MIR Performance Analysis

Damian Hazen, Jason Hick
{dhazen, jhick}@lbl.gov
National Energy Research Scientific Computing (NERSC) Center,
Lawrence Berkeley National Lab, Berkeley, CA 94720
June 12, 2012

***Abstract*—We provide analysis of Oracle StorageTek T10000 Generation B (T10KB) Media Information Record (MIR) Performance Data gathered over the course of a year from our production High Performance Storage System (HPSS). The analysis shows information in the MIR may be used to improve tape subsystem operations. Most notably, we found the MIR information to be helpful in determining whether the drive or tape was most suspect given a read or write error, and for helping identify which tapes should not be reused given their history of read or write errors. We also explored using the MIR Assisted Search to order file retrieval requests. We found that MIR Assisted Search may be used to reduce the time needed to retrieve collections of files from a tape volume.**

I. Introduction

A. Motivation

Magnetic tape continues to be an important storage medium in archival storage systems. Depending on the required capacity, data retention policies and access characteristics, tape storage can provide an economical alternative to disk based storage systems both in terms of media and management costs. When evaluating media cost, a key benefit with tape is its longevity. Properly stored metal particle tape provides a stable recording medium with a lifetime of up to 30 years[1].

Because of tape media's longevity it is often reused. Two common cases that cause media reuse are that the volume has become sparse after a large number of files have been removed, or the tape reader has reached the end of its supported life. For both cases, data is read from the old volume and written again to another volume. The old volume will then often become a candidate for receiving new data. For the case of moving to a new drive technology, it is not uncommon for the same tape cartridge to be used again by a newer drive, and at an increased bit density. In both cases, and in the case of access in general it is possible that the volume has not been read or written for multiple years.

As media is read during normal operations, or as part of clearing data from a volume and writing to another, it is important to try to identify volumes that either have or soon will have data integrity issues. For instance, in choosing to reuse media it would be very useful to be able to identify and discard the 50 or 100 worst tapes. Tape drives will report errors if unable to successfully write data to a cartridge. However, detailed media analysis from the tape manufacturer for a volume with a permanent write error frequently shows no apparent problems with the cartridge. Tape drives mark bad sections of tape to avoid bad blocks upon rewrite. So determining whether the media or drive is at fault on failed writes is difficult.

For errors encountered during writing, a conservative approach is to discard the media involved after copying any existing data to a new tape. For errors encountered when reading, the problem may not be easily solved if the data can't be read after repeated attempts. In both cases, it is important to determine what drive the media was written on and if it is available for read to rule out a drive/media interchange problem.

To attempt to further understand the quality of our media when considering reuse, The National Energy Research Scientific Computing Center collaborated with Oracle to capture hundreds of proprietary statistics recorded by their T10K series drives for each volume mount over approximately a 1 year period. Until recently, these data have not been publically available for analysis. As part of our work, we've captured over 100,000 records and have begun the first large-scale analysis of the data. The first phase of our analysis attempts to correlate permanent read or write errors with entries in the MIR and identify the statistics, which corroborate the hard failure. It is our goal to eventually be able to identify indicators in the drive or media that can be used to predict the likelihood of trouble recording data.

In addition to examining media health statistics, we also used position metadata stored for each volume which allowed us to quickly determine a more time efficient order when retrieving multiple files from the same tape volume.

B. The NERSC Archive

The National Energy Research Scientific Computing (NERSC) Center is the Department of Energy's (DOE) flagship national user facility supporting DOE Office of Science users. The range of science and applications supported by NERSC storage is broad. The allocations shown in Figure 1 are for our High Performance Storage System (HPSS), a hierarchical storage system configured to use tape as the bottom storage tier. HPSS provides scalable, high performance long-term storage to our users.

As of the writing of this paper, the NERSC HPSS contains over 20 Petabytes of scientific and backup data on about 28,000 tape cartridges. Today, two different enterprise tape drives are used in the NERSC HPSS: the access optimized 9840D and the capacity optimized T10KB. The research described in this paper focuses only on the MIR capabilities

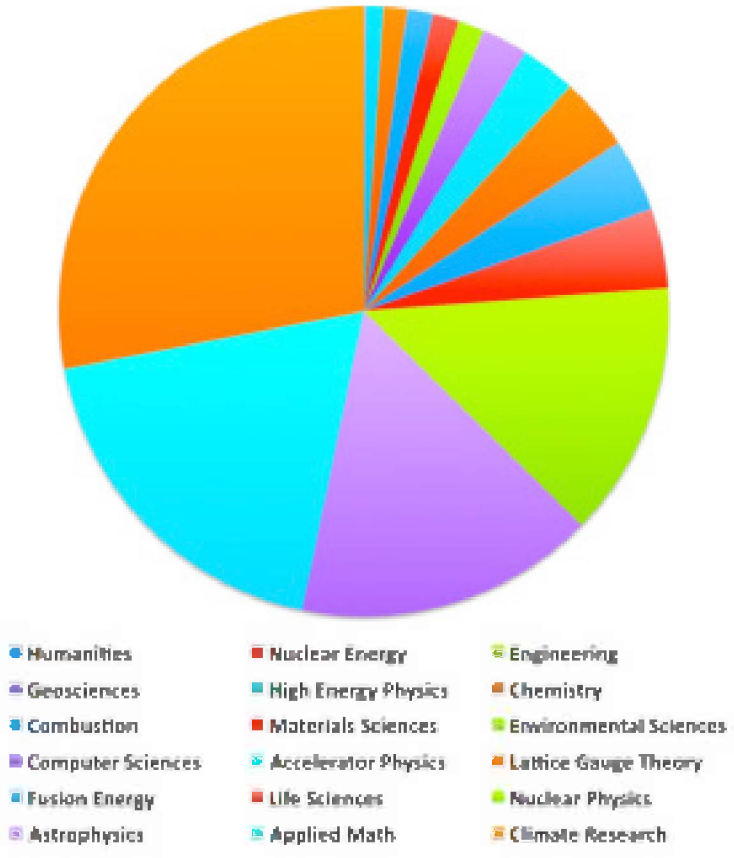

Fig. 1. Storage Allocations for 2011 by Science Program

of the T10K technology of which nearly 19,000 volumes are in use at NERSC. Additionally, at NERSC about 30% of operations to HPSS are read operations. This is an unusually high percentage for a tape archive and provides a good mix of read and write statistics for the study.

For monitoring drive health and media reliability, NERSC uses a combination of ad hoc techniques and vendor provided solutions. This has been useful in detecting drives that may be failing, as well as actively cataloging and removing problem media from the system. The efforts are largely reactive to actual failures and thus, we still experience media problems that lead to unrecoverable data. Additionally, we still have no way to identify our best media for reuse.

Recently NERSC conducted a reliability audit of data on the center's tape volumes. The study found that on over 40,000 9840A, 9940B, and T10KA tapes up to 12 years old, 99.9991% of tape volumes were 100% readable. Further, 99.99999% of all files were 100% readable.

Typically about 2% of the volumes allocated to the system have experienced one or more read errors. From these, we work to copy the active data (repack) to a new tape, and it often takes two or three attempts before getting all the data. It is rare to not be able to read a file after multiple attempts, however this approach is reactive. A better solution would be to receive early indication of volumes that may experience read errors.

II. Methods

A. Description of the Media Information Record

As part of the cartridge dismount process, the T10K series drives update the Media Information Record or MIR on the cartridge. The MIR contains two categories of data: Performance Data and Position Data. The MIR Performance data contains many different low level signal processing counters, tape usage statistics and error counters, some of which we will describe in further detail.

The Position Data is a table of the physical sectors on the volume. With the position data it is possible to map application file marks to physical location on tape. This information can be used to order application requests to reduce tape seek distances and direction changes. We developed tools to capture and display the performance data as well as an implementation of an ordering algorithm, but did not save this data for each tape mount.

The MIR performance data can be broken into several components. First it contains general information about the tape cartridge. These include the cartridge serial number, counters, such as total number of times the volume has been mounted, the validity of the MIR itself, and other identifiers such as the serial number of the first and last writing drive.

The majority of the performance data are statistics related to a certain mount. It includes per mount data for the first and last writing drives, the last 5 drives that mounted the volume, and totals for all mounts. The mount data consists of a large number of counters that record error correction statistics and other low level events that happened during read or write, metrics such as the amount of tape that crossed the drive head, the time spent reading or writing, and calculated values such as Read Quality Index and Tape Efficiency. Finally, the performance data also includes information about the 128 most recent permanent Fault Symptom Codes or FSCs, and the last 1024 temporary FSCs.

B. The Process for Capturing the MIR

The MIR Performance Data can be retrieved by issuing a SCSI Read Buffer request for the MIR Performance Data page. To capture the data, we modified the HPSS Mover so that immediately before taking the volume offline for dismount, we wrote the MIR Performance Data to a file, saving the time stamp, volume label and drive identifier. The data was later converted to ASCII and ingested into a database for analysis. When captured in binary file format, each MIR record is approximately 100KB in size.

Reading and recording the MIR Performance Data added 1-2 seconds to the tape dismount time. Over the course of 12 months, we collected over 300,000 MIR Performance Data sets for over 18,000 distinct tapes.

C. Database Schema

The database tables used to store the MIR data are divided into three primary categories: information about the performance statistics, the historical performance statistics, and the per-mount performance statistics.

For purposes of this paper, the tables of interest are the MNT_DATA, STATS, and FSC. The MNT_DATA table contains the volume identifier of each cartridge mounted, the drive serial number it was mounted on, and the count of permanent errors which generated FSC codes during the mount. The STATS table has per mount statistics showing the amount of data read and written, and the number of permanent errors encountered during the mount. The FSC table has the specific code encountered per mount, and the location on media where the error occurred.

Figure 2 shows the entire schema in an entity relationship diagram. Not all tables shown in the schema diagram were used for this first analysis of the MIR statistics. They are shown to provide a complete picture of the statistics gathered.

We explored two categories of inquiries, one of broad historical or total population statistics and the other of in-depth analysis of an individual volume or event. The total population inquiries are useful for sorting by a particular statistic over a long amount of time (e.g. a year of having collected 100,000 mount statistics) to find the most problematic volumes for reading or writing. The queries for a particular volume allows one to see the degree to which a volume is having trouble.

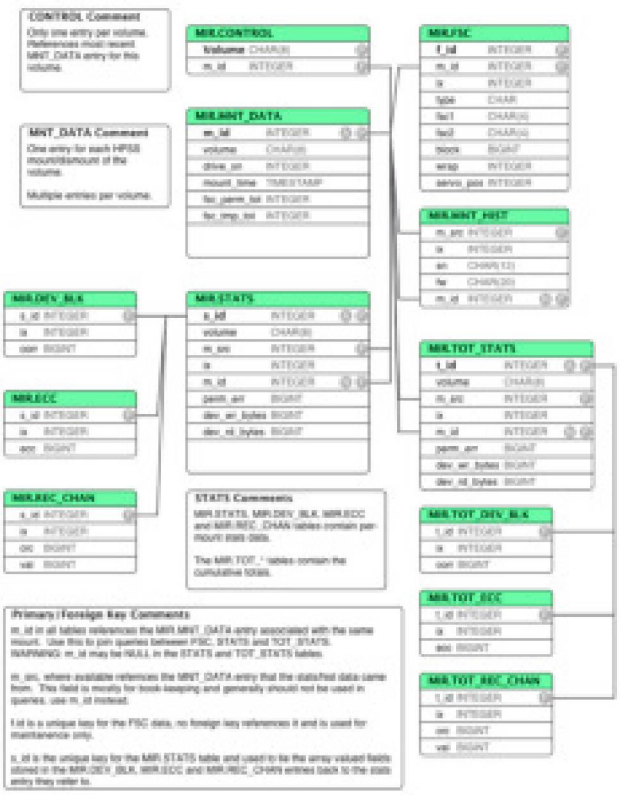

Fig. 2. Database Schema for MIR Statistics

III. MIR Results

A. Tape Population Analysis

One query we produced provided the drive serial numbers and the media labels that had permanent read or write errors on the drive. This is particularly useful for determining trends with write errors as particular drives may be of interest if producing more errors than the norm. The total count of permanent write errors per tape is provided in figure 3.

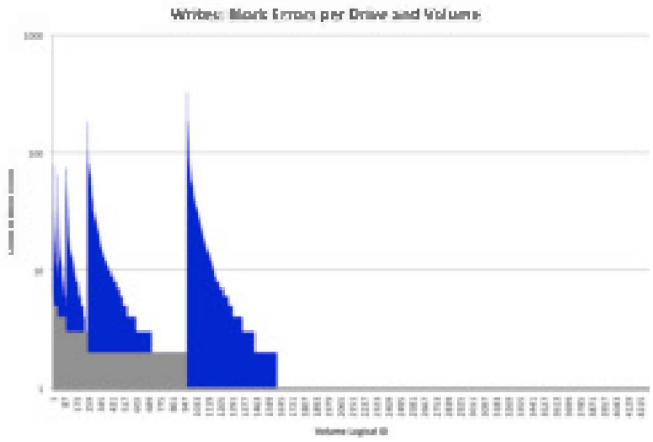

Fig. 3. Block Write Errors

Figure 3 shows each tape as a volume logical ID (e.g. tape

1, tape 2, ... tape 4215) on the X-axis with the count of distinct blocks that had permanent write errors in blue. The grey values show the distinct number of drives those write errors occurred on. Then, the results are sorted from left to right, first by number of drives and secondarily by number of block errors. Therefore, tapes 1 through 947 are of interest since they have more than one unique block on the same tape that couldn't be written by more than 1 tape drive. These 947 tapes would be candidates to remove from the system to see if write errors decrease.

Figure 4 shows similar information as figure 3 but for permanent read errors rather than write errors. The first 397 tapes in figure 4 are showing multiple read errors on more than one tape drive. This would suggest a tape problem rather than a drive problem. Data from these tapes should be moved onto new tapes to ensure data integrity.

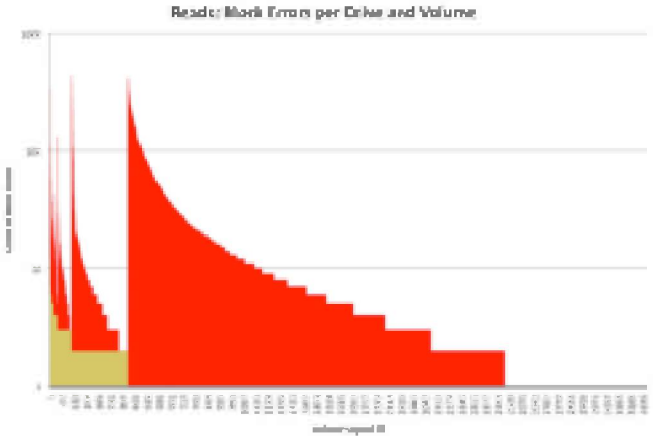

Fig. 4. Block Read Errors

We also produced a count of the number of read and write errors per tape volume, grouped by their internal serial number. The tape serial number contains a manufacturer's identification number, a tape model number, and a unique cartridge identifier. The evidence, although not confirmed by the manufacturer, is that serial numbers generally increase over time, with newer volumes having a higher value. Sorting the list, then grouping into lots of 1000 tapes gives an idea of the error trend by tape serial number.

Figure 5 shows a plot of the number of unique cartridges with permanent read and write errors sorted by serial number from smallest (left side) to largest (right side) in groups of 1,000 tapes. Lot 1 which is the lot with the lowest cartridge serial numbers shows about 140 total cartridges with permanent read or write errors.

The bar chart in figure 5 shows a definite trend downwards in the number of errors. This could simply be a function of the age of the media and the length of time the tape has been in the system. Since NERSC has a high read rate, generally older volumes have seen more accesses. Two points of interest are lots 6 and 9. Lot 6 shows significantly fewer errors than trend. Lot 9 shows a spike in errors. We believe the spike in lot 9 is

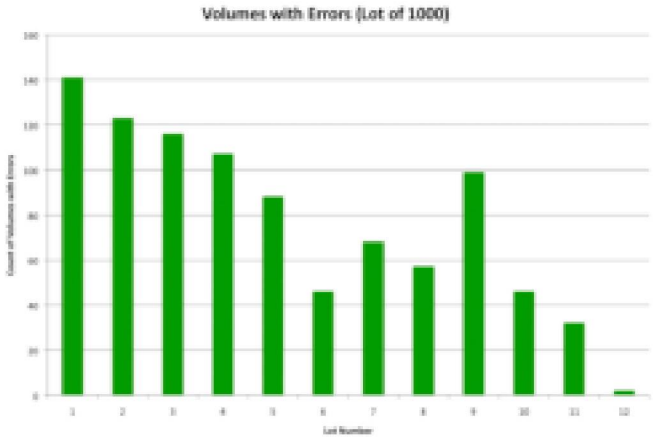

Fig. 5. Volumes with Errors Lot of 1000

due to an environmental incident that occurred at the facility which caused an increase in the number of write errors. The issue has been resolved over a four month period and errors appear to be back at trend.

An additional analysis we performed summarized the total number of fault symptom codes (FSC) occurring each day for a specified time period. This is useful in determining overall health of the tape subsystem. For example, are errors trending up or down, or are FSC errors related to IO load on the tape subsystem? It is also useful for correlating FSC errors to microcode levels since microcode is generally the same on all tape drives for a given period of time. Figure 6 shows a plot of FSC errors from August 2010 through April 2011, and includes a trendline (dark blue) showing the number of errors trending down during that time period.

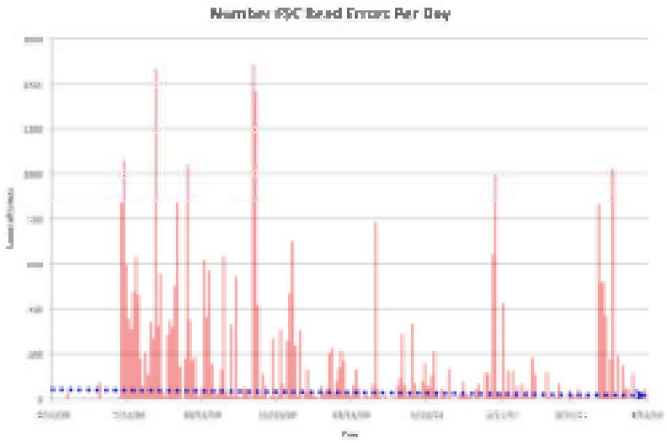

Fig. 6. FSC errors

We used the information in figure 6 coupled with install dates for different microcode on the tape drives and found that there was a strong correlation between number of errors and microcode level. It is probable that changes in microcode leads to handling of certain error conditions differently. We are

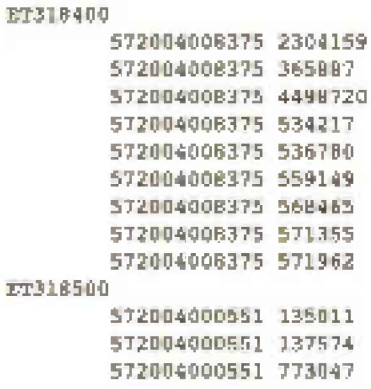

Fig. 7. Volume Read Errors

also able to determine the count of errors for each FSC code itself, so this is useful in determining the most problematic or frequent fault symptom codes occurring in our tape subsystem. We could then focus our efforts on addressing the most common FSC's to improve tape operations.

B. Tape Mount Analysis

For tape volumes that have FSCs logged in the MIR, we were able to produce a sorted list of media labels and the position on tape of any permanent read or write errors on the volume. This is particularly useful in determining the degree to which the volume might be damaged. It also helps with mapping locations to file names, providing a way to know which files couldn't be read. This could be used in recovering data from problem tapes if the need to position around the bad blocks is necessary. Sample output from the utility is provided in figure 7.

The results of the volume read error output above shows that there are two tape volumes with read errors (ET3184 and ET3185). The value in the first column is the drive serial number, and the number in the second column is MIR position information. This provides an approximate physical location on each tape where the read error occurred. This information is useful for understanding whether it's predominantly one drive or many that couldn't read the same position on tape (e.g. block of data), or whether the tape has a small or large amount of damage.

It is possible to use the MIR positioning information (block, servo and wrap) to map back to your own system's metadata to determine the actual device location (e.g. file mark position) and file attributes such as file name, assuming that your archival storage system's metadata format is well understood. In HPSS, this is possible, and the following section describes how this can be accomplished.

IV. MAS Results

A. MIR Assisted Search Position Data

In addition to performance data, T10KB MIR records contain Assisted Search fields describing the physical tape layout, which can be used for several different purposes. Referred to as MAS data, two uses we explored were locating the position of a file on the physical media and ordering retrieval requests on a volume to reduce time spent seeking.

B. Description of the MAS fields

The T10KB media is a one terabyte native (uncompressed) capacity cartridge[2]. Data is recorded in 1152 tracks in a serpentine fashion. Serpentine recording partitions the media longitudinally, recording first in the forward direction of motion, then reversing once the end of the media is reached and recording the next in the reverse direction. Thirty-six passes are made to fill a volume. Each pass is referred to as a wrap and consists of one complete traversal along the length of the tape.

The tape comes with a factory recorded servo track which encodes the longitudinal position for the media. There are approximately 73,000 longitudinal position markers spaced at 12.32mm. The MAS uses 9216 sector records to describe the tape layout. A sector is a contiguous physical tape region, laid out in a linear fashion. Sectors begin at the load point and run to the logical end of tape. Each sector record in the MAS includes a wrap, servo and tape mark field. Figure 8 shows pictorially the serpentine tape layout and provides an abbreviated description of sector records contained in the MIR.

C. Using MAS data to find the physical location of an HPSS file

Except for the case of file aggregates, HPSS writes a tape mark between each user file. Since the sector records include wrap, servo and tape mark fields, it is straight forward to find the approximate physical location of an HPSS file on T10KB media. The location is only approximate because a sector corresponds to a fairly large physical region of tape - approximately 3 meters, and the beginning of the file may occur anywhere within the region. Figure 9 provides a more detailed description of the process. The figure shows output from the HPSS lsvol utility which displays the starting tape mark, fileset and filename, and the contents of two consecutive sector records from the MAS. In sector record 768, 11,811 tape marks have been encountered to this point. By sector 769, 11,821 tape marks have been read. The lsvol output in the lower text box lists all the files that start in sector 768. Note that the file at 11,821 has data in both sector 768 and 769.

D. Using MAS data to order tape retrievals

It is often the case that many files will need to be retrieved from the same tape volume. The average seek time for a T10KB tape drive is 46 seconds, and with a native data transfer rate of 120 MB/s, often file retrieval time is dominated by seeking to the file's first byte [3]. By considering the mechanical properties of the tape drive and the layout used when writing the tape cartridge, it is possible to significantly reduce access times. Generally, stopping or reversing direction of the media and repositioning is time consuming, so ordering algorithms should attempt to keep the media moving, and in

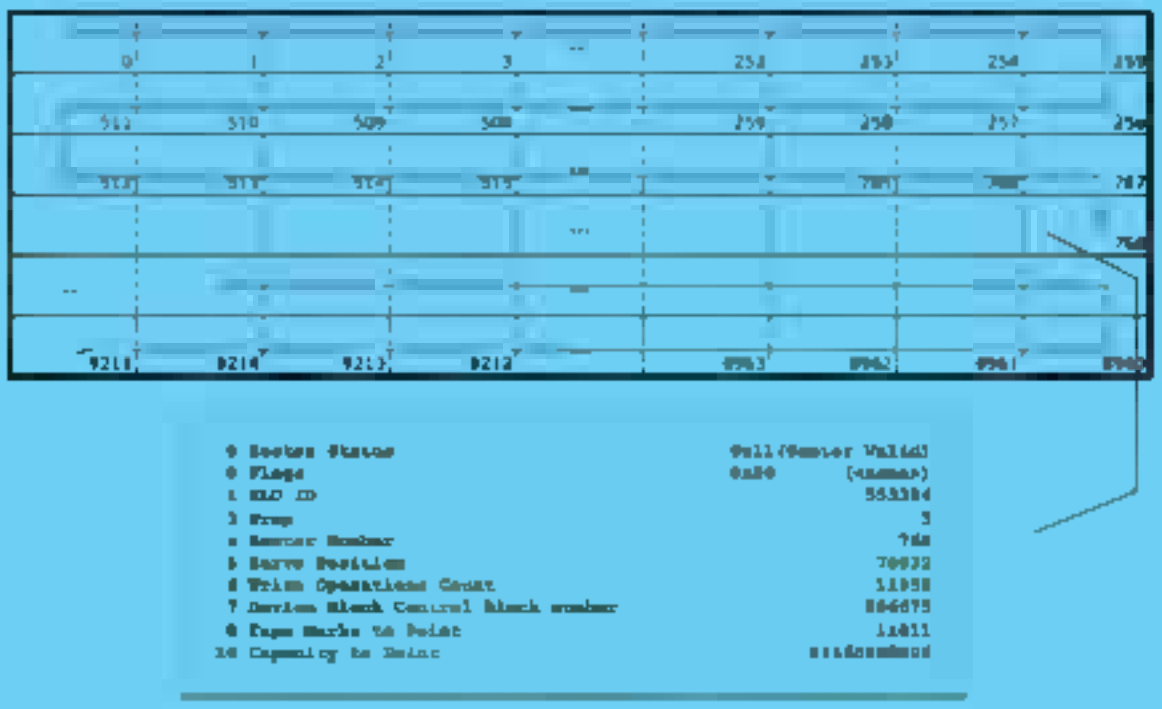

Fig. 1. Sector Layout on T10KB Media

the same direction, as much as possible. Since HPSS writes a tape mark between files, an available optimization when retrieving files is to order the requests by tape mark. Doing so keeps the tape moving predominantly in the same direction, with direction changes generally occurring after reading the file which exists nearest the end of each wrap. However, with the T10KB MAS data, it's possible to further improve access times by mapping a tape mark to a wrap and longitudinal position.

Using this information, a simple approach is to first retrieve files that were recorded in the forward direction, switching wraps as necessary, and then retrieve files recorded in the reverse direction. However since files on different wraps may overlap longitudinally, it may sometimes save seek time to reverse directions to retrieve files that are longitudinally adjacent, or to make multiple passes over the tape rather than proceed with a strict two pass approach. To deal with these situations, it becomes necessary to assign costs associated with reversing directions and changing wraps. [4] has analysis and refinements developed using a Quantum DLT2000 drive, and provides several ordering algorithms using these techniques.

For our work, we compared three of the ordering approaches used in [4] using a T10KB volume containing approximately 278 000 files. The first approach ordered requests by tape mark, and did not use the MAS data. The second approach used sector and wrap information in the MAS to order requests. A single forward, then reverse pass over the tape, was made, switching wraps as needed. For files recorded in the same direction which began on a sector on a different wrap but in the same longitudinal position on the tape, the file starting on the sector nearest the wrap of the previous request was read first, followed by files starting on the remaining wraps. Using this technique results in a total of two passes along the tape, but will require rewinding within a pass for files starting on sectors that are adjacent longitudinally.

The final ordering scheme is similar to the second but deals with the overlap problem by making multiple passes over the tape. If files start in sectors which are on different wraps but overlap longitudinally, the sector nearest the current wrap is read, and the remaining files are saved for another pass.

Fig. 9. Translating HPSS file name to physical tape location

With this approach, the tape continues to move in the same direction at the expense of possibly requiring additional passes over the tape.

Figure 10 shows a comparison of the total seek time needed by the three approaches to position to the beginning of the indicated number of files chosen at random from the tape. After positioning, the files were read and separate timings were kept for both the seek times and file read times.

The results show that ordering using MAS data performed between two and five times better than ordering by tape mark, while ordering using MAS and allowing more than two passes performed between two and four times better. Figure 11 shows the results extended out to 7,000 retrieval requests.

Here, our tests showed that for less than approximately two thousand files, ordering by MAS using either a single forward and reverse pass, or multiple passes performed better than ordering by tape mark. However for larger sets, two pass MAS took almost two times as long as ordering using multiple pass MAS and nearly three times as long as ordering by tape mark.

For further analysis, we ran several tests using different sets of 1000 files chosen at random from the tape. The timings shown for a run are summarized in Table I. The test results show that retrieval using any of the ordering algorithms is better by an order of magnitude than using no ordering at all. Additionally, ordering by tape mark took 1.77 times as long to retrieve the files compared to ordering by MAS, and 1.97 times as long compared to ordering by MAS using multiple passes. For this set of files, the average time needed to read the file data across all four runs was 29.9 seconds. For a run using MAS Multi Pass, reading the data would comprise only 1.5% of the total time needed to retrieve the files. The remaining time was spent positioning.

TABLE I
SEEK TIMINGS (SECONDS) - 1000 FILES

MAS	MAS Multi Pass	Tape Mark	Random
2208	1077	7888	38276

To understand why ordering by tape mark outperforms the other ordering schemes for large numbers of retrievals, we analyzed seek time frequency distributions for the different ordering algorithms. Table II shows frequency distributions for the number of seconds spent seeking when retrieving 1,000 files using Tape Mark Order. While 32.3% of the seek requests took between one and two seconds to complete, 11% took more than eight seconds. Table III which includes measurements for all three ordering methods shows the mean, minimum, maximum and standard deviation for time spent positioning. The table shows that there is a relatively large amount of variability in seek times when using Tape Mark Order. This reflects the fact that the files to retrieve were chosen at random, and that they are retrieved in linear order. There will be a random amount of tape that needs to be coveted to move between one file and the next, with some files being close together and others further apart.

Ordering by MAS has a distinctly different distribution as shown in Table IV. All requests complete in four seconds or less, with 48.6% finishing in between two and three seconds. Referring again to Table III, the seek times are distributed over a narrow range when compared to tape mark ordering. MAS ordering introduces additional latency needed to switch

Fig 10 Seek Times 1000 Files

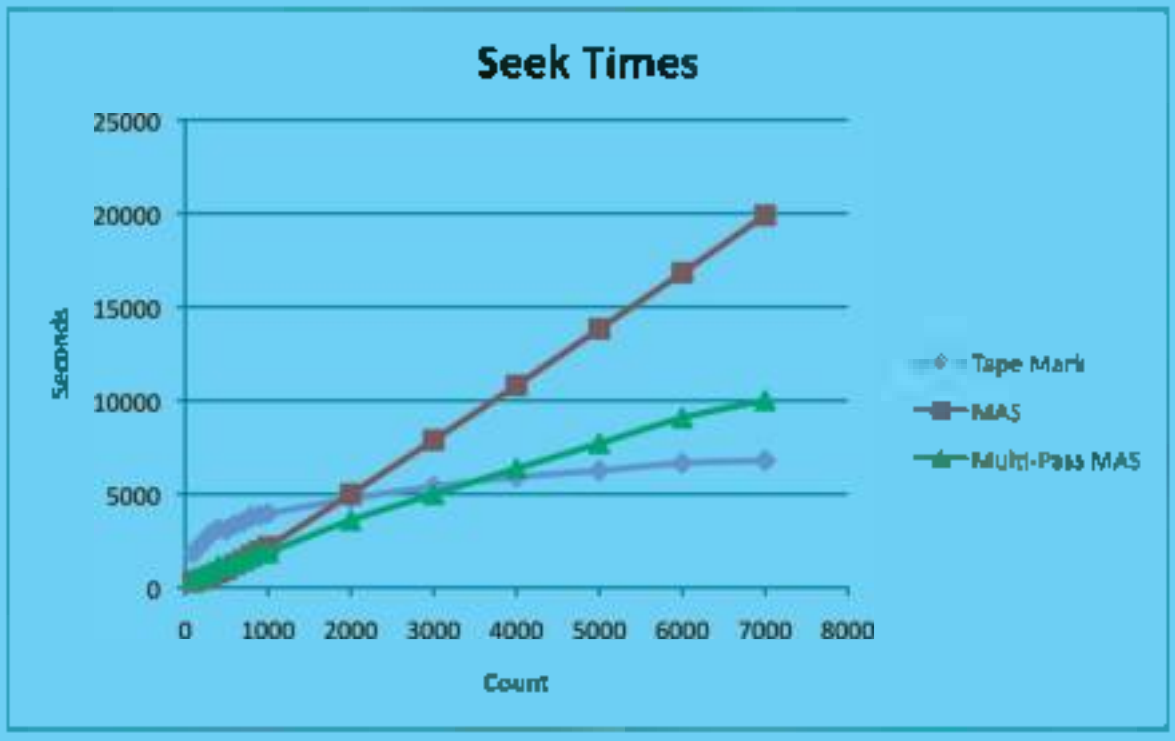

Fig 11 Seek Times 7000 Files

TABLE II
FREQUENCY DISTRIBUTION OF SEEK TIME - 1000 FILES (TAPE MARK)

Seek Time	Count	Cumulative Count	Percent	Cumulative Percent
0 To 1	149	149	14.9	14.9
1 To 2	323	472	32.3	47.2
2 To 3	168	640	16.8	64.0
3 To 4	82	722	8.2	72.2
4 To 5	65	787	6.5	78.7
5 To 6	45	832	4.5	83.2
6 To 7	27	859	2.7	85.9
7 To 8	25	884	2.5	88.4
More	116	1,000	11.6	100.0

TABLE III
POSITIONING FOR 1000 FILES

Order	Mean	Min	Max	Standard Deviation
Tape Mark	3.8997	0.0283	70.989	5.47272
MAS	2.2027	0.04975	1.82808	0.87253
MAS - MultiPass	1.97674	0.02835	61.84994	2.83114

tracks, but reduces the amount of tape that must be seeked over to position to the next file. For a relatively few number of requests, the average seek distance that needs to be covered when using tape mark ordering is significant, resulting in higher seek latencies.

TABLE IV
FREQUENCY DISTRIBUTION OF SEEK TIME - 1000 FILES (MAS)

Seek Time	Count	Cumulative Count	Percent	Cumulative Percent
0 To 1	154	154	15.4	15.4
1 To 2	178	332	17.8	33.2
2 To 3	486	818	48.6	81.8
3 To 4	182	1,000	18.2	100.0

Table V shows the frequency distribution for MAS ordering when using multiple passes to deal with overlapping files, and table III includes the average, minimum, maximum and standard deviation for this ordering run. This ordering also has a large range of seek times. [4] points out that because requests for files that coincide longitudinally to the same sector are added to later passes, the last passes may be sparse and result in long seek times. The scatter plot in figure 12 provides an illustration of this.

TABLE V
FREQUENCY DISTRIBUTION OF SEEK TIME 1000 FILES (MAS - MULTIPASS)

Seek Time	Count	Cumulative Count	Percent	Cumulative Percent
0 To 1	272	272	27.2	27.2
1 To 2	401	673	40.1	67.3
2 To 3	239	912	23.9	91.20
More	88	1,000	8.8	100.0

Tables VI provides summary statistics for seek times using a run with 5000 files. As expected, the seek times for tape mark order have decreased as the number of files requested has increased, so that switching tracks now provides significant overhead.

TABLE VI
POSITIONING FOR 5000 FILES

Order	Mean	Min	Max	Standard Deviation
Tape Mark	1.25055	0.000037	33.04072	1.19669
MAS	1.76634	0.366654	7.07619	0.59932
MAS - MultiPass	1.54107	0.0037	38.2545	1.397

V. CONCLUSIONS

The work has enabled us to demonstrate that there is indeed useful information in the MIR that can be applied to operating a tape storage system. We have the ability to identify problem tapes in a number of different ways. We can identify the data on tapes that generate errors when read or written, and are able to use MIR information to aide in recovering data. We were also able to demonstrate that the MIR Position Data can be used effectively to order tape requests to reduce seek time. The fact that the physical location of files on tape can be obtained

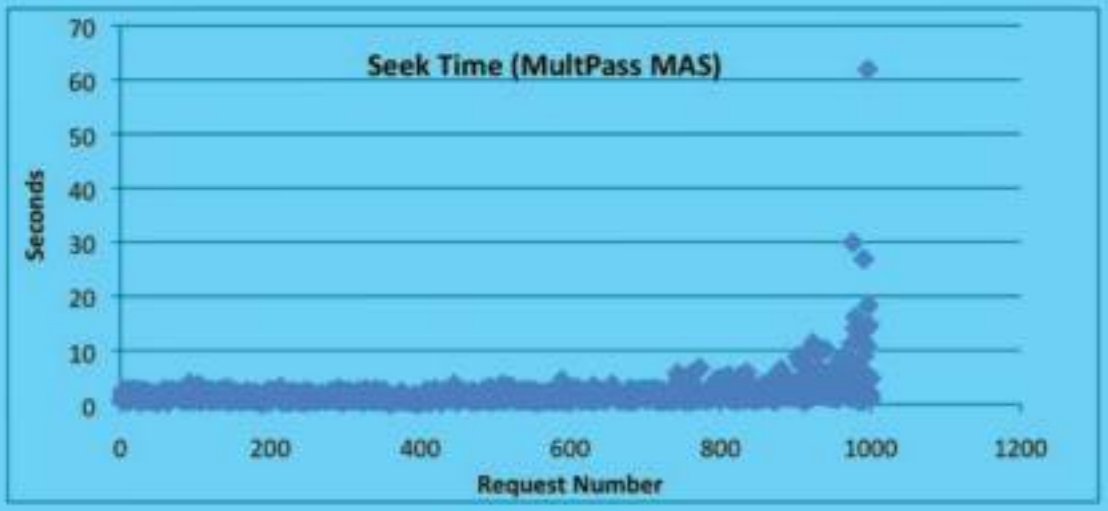

Fig. 12. By MAS

in a few seconds opens up the possibility of using the position data in existing storage systems to improve tape performance.

Unfortunately, the process of maintaining and interpreting the MIR information is arduous. It took months to develop the capability shown in this paper. We ended up using about 25 of the numerous statistics available in the MIR. Though we did do further analysis on other MIR statistics not reported in this paper, the results are inconclusive.

We look forward to future solutions that use the MIR statistics but don't require the infrastructure we needed to conduct the research.

VI. Acknowledgements

This work was funded in part by the DOE award DE-FC02-06ER25767 the Petascale Data Storage Institute and by the Advanced Scientific Computing Research (ASCR) in the DOE Office of Science under contract number DE-C02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center, under Contract No. DE-AC02-05CH11231.

We would also like to acknowledge Oracle Tape Engineering for their collaboration on this project. Without their knowledge and support, this research would not have materialized.

VII. Disclaimer

References

[1] StorageTek T10000 Family Tape Cartridge. 033617.pdf. Oracle. [Online]. Available: http://www.oracle.com/us/products/servers-storage/storage/tape-storage/

[2] "T10000 Tape Drive MIR Assisted Search," Oracle.

[3] StorageTek T10000B Tape Drive. 036556.pdf. Oracle. [Online]. Available: http://www.oracle.com/us/products/servers-storage/storage/tape-storage/

[4] O. Sandstå and R. Midtstraum, "Improving the access time performance of serpentine tape drives," in *ICDE*. IEEE Computer Society, 1999, pp. 542–551.

[5] ——, "Low-cost access time model for a serpentine tape drive," in *IEEE Symposium on Mass Storage Systems*, 1999, pp. 116–127.

[6] StorageTek T10000 Tape Drive, Fibre Channel Interface Reference. E20425_02.pdf. Oracle. [Online]. Available: http://download.oracle.com/docs/cd/E26105_01/index.html