

Protecting Your Content: Archive vs. Backup

Cache-A Corporation was formed to serve the professional media industry with the goal of making archiving easy, providing complete self-contained tools. This paper discusses how these tools can protect content for the long terms and compares the requirements and impact of using archiving solutions to backup solutions.

What is an Archive Appliance?

There are a lot of products that claim to be an appliance. Many manufacturers apply this name to any configuration of tools that serve a purpose. We use the term more specifically to mean an independent system that provides a complete solution, like a refrigerator or a dishwasher – one unit, self-contained, serving a complete function.

The specific idea of an “archive appliance” has been around for some time and many software vendors apply this term to the complete set of items used to make archives. This set typically would comprise an archival storage device (a tape drive, optical drive or library of such devices), a hardware interface to that storage device, a computer or server to talk to that device, an operating system running this computer or server, a storage subsystem to hold and migrate data (a hard disk drive or array of drives), and one or more software packages to manage data onto the archival storage device. In our view, this collection of components is not an appliance at all. Rather, it is a solution where you have to source all the parts, put them together, test, debug, develop workflow and maintain.

An archive appliance conversely would be a single device that serves the same purpose but adheres to the dictum “no assembly required.” Cache-A products meet that requirement in all regards and do it by delivering a single box solution that contains all the bits and pieces outlined above including a full high-powered Linux server, integrated into a plug-and-play box. Additionally, an appliance does not require additional support outside the system, i.e. client

software packages needed to communicate with and move data to the archiving system – in our case users need only drop data onto a network share to archive it, or access a web page to manage and access additional features of the system. While third party software can easily access the appliance, absolutely no additional client software need be used on any Windows, Mac, or Unix/Linux client system where archive data may have originated.

Cache-A Archive Appliance Constituent Components

| Component | Current | 2010 Release |
|--|---|--|
| Archival Storage Device | LTO-4 800GB Tape Drive | LTO-5 1500GB Tape Drive |
| Hardware Storage Device Interface | 4-Lane SAS Host Bus Adapter | No Change |
| Computer Server | Mainboard with 2.5GHz Core2Duo processor | Mainboard with 3 GHz QuadCore processor |
| Operating System | Fedora Core 10 Linux | No Change |
| Storage Subsystem | 1 TB HDD (Prime-Cache) 2 TB RAID 0/1 (Pro-Cache) | 2 TB (Prime -Cache) 4 TB RAID 0/1 (Pro-Cache) |
| Software Packages | Self-contained, network advertised, web served | No Change |

Appliance Maintenance

One significant consideration for any archiving solution is what it takes to keep it running over time. Not only is the appliance solution pre-assembled and ready-to-use, but also the maintenance requirements compared to a component approach are greatly reduced. Even if an end-user purchases a complete component package integrated by a VAR, all the bits and pieces must be maintained separately – Windows service packs, Archive software upgrades and any other application updates that may live on that system all need to be downloaded and installed individually as time goes on. If you don't have the in-house expertise to assemble and commission a component-based solution, hiring such an expert can cost more than the components themselves. Equally important, keeping a system up-to-date and maintained also



takes time and expertise – maintenance contracts that may be needed should be added to the solution cost considerations.

With the appliance approach, all updates are all rolled into a single package that can be updated by the user from the appliance itself, literally with just a few clicks. Cache-A appliances require only the most basic computer skills to deploy and include a one-year maintenance contract with affordable continuing support available. Best of all, our maintenance can be conducted over the Internet, providing timely and convenient service, remotely covering all but the most drastic hardware-related failures.

Roll-your-own Archiving

Many of Cache-A's customers feel our products offer a significant value as a complete integrated solution, especially when compared to the cost of a VTR, the former favored tool for archiving in our industry.

We also encounter a group of users who have long experience as do-it-yourselfers and who express dismay that we don't sell our solution for less. They frequently tote up the cost of a tape drive, bus adapter and software and tell us that they can have the same thing we offer for only \$4000... not really. The argument is frequently revisited in the RED User discussions and usually quotes something like the following:

| Archival Solution Component Costs | | |
|--|-------------------------------|--------------------|
| Component | Device | Street Cost |
| Archival Storage Device | HP1760 LTO-4 800GB Tape Drive | \$2500 |
| Hardware Storage Device Interface | ATTO SAS Host Bus Adapter | \$270 |
| Software Package | Bru Server Network edition | \$1099 |
| Total Cost | | \$3869 |

The first thing missing from their analysis is the cost of the computer they are going to use to serve this application, which, most frequently for this class of user is a high performance MacPro they are also using for editing projects. They are ignoring the fact that that computer is likely to be too busy for archiving any time except late at night and is not likely to be accessible to other users, so add in a basic MacPro at \$2500 and you are already well north of six thousand dollars.

| | | |
|--|--------------------|--------|
| Computer Server with OS & HDD | MacPro Base System | \$2500 |
| Total Cost | | \$6369 |

The other thing missing from any analysis of a home-brew solution like this is time – include the time and effort to source the components, the time and effort to assemble and test the functionality, the time and effort to install client software on other computers on the network and the time and effort to maintain this computer’s operating system, environment and backup software. The number of hours to pull this all off and their dollar value is going to vary widely depending upon the expertise of the individual and what other profitable tasks the DIY’er could be involved in, but to discount it as having no value is quite simply naïve business practice.

Even assuming all this time and effort is free, the end result of a roll-your-own solution most likely leaves the user with what is really a backup tape solution. That is, with tapes that do not permit content interchange and that are ill suited to true archive, as explained in the next section.

Backup Versus Archive

Backup is all about the protection of what is currently on your computer disks, certainly important for daily work, but archiving is really about protecting your assets. We frequently talk about how content is king in our industry; protecting that content for the long term is what we are targeting when we talk about archiving.

Everyone, without exception, wants to be sure his or her computer data is protected. By and large, only IT professionals give a lot of thought to the different needs of backup versus archiving. And by and large, most people in typical computing environments really just want protection, so what they really need is backup. But this changes in our industry – media professionals by and large, create huge amounts of data that they can’t keep on-line (see sidebar: Can’t you keep it all?) but still may need in the future, so what they really need is archive. Just what does this difference imply for hardware, software and your data policies?

Can’t you keep it all?

Some facilities have tried simply adding to their SANs and RAIDs as a strategy for accommodating an ever growing amount of data. Studies by IT experts show that the costs for doing this are deceptive – individual disk drives are cheap but not the infrastructure to support them. Add rack space, power and cooling requirements and this route is impractical in any thorough analysis. According to The Clipper Group Inc., the costs for a Terabyte stored long-term on SATA disk versus LTO-4 tape is about 23 times greater and for energy cost, it is about 290 times greater. If this doesn’t convince you, ask anyone who has tried the increased-disk drive strategy and you’ll find that their needs grew to outstrip whatever amount of storage they added far faster than they anticipated.

What makes sense is to only keep as much data on-line as you need for active and recurring projects – don’t try to keep it all, you will always run out of room

Backup, simply put, is whatever you do for disaster recovery. Backup is what covers you if your hard disk drive dies. Backup is where you go if your facility burns down or a lightning bolt takes out every solid-state box on your grid. Backup is insurance, and like most insurance, it is something you really hope you never have to use.

Archiving data on the other hand is very much like archiving any physical assets, much like what a museum does. Moving data from Hard Disk Drives to long-lived LTO media is more similar to what an archeologist might do. Shelf-stored tapes of projects (or automated libraries for that matter) with indexed catalogs are a lot like drawers full of archeological specimens with researcher’s notes about each item. The point being that not only do we have records, we want to assure that they are preserved for the future and easily accessible when any tiny bit is needed again.

The following table outlines some of the more significant differences between these two needs:

| Archive Versus Backup | | |
|------------------------------|--|---|
| Description | Backup | Archive |
| Data Characterization | Snapshot all data as it exists currently, updated frequently | Packaged as data sets, preserved for long term |
| Data Accessibility | Accessed for emergencies, need is to restore en-mass | Access need is ad-hoc, need may be for single item |
| Data Cataloging | Organized by points in time, low need for searchability | Must be searchable by file name, date and other metadata |
| Data Portability | Typically only be used on the system for which it was saved | May need to be used anywhere |
| Media Life | Must only live as long as the next full backup | May continue to be accessed long into the future |
| Performance | Large amounts of data benefit from high speeds, narrow backup windows need speed | Need to free up source space or recover projects quickly requires high speeds |
| Storage Cost | Media is reused, cost is less | Media is retained, cost per |

The implications of this impact how you should save your data and explain that the majority of backup software being used today may be inappropriate for the need to archive content.

Data Characterization

Video professionals have historically kept archives in a way that illustrates the difference from backup data. The classic scenario was where the shooter returned from the field with a stack of original tapes – those master tapes were always copied for editing and then archived on the shelf where they were held until the content was absolutely needed for another project. We now return from shooting in the field with memory cards or hard disk drives, dump the data to a workstation and often forget about that original workflow that demanded archival preservation of the source footage.

Backup software typically saves all the data in a session into a unitary data container that must be fully restored to obtain anything within it. Backup software typically retains little information about what is stored within each session and that information is accessible only to the computer server that invoked the original backup session. Backup data is typically saved in proprietary formats that can only be recovered by the same or later versions of the specific software that created it.

Backup sessions are frequently incremental, where later sessions contain newer and changed files built on top of earlier sessions containing base data that didn't change. This can save considerable time and media space in the backup process. The painful side of this is that when it comes time to do restores, users are required access many non-contiguous sessions to rebuild data sets. This can serve the needs for a facility's data protection, but is far from ideal for archives.

Archive Appliance Data Accessibility and Cataloging

Our archive appliances in contrast, save data in a file-by-file searchable and accessible format. Any single file or subfolder can be restored by itself or entire directories with all their contents can be restored as needed. The Cache-A internal catalog keeps the same kind of data about each file as the best backup software, but also allows for user metadata tags to be saved for any individual item, folder or tape; this catalog can be searched to locate any content ever seen by the device. This catalog can also in itself be backed up to tape or any other data media for protection, but even if it is lost, each tape contains its own directory that can be used to rebuild a catalog.

Archiving on Other Media

In this modern world, the idea that tape is the best archival recording media seems counter-intuitive, but it really is. All of the magnetic advances seen in the dramatic and widely recognized increases in the capacity of disk drives also apply to tape technology. Each generation of LTO nearly doubles in capacity every two years, right in step with disk. Tape also has a proven history of being recoverable over long storage periods, with formulation and manufacturing improvements continuing to extend that.

The very best hard disk drives come with a 5-year warranty for a reason. The extensive mechanical components are susceptible to freezing up and to physical shock and the electronic components, susceptible to electrical shock. Anyone who has more than a few shelf-stored drives has seen failures.

Blue laser optical media has shown promise with a claimed archival life of 50-years, but the very largest capacity is only 50GB, one eighth of current gen LTO. Also, the fastest it can be accessed is about 50Mb/s versus 80MB/s for LTO-4, over 12 times slower. Finally, this kind of optical media costs over \$1/GB where LTO-4 is about 6¢/GB or 1/16th the cost.

Some users have attempted archiving to red laser DVDs but that is a seriously flawed strategy – most DVDs have only a 5 year life, are subject to areas becoming unwritten, and typically aren't even write-verified. More costly long-life gold DVDs are available, but with only 8GB max capacity and 10Mb/s max data rates, they are a poor choice for archiving.

Archive Appliance Data Portability

Further, this tape-based directory or Table of Contents makes each piece of media itself a portable searchable archive when inserted into any other Cache-A appliance. When the tape is inserted into the new system, it automatically detects that it was written by one of our appliances and proceeds to extract the table of contents for that tape and add it to its catalog. If a Cache-A tape is transported to a location with no Cache-A device, but an LTO drive is available, the contents are still accessible by virtue of the fact that we write in a non-proprietary format that can be read by publically available free software (more about that later).

Media Life, Cost and Performance

Because the LTO tape technology was created by a consortium of IBM, HP and Quantum, some of the most sophisticated high-tech IT manufacturers, it has become the most broadly used data format in the world with over 2.5 million tape drives deployed and over 100 million tapes in use. The major players in this huge user base are high-value data users including Wall Street, medical databases, banks and insurance companies.

This level of popularity has driven performance up and costs down so far that LTO has become completely dominant as tape media for both archive and backup. All LTO tapes have been qualified for a 30-year archival life and are likely to last even longer if these tapes are stored in a cool low-humidity environment. The transfer rate performance at the current

generation of LTO-4 is faster than all but the most expensive hard disk drives and the cost per gigabyte has descended to mere pennies. LTO-5 is being introduced at the time of this writing with 1.5TB of storage per tape and even faster sustained transfer rates than any current hard disk drive. While short-term backup is increasingly moving to disk, for long term archiving, LTO is hard to beat (see sidebar: Archiving on Other Media).

Making Archives with Backup Tools

It is easily understood from this discussion that the needs for backup vs. archive are quite different and that the tools designed to meet these needs have different feature sets and a capabilities. But, can't you use one kind of tool to do the other job? The answer in our view is "yes-but." Indeed the data could still be saved to the same media with the same longevity, but making an archive with backup tools leaves your archive less secure and reduces your data accessibility because the catalog remains on the backup server and is less suited to the job of finding and restoring project-based data. Similarly, making backups with archive tools has the drawback that it requires a manual process to assure that regular backups are comprehensively made, and the user must manage not just their data, but their computing environment onto the archive. That said, we know of Cache-A customers who do use their archive appliances to make backups and we plan on adding future features to ease this process.

The "tar" Advantage

As noted above, Cache-A archive appliances write data in a format that can be read back by publicly available software. That format is called "tar" which is a Unix acronym for Tape Archive and Restore. Originating in the 1970's, tar has been the standard interface to tape for every version of Unix and Linux ever shipped. A Linux computer with tar installed can be had for literally a few hundred dollars. Any computer can be made to run Linux and tar is included with every Linux release. Open-source tar-based applications are even available for Windows. This means that any Cache-A tape can be restored on any computer with an LTO-4 drive installed and where you can find or install a copy of the tar program.

Some software manufacturers have gone to great lengths to illustrate how their software is better than tar, and in certain backup-oriented applications it may be. Those comparisons frequently ignore the constant improvements that have been made to tar as an open source solution over the years. And, those comparisons are always with respect to the raw tar program, which is admittedly an arcane command-line driven tool suitable mostly for experts. In this case however, when tar capabilities are embedded within an easy-to-use graphical interface such as

the one served up by Cache-A archive appliances with the additional enhancements we've added, those criticisms fall by the wayside.

The following chart compares features in a typical popular backup application with tar and with Cache-A's implementation of tar:

| Archive Application Comparison Chart | | | |
|---|------------------|------------|--------------------|
| Feature | Backup SW | tar | Cache-A tar |
| Open-standard Tape Format | | ✓ | ✓ |
| No Client-side Software Required | | | ✓ |
| Full Remote Technical Support | | | ✓ |
| Windows/Mac/Unix/Linux support | ✓ | ✓ | ✓ |
| Complete OEM HW/SW Package | | | ✓ |
| File level Volume Spanning | ✓ | ✓ | ✓ |
| Graphical Interface Support | ✓ | | ✓ |
| File-based Search and Restore | ✓ | | ✓ |
| Metadata-based Search and Restore | | | ✓ |

A few highlights can be seen from this table. True portability of data tapes using open-standards formats is only available with tar and tar-based solutions like Cache-A's. Some backup applications have limitations when working cross-platform and require the purchase of different client software for each operating system and some don't even do file-level restores. By contrast, Cache-A's web served and network attached solutions work with no client side software required. Some don't support spanning large files across tapes and must force media to fit leaving wasted space, but tar and Cache-A can split files as needed at media boundaries. Cache-A's complete hardware package and end-to-end hardware and software support is unmatched by other software vendors.



Where does the data come from?

In professional media environments, data is content, the very life essence of our industry and it can come from a variety of sources. In the beginning, it is acquired in the field or on-set from cameras and as we mentioned above, this has historically all been archived. But even in the simplest file based workflow, the finished product after editing will be archived – it is important to focus on archiving at this stage, but it is also important to protect source footage. And in more complex workflows, there are archival needs at multiple points through the production process – rough cuts, dailies, computer generated footage and graphical elements, a range of digital intermediate content and individual unique project creations all may need to be saved separately from source and finished product. High-end Hollywood projects always give substantial consideration to archiving at many stages, but outside of big-bucks productions, archiving disciplines vary widely. Our industry is really still learning the ins and outs of file-based workflows and how easily data can be archived... or how easily it can be lost.

Working with Asset Management

Big productions and large content producers may be using what the IT industry likes to call “heavy iron” to create their archives. This typically entails large robotic tape libraries being controlled with hundreds of thousands of dollars worth of archive management and data moving software. This is appropriate when the dollars warrant it, but is well beyond the means of many if not most media professionals in the production chain. Even when the productions are large enough to justify such expenditures, in our industry it may not make sense to invest in large libraries as the need to access this content is sufficiently infrequent that it makes more sense to shelf store such archives – users need to ask themselves the question: do I really need it all near-line in an expensive automated library rather than shelf-stored off-line?

These days, more and more affordable tools are becoming available to help you track your content. Cache-A has started working with a range of such software tool authors to marry their solutions with our appliances. While our catalog serves the needs of many users for one level of content tracking, these asset management solutions extend the searchability and accessibility of archive content. Tools for tracking the ingest of content to tools for tracking and managing the material involved in editing sessions are rapidly evolving and being used with more regularity. A variety of available third party tools can provide proxy browsing and can over a visual access complement our database catalog in managing archives.

Examples of some of the more basic tools for ingest include Imagine Product’s “Shot Put” and “HDlog” logging software or R3DDData’s “R3DDData Manager,” both of which assist users with camera content, creating additional metadata, organizing and ingesting content and creating

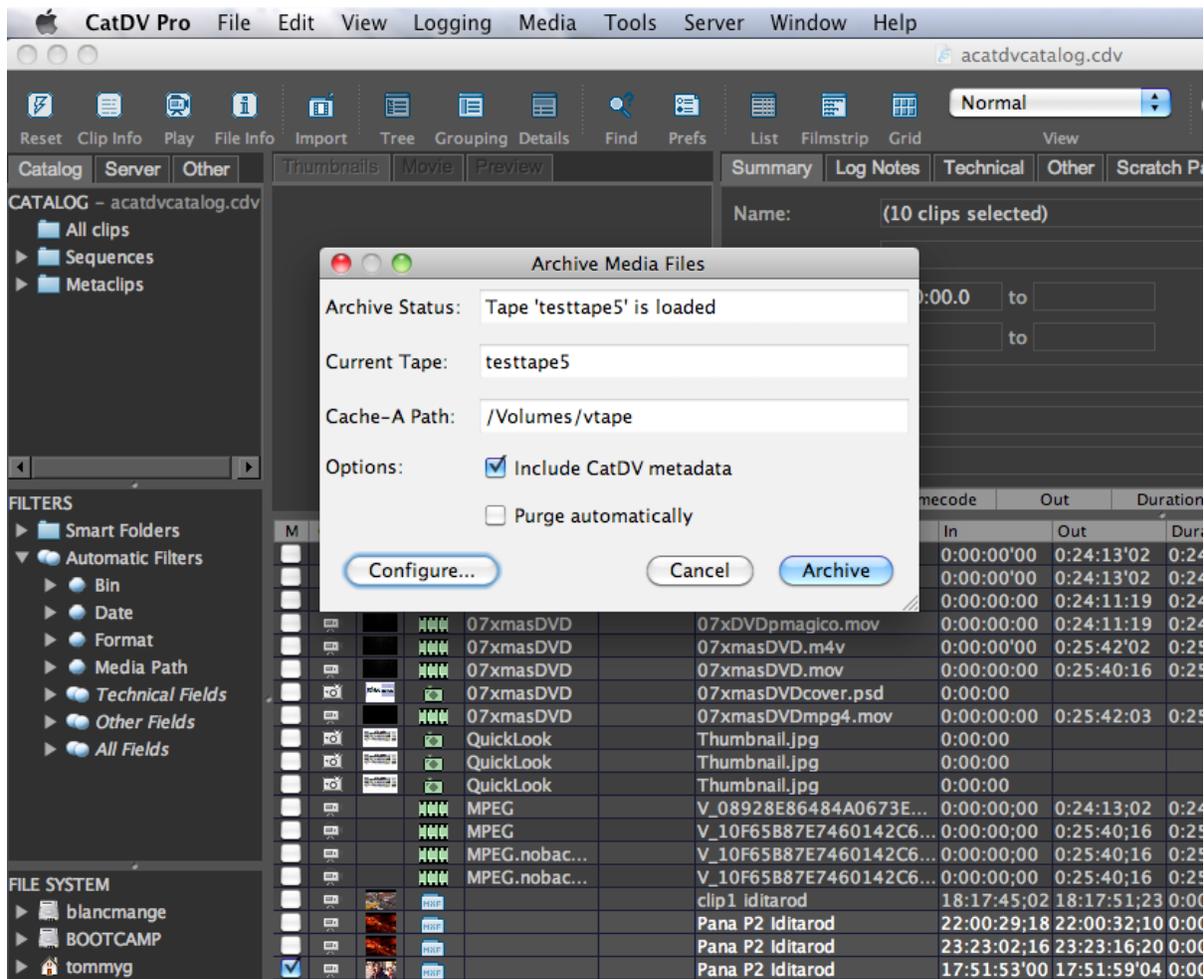
verified archives of this content. These vendors are focused on helping to migrate data from field acquisition onto workstations, but they recognize the need for archive and are integrating interfaces to Cache-A appliances into their workflows.



R3DData's Data Manager Cache-A Configuration Screen

More extensive Asset Management tools move beyond ingest to help organize and create archives of content being used in full-up editing environments. Well-known examples of such tools include Apple's Final Cut Server software and Avid's Interplay application suite. Cache-A currently offers cut-and-paste workflows with these solutions and is continuing to work on more integrated workflows with Avid and Apple - we will update this paper with progress on those fronts.

Amongst the growing selection of project asset management software includes lesser known but perhaps more capable software, Cache-A has been working with Square Box Solutions to integrate with their CatDV application. Today CatDV is a comprehensive solution that works with both Final Cut and Avid environments as well as with a wide variety of media types from RED to P2 and has integrated Cache-A's appliances into its workflows.



Square Box's CatDV PRO preparing to archive to a Cache-A appliance

Tools such as CatDV go far beyond asset management, adding capabilities including the ability to preview content, create subclips and add a wide variety of metadata within custom metadata categories. In fact CatDV even supports batch transcoding capabilities, allowing content to be repurposed or ported to alternative creative tools, all from within an asset management environment that includes archive as a central capability.

Conclusion

In this paper, we discussed the difference between backup and archive and the implications that has for solutions to protect the massive amounts of data being generated by our industry. While everyone needs to look to the day-to-day protection of normal computer data, media professionals face a unique task protecting the new file-based content that used to reside on videotape. Backup packages are poorly suited to this need; archiving tools are needed to make this content accessible, portable and to keep it useful. Cache-A's archive appliance solutions are ideally suited to this need.

We have also tried to show that build-it-yourself solutions or purchased component solutions provided by IT integrators have hidden costs to end-users. These costs are not only encountered in the initial system setup, but in maintaining the individual hardware and software components. Cache-A's plug-and-play appliance approach can deliver significant savings over time in these areas.

Cache-A archive appliances are meeting the archival needs of many users today as a stand-alone solution. Cache-A will continue to work with third parties to expand our products' usefulness across the range of archival needs from ingest to finishing, as well as to expand into the realm of content distribution from creators all the way to individual broadcasters. While additional interface developments moving forward are a given, some really excellent asset management software is available today to help deploy these solutions in a wider variety of applications.

We hope readers of this paper will seriously undertake actions to protect their data as they move into the file-based future. Those old videotapes sitting on the shelf were a form of protection that should not be forgotten, but rather adapted to the changing world our industry is experiencing. The value was built into those assets in every step of acquisition and production – you owe it to that investment to protect it.

By Tom Goldberg
Cache-A VP Product Development