

Shoah Foundation Architecture

Hriday Balachandran, Sam Gustman, Christopher Ho, Luke Sheppard
Shoah Foundation Institute for Visual History and Education
University of Southern California

September 21, 2009

1 Abstract

In anticipation of the impending hydrolysis¹ of nearly 244,000 Betacam SP tapes, the Shoah Foundation needed a way to (a) digitize 107,500 hours of video testimony, (b) create a filesystem large enough to archive the resulting 8 petabytes (PB) of data while still keeping the files locally available on the network, and (c) complete the digitization project within 5 years, which represents the end of the worst-case-scenario 10 year shelf-life[1] of the video tape media.

The foundation's mission² depends upon preserving the testimonies in perpetuity.

As with most organizations, our design decisions were a product of the equipment we already had and the funding available for new equipment. We designed an architecture based on our load assumptions. After installation, configuration, and testing, a few modifications were needed to achieve efficiency and stability. Initially we had difficulties with managing disk capacity, CPU load, LAN bandwidth, and even the number of tapes in the digital archive library pool. But six

months after starting the project we achieved a digitization rate of about 80 terabytes (TB) per month, or over 5000 Betacam SP tapes. Having attained this rate we are well on track to completing our project within the original timeline.

The expectation at the beginning of the project was simple enough—digitize all the testimonies in five years. That meant building a system that can digitize at least 20% of the archive, or at least 20,000 hours of video per year. Our current digitization process can handle over 30,000 hours per year.

2 Problem Definition

The Shoah Foundation has 234,986 Betacam SP[2] master video tapes, containing 107,500 hours of testimony from 51,682 holocaust survivors. These testimonies were recorded in the mid-1990s. The master tapes are expected to physically deteriorate³ beyond readability by 2014, five years from the date of this paper. Sustaining and expanding the Foundation's mission

¹*"The polymer used as binder in the tape structure is subject to chemical decay through a process known as hydrolysis. As the deterioration progresses, long polymer molecules become shorter, and the integrity of the binder is reduced. The process may lead to softening, brittleness, loss of cohesion, and the formation of sticky products, and the tape may become unplayable."*[1]

²To overcome prejudice, intolerance, and bigotry—and the suffering they cause—through the educational use of the Institutes visual history testimonies

³See the footnote on hydrolysis, above.

rests upon the ability to preserve the integrity of the testimonies in perpetuity. Digitizing the testimonies—replacing the master tapes with Motion JPEG 2000 digital master files—expands the possibility of providing access to the testimonies via digital media and access technologies in the future. But, storing and backing up what is expected to become 8 PB of data presents technical and budgetary difficulties. Considering current hard disk capacities and prices, the size of a purely disk-based filesystem would be beyond what is economically or technically reasonable.

Today’s hard drive capacity and rack density maximums hinder the scalability of a purely disk-based filesystem of this size. When designing storage arrays, we have set a theoretical maximum number of hard disks per rack at about 150 (assuming the standard 3.5 inch form factor). Beyond that density, we expect unmanageable cooling, power, and weight problems. The largest disks available from established and proven manufacturers are 2 TB today. In practice, disks available in storage arrays are rarely the highest density available on the market. Overhead for RAID 5 formatting and SamFS formatting consumes about 20%, leaving approximately 1.6 TB of capacity per disk. Our data is 4000 TB but becomes 8000 TB total when we include backup copies. So, for the formula:

$$\frac{TotalData}{(RackDensity * DiskCapacity)} = NumberOfRacks \quad (1)$$

we have

$$\frac{8000TB}{(150 * 1.638TB)} \approx 33 \text{ Racks (rounded up)} \quad (2)$$

Thirty three racks would have consumed about 18% of our data center’s available rack space. Other problems with this approach are the power requirements, capital cost, support cost, patching and rebooting dozens of disk controllers, etc.

3 Scope

This paper only deals with the hardware and software architecture of the project. We touch cursorily on networking issues where needed for clarity. And we have not at all delved into the issues of data center footprint and power consumption of the infrastructure.

4 Solution

Our solution involved the integration of three components—the infrastructure, hardware and the software. Each component is detailed below.

4.1 Infrastructure

For archiving we used two existing Sun StorageTek SL8500 robotic tape libraries[3] (referred to here as silos) already in place in the USC data center. When we began however, the tape drives in the silos were the 9940Bs[4], which were a generation old, could only support 2 gigabits per second (Gbps) of network throughput and could address tape media with a capacity of only 200 gigabytes (GB). We decided to go with Sun’s latest T10000-B[5] drives that were 4 Gbps compatible, and could write to media that had a capacity of 1 TB. These drives were designed to be compatible with our SL8500 silos, which was a critical aspect of our requirements.

For high speed communication between the disk arrays (described below), filesystem control nodes (described below) and the silos, we used a fibre channel[6] (FC) network. At the time, USC had an existing 2 Gbps FC backbone in place. This legacy FC network, designed for redundancy by way of two separate rails, interconnected through twelve Cisco 9140[7] switches. However, the FC backbone was aging, and it became apparent during the planning stage of

this project that this infrastructure would, in the least, have to move to 4 Gbps in order to take advantage of modern FC cards. The FC switches we settled on were QLogic's 9200s[8] (QL9200), which had 4 Gbps capabilities. The QLogic switches also had failover CPUs, were enterprise grade and easily stackable for future expansions.

We were able to use USC's existing copper and fiber Ethernet network, with very few modifications. Host to host interconnections are at 1 Gbps and in some places 10 Gbps. The details of the configurations will be discussed later in this document.

4.2 Hardware

The first component written into our hardware design was the videotape to digital file migration system. This equipment reads the Betacam SP tapes and converts the signal to the various digital derivatives. We chose a packaged solution from Samma Systems (now Front Porch digital). We have two Samma Robots[9], each housing a 60-slot rotating tape library, six Sony J-10SDI tape playback decks[10], and a Samma Clean videotape inspector/cleaner. Each of the six decks is attached to a Samma Server[11] mounted in an adjacent rack. Each Samma Server has two 1 Gbps RJ-45 Ethernet cards used for remote management and for transferring digital files to the quality assurance (QA) directories on the SamFS[12] filesystem (described below).

We used Sun Microsystem's SunFire x4600 M2[13] servers. The x4600 had impressive benchmarks for performance[14], had eight quad-

core processors, and was enterprise grade, having several PCI-X/E slots for our FC and network cards.

For our FC attached storage, we settled on the 6540s[15] from Sun's StorageTek line. The 6540's had 4 Gbps support and had good benchmark numbers. Additionally, their configuration was through the Common Array Manager[16] (CAM), and USC had several other disk devices that used the same software, so there were skills in house for fine tuning. We procured two 6540 head nodes, and a total of 12 JBODs⁴.

4.3 Software

In order to address over 8 PB of data, we had to use a filesystem that could handle such a large addressing⁵ capacity. Coupled with the fact that we needed hierarchical storage, the choice of filesystem was invariably going to be SamFS. We use SamFS for various projects at USC and all the experience gained made it a very easy choice for us. Apart from handling our vast data, and handling hierarchical storage, SamFS is also very robust and can be made to adapt to our fluid design very easily. Also, the massive scalability of SamFS is important to our organization; as testimonies and other digital assets are collected and generated as part of new projects, the existing SamFS filesystem can easily scale beyond the initially planned 8 PB.

For data tracking and archive management, we have a custom built application we call the Visual History Media System (VHMS). VHMS tracks testimonies and the multiple files that make up a complete testimony, as well as cal-

⁴At USC we use the term JBOD (Just a Bunch of Disks) in a slightly different way than others might. We configure a group of disks as a RAID array, then, we configure the RAID array(s) to be part of a SamFS, filesystem. To SamFS, these RAID arrays appear to be JBOD, since the hardware disk controllers handle the RAID striping and disk management. This differs from another common use of the term JBOD, which is as an alternative to RAID altogether.

⁵In this context we use "addressing" to mean the location of data (whether on disk or tape). This is analogous to, but different from, the traditional Unix path location notation: *volume + full directory path + filename = file location*. In SamFS addressing information in the form of path, inode, etc. is stored locally in the form of metadata. SamFS refers to the metadata in order to retrieve the file from either disk or tape.

culating (SHA1) checksums for each file. Additionally, the VHMS distributes media derivatives to the various caches around the world. This latter function, however, is beyond the scope of this paper.

Our QA is handled by another custom built application called Media Problem Manager (MPM), which plays back a file on demand, then handles the release of the file to the VHMS, based on whether a file passes QA or not.

5 Configuration

We have eight x4600s - one serving as the SamFS master and another as the SamFS secondary (failover) server. The other six are QFS[17] (Quick File System) clients. The two 6540s combine to form the SamFS disks. The 192 disks have an aggregate capacity of 109 TB, after formatting.

Each 6540 has two disk controllers. Each controller has two FC ports, each of which are connected to one of two QLogic 9200 FC switches, one for each rail of the FC network. Each QL9200 is outfitted with four blades of 16 ports each. In order to balance the load, each FC port on the 6540s goes to a different blade on the QL9200. This setup is illustrated in Figure 1.

The SamFS master server and the secondary failover server are each outfitted with four dual-port FC cards. Each card connects to both FC rails, and similar to the 6540 configuration, each port is connected to a different switch blade on the QL9200 switches. The other six servers each have two dual-port FC cards, so they connect to only two FC cards on each rail. This setup is illustrated in Figure 2.

We purchased 14 T10000-B tape drives for this project, and installed seven in each SL8500 silo. Each of these drives connects to one of four 16-port FC blades on a single QL9200 switch. The

cabling between these tape drives and the FC switches has been carefully designed to limit the impact of a failure of any single blade or an entire FC switch. In order to protect against losing connectivity to multiple tape drives in the event of a failure of one of these blades, we have a maximum of two tape drives cabled to any single blade on a QL9200 switch. Another way to think of this is that the failure of a single FC blade (on either switch) would disable only two of our tape drives. And the failure of one of the switches would only disable seven of our 14 tape drives.

The disk drive numbering range, known to SamFS as an “ordinal family”, is arbitrary and is configured in the mcf file. Our own number system allows for expansion while still maintaining the differentiation between drives installed in silo 1 or silo 2. This configuration is tabulated in Tables 1 and 2, and illustrated in Figure 3.

Table 1: SL8500 Silo # 1: Cabling Between Tape Drives & QL9200 FC Switches

T10000B Drive Number	QL9200-0 Blade Number	QL9200-01 Blade Number
2901		Blade 0
2902	Blade 0	
2903		Blade 1
2904	Blade 1	
2905		Blade 2
2906	Blade 2	
2907		Blade 3

Table 2: SL8500 Silo # 2: Cabling Between Tape Drives & QL9200 FC Switches

T10000B Drive Number	QL9200-0 Blade Number	QL9200-01 Blade Number
3001	Blade 0	
3002		Blade 0
3003	Blade 1	
3004		Blade 1
3005	Blade 2	
3006		Blade 2
3007	Blade 3	

For Ethernet connectivity, the two SamFS servers (master and secondary) are each outfitted with an additional 10 Gbps network card. This allows the servers to have additional, dedicated bandwidth. The SamFS servers also use one of the built-in 1 Gbps interfaces, and are dual homed. The other six QFS servers only use the in-built interface.

6 Workflow

After the preservation team determines which testimonies are going to be migrated, the bar codes of each Betacam SP master tape are scanned. The Samma Robots are then loaded with up to 60 of these Betacam SP master tapes. Loading the Samma Robots consists of a human operator manually inserting the tapes one by one into a revolving multilevel tape carousel. Once the Samma systems are powered up, the robot loads each tape into a tape cleaning machine, then removes the cleaned tape and inserts it into one of the six tape decks. When playing back a Betacam SP master tape, the Sony J-10SDI

tape playback deck outputs a 10-bit serial digital interface (SDI) signal which is captured on the connected Samma server, which then converts the stream into five digital derivatives:

- Archive quality Motion JPEG 2000 (.mxf) at a VBR (variable bit rate) of ≈ 75 Mbps
- Mpeg-2 (.mpeg) at 5 Mbps
- Quicktime (.mov) at 1 Mbps
- Flash (.flv) at 1 Mbps
- Windows Media Video (.wmv) at 1 Mbps

The Motion JPEG 2000 file then serves as the new "digital master". The digital master is not intended for direct viewing by remote subscribers via the Internet (It would require end-to-end bandwidth of ≥ 75 Mbps). Additionally, a PDF document containing textual and graphical analysis of the transfer process is generated by the Samma Server for each testimony. The Samma Server also creates an XML file containing all the metadata for the testimony. Finally, a checksum

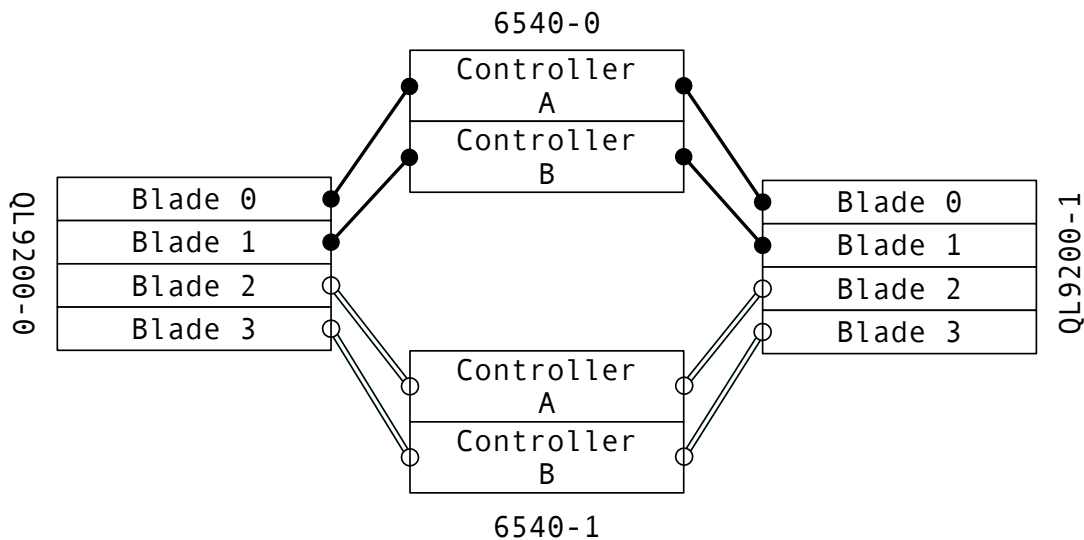


Figure 1: Fiber Channel cabling configuration between Sun 6540 disk controllers and the QLogic 9200 Fiber Channel switches

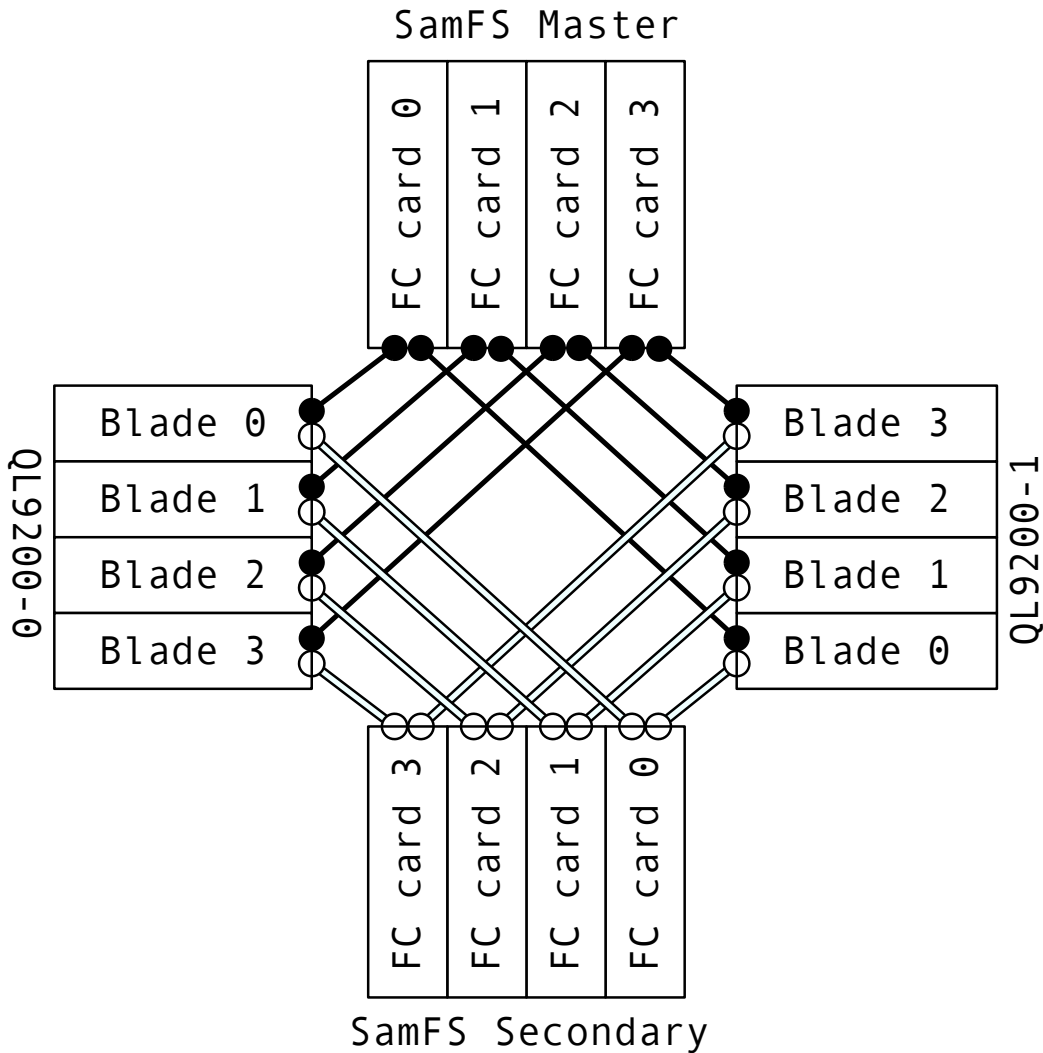


Figure 2: Fiber Channel cabling configuration between SamFS master and secondary servers and the QLogic 9200 switches

is calculated on each of these files, and the output is stored in a separate XML file containing only the checksums for each (new) derivative of that testimony. So in addition to the original five digital video files, we now have a PDF and two XML files. All eight of these files are initially stored on the Samma Server where they were generated.

On the Solaris servers, VHMS runs a process that constantly monitors the Samma servers looking for the presence of a new checksum file. The VHMS process does not have to parse the XML files to know which one contains the checksums. The checksum file is kept in a separate directory from the seven other files derived from each testimony. Also, to avoid filename collisions or any other confusion with the other XML file for that testimony, the checksum file ends

⁶However, this file extension does not indicate any relationship to Voice XML or the W3C standard VXML.

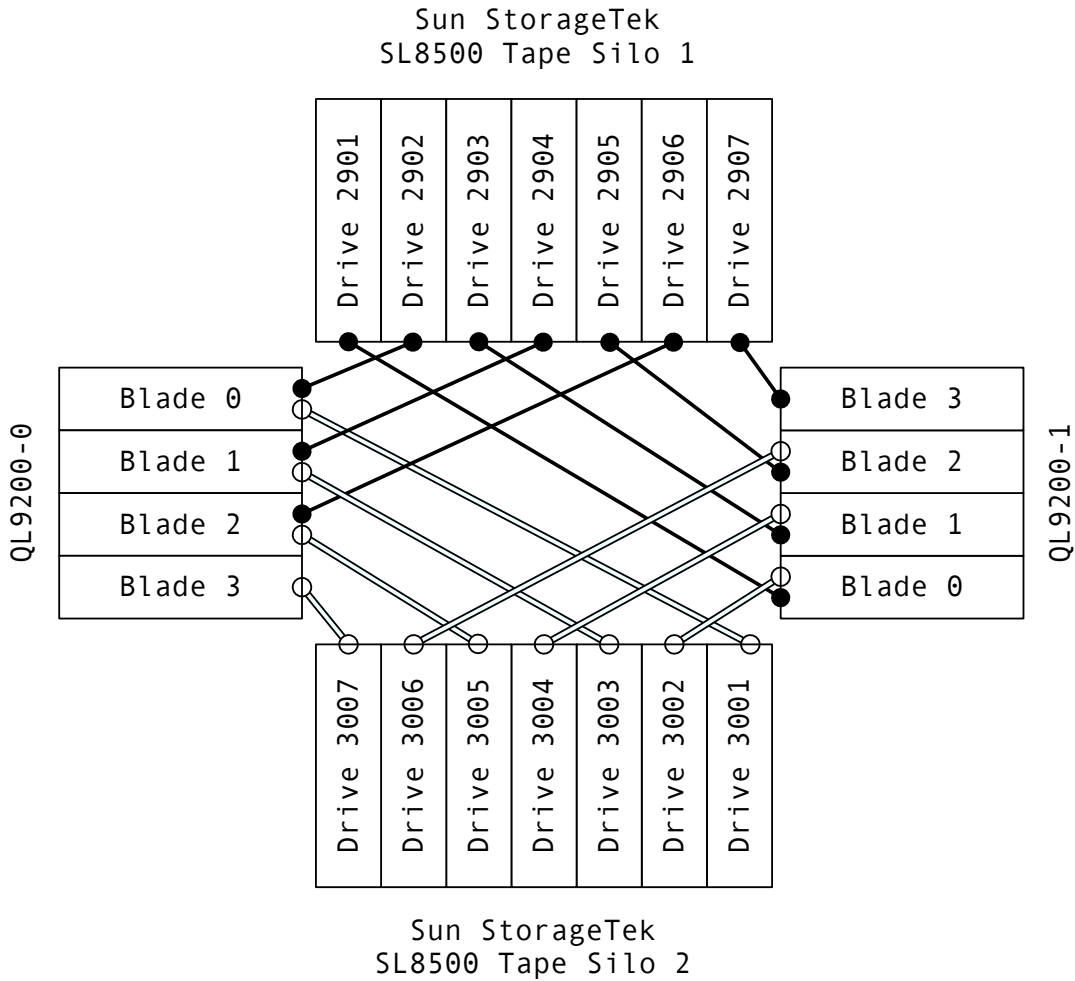


Figure 3: Fiber Channel cabling configuration between the T10000B tape drives and the QLogic 9200 switches

in .vxml⁶. The checksum file acts as a trigger, indicating that VHMS should pull the associated digital files from the Samma server to the SamFS filesystem. The monitoring is done by sending an FTP "ls"⁷ to each Samma Server every 60 seconds. When VHMS does see a checksum file, it spawns an FTP pull of the checksum file and all the associated files from the Samma Server. Once each file is transferred, VHMS runs a checksum on the file, and if it does not match the value in the original checksum file, the transferred file is discarded and retransmitted (specifically a new FTP GET is issued from the So-

laris server to the Samma Server). We have set the retry limit for this file transfer activity to eight tries per file, after which VHMS will shut-down after alerting the system administrators by email. The transmitted files are then stored on the SamFS filesystem, in a directory that is not archived to tape (because those files have not yet been QA'ed).

As mentioned previously, the QA operations are handled by human operators running the MPM software. Once the video files are QA'ed and marked as good, the MPM moves the vxml file

⁷NcFTP is used for performance reasons. See <http://www.ncftp.com/ncftp/>

from the sub-directory to the main directory. VHMS watches this directory for the presence of vxml files, and the moment it finds one, it takes it as an indication that the testimony is good and is ready to move to tape. VHMS then updates the testimony database and moves the various derivatives to another directory that has archiving turned on. We have configured our SamFS archiving daemon to send data to tape when the age of the data is about 20 minutes.

7 Administration and Pitfalls

One of the first issues that we ran into was that there were a lot of checksum errors in the files after transfer to the SamFS filesystem. Upon investigation, we were able to isolate these errors to the 10 Gbps cards, and we have a case open with Sun regarding this issue. The 10 Gbps cards introduced checksum errors into about 25% of our transfers. Disabling the 10 Gbps cards and using only the onboard 1 Gbps cards eliminated the checksum errors completely.

Once the 10 Gbps cards were disabled, network congestion became a concern. To avoid congestion and the productivity delays it would cause, we worked on a strategy of sharing the load amongst the different SamFS and QFS servers. Our solution was to have each QFS server pull a maximum of two FTP streams, one from each of its assigned Samma servers. The SamFS master would play no part in the VHMS operation and would concentrate its effort solely on managing the data being written to tape. The SamFS secondary server would solely be responsible for handling the move from the non-archive directory to the archive directory, in preparation for that testimony's move to tape.

One word of caution: while employing a distributed strategy, it is important to keep an eye on several aspects. These aspects include the available (unused) capacity of the Samma server disks, the load on the Samma server CPUs, the

size of the free tape pool, and the utilization of the SamFS disks. If any of these resources becomes depleted or unavailable, the whole digitization process grinds to a halt.

A more worrisome problem is silent data corruption. While silent data corruption is well documented for disk[18][19], there is little research on bits changing on tape. Regardless, the fear is that when a tape is recycled, the data is moved from tape to disk and then onto new tape, and somehow, a corruption is introduced in the process. We are exploring the checksum flags that SamFS has to alleviate the problem, and also designing custom scripts that will run checksums on all files that have moved due to a tape recycle.

8 Results

After ironing out the initial bugs in the system, we have a digitization rate of about 80 TB per month, or over 5,000 Betacam SP tapes. At this rate we are well on track to completing our project within the 5 year timeline.

9 Acknowledgements

We would like to thank Jon Fields, Ryan Fenton-Strauss, Toan Nguyen and Asbed Bedrossian for their contributions towards the architecture described in this paper.

References

- [1] The Preservation of Magnetic Tape Collections: A Perspective
§Magnetic Tape Longevity, page 6
Image Permanence Institute,
Rochester Institute of Technology,
December 22, 2006
http://www.imagepermanenceinstitute.org/shtml_sub/NEHTapeFinalReport.pdf
- [2] Betacam SP video tape format, Sony Electronics, Inc.
<http://b2b.sony.com/Solutions/subcategory/recordable-media/professional-media/betacam-sp>
- [3] Sun StorageTek SL8500 Modular Library System
Sun Microsystems, inc.
http://www.sun.com/storage/tape_storage/tape_libraries/sl8500/index.xml
- [4] Sun StorageTek T9940B Tape Drive
Sun Microsystems, inc.
http://www.sun.com/storage/tape_storage/tape_drives/t9940/
- [5] Sun StorageTek T10000B Tape Drive
Sun Microsystems, inc.
http://www.sun.com/storage/tape_storage/tape_drives/t10000b/
- [6] Fibre Channel Industry Association
<http://www.fibrechannel.org/>
- [7] Cisco MDS 9140 Fabric Switch
http://www.sun.com/storage/storage_networking/switches_directors/cisco_9140/
- [8] QLogic SANbox 9200 Stackable Chassis Switch
http://www.sun.com/storage/storage_networking/switches_directors/qlogic_9000/
- [9] Front Porch Digital, SAMMA Robot Technical Specs
http://www.fpdigital.com/Products/Downloads/RobotTech_Low.pdf
- [10] Sony J Series Compact Players
(brochure MK10116V1IW04MAR)
Sony Corporation, 2004
<http://ws.sel.sony.com/PIPWebServices/RetrievePublicAsset/StepID/SEL-asset-46298/original/j30\%20brochure.pdf>
- [11] Front Porch Digital, SAMMA Server Technical Specs
http://www.fpdigital.com/Products/Downloads/Solo_Spec_Sheet.pdf
- [12] Sun Storage Archive Manager software
http://www.sun.com/storage/management_software/data_management/sam/
- [13] Sun Fire X4600 M2 Server
<http://www.sun.com/servers/x64/x4600/index.xml>
- [14] JavaSun press release
Sun Microsystems, August 4, 2008
<http://www.sun.com/aboutsun/pr/2008-08/sunflash.20080807.1.xml>
- [15] Sun StorageTek 6540 Array
http://www.sun.com/storage/disk_systems/midrange/6540/
- [16] Sun StorageTek Common Array Manager Software
http://www.sun.com/storage/management_software/resource_management/cam/get_it.jsp
- [17] Sun QFS Software
http://www.sun.com/storage/management_software/data_management/qfs/
- [18] Data Integrity, Bernd Panzer-Steindel, CERN/IT
<http://indico.cern.ch/getFile.py?access?contribId=3&sessionId=0&resId=1&materialId=paper&confId=13797>
- [19] Silent Corruptions, Kelemen Péter
http://fuji.web.cern.ch/fuji/talk/2007/kelemen-2007-C5-Silent_Corruptions.pdf