. To provide faster response for the interactive users, the vendor has modified its system to handle the interactive users’ requests first. This modification with other several changes mitigates this slowness.

Another issue comes from open source codes that are not well programmed for file I/O. The users

believe that low performance of such a code results in slow file I/O access, but this is sometimes not true. For example, a bio informatics tool, used by our bio informatics users, consists of two processing modules, genome alignment and data format change. In the genome alignment processing, there are so many critical regions, and the low performance results in those regions. After reducing the number of critical regions, the program is twice faster. In data format change processing, it opens and reads the same file about 1000 times and eventually reads 1 TB data in total. To eliminate this silly code, the program is twice faster. The users have thought if the file system is twice faster, the program would run twice faster. But, though the file system is twice faster using SDD, the performance is only 10 to 20 % improvement.

### **Usability Issues in HPCI storage system**

This workshop may not consider distributed shared storage systems shared by different organizations. But we would like to address this kind of storage systems because data-sharing is important in the data-intensive science. One good example is ILDG (International Lattice Data Grid) where lattice QCD (Quantum chromodynamics) data generated by supercomputers are shared by international organizations. Another example is the climate

simulation field. As far as we understand, the research group develops their simulation code, obtains data generated by the simulator, and after the generated data are examined and new information for them is obtained, the data are open to others who are interested in data for other purposes. Thus, the data is eventually shared by others.

The Japan HPCI tends to provide storage resource for not only traditional computational sciences but also data-intensive sciences including life science/drug manufacture, new material/energy creation, global change prediction for disaster prevention/mitigation, manufacturing technology, the origin of matters and the universe. To provide better usability for those users, issues are listed below. All issues arise because many research fields use the storage system, it is not yet predicted that their peak and sustained demands.

- ✓ Prediction of storage capacity
- ✓ Prediction of amount of file transfers

### **CONCLUSIONS**

Dusty code must be replaced with modern code to provide usability for such application users. We have to much pay attention of distributed shared file systems with local file systems.

# Appendix B: Workshop Agenda

**5th Best Practices Workshop for High-Performance Center Managers  
San Francisco, CA  
Marriott Marquis Hotel  
September 26-27, 2011**

**Monday, September 26**

- 7:30-8:30 Breakfast and registration
- 8:30-8:45 Welcome (Club Room)  
*Jason Hick, Lawrence Berkeley National Laboratory, and Yukiko Sekine, U.S. Department of Energy, SC*
- 8:45-9:00 Thuc Hoang, U.S. Department of Energy, NNSA
- 9:00-9:15 Yukiko Sekine, U.S. Department of Energy, SC
- 9:15-9:30 Instructions for breakout sessions
- 9:30-10:00 Breakout Sessions
- Business of Storage Systems (Club Room)
  - Administration of Storage Systems (Foothill B)
  - Reliability and Availability of Storage Systems (Foothill D)
  - Usability of Storage Systems (Foothill E)
- 10:00-10:30 Morning break
- 10:30-12:00 Breakout Sessions Continued
- 12:00-1:00 Lunch (Foothill G)
- 1:00-3:00 Breakout sessions continue
- 3:00-3:30 Afternoon break
- 3:30-4:00 Breakout sessions continue
- 4:00-4:15 Business Breakout: collection of thoughts/outbrief to entire group
- 4:15-4:30 Administration Breakout: collection of thoughts/outbrief to entire group
- 4:30-4:45 Reliability Breakout: collection of thoughts/outbrief to entire group
- 4:45-5:00 Usability Breakout: collection of thoughts/outbrief to entire group

**Tuesday, September 27**

- 7:30-8:30 Breakfast
- 8:30-9:00 Checkpoint and directions to breakout leaders
- 9:00-10:00 Breakouts continue
- 10:00-10:30 Morning break
- 10:30-12:00 Breakouts continue
- 12:00-1:00 Lunch
- 1:00-2:00 Breakouts continue
- 2:00-2:30 Business Breakout: collection of thoughts/outbrief to entire group
- 2:30-3:00 Administration Breakout: collection of thoughts/outbrief to entire group
- 3:00-3:30 Afternoon break
- 3:30-4:00 Reliability Breakout: collection of thoughts/outbrief to entire group
- 4:00-4:30 Usability Breakout: collection of thoughts/outbrief to entire group
- 4:30-5:30 Plenary workshop summary and next steps (report)

## Breakout Sessions

### The Business of Storage Systems (Club Room)

*Sarp Oral, Oak Ridge National Laboratory and  
David Cowley, Pacific Northwest National Laboratory*

After reviewing position papers, here are the topics of discussion towards identifying best practices:

- Combining in-house expertise with open source and proprietary solutions and managing the vendor relationship
- Using COTS in HPC storage
- Planning and implementing center-wide file systems
- Establishing I/O requirements
- Dealing with exponential data growth
- Evaluating and integrating new technologies
- Making effective use of storage hierarchies

### The Administration of Storage Systems (Foothill B)

*Susan Coghlan, Argonne National Laboratory and  
Jerry Shoopman, Lawrence Livermore National Laboratory*

#### Plan

We have selected 5 position papers and a single topic from each paper. We are asking the authors to prepare 2-3 slides on that topic from the paper. During the breakout session, for each paper, an author will present their slides and the group will discuss. We chose to do it this way because most of the papers had a lot of topics and we didn't feel there would be time for each author to present all of them.

#### Administration Breakout Agenda

1. Review plan for the day, scope of discussion, get feedback for modifications, finalize plan (15 mins)
2. Position paper presentations and discussion (1.5 hrs - {10 mins presentation, 10 mins discussion} x 5 position papers)
3. Free discussion (30 mins)
4. Finish preparing report (30 mins)

#### Topics/Papers

1. Tiered solutions (Performance improvements) - Torres/Scott paper [LANL]
2. Data integrity/Availability - Heer paper [LLNL]
3. Disk quotas (managing growth) - Cardo paper [NERSC/LBNL]
4. Performance over time - Harms paper [ALCF/Argonne]
5. Configuration management and change control - Hill/Thach paper [OLCF/ORNL]

## **The Reliability and Availability of Storage Systems (Foothill D)**

*Mark Gary, Lawrence Livermore National Laboratory and  
Jim Rogers, Oak Ridge National Laboratory*

- Resilient and fault tolerance - RAID, backup, multi-path
- Data integrity - checksums
- Daily operations - software maintenance
- Off-hour support and availability
- Monitoring and tools
- Other - leveraging procurements, contractual reliability commitments

## **The Usability of Storage Systems (Foothill E)**

*Shane Canon, Lawrence Berkeley National Laboratory and  
John Noe, Sandia National Laboratories*

Scope of the session and boundaries

- Between usability and administration
- Between usability and data management and data formats
- Position paper authors present (slides or talk) also a few others with contributions. (1st hour)
- Free form discussion after lunch to pull forth ideas
- Last half hour to firm

## Appendix C: Workshop Attendees

John Bent, Los Alamos National Laboratory  
Ryan Braby, National Institute for Computational Sciences  
Jeff Broughton, National Energy Research Scientific Computer Center  
Shane Canon, National Energy Research Scientific Computing Center  
Nicholas Cardo, Lawrence Berkeley National Laboratory  
Geoff Cleary, Lawrence Livermore National Laboratory  
Susan Coghlan, Argonne National Laboratory  
Roger Combs, HPC Navy Program  
David Cowley, Pacific Northwest National Laboratory  
Kim Cupps, Lawrence Livermore National Laboratory  
Philippe Deniel, Center for Atomic Energy, Military Applications Division  
Mike Farias, Sabre Systems  
Evan Felix, Pacific Northwest National Laboratory  
Naoyuki Fujita, Japan Aerospace Exploration Agency  
Mark Gary, Lawrence Livermore National Laboratory  
John Gebhardt, AFRL DSRC  
Pam Gillman, National Center for Atmospheric Research  
Kevin Glass, Pacific Northwest National Laboratory  
Stefano Gorini, Swiss National Supercomputing Centre  
Stephan Graf, Forschungszentrum Jülich GmbH/Jülich Supercomputing Centre  
Gary Grider, Los Alamos National Laboratory POC  
Kevin Harms, Argonne National Laboratory  
Richard Hedges, Lawrence Livermore National Laboratory  
Todd Heer, Lawrence Livermore National Laboratory  
Cray Henry, High Performance Computing Modernization Program  
Jason Hick, National Energy Research Scientific Computer Center  
Jason Hill, Oak Ridge National Laboratory  
Thuc Hoang, U.S. Department of Energy, National Nuclear Security Agency  
John Hules, Lawrence Berkeley National Laboratory  
Wayne Hurlbert, National Energy Research Scientific Computer Center  
Yutaka Ishikawa, The University of Tokyo  
M'hamed Jebbanema, Los Alamos National Laboratory  
Thomas Kendall, U.S. Army Research Laboratory  
Dries Kimpe, Argonne National Laboratory  
Ruth Klundt, Sandia National Laboratory  
Rei Lee, Lawrence Berkeley National Laboratory  
John Noe, Sandia National Laboratory  
Lucy Nowell, U.S. Department of Energy, Advanced Scientific Computing Research  
Jack O'Connell, Argonne National Laboratory  
Sarp Oral, Oak Ridge National Laboratory  
Alex Parga, National Center for Supercomputing Applications  
Norbert Podhorszki, Oak Ridge National Laboratory  
Mark Roberts, Atomic Weapons Establishment  
James Rogers, Oak Ridge National Laboratory  
Mark Roschke, Los Alamos National Laboratory  
Tim Scott, Northrop Grumman

Yukiko Sekine, U.S. Department of Energy, Office of Science  
Jerry Shoopman, Lawrence Livermore National Laboratory  
Mike Showerman, National Center for Supercomputing Applications, Innovative Systems Lab  
Kazuhiro Someya, Japan Aerospace Exploration Agency  
Bert Still, Lawrence Livermore National Laboratory  
Osamu Tatebe, The University of Tokyo  
Kevin Thach, Oak Ridge National Laboratory  
Erich Thanhardt, National Center for Atmospheric Research  
William Thigpen, National Aeronautic and Space Administration, Ames  
Aaron Torres, Los Alamos National Laboratory  
Dominik Ulmer, Swiss National Supercomputing Centre  
Andrew Uselton, Lawrence Berkeley National Laboratory  
Dick Watson, Lawrence Livermore National Laboratory  
Charlie Whitehead, Massachusetts Institute of Technology, Lincoln Lab  
Yushu Yao, Lawrence Berkeley National Laboratory