

A large, abstract, light gray graphic on the left side of the page, consisting of overlapping curved shapes and a thin white line that curves from the top left towards the bottom right.

ARCHIVING MOVIES IN A **DIGITAL WORLD**

White Paper
January 10, 2007

Dave Cavena
Chris Wood
Jeff Bonwick
Guy Steele
Jay Seaman
Michael Selway

Table of Contents

Executive Overview	3
Introduction	3
Archiving Films	4
Capturing and Archiving the Image	5
A Changing World	6
Capturing, Digitizing, and Storing the Image	6
The Heart of the Debate – Ensuring the Bits are Still the Bits	7
Ensuring Data Integrity Algorithmically	7
Bit Error Detection	7
Bit Error Correction	8
The Time Factor	8
The Software to Archive the Content	9
Other Strategies	9
Weaknesses	10
Retrieval from Archive	10
Reuse of Image	11
Costs	12
Can Digital Compete with Celluloid?	12
What About Celluloid?	12
Conclusion	13
Notes	15
About the Authors	16
Acknowledgements	17
Copyrights, Trademarks, Etc.	17
Appendix A: Assumptions and Methodology	18
Appendix B: Analysis of 100-yr Archiving Costs	19

Executive Overview

Filmed entertainment, the most visible of the Intellectual Property industries, quickly is moving to digital files from analog film for capturing and displaying images. When completed, this move will catch-up to the decade-old move to digital for intermediates and editing.

Archiving of this invaluable, irreplaceable content, however, still is done by outputting the images to film and storing the film in a vault under controlled environmental conditions, a process little changed by technology over a century and unyielding to current repurposing business models and market pressures.

Movie archiving, however, is coming under increased scrutiny as analog film quickly is being replaced by digital bits in all phases of the imaging process, from capture to post-production, to projection in the theater, to distribution to the consumer over the internet, in packaged media or via digital television, whether broadcast, satellite or cable.

Archiving for long-term preservation remains the only part of the workflow still reliant on film.

Using current technology and defined processes, recorded images and sound now can be archived digitally at a cost measurably lower than that of archiving film. With a faster recall from archive this more responsive archive infrastructure can be used to create a more viable on-going market for content repurposing and re-monetization.

In this paper a viable, scalable, cost-effective method is presented for using digital technology to archive filmed entertainment for the century-long duration now expected from, and achieved with film. This method is based on Commercial Off-The-Shelf (COTS) technology; no special application software or systems are required.

Acceptance of this paradigm shift will not come overnight. The hope in presenting this paper is to initiate serious discussion, examine and improve the model, and assist content creators in creating further revenue opportunities for the ever-more expensive content they create.

Introduction

Feature films, television broadcasts and other image and audio content are archived for two reasons.*

Firstly, the content owner has a fiduciary responsibility to retain the assets with which he or she has been entrusted, making additional profit from them, if possible.

Secondly (though the primacy changes with the role of the person answering the question), films are archived because they are the stories of our culture, our civilization, our world. They are the recorded traditions of an age told by many of the best artists of their time or any other. They should not be lost to posterity.

With the average cost of a feature film original camera negative approaching \$100M, the desire increases to repurpose that content over longer time periods and larger markets. The current archive model tied to analog film does not readily support the efficient repurposing of content more than a few years from its release date.

* In this paper the authors do not deal with archives of television content: sports, news, etc. The issues and costs are similar. The benefits of a digital content archive increase quickly by adding this content to it.

Celluloid archives of these irreplaceable cultural artifacts and revenue vehicles do not enable efficient re-use of these stories, nor do they allow for an easy access to our past, whether the motive is profit or understanding. They should do both.

A digitally-formatted archive can allow that access, and can do so by providing pristine copies of those stories efficiently in time, space and cost. A digitally-retained copy provides future generations with the view into our collective past that is required for them to be informed and understanding peoples, and provides content owners with long-term returns on the increasingly expensive stories we have to tell.

The content stored in the vaults of the movie industry is valuable but very difficult to access.

How valuable is Hollywood's content?

“The copyright industries—which make movies, TV programs, home video, books, music and computer software—comprise an awesome engine of growth that nourishes the national economy. Not only is intellectual property America's largest trade export, and arguably its most influential, the industry sector is creating jobs three times faster than the rest of the economy. This sector generates more than 5 percent of U.S. gross domestic product.”⁽¹⁾

In 2005 the domestic receipts for MPAA films were US \$8.99B; the global receipts US \$23.24B.⁽²⁾

That's how valuable.

This content retains value over time. In order to capture this value, however, the content must be archived⁽³⁾ in such a manner that when recalled, it is usable for its intended purpose of capturing new revenue or telling a story.

It is the opinion of the authors, and of experts in the described technology spheres, that digital archiving of filmed entertainment has reached the maturity necessary to be put into general use.

Archiving Films

Filmed entertainment currently is archived, interestingly enough, on film. The standardization of 35mm film is such that any theatrical film recorded in 1907, properly archived, can be retrieved from that archive and played in any theater projecting feature films in the world today in 2007.

As movies increasingly are captured digitally and projected digitally, some major questions arise: Will it be possible in 2107 to retrieve from a film-based archive any movie placed there in 2007, and play it out in any theater in the world? If the movie was archived as celluloid, will there be any projectors around on which to project it or scanners to scan it? Projectors and scanners can be rebuilt, they're just electromechanical devices, but is this the desired future of this very valuable industry?

Or should the current—and future—digital world be embraced fully by the film industry now that the technology exists to do so?

A digital archive will ensure the ability in 2107—or in 2207 and beyond—to see the stories placed there in 2007, and to profit from them.

Capturing and Archiving the Image

Historically movies have been shot on film. The Original Camera Negative (OCN) is shot and developed, an InterPositive (IP) created, an InterNegative (IN), and finally a Release Print. Whether archiving the IP or the Release Print, the version of the film archived is at least one image-degrading generation, and perhaps more, away from what the Cinematographer and Director captured, away from the story they wanted to tell. As that celluloid copy sits in an archive, it continues to degrade even under the most stringent conditions, removing the presentation of the story still farther from the intent of the artists.

In a digital archive the images shot and projected, the exact story the Director and Cinematographer wanted to tell, can be stored perfectly in a digital format for presentation to future generations.

The properties of polyester or celluloid film stock are well-understood by Archivists trained in the current processes. When a new film stock comes along, it is tested and, over time, may be accepted as an archive medium for use with processes and environmental conditions similar to the well-understood ones of today. There have been glitches using this method, however, as testing occasionally has been found not to reflect adequately the real-world experience of new films stocks over time. The industry thus has become wary of new archive technologies.

When a movie today is shot, processed and projected digitally, but archived on film, it must go through a film recorder⁽⁴⁾ / film-out stage to create the archive copy. This film-out step may cost as much as \$50,000 - \$70,000 per film, and creates something both different from the Director's intent and that may never be used again — an analog version of the digital movie.

The films are archived in cans in controlled temperature and humidity environments, often deep underground in salt mines. It is this well-understood archival process that results in the ability to recall films properly archived in 1907 and project them today.

When a film is retrieved from archive, it must be located, pulled from the archive inventory, brought back up to standard temperature and humidity conditions from the conditions in which it has been stored, and then packaged, insured and shipped to the requesting party. This process often requires as many as 60 days and can cost as much as \$10,000.

When the film has been delivered to the requestor a visual Quality Control (QC) is performed to assess the status of the images and whether restoration is required prior to the planned re-use. Once this decision is made, the film is scanned and digitized, as the restoration (if required) and intermediate work will be performed digitally. The reuse itself is increasingly likely to be digital, whether the reuse is to deliver the movie, or a new cut of the movie, on packaged media (DVD, HD-DVD, Blu-ray Disk, etc.), to digital television/satellite/cable or, increasingly, to deliver it as a DCI⁽⁵⁾-compliant package to a digital exhibition location. Re-digitizing can cost \$50,000 or more.

Once the use of the retrieved asset has been completed, it is packaged, insured, shipped to and returned into the archive at the proper temperature and humidity conditions. If the movie has undergone digital restoration, again it will need to be recorded to film for replacement in the film archive.

As can be seen, several opportunities exist for use and preservation problems along the way. The film may have been misplaced in the archive, requiring extra time to find it, slowing time-to-market for the new project. Time is required to return the film to standard temperature and humidity conditions, slowing time-to-market. A new film stock may have degraded faster than modeled. The quality of the film as observed in QC may preclude entirely the revenue opportunity for which the film was retrieved, or diminish the return on the project due to necessary restoration costs and time.

A Changing World

While the ability to keep digital bits pristine over time has been the subject of much discussion for decades, and digital technology and processes often are assumed not to have kept pace with the need to archive irreplaceable content, it is time to revisit those assumptions.

The assumed but unknown qualities of each movie in current film archives, and the time and monetary cost of recalling a film from its archive, are such that one wonders at the business case for preserving these assets for commercial reuse.

The archive business case is made viable through the implementation of digital archiving.

With the maturation of the digital age, and the growth of the internet and consumer devices on which to watch content, movie archives are natural places to go to create new profit opportunities. As shown above, however, the cost in time and money for retrieving an asset from analog archives is immense, and the quality of the retrieved asset questionable absent the further, and unknowable, cost of restoring the recalled asset.

To capture revenue efficiently from previously-recorded content requires a workflow amenable to future repurposing of that content. A digital archive can provide this at a cost lower than that of a conventional film-based archive, and do so more quickly. This archive can provide quick and easy access to new revenue flows, in a time-to-market model that works in our fast-paced world of ever-new-devices on which consumers want to consume content — and on which content owners want to deliver and monetize their assets.

Capturing, Digitizing, and Storing the Image

Although film capture still reigns in Hollywood, this is changing. As with the adoption of new technology in many industries, the changeover is moving in stutter-steps. Magnified by the influence of personalities in Hollywood as the opinions of the quality of the new systems are debated, the adoption curve is not smooth, but it is moving upwards at an accelerating rate.

Several factors are driving this acceleration. The advent of new digital cameras suitable for the capture of feature films is causing more films to be shot digitally. New generations of cinematographers and directors graduating from film schools are eager to use these new technologies. Shooting digitally can cost less. All intermediate work is done digitally. The newly-agreed-upon Digital Cinema specification is resulting in a quick up-tick in digital projection at theaters.

All of these mean that at some not-too-distant date the first and only time a movie will exist on film will be for archiving. This means, of course, that the archive print no longer will be the image and medium that the Cinematographer and Director captured and through which they told their story.

Once shot and processed (meaning as the debate progresses and film still is shot — a process that likely will continue among some artists for some time) film is digitized via a film scanner, generically a telecine⁽⁶⁾, and turned into bits stored as files on magnetic storage disks and/or computer (as opposed to video) tape cartridges. (In a digital archive world, bits are archived most cost-effectively on magnetic tape. Tape is an inexpensive data storage medium, requiring neither electricity nor cooling when not in use, and lasting for a period of years when handled correctly). This intermediate digital representation of the image is called a “Digital Intermediate,”⁽⁷⁾ or “DI”. The DI has become the standard input to all systems used in the

post-production process: editing, sound, CGI, conforming, etc. Once the movie has been completed and is ready for theatrical release, it is sent to a film recorder and output to film by reversing the scanning process and turning bits back into film.

It is film produced by this recorder process that is developed and sent to an archive.

The Heart of the Debate – Ensuring the Bits are Still the Bits

Hollywood is accustomed to an archive duration for their irreplaceable content in excess of 100 years, far longer than archives required by federal and state authorities for various other purposes, such as tax audits. One can re-build bank records from general ledgers, retype or re-scan printed material from paper (which has millennia-long archival periods when properly stored). One cannot, however, re-record Louis Armstrong or Buddy Holly, or re-shoot Henry Fonda, Jimmy Stewart, Audrey Hepburn or Greta Garbo.

Ensuring Data Integrity Algorithmically

Fortunately, algorithms and procedures exist to ensure bits stored digitally remain the same bits over extended periods of time.

When writing bits to a computer tape, tape drive hardware uses an Error Correcting Code, or ECC⁽⁸⁾, scheme to encode the data so that if fingerprints get on the tape or the emulsion flakes off, or even if holes are punched in it, the drive will detect this and reconstruct the data from redundant bits recorded elsewhere on the tape.

The Sun StorageTek 10000⁽⁹⁾ 500GB data cartridge, for example, has an actual data capacity, including overhead, of approximately 843GB. Of this total capacity, 59% (500GB) is provided for user data space, 18% (148GB) is reserved for servo track data, and 23% (195GB) for ECC information, allowing lost data to be rebuilt.⁽¹⁰⁾ Tape cartridges from other technology vendors are similarly formatted and recorded.

When the tape drive actually writes the data to tape, it reads back that data immediately to ensure that what it was told to write is what it really did write. The drive software writing the file uses the ECC to accomplish this.

To have recourse in the event that data on a particular tape cannot be read in the future, it should be ensured that two copies of each Digital Intermediate are made and stored in a tape library system. For further reliance, another library archive should be maintained in a different geographical location (for Hollywood, preferably on the other side of the San Andreas Fault), and that another two copies of each DI are kept in that remote library. Four copies of each DI thus are archived.

To ensure that the movies in the archive are as close to “perfect” as possible, two things must be accomplished with the tapes: Bit failures must be detected, and then they must be corrected.

Bit Error Detection

Sun’s 9940 tape drive, has an undetected bit error rate of 10^{-33} ⁽¹¹⁾. In our archive model we have four copies of each archived movie. The 10^{-33} BER is for one file. These nines are additive, however, so this ECC alone provides 10^{-128} ($10^{-32 \times 4}$) protection against undetected errors. This is 128 nines of protection against undetected errors across the four copies of the archived movie.

If desired and used, a 256-bit Secure Hashing Algorithm, such as SHA-256⁽¹²⁾, can provide additional error detection to this archive. The probability that a SHA-256 checksum will fail to detect an error is 2^{-256} . The usual caveats about birthday collisions don't apply because the implementation is not one requiring defense against traffic analysis; SHA-256 rather is being used as a good pseudo-random hash function to ensure data integrity. And it's very good indeed: 2^{-256} is roughly 10^{-77} , i.e. SHA-256 is good to 77 nines.

This assurance is such not just that the movie will be of "good visual quality", but that it will be perfect.

(Note that the main limitation here is that only one checksum is maintained for the entire movie. If SHA's 10^{-77} isn't good enough, an even better job could be done by breaking the movie into parts and keeping a checksum of each part, and a checksum of the checksums. This is, incidentally, exactly what Sun's new file system ZFS⁽¹³⁾ does for disk storage.)

Bit Error Correction

To solve the digital archiving problem, however, it's not enough to detect errors—it must be possible to correct them as well. This is quite a bit more challenging.

The ECC scheme for Sun's T10000 Tape has an uncorrected bit error rate drive of 10^{-19} ⁽¹⁴⁾. (IBM and Hewlett Packard's Ultrium LTO-3 drives have a corresponding uncorrected BER of 10^{-17} .)

This means that a 1 in 10^{19} chance exists of getting an error that the tape drive's ECC cannot repair per movie archived. As above, the nines are additive. With four copies of the DI, the chance of an uncorrectable bit error across all four copies in the archive is 10^{-76} (10^{-19} to the 4th power), which seems quite reasonable—that's 76 nines. Given that a DI is 8×10^{13} bits (10^{13} Bytes), that is one uncorrectable bit error per 10^{63} movies.

The Time Factor

Like film, data tape tends to degrade over time, and a century is a long time. Because of the irrepressible dynamic interactions of electromagnetic signals and the physical tape medium, degradation of the on-tape signal (bits) occurs somewhat more quickly than images on film. A tape cartridge in and of itself is not conducive to a century archive.

Additionally, the form factors that digital media take evolve at a rate requiring one to move digital files "forward" to the new media and form factor, a need that must be met to ensure long-term ability to have equipment capable of reading the media. These physical problems and technological realities create the need to re-write each digitally-stored DI on a periodic basis.

The current top-of-the-line tape cartridges, for example Sun's T10000⁽¹⁵⁾, and IBM's 3590 Extended High-Performance Cartridge Tape⁽¹⁶⁾, have published archive lives of 30 years. Additionally, the National Media Lab uses a duration of 30 years as the usable life of magnetic tape⁽¹⁷⁾.

To be very conservative, in our model we propose to rewrite each DI on new tape cartridges once every 10 years and whenever cartridge density changes significantly. (We propose a full tape audit every six months, however, to validate data integrity.)

By rewriting the DI periodically, we are accomplishing two major tasks: we are ensuring the currency of the tape cartridge and drive (i.e. we never will have a 100-yr-old cartridge and go looking for a machine in which to read it, nor a 100-yr-old tape drive looking for a library) and, as the tape software changes and upgrades over time, we are keeping current with

that software as well (ensuring we never have a 100-yr-old tape file format without any tape software to read it). (The application software that determines the file format is a different issue and is dealt with below.)

With a 10-year rewrite period, each recorded DI will need to be rewritten ten times during the course of a century archive. What will be the Bit Error Rate, BER, at the end of ten generations of rewrite, and will that result in a usable movie capable of meeting a revenue-creation opportunity?

Here's the interesting part:

For this application, it doesn't matter how many times the data is accessed.

Here's why:

The probability that the ECC's undetected BER will fail to detect damage during any given access is 10^{-19} . The probability that it will fail one more times during N accesses is 1 minus the probability that it will succeed N times in a row, i.e. $1 - (1 - 10^{-19})^N$. For N less than 10^{19} , this is well approximated by $N \times 10^{-19}$.

(For the sake of an exaggerated example, let's say a movie is accessed a million times. Assuming the interpretation of "undetected bit error" as above, for a 10TB movie the chance of misreading the movie (e.g. having one uncorrectable bit error) on one attempt is 10^{-19} , so the chance of misreading on any of the one million accesses is $10^6 \times 10^{-19} = 10^{-13}$. If we assume 500 major studios producing 20 movies per year, that's a million movies per century to be preserved, and the chances that any will be misread in the course of the century, even under the unrealistic assumption that each is accessed one million times per century (that is, more than once per hour), is 10^7 .)

The point being: it reasonably can be assumed, for the purposes of this application, that the ability to detect errors in transcription is perfect.

The Software to Archive the Content

But the software performing the moves, reads, copies, and migration does matter. Long-term storage of static data is the reason archive management software exists. Sun's Storage Archive Manager File System, SAM-FS⁽¹⁸⁾ provides the capabilities required by an archive of this type: automatic copy creation, open tarball format (to ensure readability even without the archive application), ability to create and manage file subsets, improving read/write throughput, etc. Other technology companies have similar archive software.

Other Strategies

While the above strategies are in use by Sun Microsystems in the ongoing archive preservation project at the Library of Congress (see below), other strategies exist for preserving high-value content over extended periods.

As described in this paper, the redundancy afforded by having four physical tape copies of the data is used only in an all-or-nothing manner: either a copy is good or it isn't. In fact, there are lots of interesting strategies involving majority voting (on each bit, if necessary) among three or four copies that can greatly improve the reliability.

Moreover, there are esoteric mathematical methods that could be applied to further increase reliability⁽¹⁹⁾. One could, for example, construct from a 10TB movie file 40 files of 1TB each with the property that the original file can be reconstructed

from any 10 of the 40 files. This means that one could protect each 1TB file with a 256-bit checksum (or just rely on what the T10000 tape drive can do, e.g. 10^{19}), and be able to reconstruct the original file if at least 10 of the 40 files were uncorrupted. Such strategies are computationally intensive, but in an archive environment, i.e. one not driven by high transaction rates, the computational costs may be acceptable.

Weaknesses

The biggest weaknesses in any archiving scheme are external agents: human error, sabotage, network-based attacks, and natural disasters. Therefore our model suggests the following:

1. Find ways to create institutional memory of the archive's existence and of the need to refresh it periodically. This seems obvious, but history suggests it's the weakest link.
2. Have two or more sites, geographically far apart, maintained by different people. Particularly in the case of Los Angeles, where the major domestic studios are located, having a disaster recovery (duplicate) archive on the east side of the San Andreas Fault, whether in Phoenix or Philadelphia, would make enormous sense. Additionally, a fire localized to a movie studio lot, on which copies of archives traditionally have been kept, would not destroy the entire archive even if it overwhelmed local fire suppression, destroying the local site.
3. Have absolutely no network connectivity. Zero network connectivity would be desirable to preclude, with the exception of the Archivist, any type of electronic tampering or theft, internal or external. Lack of connectivity would be balanced against the costs of physical transfer of these tapes and the security involved therein, but films now are transferred via insured physical transfer. (The intent of proposing a digital archive isn't to break working, extant models. It is rather to enhance those models. In fact, however, while encrypting an archive would not be a wise decision, i.e. the archive files always should be in the clear (losing the key or the algorithm or the encryption application could render the entire archive useless), a tape recalled from archive certainly could be encrypted for transport and then decrypted once it reached its destination. Hardware encryption on the tape drives is a reality now, adding very little performance cost to secure the content during transit.)
4. Each time a new copy of a movie is generated, send the old copy to another location. It costs almost nothing; doing so verifies the movie still is intact; and it may remain intact for quite a bit longer.

Retrieval from Archive

If, based on the above, it is decided that films can be archived digitally, what about the recall and reuse of those assets?

As noted above, recalling a single film archive copy can cost as much as \$10,000, and take as long as 60 days. Recalling a digitally-stored archive copy will cost a small number of hours (which will be decreased as tape density increases, throughput being directly related to data density on the tape), and can be delivered within a day (from the local archive) to a disk drive subsystem available to the person requesting the content.

(The latency involved in local delivery is determined by the throughput of the tape and disk subsystems. With current-year (2007) throughputs of approximately 120 Mbytes per second (uncompressed) on the Sun/StorageTek T10000 tape drive running in a Sun/StorageTek SL8500⁽²⁰⁾ tape library, and using 500-gigabyte cartridges, it may take as long as 1389 minutes, or 23 hours, to read a 10 TB DI file.

(Multi-threading can decrease this time, but not to insignificance. It may be tangential to the main argument of the paper, but decreasing fetch latency by a factor of four or even two, might be of great value to consumers of the archived data. Given the current 500-GB capacity of a data cartridge, and even assuming that cartridges are not shared among movies, if one could read all 20 cartridges for a 10TB movie simultaneously, one could decrease fetch time from a day to an hour. This might make a big difference to a studio consumer. And this concurrent read can be done using SAM-FS⁽²¹⁾).

A copy of the movie from the remote archive never should be requested by a user. The remote archive exists to ensure the currency of the local archive, which should be used to respond to all requests.

In a digital archive, no time is spent on temperature and humidity standardization, finding a single physical asset in a large salt mine in Pennsylvania holding tens of thousands of identical film cans, insuring and shipping it, and getting a usable image to the desk of the requestor.

Reuse of Image

Nor is any time or money spent doing a new scan of the old film, a cost on the order of \$50,000 today. The delivered asset already is in the digital form.

But is it in the digital form required? What about the software used to write-out the DI to begin with? Is that software still on the market? Can it still read the down-level file created ten (or more) years ago? Has the company which marketed the file format application software gone out of business or been acquired and the software no longer available? Is a computer on which it runs still available?

Fortunately each of these can be dealt with within the normal functioning of the technology world.

As a new file format is released by a software vendor, backward compatibility is normal. In the few instances in which this has not been the case resourceful people have written and marketed backward-compatibility tools. Or the software in question has become so standard that similar, competing applications have created import tools. One major studio for a decade has been escrowing the source code for the software used to write their digital movie files as further assurance of availability.

Assuming the availability of the operating system under which this software runs, and a computer on which the operating system can be executed (see below), these and the above file format issues can be dealt with in the above-described generational refresh of the tape drive.

Above we proposed that the archived content is re-written every ten years. Should the future of the legacy application software (or operating system or computer hardware running the operating system running the application) be tenuous at any time, the content can be ingested into a then-current application and system, and re-archived in that new file format. In fact, this process logically would be tracked and planned to ensure that as the industry migrates to new file formats, the appropriate steps are taken to archive the content in that new format. This is feasible because the underlying data format of the film remains digitally fixed; only the storage medium changes.

Contrast this, for example, with the problem of nitrate film stock, which can be viewed as an analogous situation. Nitrate-based⁽²¹⁾ film stock can be viewed as a file format that went out of favor, requiring either a re-copy of that film (with consequent generational image loss), the loss of the film, or a gamble that the film would be good enough if and when required at an indeterminate future time. No "Plan B" exists for that film recorded on nitrate stock that does not result in a degradation or loss of the content.

Costs

If we assume from the above that one concurs that a film reliably can be archived digitally for the required durations, what are the costs? After all, the scenario above posits that many generations of tape cartridges and tape drives will be required, a tape library to hold the drives, and computing power and front-end disk storage to run and manage the whole process. The disk and compute front-ends for the tape systems will reach end-of-life approximately every five years and need to be replaced or upgraded, the tape drives every ten years (or when density changes), and the tape libraries holding those drives and cartridges will require replacement every twenty years or so. What are the impacts of these recurring costs?

Can digital compete with celluloid?

The short answer is that, at list price, an archive as described in this paper provides an archive cost for 2,000 movies (10TB DIs) for 100 years (an annual output of 20 features at a major studio) of approximately \$45,000 per movie. The cost is higher at the beginning of this archive as infrastructure is implemented, and fluctuates with tape density and replacement of infrastructure, but remains below the film threshold of \$100,000 per movie once about 670 movies are contained in the archive, or within approximately thirty-two years of initiation.

Put another way, for 68% of the lifetime of a century-long movie archive, it is paying for itself—at list price.

And, 100% of the time it is providing easy access to valuable content, enabling efficient repurposing to meet new business models and opportunities for that content, quickly, effectively and easily.

In reality, however, street prices are not list prices. For the purposes of Hollywood's archiving, Sun will offer unique licensing and support agreements tailored specifically for long-term entertainment content digital archive requirements. A fixed cost, fixed term "site license" will be offered that is customized to the specific implementation, capacity, low-frequency access requirements and availability requirements of any given customer. Support services will be priced as a fixed uplift to the negotiated site license.

Appendix B is a series of four charts on which are shown the cost curve for the archive and a pie chart of the cost breakdown by expense type (computer, disk, tape, software, etc.). Archive objects of both 10TB and 100TB are shown, as are list and discounted price. Discounted price are based on assumed discounts that may be based on a number of factors and may not be generally applicable across industries.

What about celluloid?

An unaddressed question remains, however: What about celluloid as a backup medium? What will be its costs? Will it even be available?

Underlying any assumption of celluloid archives and their costs is the assumption that film media will remain available over time. Is this a valid assumption, and will film remain available at a cost at which it can compete with digital? Will any labs remain employing people with film competencies, or stocking the environmentally-unfriendly chemicals required in film processing?

These are questions the authors have not seen addressed. It seems reasonable to conclude that as shooting and projection join post-production as digital activities, the price of film and film processing will skyrocket, if they remain available at all.

Kodak, Fuji and the other major film manufacturers have been eliminating jobs and changing business models for years as digital photography has overwhelmed the demand for film. This trend shows little sign of slowing and none of reversal.

Commercial photographers in journalism have completed the switch to digital. Within a decade of its introduction digital photography has displaced film globally with the exception of feature films. The market for film faces a precipitous and perhaps final decline. What remains for feature archiving, if anything, will be priced as is any product with low demand, very high manufacturing and environmental costs, and a limited market.

Regardless of the costs of digital as we look at it competing with film for the movie archive business, it is far more likely that no film will exist with which digital will be forced to compete.

Conclusion

The consumer wants their content anywhere, on any device at any time. Creating a digital archive as described here will enable the content owner to provide the consumer with the content they desire in a format they want. The current analog archive model simply can not meet this need. Indeed it is likely and quite probably that film no longer will exist as a recording or archive medium long before this century is out.

The studios have archived analog content to license, but it is stored nearly inaccessible to the market and consumer. This makes questionable at best the ability of the content owner to realize added revenue on archived content. Continuing with this model may not be in the best interests of content owners.

The consumer wants content. As was noted by several presenters at IBC in 2006, content no longer is "King." The mantra now has become, "The Consumer is King." To meet the growing and varied means in which a consumer will demand content (from in-theater digital prints to VoD to cell phones and likely more ways than are dreamed of today), only a digitally-managed file can be reformatted and reused with mathematical purity and accuracy.

The ease-of-access to digitally-formatted movies for recall and repurposing is far higher than access to celluloid archives. The steps to repurpose digitally-stored content are reduced to one. The time and cost to recall a movie from the archive are far lower. All of these factors result in increased use of and revenue from that archived movie.

Using a digital content archive to make available studio content quickly and efficiently can create an entirely new business model based on the speed of the market adoption and the velocity with which content will flow through this new market to consumers in theaters, in homes and on mobile devices.

The archive can become a part of the revenue flow of a studio, rather than simply a large and ever-increasing cost to maintain.

Technology has matured to the point at which this digital archive is feasible. The Library of Congress has embarked on a similar long-term digital preservation and archive project for all of their media assets: "Storing National Treasures".

<http://www.enterprisestorageforum.com/sans/features/article.php/3586066>

The technology and processes chosen for this LoC project are the same Sun Microsystems technology drawn on for the discussion above: "Sun Rises at the Library of Congress".

<http://www.enterprisestorageforum.com/sans/features/article.php/3619646>

The pivotal and immutable point is that this can be done beginning today, and Sun is the best vendor for the solution. The experience we bring to the project already has been recognized, and is being broadened by, the Library of Congress and other locations undertaking digitizing their media assets around the world using solutions from Sun Microsystems.

The time to begin serious efforts on testing and implementing studio digital archives is now.

Notes

1. "MPAA's VALENTI OFFERS SUPPORT FOR INDUCING INFRINGEMENT OF COPYRIGHTS ACT OF 2004" Statement by Jack Valenti, President and CEO, MPAA, June 23, 2004
2. MPAA Snapshot Report, 2005 International Theatrical Market
3. In this paper the term "Archive" is used distinctively. It does not mean long-term storage alone, but encompasses both technology and processes to ensure that content stored digitally is preserved for generations, and can be recalled efficiently so as to be available to its owners for additional repurposing revenue opportunities far into the future.
4. http://en.wikipedia.org/wiki/Film_recorder
5. http://en.wikipedia.org/wiki/Digital_Cinema_Initiatives
6. <http://en.wikipedia.org/wiki/Telecine>
7. http://en.wikipedia.org/wiki/Digital_intermediate
8. http://en.wikipedia.org/wiki/Error-correcting_code
9. http://www.sun.com/storagetek/tape_storage/tape_drives/t10000/specs.xml
10. See notes, slide 21, of the following presentation:
<http://www.thic.org/pdf/July06/sun.raymond060718.pdf#search=%22ecc%20tracks%20overhead%20storagetek%22>
11. http://www.storagetek.com/products/product_page38.html
12. <http://csrc.nist.gov/publications/fips/fips180-2/fips180-2.pdf#search=%22sha-256%22>
13. <http://www.sun.com/2004-0914/feature/>
14. http://www.sun.com/storagetek/docs/TC0049D_T10Kdrive_DS.pdf
15. http://www.sun.com/storagetek/tape_storage/tape_drives/t10000/index.xml
16. ftp://ftp.software.ibm.com/common/ssi/rep_sp/n/TSD00259USEN/TSD00259USEN.PDF
17. <http://palimpsest.stanford.edu/bytopic/electronic-records/electronic-storage-media/bogart.html>
18. http://www.sun.com/storagetek/management_software/data_management/sam-fs
19. See, for example, M. O. Rabin, "Efficient dispersal of information for security, load balancing, and fault tolerance," *Journal of the ACM*, 36(2):335-348, ACM, April 1989.
20. http://www.sun.com/storagetek/tape_storage/tape_libraries/sl8500/
21. <http://www.loc.gov/preserv/care/film.html>

About the Authors

Dave Cavena is an Engagement Architect with Sun Microsystems, supporting the digital content space for studios and postproduction companies. Prior to coming to Sun he held positions as Project Director, Blu-ray Authoring System for Sony Pictures Entertainment, Program Manager for The Walt Disney Company's MovieBeam, and as IBM's Digital Cinema Executive Project Manager. Dave can be reached at: david.cavena@sun.com

Chris Wood, Director and Chief Technology Officer for Sun's Data Management Storage Practice, Client Services Organization is responsible for identifying and delivering the best solutions available that can address our customer's complex data management problems. He joined Sun Microsystems when his prior company, MaxStrat, was acquired by Sun early in 1999. Mr. Wood has held prior positions at IBM, Litton Industries and other computer-related firms. He can be reached at: chris.wood@sun.com

Jeff Bonwick, DE and CTO, Storage, Sun Microsystem. Jeff can be reached at: jeff.bonwick@sun.com

Guy Steele is a Sun Fellow with Sun Microsystems Laboratories, conducting research in programming languages, algorithms, and processor architecture. He is a well-known author or co-author of books about the programming languages C, Common Lisp, High Performance Fortran, and Java. He is also an ACM Fellow and a member of the US National Academy of Engineering. Prior to coming to Sun he was a Senior Scientist at Thinking Machines Corporation, a pioneering manufacturer of massively parallel supercomputers and of the DataVault, the first commercial RAID disk array. Guy can be reached at: guy.steele@sun.com

Jay Seaman is an Engagement Architect with Sun Microsystems, based on the east coast, responsible for a large corporate account. During his 11 years at Sun, Jay has been responsible for the design and implementation of several high profile projects such as NBCOlympics.com. Jay can be reached at: jay.seaman@sun.com

Mike Selway is a Consulting Systems Engineer with Sun Microsystems supporting the crafting of data management tiered storage solutions around Sun Microsystems' high performance storage management file system, SAM-QFS. He has been a member of several data storage organizations integrating a wide variety of hardware and software technologies for creating film, audio, video, and post-processing market-specific data management solutions. Mike can be reached at: michael.selway@sun.com

Acknowledgements

This paper was a collaborative effort involving many Sun Microsystems professionals. The authors would like to thank the following contributors for their assistance in reviewing and improving this paper.

Richard Dee, Sun Fellow
Jim Cates, Director, Tape Drive Development, Sun StorageTek
Ian DelBlaso, SAM-FS Marketing
Margaret Hamburger, SAM-FS Marketing
Jason Kranitz, Account Executive
Chuck Wenner, Systems Engineering Manager
Scott Matoon, Sr Systems Engineer, Western Region
James E. Brennan, Senior Systems Engineer
Kristen Powers, Account Manager

Copyrights, Trademarks, etc.

The following names and marks are the property of their owners:

IBM Corporation

- IBM
- 3590
- Ultrium

Litton Industries

MovieBeam, Inc.

Sun Microsystems

- 9940
- JAVA
- MaxStrat
- SAM and SAM-QFS
- SL8500
- StorageTek
- Sun
- T10000
- Titanium
- ZFS

Sony Pictures Entertainment

The Walt Disney Company

Thinking Machines Corporation

- DataVault

Appendix A: Assumptions and Methodology

The assumptions and methodology for the included Tables are listed below.

The purpose in disseminating this document is to assist Industry in the investigative process by gaining insight and input from those in the field who have participated in the development of the model, or who may be interested in expanding it.

It is important to remember that this is an investigative work-in-process. Pricing and costs are shown only to provide a relative cost comparison with film-based archiving, and to show that a digital archive is, indeed, financially feasible today. Prices are not meant to be an offered quote or offer to sell of any kind.

The methodology reflected in this archive model is as follows:

1. Two libraries exist in geographically separated locations for disaster recovery (DR) purposes. Both libraries are identical in drives, slot count and front-end compute and disk configurations.
2. A DI at 10TB is ingested to the archive at the primary location.
3. At ingest both ECC and a 256-bit Secure Hashing Algorithm (e.g., “SHA-256”), are run, creating two unique hashes for that DI. This hash is stored with the data for future access and compare operations.
4. On read-back after write, the ECC and hash are compared to ensure that the DI sent to tape is, indeed, the bit-wise replicate of the DI recorded on the tape.
5. A hierarchical storage manager, in this case Sun’s SAM-FS, creates and places four copies as previously defined:
 - a. Two copies remain in the primary library
 - b. Two copies are sent to the secondary (disaster recovery, DR) library.
6. All tapes are re-read and audited for hash validity approximately every six months to reduce the incidence of physical degradation, and to refresh the file format if required.
7. When a tape is audited and determined to be degrading over time, it is ejected and a backup copy, after checking the hash and ensuring it still is valid, is used to create a new copy on that library, ensuring two copies always exist in each library. These steps are done automatically via the HSM, reducing human involvement in overall media maintenance.
8. Should both copies in one library hash incorrectly, two new copies can be made from a determined good copy in the other library and then these tapes transferred to the first library. The intent is always to have two good copies in each library.
9. Content is copied to new tape media with the above hashing and checking, every ten years, discarding tapes older than that. The published media life for the T10000 tape cartridge is 30 years. Ten years has been used as a very conservative tape rewrite period — one-third the published specification — to ensure further the pristine quality of the archived content.
10. As additional cartridge slots are needed they are added in increments of 1500.
11. Vacant slots are kept in the library to ensure space for blank cartridges to be used when new copies are required.
12. Tape drives are assumed to reach end-of-life and are replaced in ten-year intervals.
13. When tape is re-written it is written to then-current density cartridges; Tapes are not reused.

14. Disk and compute front-end infrastructure is replaced every five years.
15. Each library is assumed to reach end-of-life at twenty-year intervals and is replaced.
16. The model assumes a constant price of drives, tape media and libraries over time under the assumption that as technology drives down the price, added features and inflation will keep it constant.

Appendix B: Analysis of 100-yr archiving costs

In the spreadsheets on these pages two different models are analyzed based on the size of the content object being archived. A 10TB and a 100TB archive object are used in the analysis.

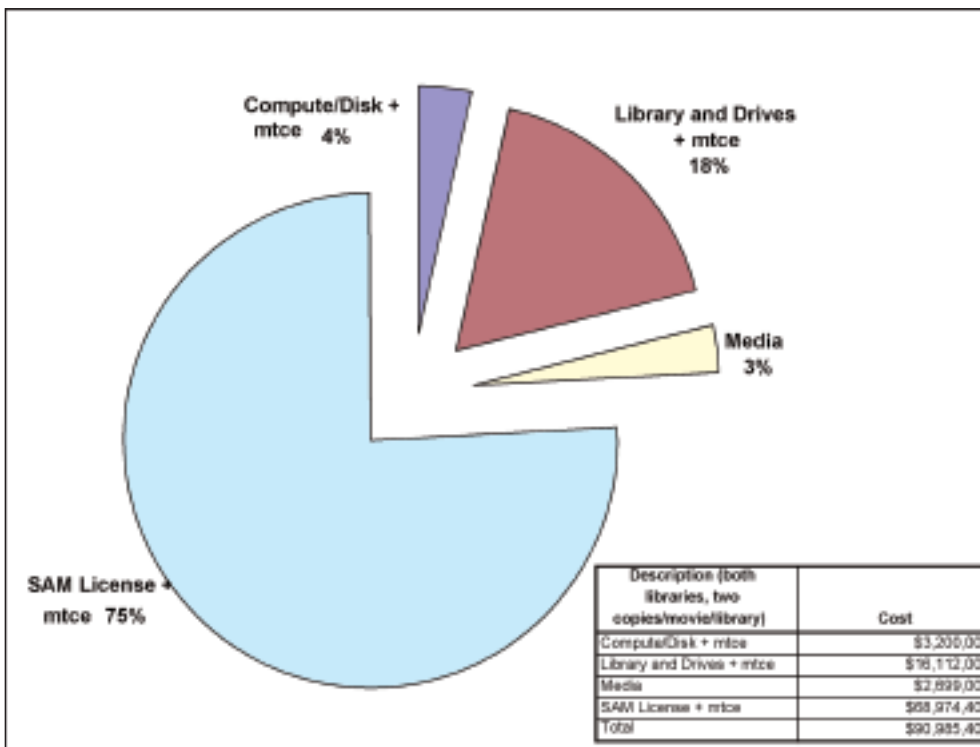
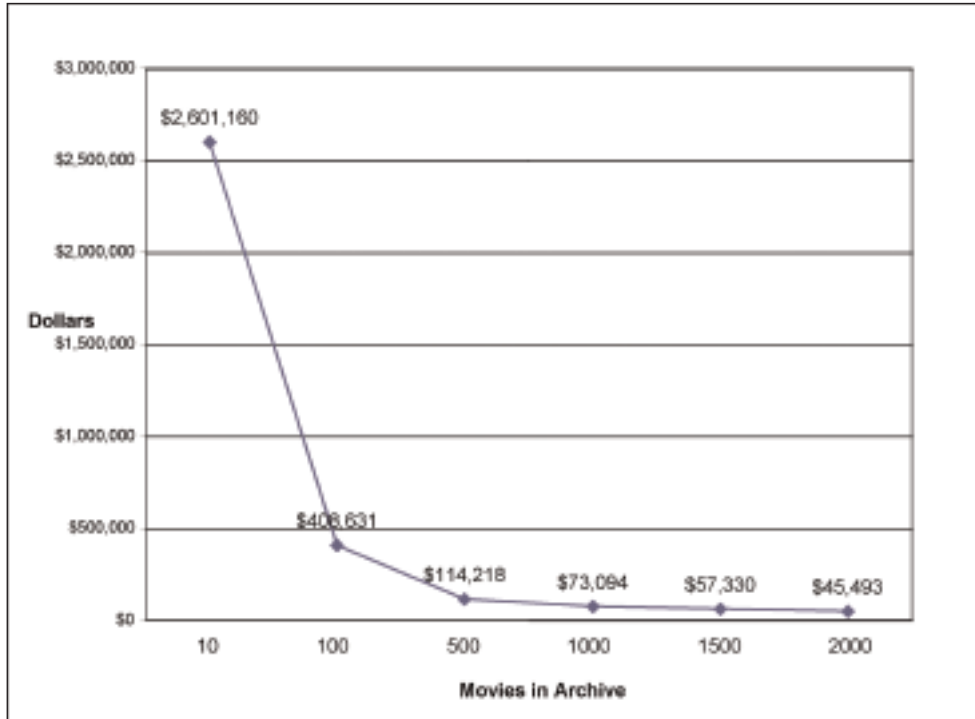
Both list and hypothetical discounted prices are shown. The cost curves show the cost to archive for a century a given number of movies when filled at a rate of 20 movies per year for a century.

Although the list-price cost curve for the 100TB object has some interesting values, in reality it can be discounted. For a 100TB object, list price shows the price of 100-yr maintenance to be about \$4B. In reality, per-call maintenance at a list price of \$285/hr would yield a price of only about \$250M, or 1/16th the list price, even if every one of the approximately 876,000 hours of the century were billed.

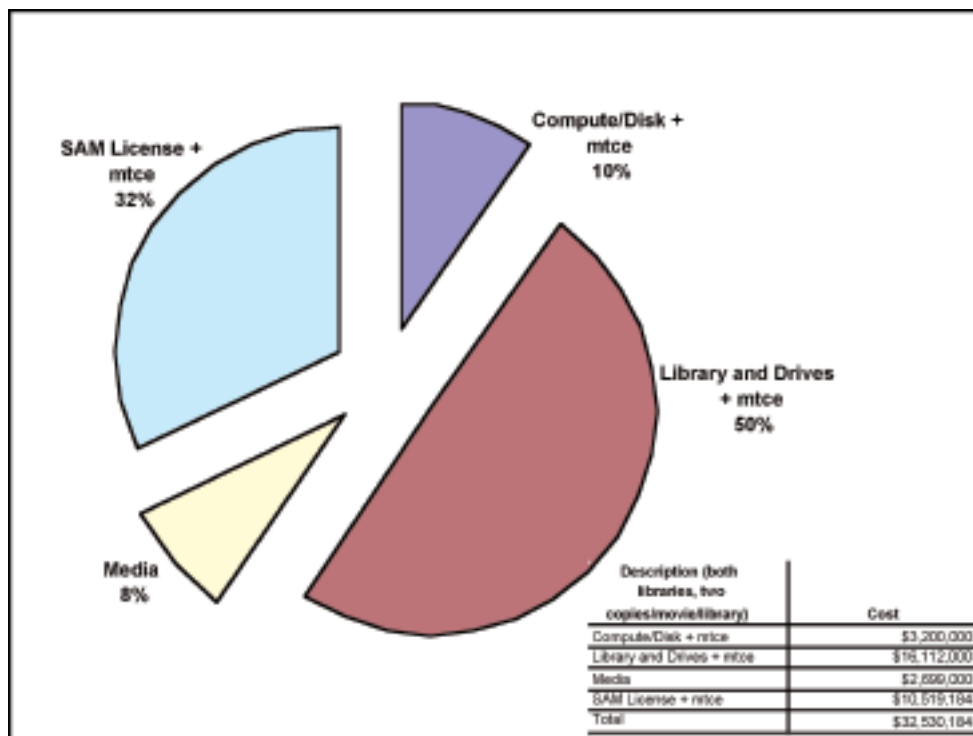
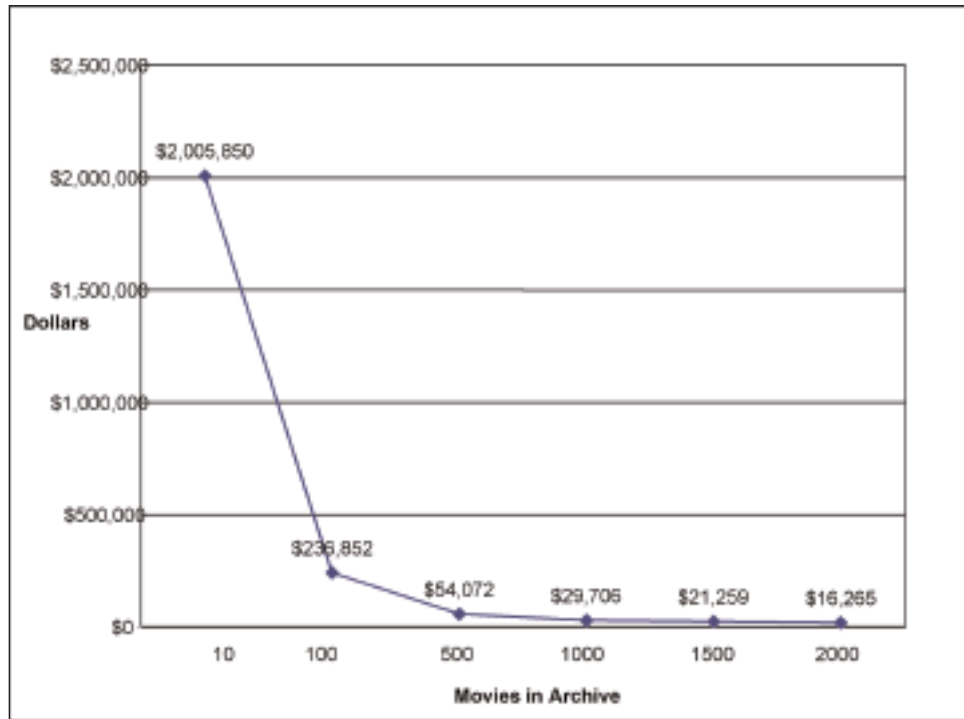
A pie chart also is included for each cost curve in which is shown the price by segment of the archive (HW, SW, etc.).

The list price analysis was done with list prices current as of the date of this paper.

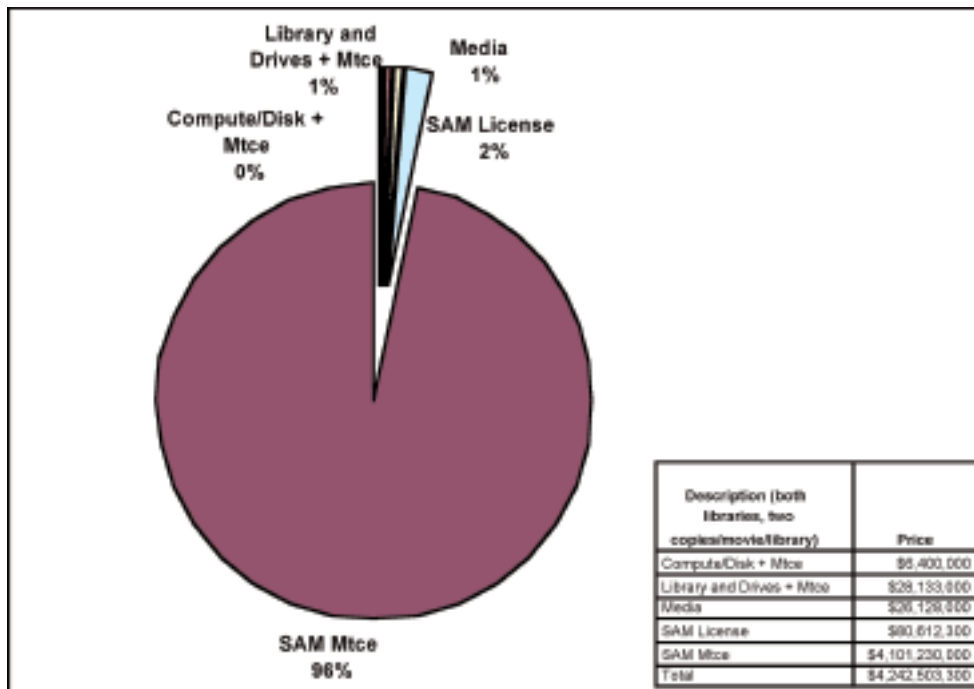
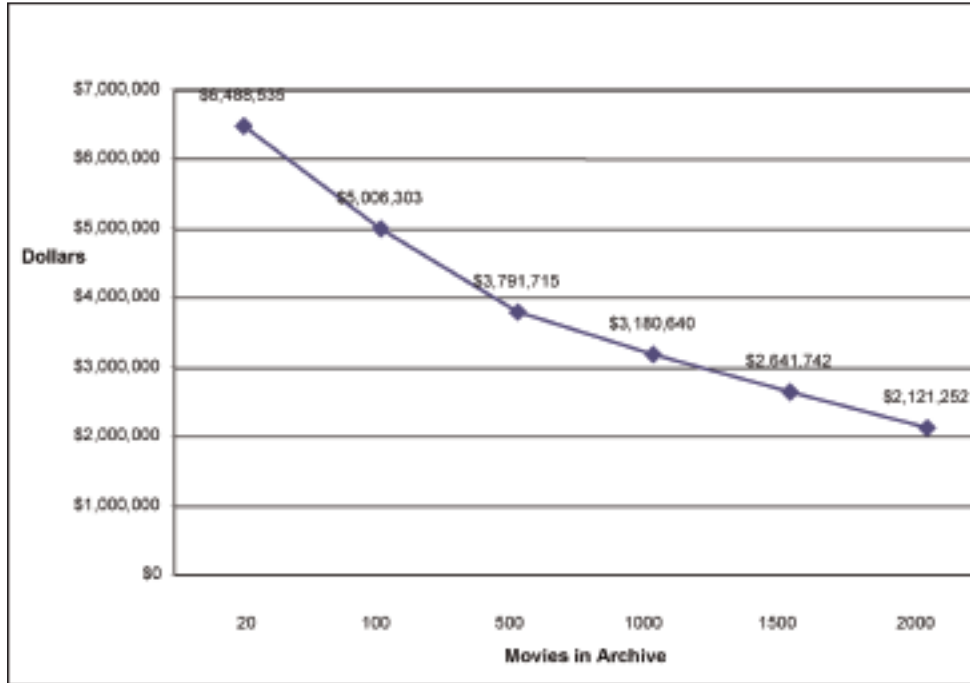
Analysis of 100-yr archiving price (list), 10 TB DI



Analysis of 100-yr archiving price (archive pricing), 10 TB DI



Analysis of 100-yr archiving price (list), 100 TB DI



Analysis of 100-yr archiving price (archive pricing), 100 TB DI

