The Century Archive Project "CAP"

Technology-Independent Information Storage

Steven H. McCown & Michael Leonhardt

Storage Technology Corporation

4 April 2002

What is a "Document"?

- A document is:
  - Letter, check, picture, plot, report, birth certificate, deed...
- A document is NOT:
  - Database element
  - Encoded record
  - Encoded object
- Perhaps:
  - ASCII record of transaction
  - Image of database table
  - etc.

Documents in a *"Paperless"* Environment

- 4.4 M Tons of Paper Printed in 1995 … to 5.9 M in 2000
- 790 B Sheets Laser Printers in 1996 … to 1.2 T sheets in 2001
- 810 B Sheets From Office Copiers in 1996...to 1.1 T Sheets in 2001
- 21 Billion Letters Sent
- 170 Billion Pages of Fiche
- 60 Billion Checks Processed Each Year
- E-Mail has created 40% more (personal) printing
- +$100 M in corporate revenues adds 8.8 M sheets printed

McCown & Leonhardt – 4/4/2002

"To ensure that the media will be readable far into the future, it may be necessary to archive the system along with the media. For a 100-year life, recording systems and sufficient spare parts will need to be archived along with the data storage media. Media with life expectancies greater than 20 years are capable of out-surviving existing recording system technologies."

-- John Van Bogart, NARA 11th Annual Preservation Conference, "*Magnetic Tape Storage*", 1996

Information Management

- Long-term storage
  - Defined: in excess of 100 years
  - Inherent to many domains such as genealogy
- Information Management strategies
  - Usually based on frequent data migration
  - Poor incorporation of long-term storage
- Problem:
  - How to access today's archives in 100 years or more

Long-Term Storage Wish List

- Easy integration with data processing environments
- Easy data access
- Migration free
- Long-life media – “no maintenance”
- Reader technology independence
- Human readable data
- Low cost

Current Options

- Encode the data and record digitally
  - Magnetic media
  - Optical media
- Store unencoded, human readable images
  - Microfilm
- Something new - “CAP”

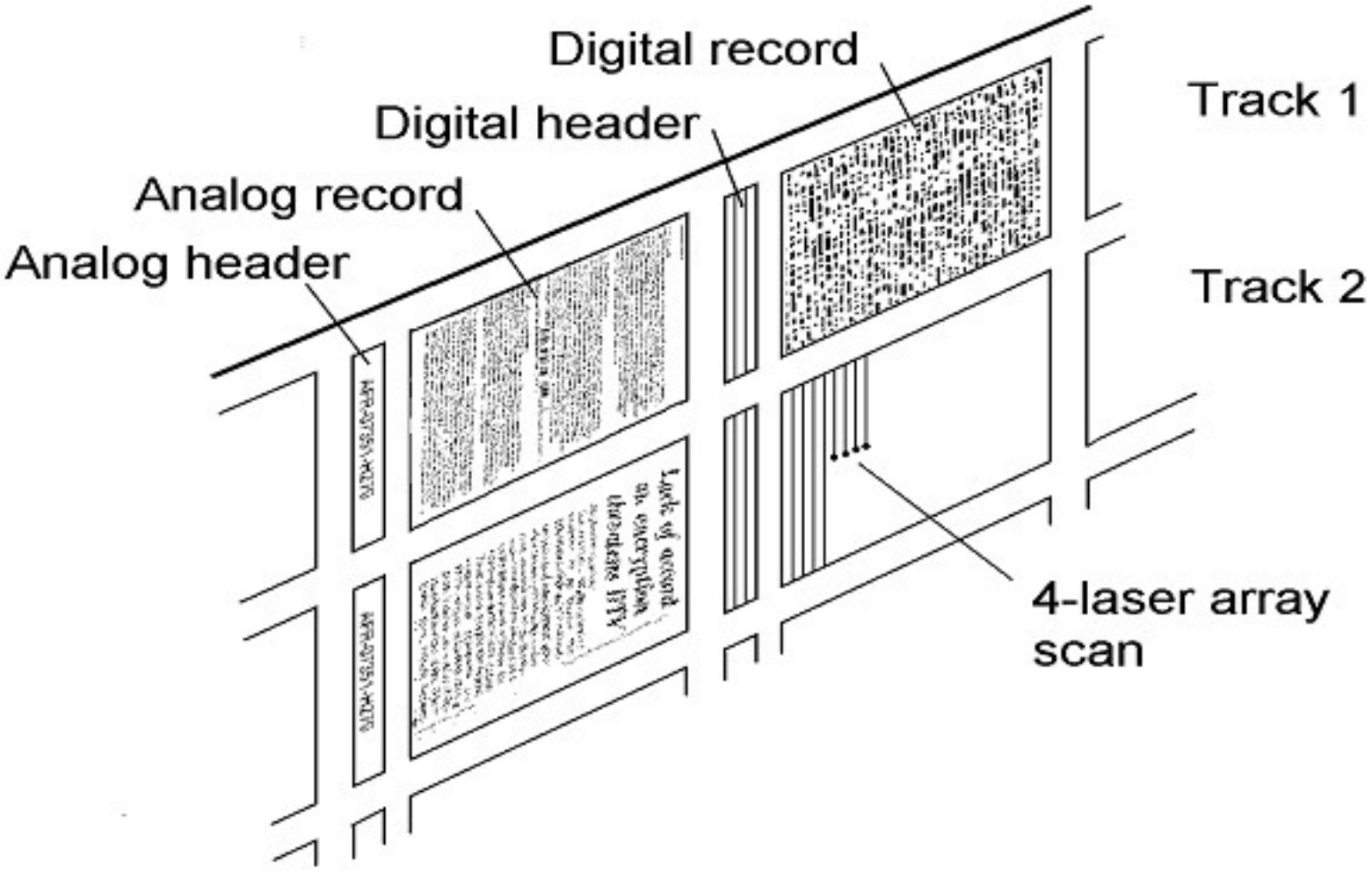
Century Archive Project

- Features
  - High density storage of human-readable document images
  - Storage of digitally encoded documents
  - Metadata ascription to aid retrieval
  - Industry standard physical media form factor
  - Patent on concepts and format filed

CAP Operations

- Scan document to create electronic (e.g. TIFF) file
- Write de-magnified image on optical tape with scanning laser
  - Use WORM optical media
  - Create analog record (human readable)
  - Append new documents as needed
- Write digitally encoded document file
- Read
  - View magnified image
    - **Direct or CCD camera & monitor**
  - Recover digital file

Tape Record Layout



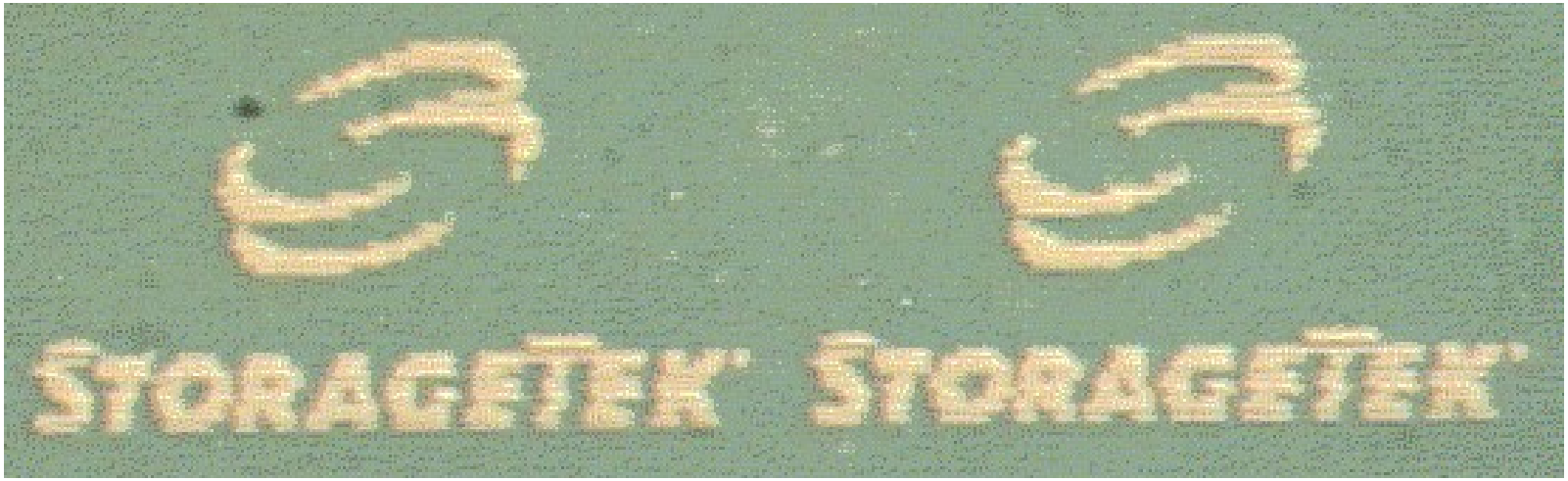
Features

- Record header with document index and metadata
- Updatable Table of Contents
- Digital record in addition to analog record
  - Retrieve digital version if compatible reader available
  - Include digital header and TOC
- Gray scale Documents
  - Use half-toning technique
- Color Documents
  - Store separate images for red, green and blue breakdown
  - Requires three-beam optics for direct color viewing
- Stereoscopic images

Adjacent Digital File

- TIFF file is reformatted with ECC and bit encoding
  - Image file compressed using lossless compression
- Digital record format on tape:
  - Width same as analog record
  - Track spacing doubled to reduce crosstalk on read-out
  - Length 1.5x to 2x analog record

Retrieved Image



Tape Format Example

- For 8 1/2” wide documents
  - Scan images at 300 dpi
  - Write images on tape at 25,400 dpi
    - **85x reduction in size**
  - Document length parallel to tape
    - **Accommodates different lengths**
  - 5 document tracks across tape
- 1/2” tape in 3480 cartridge
  - 200m of tape
  - 220,000 image documents
  - 80,000 documents with both image and digital records

Storage Costs ($/MB)

- Manually intensive
  - Paper - $10.00
  - Microfiche (volumetric improvement only) - $1.20
- "Semi-automated" (manually mounted media)
  - Non-automated magnetic tape
  - Microfilm - $.005 (media only, 16 mm)
- Full automation
  - Magnetic tape - $.004
  - Optical disk - $.03
  - CAP - $.002

Century Archive Summary

- Provides alternative storage method for valued documents, images
  - Direct optical viewing
  - Eliminates drive, media technology migration
  - Robust media options for relaxed environmental storage conditions
- Provides digital storage
  - Faster availability
  - Data integration
- Complements magnetic tape storage of bulk records

Questions?