Helical Scan Reliability: Lessons Learned from the Exabyte 8500

Angela Schuett

November 7, 1996

1 Abstract

The workloads of large database systems such as the EOSDIS global information system require not only high performance but also reliability over repeated reads from tape. To evaluate tape reliability, we have performed extensive measurements of tape error behavior and have discovered several facts that defy conventional wisdom. First, we found that Exatape 8mm helical scan tapes can sustain many more passes than usually thought; most tapes had few errors for up to 10,000 passes, even though the tapes are rated at only 1,500 passes [Exa94]. We also found that hard errors (errors that cannot be corrected by repeated retries at the location) are surprisingly transient; virtually all disappeared on subsequent passes over the tape. We also observed that, while cleaning the tape heads is essential to prevent catastrophic drive failure, cleaning itself may, in some cases, cause hard errors. This paper gives a full description of our experiments and results.

2 Introduction

In this paper we re-examine the reliability of magnetic tape systems, and present some findings that defy conventional wisdom. It has been claimed that helical scan tapes are "inferior" because they can not sustain many passes, as compared with linear tapes. We show that Exatape helical scan tapes can be accessed repeatedly over many more passes than previously thought, that many errors are transient, and that the cleaning operation that prevents long term reliability problems can in some cases cause short-term ones. Tape reliability problems arise when tapes are written or read repeatedly. We argue that an evolving class of applications will require repetitive access to tapes in ways that do not occur in conventional backup applications.

Magnetic tape systems, which typically have the lowest cost per megabyte of any storage media, have found wide use as offline storage (accessible only through human intervention), as a means to backup secondary storage devices. They have also found increasing use as a level of online, or "nearline" storage (accessible through an autoloader or robot arm) for hierarchical storage systems, digital libraries and very large databases. The Earth Observing System Data and Information System (EOSDIS) will be made of a number of very large tape libraries. This system will be supporting database style queries from a large number of users: thousands of scientific users [Far95]. This will stress not only throughput of the system, but reliability, since queries will be spread throughout the system, and difficult to cache.

For scientific workloads, where very large data sets are too expensive to be kept on disk, reliability over repeated reads is important. Ethan Miller, in [Mil95], found that 5% of all files on the National Center for Atmospheric Research (NCAR) mass storage system were accessed more than 10 times. He also found that .02% of tape accesses resulted in media errors. Reagan Moore, in [Moo96], suggests that in the past, because of the low throughput of mass storage systems, scientific researchers tailored their experiments to rarely access the full data set, and to use a compressed version or a subset for most computation. However, as parallel access techniques and faster tape drives increase throughput, new techniques will be used by researchers, including data mining and other database style techniques [MFW+96]. An example given in [Moo96] is the use of remote sensing data from satellites to simulate the results of future policies of water management or land use. Since data from the Earth Observing Satellite will be of a geographic nature, some queries will span hundreds of tapes [DIS96]. Overlapping queries from the large number of users are unlikely to share much locality and will cause tapes to be accessed numerous times, stressing tape reliability far more than for previous scientific workloads.

This paper first provides a brief overview of magnetic tape technology, focusing on causes of tape errors and the difference in typical tape and disk errors, in Section 3. The Exabyte 8500 drive and the diagnostic information available during operation is covered in Section 3.1. Section 4 discusses other tape reliability tests done by the National Media Lab and by Exabyte. Section 5 begins the discussion of our actual tests. The initial test which used the **tar** utility to repeatedly read data from tape was used to familiarize the researchers with the device. Section 6 describes a series of shorter tests that were used to help us understand the buffering and streaming operation of the tape drive. The results of these tests were used to formulate long term repeated read and write tests, described in Section 7. We ran 3 read tests and 2 write tests, each test covering from 1 to 2 weeks of tape drive operation. Error patterns, and the location, frequency and cause of hard errors are covered in this section. Section 8 covers our future work, and conclusions are presented in Section 9.

3 Magnetic Tape Characteristics

Since magnetic disk and magnetic tape use the same technology, error rates are very similar. Both media advertise a corrected bit error rate of around 1 error per 10^{13} bits read. However, since magnetic tape uses a flexible recording media, with contact between the read/write head and the media, and since the drive unit is not sealed, error behavior is different. Disk errors may have several different causes: defects in the recording surface, improper head alignment during seek, abnormal head elevation, and damage to the disk surface caused by a head crash [MT95]. Data can typically be reconstructed after 1 or 2 retries, as in the case of improper head alignment during seek, or it can never be reconstructed, as in the case of damage to the disk surface.

For tapes, however, errors are often caused by debris that clogs the tape head or increases the separation between the head and the tape [Che94]. These types of errors are transient, since the debris may be pushed aside, to another section of tape, or off the tape. Debris may occur from the contact between the head and the tape, as bits of the tape break off and are pushed by the tape head, or debris may be left from the manufacturing process. Cleaning tapes are used to remove debris from the tape head, but no process exists for cleaning the tapes themselves.

Finally, errors can be caused by breakage or flaws in the tape which cause the magnetic recording material to flake off or be damaged. Tapes that are 5-10 years old are often no longer viable because of this problem, and similarly, tapes that have been read and rewound a large number of times may experience damage of this sort because of the stress on the tape of being stretched tight. For a report on degradation of tapes over time see [Bog95].

Transient tape errors may require many more retries to reconstruct data than transient disk errors. For this reason, failure modes are more complicated, and it may be difficult to distinguish a permanent tape error from a transient error. This paper seeks to explore these types of errors and understand how many passes are possible before unrecoverable, tape stress errors occur.

A variety of formats and form factors are in use for magnetic tape. This project used the 8mm helical scan tape format. Helical scan uses a rotating drum which writes tracks diagonally across

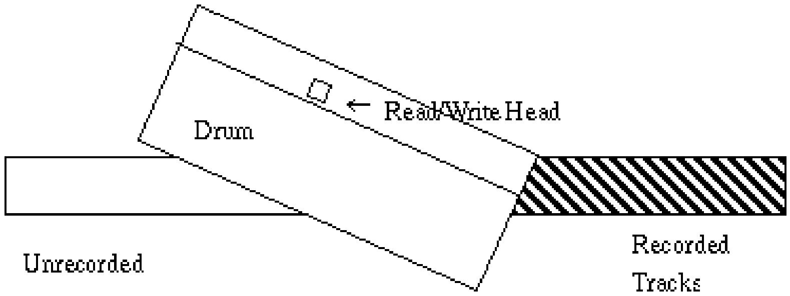

Figure 1: Helical Scan Tape Drive - Side View

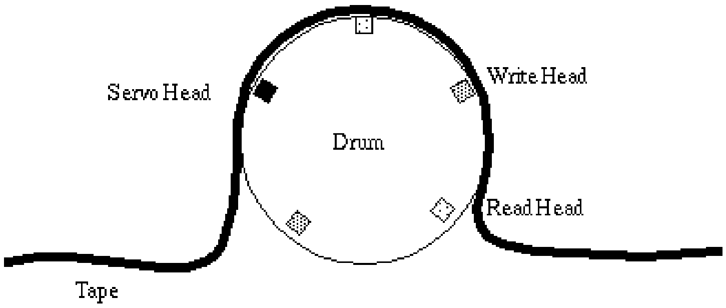

Figure 2: Helical Scan Tape Drive - Top View

the width of the tape, as seen in Figure 1. This setup causes the relative head to tape speed to be high, and the actual tape speed to be low, since both the tape and the heads are moving. The slower tape speed is less stressful for the tape. However, this advantage may be offset by the additional stress caused by wrapping the tape around the drum. Multiple read and write heads increase read and write speed, (see Figure 2) and allow verification of writes, since the read head follows the write head, data may be read after writes without rewinding the tape.

The helical format is in wide use and is valued for its high speed search and high areal density, but is generally considered to be good for fewer passes than serpentine or longitudinal tape. Those formats use a stationary read/write head that writes tracks which extend the length of the tape.

3.1 The Exabyte EXB-8500 8mm Cartridge Tape Subsystem

The EXB-8500 is packaged in the industry-standard 5.25-inch form factor. It uses dual read and write head pairs with helical-scan recording technology enabling it to achieve a transfer rate of up to 500 kilobytes per second with peak transfer rates of up to 4 megabytes per second. To ensure data integrity, it uses a Reed/Solomon error correction code (ECC) and full read-after-write verification. The EXB-8500 includes a SCSI controller and supports the SCSI command set [EXA91].

The EXB-8500 uses a 1 megabyte data buffer for both read and write operations. On a read, bits are read from the tape and transferred to the buffer. These bits are then transferred directly from the buffer to the host. On a write, bits are first written to the buffer and then written to tape from the buffer. The data buffer enables the EXB-8500 to operate in **streaming** or **start/stop mode**. A tape drive is in **streaming** mode if tape media continuously moves forward across the read/write head during operation [Com95]. In **start/stop** mode, not only does tape motion start and stop but the tape also has to reverse direction in order to cue properly. The preferred mode is streaming because start/stop reduces media life, increases read/write head wear and lowers the overall performance of the EXB-8500. Consequently, the EXB-8500 tries to operate in streaming mode as much as possible. However, because of the need to gather tape drive statistics at short intervals during operation, we have observed that the EXB-8500 has generally not been able to maintain streaming mode during most of our tests.

The EXB-8500 provides a parameter, called the **motion threshold**, to allow users to control the starting and stopping of tape motion. The default motion threshold value is 512 kilobytes. On a write, bits are first written to the data buffer. Once the amount of data written to the buffer exceeds the motion threshold value, tape motion starts and bits are written to tape until the buffer is emptied. At this point, tape motion stops again. On a read, tape motion starts when the amount of free space in the buffer exceeds the motion threshold value. Tape motion continues until the buffer is full. Users can set the motion threshold value to have better control over tape motion. Section 6 describes our experiments that varied the motion threshold value.

The EXB-8500 provides another parameter, called the **reconnect threshold**, to allow users to control the rate of disconnects and reconnects between the EXB-8500 and the host. Its default value is 512 kilobytes. In streaming mode, the EXB-8500 disconnects from the host when the data buffer becomes full but continues to write data from buffer to tape. When the amount of free space in the buffer exceeds the reconnect threshold value, the EXB-8500 will reconnect to the host to receive more data. On a read, the EXB-8500 disconnects from the host when the data buffer becomes empty but continues to read data from tape to buffer. When the amount of data read to the buffer exceeds the reconnect threshold value, the EXB-8500 reconnects to the host to transfer data from the buffer. Our experiments varying the reconnect threshold are also described in Section 6.

To be able to gather error and other relevant statistics, request sense data is obtained from the EXB-8500 during the tests. Request sense data relevant to our tests are as follows:

- **Sense key:** specifies any error or significant status sensed during an operation, e.g., *Not Ready, Medium Error, Hardware Error, Illegal Request.*
- **Additional sense code:** provides additional information about the sense key value returned.
- **Additional sense code qualifier:** also provides additional information about sense key value returned.
- **Fault symptom code:** used to provide more specific information about events that occurred during an operation such as hardware and software errors.
- **Read/write data error counter:** keeps track of the number of read errors that occur during a read operation and the number of write errors that occur during a write operation.
- **Remaining tape:** indicates the amount of tape remaining in 1 kilobyte blocks.
- **Read/write retry counter:** keeps track of the number of read/write retries performed.
- **Underrun/Overrun Counter:** is incremented whenever streaming is not maintained. For writes, an underrun is when the buffer is empty. For reads, an overrun is when the buffer is full.

For data integrity, the EXB-8500 uses read-after-write checking. Once a block of data is written to tape, it is immediately read and checked to see if the block was correctly written. If an error is detected, the block is rewritten without repositioning the tape. Instead of writing to the same tape location, the tape is kept streaming so the block is rewritten in the next available location on tape. When reading from tape, the EXB-8500 uses the data from an error correction code (ECC) to recover from read errors.

Read/write errors are classified into **soft** and **hard** errors. **Soft errors** are recoverable errors. By using ECC and retries, the EXB-8500 is able to recover from a large number of errors. **Hard errors** are non-recoverable errors. They may be hardware errors, tracking errors, or errors caused by not being able to read a block within the specified number of retries. The default number of retries for a read operation is 10, for a write operation 5 retries are the default. A hard error causes the EXB-8500 to abort the current read or write operation.

We ran all of our tests using three Exabyte EXB-8500 8mm Cartridge Tape Subsystems connected to three different HP 9000/700 series workstations running HP-UX V9.03. We used Exatape brand 8mm 112-meter cartridges. Each tape has a capacity of approximately 5 gigabytes. We used cleaning cartridges from Exabyte. All equipment and materials were chosen based on availability.

4 Related Work

"Input/output has been the orphan of computer architecture [PH96]." This statement applies doubly to magnetic tape. However, there are several durability and reliability studies in the public domain, albeit with an emphasis on the backup workload, rather than the random access repeated read workload of the EOSDIS system.

A study performed by the National Media Lab examined many aspects of the Exabyte 8505 tape drive [JS94]. This drive is very similar to the Exabyte 8500 drive used in our studies, but the 8505 is specified to have a higher mean time before failure. Of most interest to our work are the tests that examined tape durability. The durability test consisted of 300 read cycles at 3 locations on each cartridge. This test found an average of .0005 corrections per kilobyte of data read. This is consistent with our results for the first 300 passes over a tape. These tests were performed with two Exabyte tape drives on 4 Exabyte tapes (Exatape) and one Sony video tape.

Another durability test, performed by Exabyte [EXA93] took a different approach. Exabyte 8200 drives were programmed to dwell on a single track (8 kilobytes of data). Tapes from 9 manufacturers were tested to see how long before the samples degraded to 50% of initial readback amplitude. The Exabyte tape performed the best, lasting 28 days. The poorest sample only lasted a fraction of a day. It is difficult to generalize from this data to a model of tape reliability, because it is not clear how additionally stressful it may be to cause a tape to dwell on a single track. In addition, we believe it is more useful to have a measure of errors than of amplitude. This paper also includes graphs that show individual errors' location by track for the first pass of various tapes. We found this to be consistent with our data.

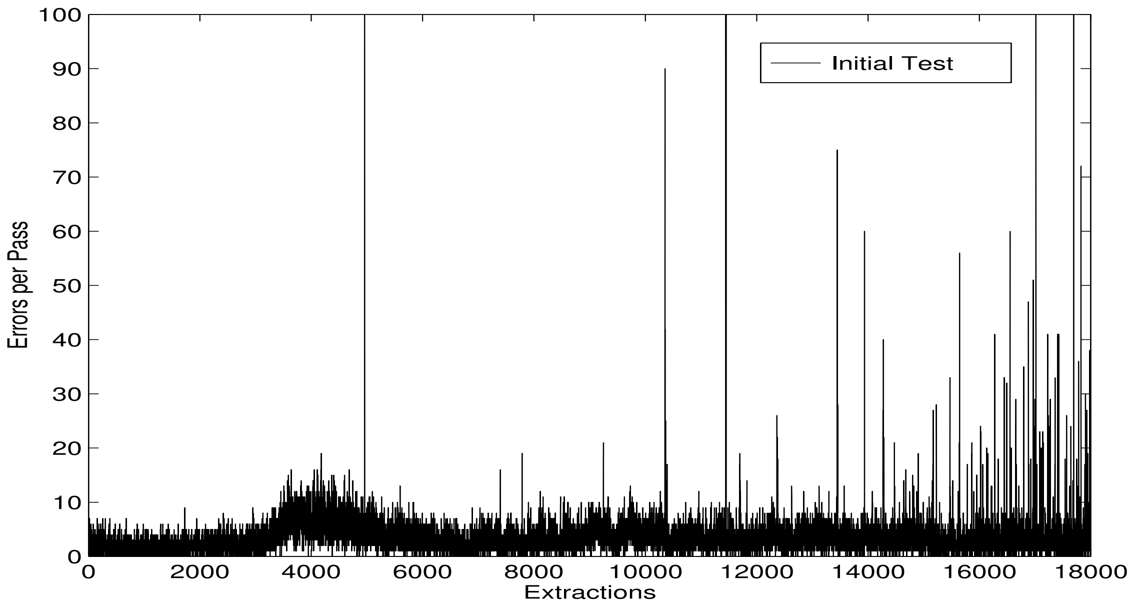

Figure 3: Read errors for all 18,000 extractions done in the initial test. Error rate remains fairly flat, but with increasingly frequent bursts after 8000 extractions. Note: this test does not differentiate between soft and hard errors.

5 Initial Test

The main goal of this test was to familiarize us with the operation of the drive, and to see what types of errors and error patterns occur when reading a magnetic tape for thousands of passes without cleaning the read/write heads of the tape drive. The test consists of using the UNIX **tar** utility to write three postscript files to tape, totaling 1,020,413 bytes. The test then proceeds to extract all three files from the tape repeatedly using the UNIX **tar** utility. After each 1 megabyte extraction, request sense data are obtained from the tape drive and logged. We ran the test for 18,000 extractions, which took around 160 hours.

Figure 3 shows the overall read error pattern for all 18,000 extractions. Each data point represents the number of **correctable** read errors obtained after each three-file extraction, so the scale is errors per megabyte. The errors are bursty rather than steadily increasing, but are fairly flat. The majority of extractions have a low level of errors, which doesn't grow much throughout the test. However, the extractions that have high error levels become more common throughout the test.

Figure 4 shows a close-up of the graph for the first set of 1,000 extractions. The absolute error

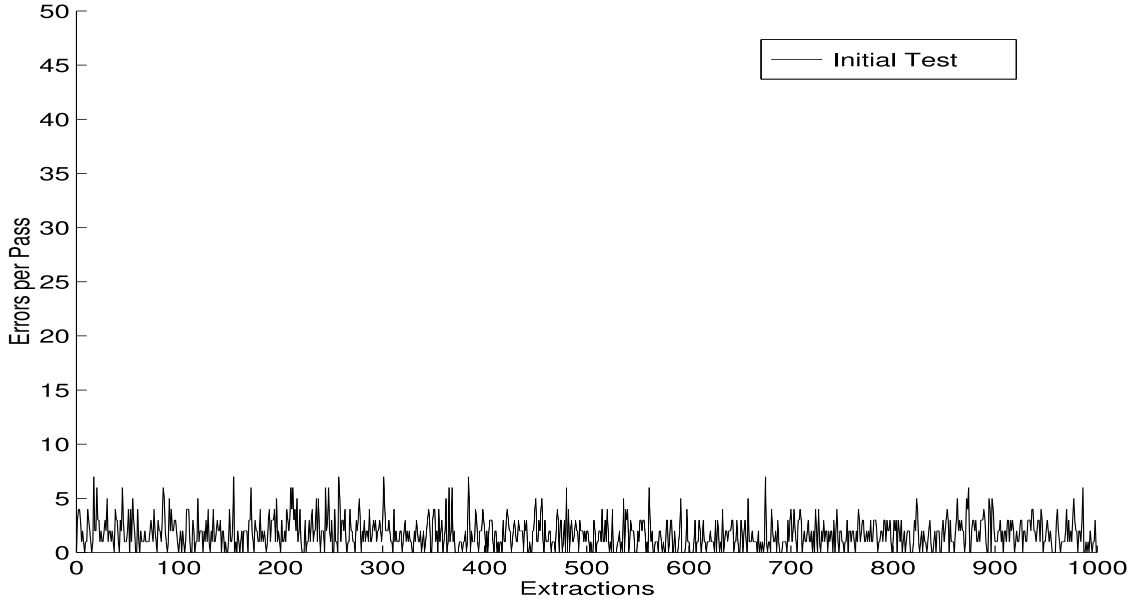

Figure 4: Read errors for passes 1 to 1000 of the initial test. In these early passes, error rates are very low.

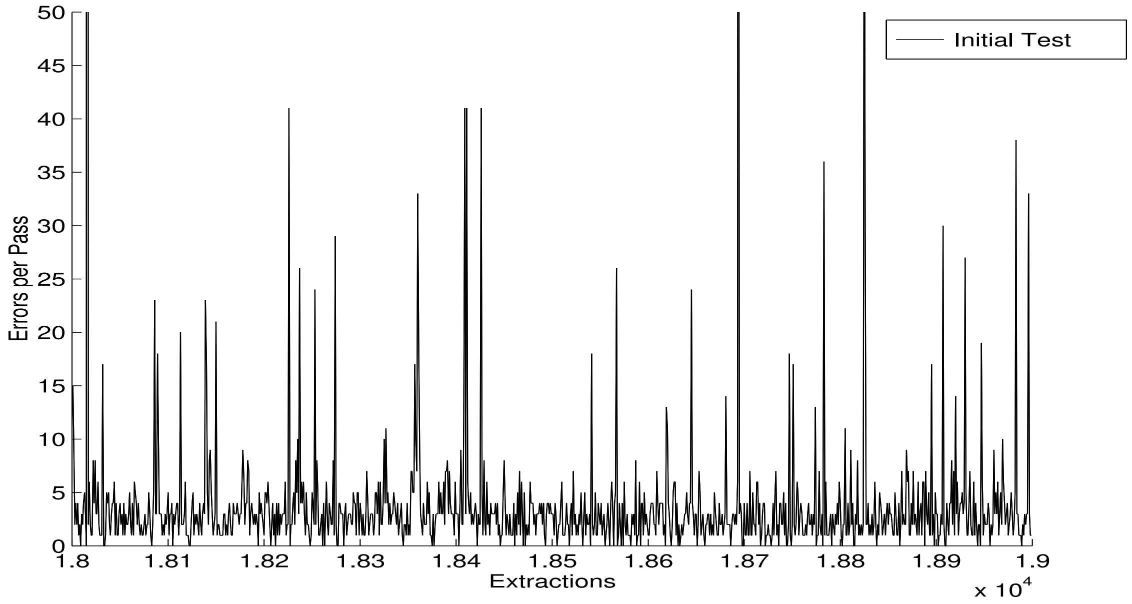

Figure 5: Read errors for passes 18,000 to 19,000. Bursty errors have increased significantly compared to passes 1 to 1000.

values are quite low, ranging from 0 to 9 errors per megabyte. Figure 5 shows the data for the last set, passes 17,000 through 18,000. The graph is cut off on the y-axis at 50 errors per extraction, to show detail, but some of the bursts of errors reach above 400 errors per megabyte. We gathered limited error information for this test, so data is not available on retries, hard errors, or how soft errors are clustered within the extraction. These problems were corrected in future tests.

We were surprised by the generally flat rate of soft errors since we expected an increase in soft errors as the number of tape passes increases. However, after about 8,000 extractions, the frequency of error bursts began to increase. We were also surprised by the small peak around 4,000 extractions, i.e., an increase in errors followed by a decrease.

Given the data that we gathered from this initial test, we wanted to have a better understanding of the error bursts that are seen and we also wanted to know where on the tape the errors actually occur. Do error bursts recur on the same locations on the tape or are they randomly distributed across the tape? Do soft errors become hard errors as the tape becomes worn? This led us to the definition of the next set of tests that allow for finer granularity in error data gathering and better control over read and write operations.

6 Understanding Tape Drive Performance and Error Behavior

The initial test raised questions for us about the low-level operation of the drive. This series of tests was designed to gather error information at a smaller granularity. Since request sense data is only reported after each read or write operation, we needed to use much smaller individual blocks. Instead of gathering request sense data at the end of each pass of the tape, we wrote and read a number of smaller blocks, gathering data after each block was read or written.

For this test, the block size is set to 4 kilobytes. We wrote 1,000 consecutive 4 kilobyte blocks, using 4 megabytes of the tape. The test reads these blocks repeatedly. Figure 6 shows the soft errors for the first 50 passes over the data. Each pass has its own line on the y-axis of the graph. Kilobytes read is shown on the x-axis, showing where errors occurred in the tape, and the number of errors per 4 kilobyte block is shown on the z-axis. The most noticeable feature of this graph is the periodicity. The errors occur approximately every 512 kilobytes of data read.

This periodicity is primarily due to the data buffering done by the EXB-8500, and the way the errors are reported. When the first 4 kilobyte read request is received by the drive, it does not just

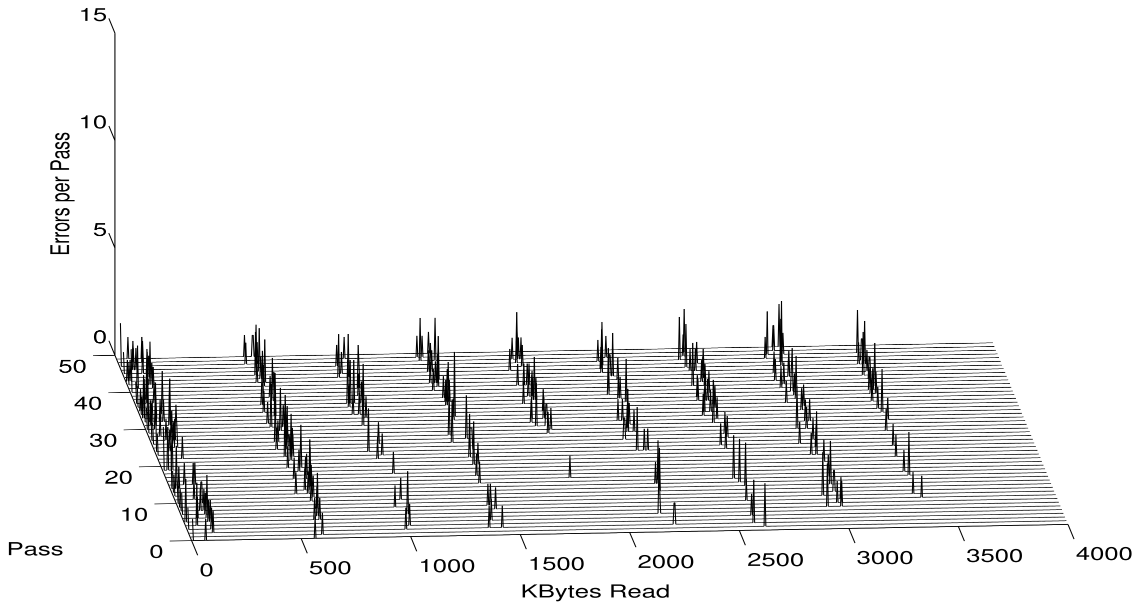

Figure 6: Repeated Reads, 4K block size. Shows the errors per kilobyte read for passes 0 to 49, where each pass reads 1000 4K blocks. Error clusters every 512K are caused by data buffering and error reporting protocols.

read 4 kilobytes of data from the tape but fills up the data buffer to the current threshold value. For this test, the threshold was set to the default, 512 kilobytes. If the EXB-8500 is in streaming mode, it does the following: when the data buffer is empty, it disconnects from the host but starts reading from the tape. When more than 512 kilobytes of data have been transferred from tape to buffer, it reconnects to the host to begin the transfer of data from buffer to host. Therefore, the request for the first 4 kilobytes of data starts the transfer of data from tape to buffer. Subsequent requests for data will be satisfied using data already read into the buffer. Read errors occur and are logged only when data is actually being read from the tape. Read requests that are satisfied using data already read into the buffer do not cause any new read errors to be logged. Therefore, as shown in Figure 6, read errors occur only at the beginning of each 512 kilobytes set of data since it is at this point that actual reading from tape occurs. Subsequent requests are satisfied using data already read into the buffer thereby causing no errors. Unfortunately this also means that the location of errors on tape is not discernible at a fine-grain.

We tested the hypothesis that the periodicity was caused by data buffering by changing the motion and reconnect thresholds. Recall that the motion threshold specifies how much space must be in the buffer (or, in the case of writes, how much data must be in the buffer) before starting tape

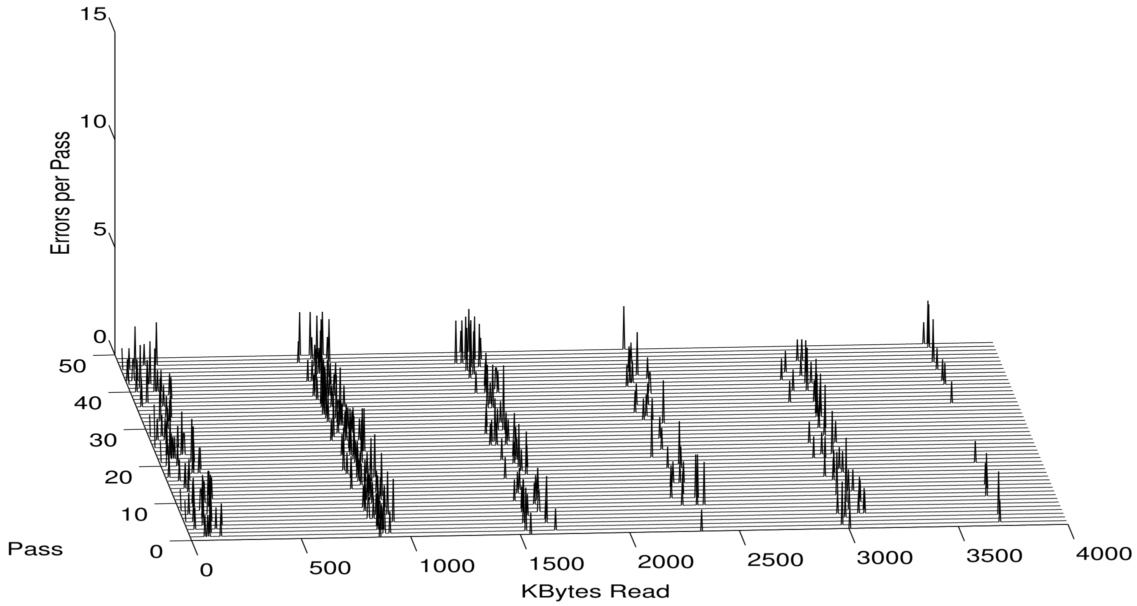

Figure 7: Repeated read test with 768K motion and 128K reconnect threshold. Shows the errors per kilobyte read for passes 0 to 49, where each pass reads 1000 4K blocks. Errors occur at a period of 768K because of buffering based on the 768K motion threshold.

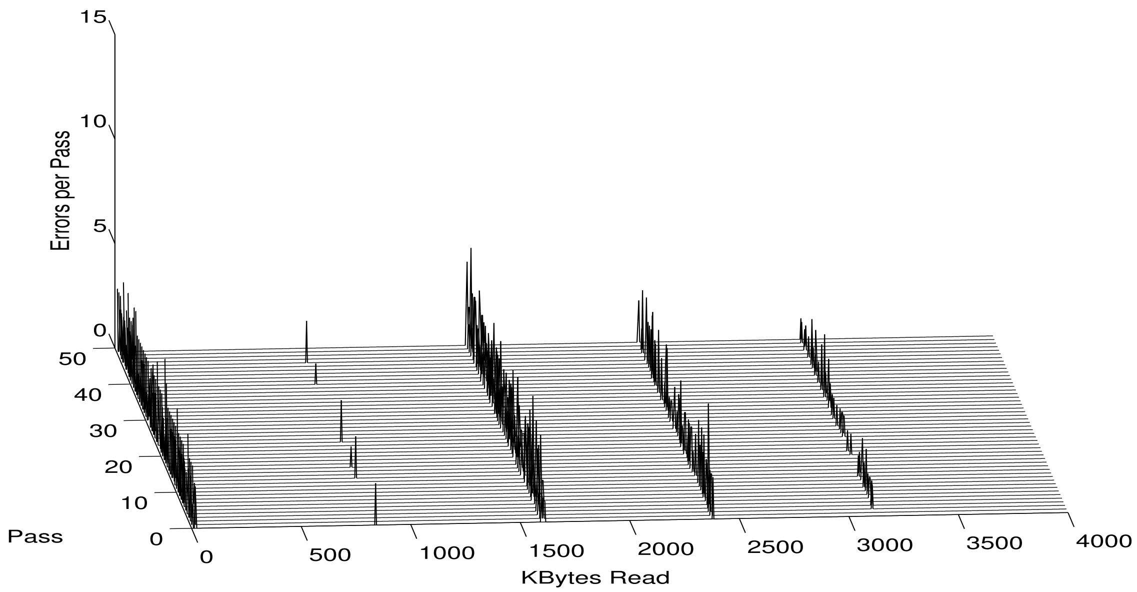

Figure 8: Repeated read test with 768K motion and reconnect thresholds. Shows the errors per kilobyte read for passes 0 to 49, where each pass reads 1000 4K blocks. Errors again occur at a period of 768K.

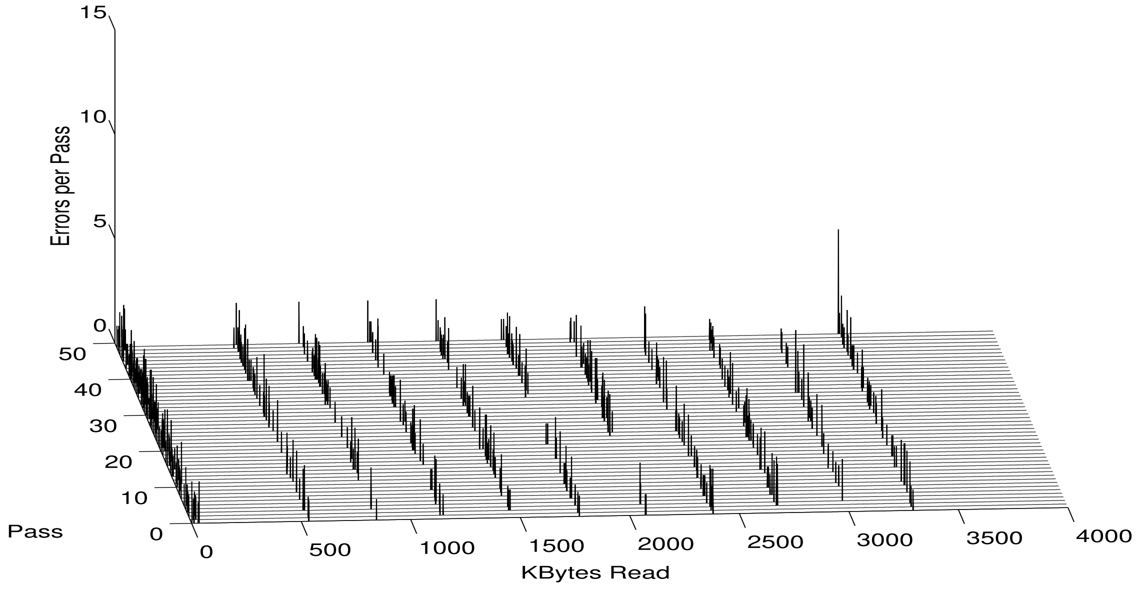

Figure 9: Repeated Reads, 1 K block size. Shows the errors per kilobyte read for passes 0 to 49, where each pass reads 4000 1K blocks. Despite more frequent error reporting, the graph is very similar to Figure 6.

motion, and further reads. The reconnect threshold specifies how much data (or, in the case of writes, free space) must be in the buffer before reconnecting to the host computer to transfer that data out of the buffer. Figure 8 shows error data from 50 passes of a read test using motion and reconnect thresholds both set to 768 kilobytes. It shows that read errors occur at approximately 768 kilobyte intervals, as expected. Figure 7 shows error data from 50 passes of a test using a motion threshold set to 768 kilobytes, and a reconnect threshold set to 128 kilobytes. The period is approximately 768 kilobytes. Apparently the motion threshold is the dominant factor in our case.

We also ran tests using larger and smaller block sizes. Larger block sizes have less overhead in bus traffic and error logging, and use the tape drive more efficiently. In normal operation, larger block sizes are more efficient. However, because request sense data can only be gathered after each write or read is complete, smaller block sizes offer the advantage of finer granularity data.

Figures 9 and 10 show error data for read tests using 1 kilobyte and 128 kilobyte block sizes, respectively. In both cases 4 megabytes of the tape were used; the 1 kilobyte test repeatedly read 4,000 consecutive blocks, and the 128 kilobyte test repeatedly read 32 consecutive blocks.

The 1 kilobyte test shows a granularity very close to that of the 4 kilobyte test, but it ran

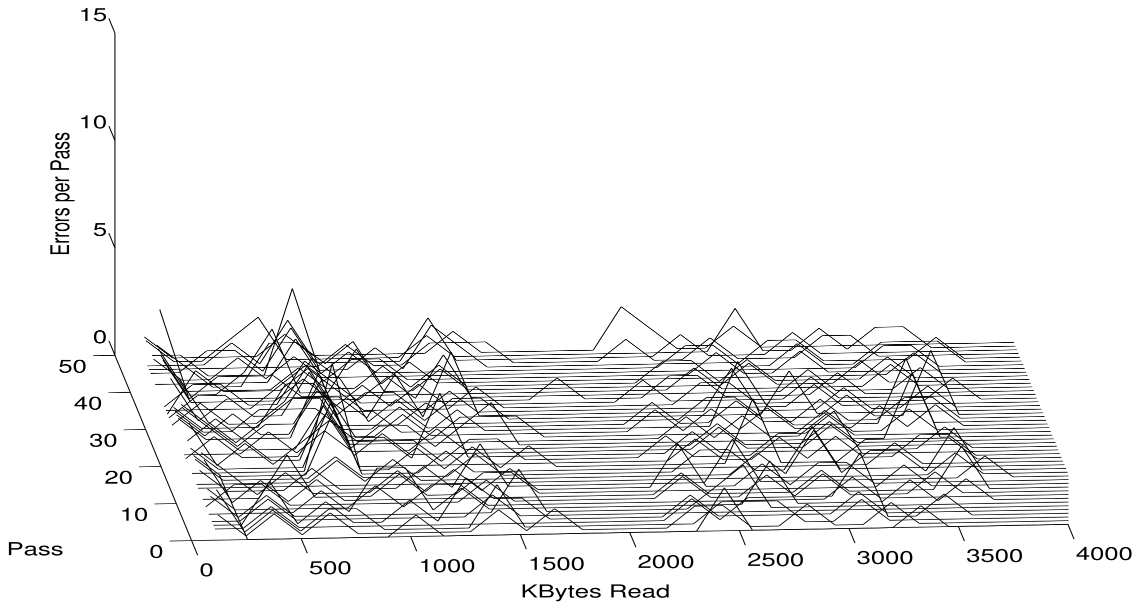

Figure 10: Repeated Reads, 128K block size. Shows the errors per kilobyte read for passes 0 to 49, where each pass reads 32 128K blocks.

considerably slower, partly because it had many more individual requests to send over the bus, and partly because so much more error information was being sent back to the host computer. The 1 kilobyte test took 38 hours to complete 1,000 passes, the 4 kilobyte test took 17 hours, and the 128 kilobyte test took 10 hours to complete 1,000 passes. Keep in mind that all three tests read the same amount of data. While the 128 kilobyte test ran much faster, we decided that the granularity of error information was too coarse for our purposes. For all of our long term read and write tests, described in the next section, we decided to use a 4 kilobyte block size, and the default 512 kilobyte motion and reconnect thresholds.

We also ran similar tests for repeated writes. Figure 11 shows the write errors found for the second set of 50 passes. Contrary to our expectations, errors start occurring after 1 megabyte of data have been written instead of after 512 kilobytes. Exabyte Technical Support staff said that this is probably due to the fact that tape is kept slack when not being used. Before the drive can start writing to tape, the EXB-8500 must first wind the tape around the drum, locate the beginning of the tape, and write a logical beginning of tape (LBOT) pattern. By the time the EXB-8500 has finished with these operations, the host has already filled up the buffer. It is only at this point that tape motion actually occurs therefore it is only at this point that errors are

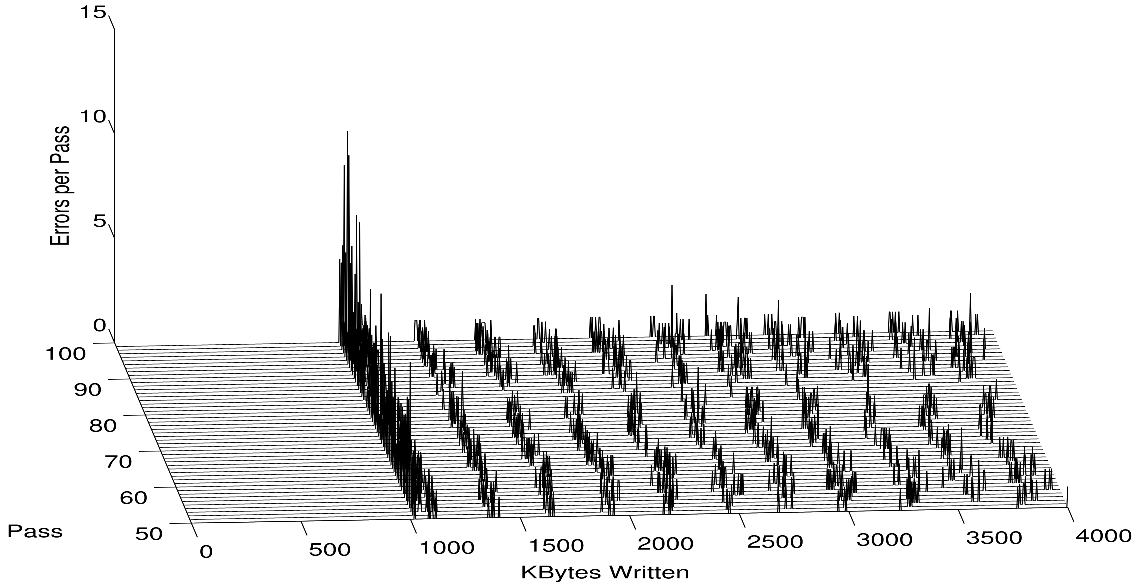

Figure 11: Repeated writes, 4K block size. Shows the errors per kilobyte written for passes 50 to 99, where each pass writes 4000 1K blocks. Errors are not reported until after 1M buffer is filled.

detected.

Figure 12 shows the write errors for a repeated write test with a 4 kilobyte block size and motion and reconnect thresholds set to 768 kilobytes. Unlike the situation we saw in the read tests, the period between errors has not changed significantly compared to the period shown in Figure 11. This leads us to believe that either the threshold values are not used during write operations or that factors other than the threshold values affect the periodicity of write error behavior.

These short term repeated read and write tests gave us new insight into the behavior of the Exabyte 8500 tape drive. The effects of the data buffer on error reporting, the impact of block size on the time to run a test, and the impact of the motion and reconnect threshold were all seen in these tests. This helped us to formulate our next series of tests, the long term read and write tests.

7 Long Term Read and Write Tests

The purpose of the long term read and write tests was to gather information on what happens when a tape is worn out. Information was gathered on soft errors, hard errors and retries. We varied the tests over a number of parameters, including: which of the three drives the test was run

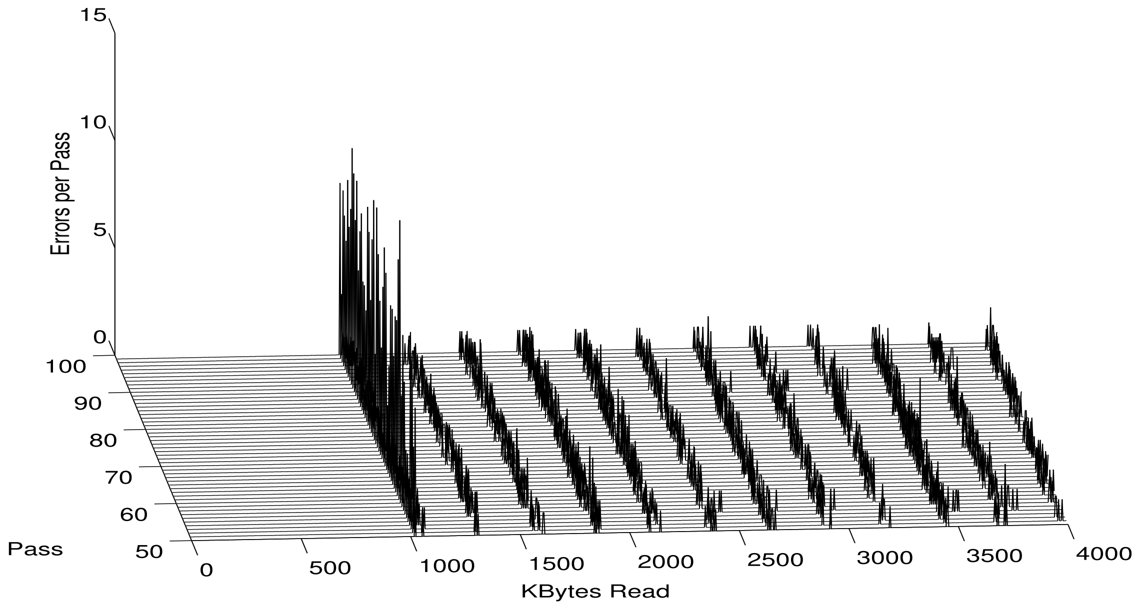

Figure 12: Repeated writes, 768K motion and reconnect threshold. Shows the errors per kilobyte written for passes 50 to 99, where each pass writes 4000 1K blocks. Despite the changed thresholds, this graph is very similar to Figure 11.

on, how often cleaning operations were run, if at all, and whether the test repeated read or write operations. Because tests take such a long time to run, it was not possible to run every variety of test. We decided instead to rapidly probe the design space.

Table 1 shows some of the parameters for the 7 long term read and write tests. All tests were run with motion and reconnect thresholds set to 512 kilobytes. All tests used a 4 kilobyte block size, with 1,000 consecutive blocks, so that 4 megabytes of data were read or written in each pass. The tests differed in what drive they were run on, whether they were repeated read or repeated write tests, and whether tape head cleaning was used. Tests 1A and 1C were run without any tape head cleaning. This type of operation is not recommended by Exabyte, but we wanted to be able to separate the effects of cleaning from the effects of normal tape wear. Read tests 1B and 2C had tape head cleanings every 2,000 passes with an Exabyte 8mm 3c cleaning cartridge. This works out to one cleaning approximately every 28 hours, or every 8 gigabytes of data read. Write test 2A had tape head cleaning every 1,000 passes, which is one cleaning approximately every 26 hours, or every 4 gigabytes of data written. Exabyte recommends cleaning after every 30 hours of tape motion. Tests were run until drive or tape problems became catastrophic. Section 7.1 gives a high-level view of the results of the long term tests, showing both correctable and hard errors.

Test Name	Tape Drive	Test Type	Cleaning	Passes	Hours
Test 1A	A	Read	No	13,000	220
Test 2A	A	Write	Every 1,000 passes	8,000	190
Test 1B	B	Read	Every 2,000 passes	10,000	170
Test 2B	B	Read	Every 2,000 passes	14,000	246
Test 3B	B	Read	Every 2,000 passes	7,000	126
Test 1C	C	Write	No	7,000	150
Test 2C	C	Read	Every 2,000 passes	12,000	200

Table 1: Parameters for long term read and write tests.

Section 7.2 examines the results at a finer granularity, discussing the types of hard errors, retry patterns, and the performance implications of errors.

7.1 Overview of Results

Results of the 7 long term tests are shown in Figures 13, 14 and 15, divided according to which drive the test was run on. Test names indicate the drive the tests were run on, Test 1A ran on drive A, and the order in which they were run, Test 2A ran after Test 1A. The y-axis shows errors per kilobyte, which was calculated by totaling errors per pass and dividing by 4,000, since each pass read or wrote 4,000 kilobytes of data. All errors shown are soft, or correctable errors, except for the specifically marked hard errors. Hard errors do not have an errors per kilobyte value, but are shown on the y-axis at the value of the last soft error before the hard error occurred. These indicate hard errors, either a tape medium error, or a hardware error. These errors will be examined individually in Section 7.2. These graphs show a high level view of the data, each point indicating the average of 50 passes.

These three graphs show which drive a test ran on is a more dominant effect on soft errors than other parameters. Note especially the difference between the tests run on drive A, shown in Figure 13 and the tests run on drive C, shown in Figure 15. We hypothesize that drive C has "better" read and write heads than the other two drives. This may be due to differences in manufacturing, or to head wear that occurred before of our study.

An interesting avenue of research is to examine how error behavior changes as tape heads are

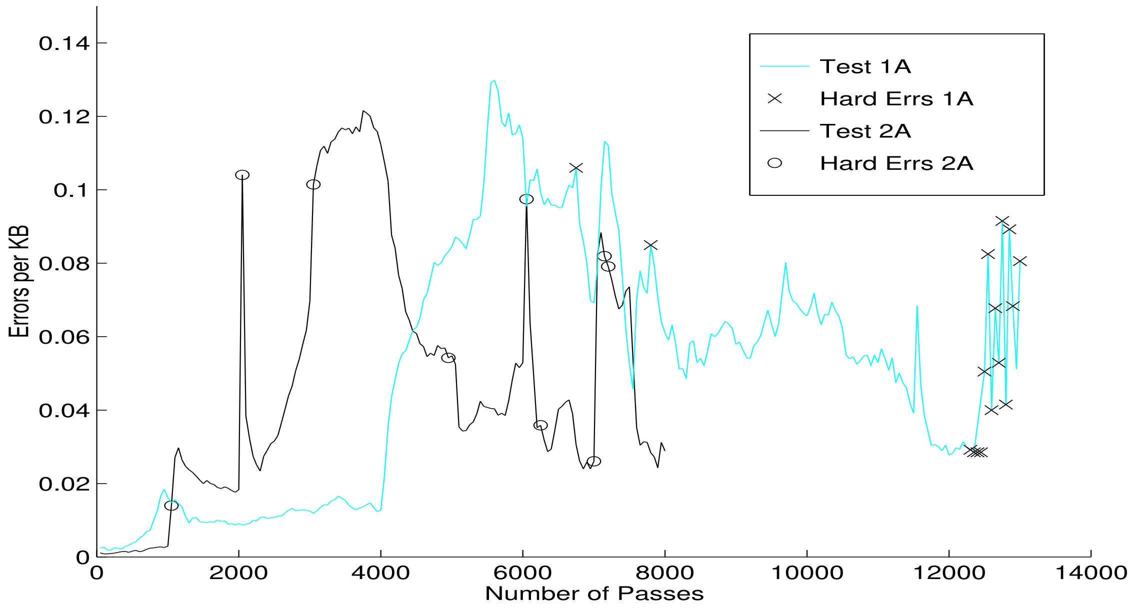

Figure 13: Read test 1A and write test 2A, performed on drive A. Read test 1A had no cleaning, write test 2A had cleaning every 1,000 passes.

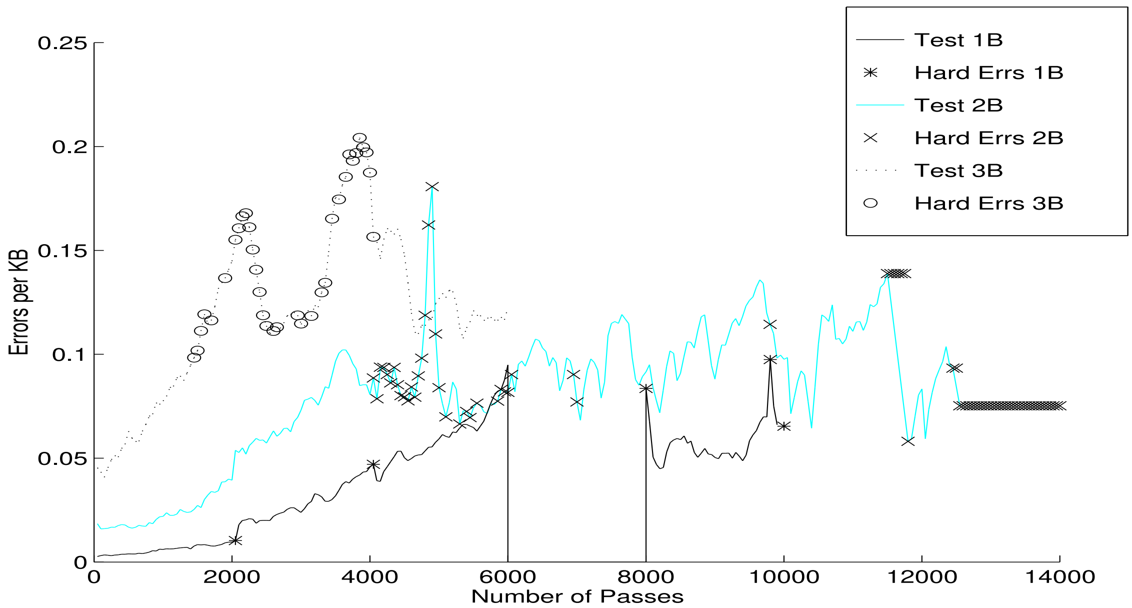

Figure 14: Read test 1B, 2B and 3B, performed on Drive B. All three tests had cleaning every 2,000 passes. The data for passes 6,000 to 8,000 of test 1B was lost in a disk crash.

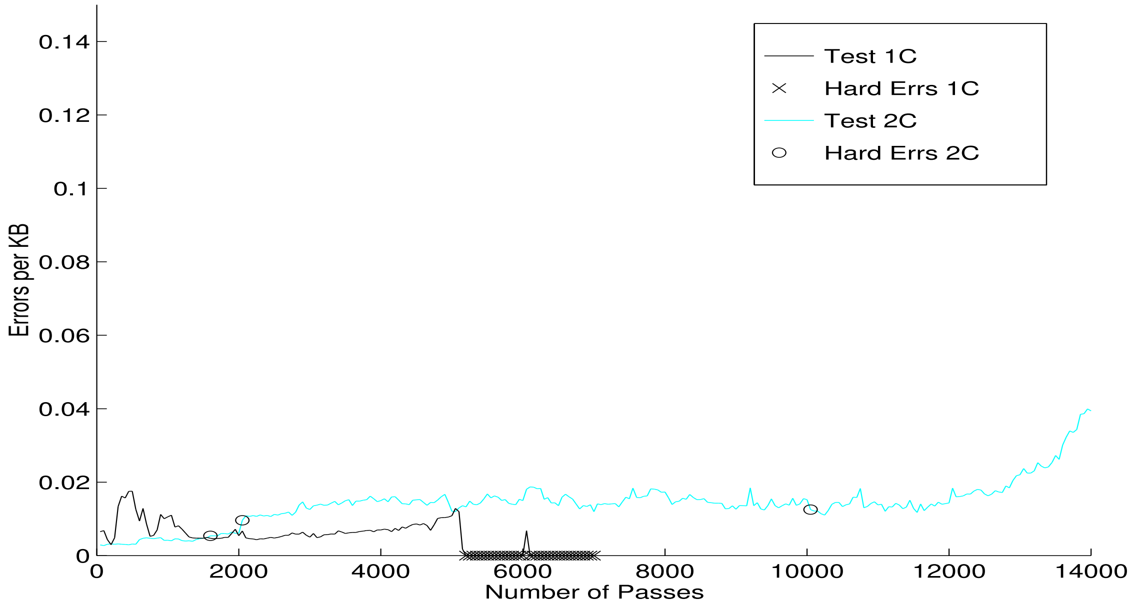

Figure 15: Read test 2C and write test 1C, performed on Drive C. Read test 2C had cleaning every 2,000 passes, write test 1C had no cleaning.

worn. Figure 14 shows three tests which were run on drive B. Error performance worsens with each successive test. Soft errors start at a higher rate, and hard errors occur sooner. We did not design our tests to measure head degradation, but clearly the state of the tape head is a very important factor in error rates, and characterizing this is an interesting area for future work.

Also notice that soft error rates are not steadily increasing. This is especially visible on Figure 13. It may be that debris that has accumulated reaches a mass large enough to cause it to dislodge from the head, improving read performance and causing soft error rates to drop. Or, debris may be pushed off the side of the tape. The National Media Lab study [JS94], which ran tapes for 300 passes, also noted unexplained fluctuations in error rates, which they attribute to debris.

7.2 Analysis of Long Term Tests

The aspect of error behavior that is most visible to magnetic tape users is hard errors. Since our tests ran for so many passes we were able to gather a significant amount of information on hard errors. We found several surprising results. First, hard errors occur later in the life cycle than we expected. Only one of our tests had a hard error before 2,000 passes over the data. Also, hard errors were most often transient, occurring for a few blocks within a pass, and then not re-occurring.

Test Name	Cleaning	Pass	Hard Error Description	Number of Hard Errors
Read Test 1A	None	7,729	Uncorrectable block encountered	4
		12,289	Unable to achieve tracking	50,000+
Read Test 1B	Every 2,000	2,000	Uncorrectable block encountered	36
		4,000	Uncorrectable block encountered	14
		8,000	Uncorrectable block encountered	27
		9,757	Uncorrectable block encountered	1
		10,000	Uncorrectable block encountered	4,000+
Read Test 2C	Every 2,000	1,594	Already at blank tape	1
		2,000	Uncorrectable block encountered	6
		10,000	Uncorrectable block encountered	2
Write Test 1C	None	5,107	Servo hardware error	50,000+
Write Test 2A	Every 1,000	1,000	LBOT write failure	9,000
		2,000	LBOT write failure	39,000
		3,000	LBOT write failure	1,000
		4,923	Retry limit exceeded	390
		6,000	LBOT write failure	2,000
		6,223	Write EOD failure	1
		6,977	Rewrite threshold exceeded	395
		7,139	Rewrite threshold exceeded	612
		7,157	Rewrite threshold exceeded	608

Table 2: This chart lists hard errors in 5 of the long term read and write tests. Notice the cleaning related errors which occur after each periodic cleaning. Also notice the catastrophic errors which ended read test 1A and write test 1C.

Cleaning also affects the hard error rate. All of these results are shown in this section.

Table 2 enumerates all of the hard error clusters found throughout the long term read and write tests. Tests 2B and 3B have too many hard errors to enumerate in this format. The first column gives the test in which the errors occurred, the second column gives the pass number at which the first error in the group occurred, and the third column gives the descriptive name for the error. The last column gives the number of hard errors in the group. We refer to error groups because most hard errors were not single errors, but persisted over several blocks within a pass, and sometimes over several passes. Recall that each pass reads or writes 1,000 blocks, giving a possible 1,000 hard errors per pass. Hard errors that occurred during write tests cause the rest of the write to be aborted, so if the error occurs at the beginning of the pass, 1,000 block writes will have a hard error.

When the number of hard errors ends with a "+", this indicates that the test was ended because of the high number of hard errors. These are considered catastrophic hard errors.

7.2.1 Errors caused by lack of cleaning

Catastrophic hard errors, hard errors that persist over a large number of passes, were seen in read test 1A, read test 1B and write test 1C. Read test 1A and write test 1C had no tape head cleanings during the test. We believe that the buildup of debris on the tape drive head caused the catastrophic errors in these tests. The error type for test 1A was "Unable to achieve or maintain tracking", for test 1C it was "Servo hardware failure". In both cases, the problem persisted when a new cartridge was inserted and was not solved by cleaning the drive with a standard Exabyte cleaning cartridge. Exabyte technical support said that problems of this type occur when the drive is not cleaned often enough, because debris can be burnished into the tape drive head, forming a durable layer that requires a special cleaning tape to remove it. The problem was solved by cleaning both drives with a "Harsh Environment Maintenance Cartridge", supplied by Exabyte. We do not know why test 1C failed after only 5,000 passes, while test 1A failed after 13,000 passes. However, it is clear that a regular cleaning schedule of some type is necessary for uninterrupted operation.

7.2.2 Errors caused by cleaning

Regular tape drive cleaning was done every 2,000 passes for read tests 1B, 2B, 3B, and 2C, and every 1,000 passes for write test 2A. In all three of these tests, hard error groups were often seen directly after cleanings. For example, read test 1B had hard errors following all but one of its cleanings, at passes 2,000, 4,000, 8,000 and 10,000. Only the cleaning at pass 6,000 did not cause a hard error. Lance Blumberg, an engineer at Exabyte said that this was caused by a contamination problem in certain lots of cleaning cartridges. The new cleaning cartridge "Exabyte Premium Cleaning Cartridge" is said to be free of this problem. An area of future work is to conduct tests which use this cleaning cartridge, to see how error rates change.

7.2.3 Retries

Hard errors that are not directly attributable to cleaning or lack of cleaning, may be caused by transient debris, tape wear, or a combination of both. Since the most common type of hard error is

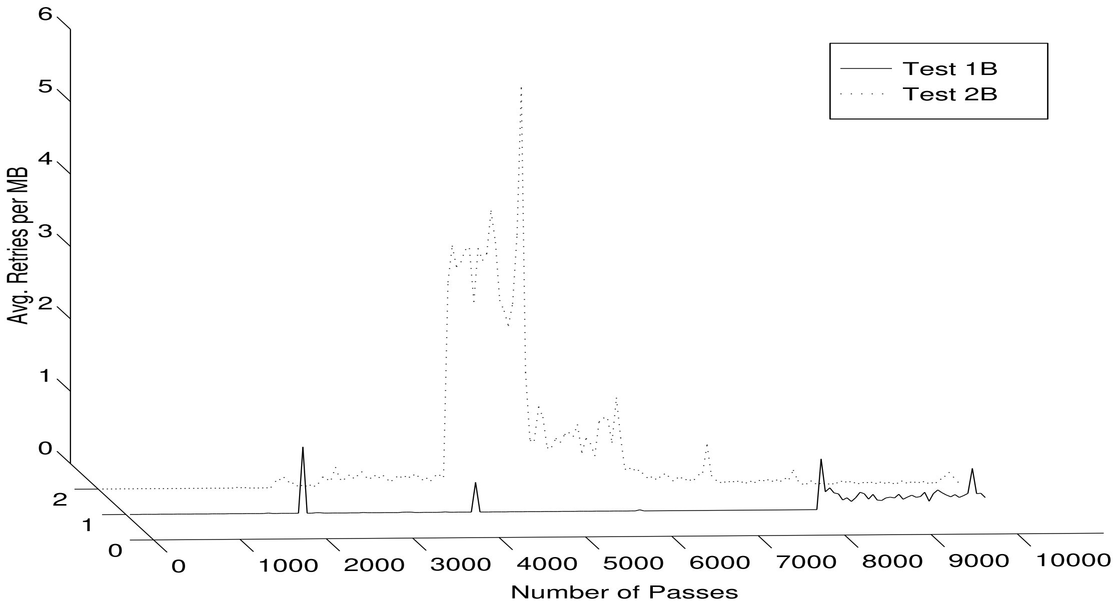

Figure 16: Average retries per MB for read tests 1B and 2B. The rising retry rate in test 2B from passes 2,000 through 4,000 predicts the hard errors which occurred from passes 4,000 through 6,000.

caused when the retry limit is exceeded, monitoring the retry rate may be a useful way to predict when a tape is too worn, allowing data to be recopied safely to a new tape. Figure 16 shows the retries for read tests 1B and 2B. Test 1B had only transient hard errors, caused by the cleaning tape contamination. The retry rate for test 2B begins to rise after 2,000 passes, and hard errors began to occur after 4,000 passes. More study is necessary to provide a model of retries, but our work does indicate that the retry rate may be a more accurate predictor of hard errors than the soft error rate, which can rise and fall, as seen in figure 13. Also notice that, in figure 16, the retry level is very low on early passes. This indicates that retries do not affect performance until the tape is quite worn.

8 Future Work

It should be emphasized that all of our tests were run using the Exabyte EXB-8500 8mm Cartridge Tape Subsystem, and Exatape brand tapes. Therefore, all of our data only explains error behavior with respect to the EXB-8500. We believe that is is generalizable to other helical scan tape drives. It would be instructive to run the same tests using magnetic tape drives from different manufacturers.

On the average, it takes approximately 16 hours to run a thousand passes of a read test. Write tests average much longer, about 26 hours for one thousand passes. Because of this, we were limited as to the number of tests we ran. A variety of other tests would be useful in gaining a better understanding of how various interacting parameters affect error behavior in the EXB-8500. Among these are:

- *Tests which used a larger percentage of the tape.* Our tests only use a small portion of the tape. While this was necessary in order to limit the time for each pass, it would be interesting to run longer tests. It might be that we would see more debris since more of the tape would be in use.
- *Running our tests multiple times.* Many of our tests were only run a few times in their particular configuration. Running tests additional times would allow us to characterize the variability due to individual tapes.
- *Running additional long-term tests to determine the impact of tape head cleaning on error behavior.* We did not have enough time to run tests that vary cleaning intervals. It is possible that we cleaned the tape heads too soon or too late, causing error rates to rise.
- *Running tests exclusively in streaming mode.* Because our tests gather error information at such a fine granularity, the drives were not able to maintain streaming. It would be interesting to run tests which maintain streaming to understand the effect of the additional wear on the tape from start/stop operation.
- *Tests which fast-forward.* It would be interesting to access the tape with a more sophisticated pattern, rather than just reading sequentially. This would help us understand the stress of fast forwarding versus reading a section of tape.
- *Testing head wear.* The disparity between results of the same tests on different drives indicate that head wear plays a large factor in error behavior. A test that tracked errors over a small number of passes on a series of new tapes could provide useful information on the effects of head wear.
- *Testing with a larger group of tape drives.* All tests were run using only three tape drives. It is always preferable to have more drives to verify testing, but outside of an industry setting it is not always practical. The National Media Lab study was run with only two drives [JS94].

We feel that for the broad conclusions we draw, three drives provides an acceptable level of confidence, but it would be interesting to verify our results using more tape drives.

9 Summary and Conclusions

New applications such as the EOSDIS archive and data mining over very large databases will require magnetic tape systems to support many more passes than required in previous workloads. We sought to add to the understanding of error behavior by running read and write tests on the EXB-8500 8mm tape drive on the same tape for thousands of passes. Previous studies ran tests for only a few hundred passes [JS94], or only on a few block of the tape [EXA93]. Running our tests over a more lengthy time period has enabled us to gather information on the progression of soft errors, retry behavior, and hard errors. The fine granularity of the data allowed us to examine where on the tape such errors occur.

After analyzing the data gathered from our tests, the major conclusions we have reached are as follows:

- Error behavior is not as clear cut as we initially hypothesized. Various factors contribute to error behavior and it is difficult to isolate each factor's effect on error behavior.

- Hard errors are more rare and occur after many thousands of passes of a tape. Hard errors are also transient. In every case, data was able to be read on later passes even when hard errors had occurred several times consecutively. This suggests that a lazy approach to data re-copying, that is, do not copy to new tapes until a hard error is actually seen, may be practicable.

- Head cleaning does not seem to reduce the number of media errors, but prolonged avoidance of cleaning does have catastrophic effects on tape drive operation, exhibited as hardware errors and tracking errors.

- Retries become more frequent and have higher bursts as the tape becomes worn. Monitoring of the level of retries could be used to decide when it is necessary to re-copy a tape that is in danger of data loss.

In order to get the best performance out of magnetic tape systems, especially large robotic library systems, it is important both to protect data from loss, and to maximize availability.

The information we have gathered should be useful to formulate more efficient policies for data protection.

10 Acknowledgements

Many thanks to my advisor, Randy Katz, for his guidance throughout this project. Thanks to Ann Chervenak and Dave Patterson for reading this paper and offering many helpful comments. Also, thanks to Marylou Orayani who conducted the long term tests 1A and 1C and the short term tests. We wrote together an early version of this paper for a class project.

References

[Bog95] John W.C. Van Bogart. Magnetic tape storage and handling: A guide for libraries and archives. Technical report, National Media Laboratory, June 1995.

[Che94] Ann Louise Chervenak. *Tertiary Storage: An Evaluation of New Applications.* PhD thesis, University of California, Berkeley, December 1994.

[Com95] Compaq Computer Corporation. *Tape Drives, Media, and the Importance of Cleaning*, September 1995.

[DIS96] James Demmel, Melody Y. Ivory, and Sharon L. Smith. Modeling and idenitfying bottlenecks in EOSDIS. Technical report, University of California, Berkeley, 1996.

[EXA91] EXABYTE Corporation, Boulder, Colorado. *EXB-8500 8mm Cartridge Tape Subsystem. User's Manual*, April 1991.

[EXA93] EXABYTE Corporation, Boulder, Colorado. *8mm Data Grade Tape. A Comparative Analysis.*, January 1993.

[Exa94] Exabyte Corporation, Boulder, Colorado. *Media Guide - Digital 8mm Media*, July 1994.

[Far95] Giulietta S. Fargion. EOSDIS product use survey. Technical Report 161-TP-001-001, Hughes Information Technology Corporation, September 1995.

[JS94] Tony Julik and Robert Schlentz. Systems reliability study - 8mm test. Technical Report TR-0029, National Media Lab, St. Paul, Minnesota, November 1994.

[MFW$^+$96] Reagan W. Moore, Richard Frost, Mike Wan, Joe Lopez, and Richard Marciano. The design of a parallel data handling system for scientific data management and mining. Technical report, San Diego Supercomputer Center, 1996.

[Mil95] Ethan Leo Miller. *Storage Hierarchy Management for Scientific Computing.* PhD thesis, University of California, Berkeley, April 1995.

[Moo96] Reagan Moore. High performance data assimilation. Technical report, San Diego Supercomputer Center, 1996.

[MT95] Manish Malhotra and Kishor Trivedi. Data integrity analysis of disk array systems with analytic modeling of coverage. *Performance Evaluation*, February 1995.

[PH96] David A. Patterson and John Hennessey. *Computer Architecture: A Quantitative Approach.* Morgan Kaufmann Publishers, Inc, 1996.