An Analysis of Latent Sector Errors in Disk Drives

Lakshmi N. Bairavasundaram†, Garth R. Goodson, Shankar Pasupathy, Jiri Schindler
†University of Wisconsin-Madison, Network Appliance, Inc.
†laksh@cs.wisc.edu, {goodson, shankarp, jiri}@netapp.com

ABSTRACT

The reliability measures in today's disk drive-based storage systems focus predominantly on protecting against complete disk failures. Previous disk reliability studies have analyzed empirical data in an attempt to better understand and predict disk failure rates. Yet, very little is known about the incidence of latent sector errors i.e., errors that go undetected until the corresponding disk sectors are accessed.

Our study analyzes data collected from production storage systems over 32 months across 1.53 million disks (both nearline and enterprise class). We analyze factors that impact latent sector errors, observe trends, and explore their implications on the design of reliability mechanisms in storage systems. To the best of our knowledge, this is the first study of such large scale – our sample size is at least an order of magnitude larger than previously published studies – and the first one to focus specifically on latent sector errors and their implications on the design and reliability of storage systems.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Reliability, availability and serviceability

General Terms

Measurement, Reliability

Keywords

Latent sector errors, disk drive reliability, MTTDL

1. INTRODUCTION

Hard disk drives are the primary storage media both in enterprise environments and in personal computers. Like other hardware components, understanding the nature of their failures is essential for building more reliable and highly available storage systems. Previous studies have focused on complete disk failures in order to estimate their expected life [6, 7, 12, 17, 19, 20]. However, factors other than complete disk failures influence the reliability of data, often expressed as the mean time to data loss (MTTDL).

Most storage systems today make certain disk drive reliability assumptions and build redundancy mechanisms such as RAID [5, 11] to compensate for complete disk failures. However, the inability to either temporarily or permanently access data from certain sectors can also affect the MTTDL. Such incidents are often referred to as *latent sector errors* because the disk drive does not report any error until the particular sector is accessed. Such errors can usually be repaired by rewriting the data to a spare sector without having to replace the entire disk drive.

The impact of latent sector errors on the MTTDL in RAID systems is well known. Hafner et al. pointed out that a single latent sector error can lead to data loss during RAID group reconstruction after a disk failure [8]. Similarly, Baker et al. developed new RAID equations that account for latent sector errors when calculating the MTTDL [4]. However, very little data is publicly available that quantifies the incidence rate of latent sector errors in deployed systems.

The collection and analysis of such data can help us influence how systems are built. For example, understanding how latent sector errors are spatially clustered on a disk can determine the placement of important file system metadata structures like its superblock or metadata logs [14, 21]. Similarly, understanding the incidence of latent sector errors can aid in choosing the most effective execution schedule of proactive reading and data verification, called disk scrubbing [18]. An overly aggressive rate of such background operations may negatively impact performance [3], while a low rate may adversely affect MTTDL.

To address the described gap in storage systems design, we collected and analyzed error logs from production systems that use both enterprise class and nearline (commodity) disks. Typically, mission- or business-critical applications use systems configured with enterprise class disks, while nearline disks are deployed in cost-efficient, archival, or backup storage systems. The collected data covers a period of 32 months and includes about 1.53 million disks from over 50,000 systems deployed in the field at various customer sites. To the best of our knowledge, this is the first study of such large scale – our disk sample size is at least an order of magnitude larger than any previously published studies – and the first one to focus specifically on latent sector errors and their implications on the design and reliability of storage systems.

SIGMETRICS'07, June 12–16, 2007, San Diego, California, USA.

1.	A total of 3.45% of 1.53 million disks developed latent sector errors over a period of 32 months.
2.	Enterprise class disks are less likely to develop latent sector errors than nearline disks. Surprisingly, enterprise class disks with at least one error are as likely to develop additional errors as nearline disks with at least one error.
3.	The fraction of disks affected by latent sector errors increases linearly with time for enterprise class disks and super-linearly for nearline disks.
4.	For most disk models, the ASER (annual sector error rate) increases between the first and second year of disk use. This increase is very sharp for nearline disks.
5.	The fraction of disks affected by latent sector errors increases as disk capacity increases.
6.	For most disk models, more than 80% of disks with latent sector errors have fewer than 50 errors.
7.	Latent sector errors are not independent of each other. A disk with latent sector errors is more likely to develop further errors than a disk without a latent sector error.
8.	Latent sector errors exhibit a significant amount of spatial and temporal locality.
9.	Disk scrubbing is very useful for proactively detecting latent sector errors. More than 60% of these errors are discovered through scrubbing.
10.	Enterprise class disks show a high degree of correlation between recovered errors and latent sector errors.
11.	Nearline disks show a high degree of correlation between not-ready-condition errors and latent sector errors.

Table 1: Overall Observations. This table summarizes the important observations of our study.

This paper examines latent sector errors from a variety of angles. First, we examine latent sector errors for dependence on two factors: disk age and disk size. Second, we examine three characteristics of latent sector errors: the number of errors per disk, for disks with at least one error, spatial locality of errors, and the temporal behavior of errors. Third, we examine the types of requests that encounter latent sector errors. Finally, we examine correlations between latent sector errors and two other errors: recovered errors and not-ready-condition errors. Our most important observations are summarized in Table 1.

The rest of the paper is structured as follows. We give some background on the causes of disk errors (Section 2) and describe our storage system architecture (Section 3). Section 4 describes the methodology used to analyze our field data. Section 5 discusses our results in detail while Section 6 presents trends and lessons from our data. Section 7 discusses previous work and Section 8 concludes.

2. BACKGROUND

We now present a brief background on disk technology and the types and causes of latent sector errors in order to establish the appropriate context within which to interpret our observations.

2.1 Disk Drives

Disk drives are composed of a variety of complex mechanical and electrical components where the reliability of each component contributes to the reliability of the device. They store information on *platters* coated with a magnetic material, called *magnetic media*. Disks typically have multiple platters, where each platter usually has two surfaces, each accessed by a dedicated read/write head. A single surface is divided into tens of thousands of concentric circular tracks and each track is subdivided into *sectors*, the smallest addressable unit of data access, usually 512 bytes in size. Each sector is protected by error correcting codes (ECC).

The disk interface abstracts the disk as a linear array of equal sized blocks each identified by a logical block number (LBN). Internally, the disk reserves a small portion of sectors called *spares*, which are not initially mapped to a particular LBN. The disk firmware can map a spare sector to the LBNs of failed sectors. Today's disk drives allocate a few thousand spare sectors for re-mapping.

2.2 Disk Errors

Disk access errors can occur due to various reasons. Shah and Elerath describe the dominant failure mechanisms that occur in today's disk drives [20]. The contributing factors include: (a) media imperfections, (b) loose particles causing media scratches, (c) "high-fly" writes leading to incorrect bit patterns on the media, (d) rotational vibration, (e) read/write head hitting a bump or media, and (f) off-track reads or writes. Other factors including design, manufacturing, and operational environment can have a great impact on disk drive reliability. Anderson et al. provide a detailed description of the differences between nearline and enterprise class disks and their impact on reliability [2].

We distinguish three mutually exclusive error types.

Latent Sector Errors This error occurs when a particular disk sector cannot be read or written, or when there is an uncorrectable ECC error. Any data previously stored in the sector is lost.

Not-Ready-Condition Errors These errors may imply that the disk drive is not ready to handle a command from the host. This error could also indicate that the disk itself is not accessible. They are often resolved by waiting and retrying.

Recovered Errors These "errors" are returned by disks when an access to a sector required disk-level retry or error-correction to retrieve the data. Although the operation completed successfully and returned the data, they may serve as a warning.

The disk interface identifies each error type by a specific error code. For example, upon discovering a latent sector error during a read or write operation, the Small Computer Systems Interface (SCSI) returns status code *Check condition* with the sense key *Medium error* (0x03), specifying medium error as the reason why the last read or write command failed [23]. Similar mechanisms and error codes exist for nearline drives.

2.3 Recovery Mechanisms

Before reporting media access errors, disks typically perform error correction with multiple retries of a given operation. Additionally, after a (configurable) number of unsuccessful retries, disk drives can automatically re-map failed writes to *spare* sectors. More precisely, an LBN is reassigned from the failed sector to a spare sector, and the content is written to that new location. Nearline disks typically perform re-mapping automatically, while enterprise class disks can be configured to allow system software to perform the re-mapping on demand. Sparing and re-mapping can only occur on detected write errors; read errors require higher-level mechanisms such as RAID reconstruction to obtain the lost data. Different disks may have different algorithms for recovery. Some are more "programmable" than others; e.g., enterprise class disks generally allow finer control over retries and re-mapping.

2.4 Terminology

We use the following terms in the remaining sections.

Disk class Enterprise class or nearline disk drives with respectively Fibre Channel and ATA interfaces.

Disk family A particular disk drive product. The same product (and hence a disk family) may be offered in different capacities.

Disk model The combination of a disk family and a particular disk size. Note that this term does not imply an analytical or simulation model.

Disk age The amount of time a disk has been in the field since its ship date, rather than the manufacture date. In practice these these two values are typically within a month of each other.

Error disk This term is used to refer to a disk drive that has at least one latent sector error.

3. STORAGE SYSTEM ARCHITECTURE

We analyze data from production storage systems installed at many customer sites. We now briefly describe our system architecture, focusing on error handling policies and detection mechanisms for latent sector errors. In the rest of the paper, we refer to this storage system as our system. In general, we gather data from a variety of different NetApp Filers running our Data ONTAP^{TM} software [10].

3.1 Storage Software Stack

At a high level, the software stack of our system is composed of three layers: the WAFL^{TM} file system, the RAID layer, and the storage layer. The file system layer processes client requests by issuing read and write operations to the RAID layer. The RAID layer transforms the file system requests into disk logical block requests and issues them to the storage layer, which includes a set of customized device drivers. The RAID layer also generates parity for writes and reconstructs data after failures. The storage layer communicates with physical disks using the SCSI command set [23].

The storage layer writes checksum information along with each file system data block. For enterprise class disks, our system uses 520-byte sectors. Thus, a 4-KB file system block is stored along with 64-bytes of checksum in eight 520-byte sectors. For nearline disks, we use the default 512-byte sectors and collocate several checksums into a separate sector. The storage layer also handles various disk errors including latent sector errors, transport errors, recovered errors, and not-ready-condition errors. Individual disk drives are housed in storage shelves connected to the CPU complex through two independent Fibre Channel (FC) loops. Nearline drives use hardware attachments to convert the ATA interface to the Fibre Channel protocol. Hence, our system views them as FC drives, but can distinguish them by their model names provided by the SCSI INQUIRY command.

3.2 Error Handling

Latent sector error handling depends on the type of disk request and the type of disk. For enterprise class disks, the storage layer re-maps the bad sector to a spare sector. If the request is a write, the storage layer re-issues the write to the re-mapped sector. If the request is a verify or a read, the RAID layer reconstructs the sector and passes it to the storage layer for rewrite. It is important to note that such system-level remapping or RAID reconstruction and recovery is not counted as a recovered error; the term recovered error denotes only disk drive-level recovery (i.e., errors recovered internally by the drive). For nearline disks, sector re-mapping on failed writes is automatically performed by the disk and not reported to the storage layer. Our system handles read and verify errors in the same fashion for both nearline and enterprise class drives.

Our storage systems use proprietary heuristics for determining when to fail a disk drive. These heuristics are threshold-based and take into account the time between latent sector errors, as well as the total number of latent sector errors encountered. Other systems use similar heuristics to predict disk failures based on observed errors; e.g., Linux systems often use SMART [1]. Our study enables us to tune the thresholds used to predict disk failures.

Similar to latent sector errors, our systems proactively remap sectors associated with recovered errors. Proprietary heuristics are used to fail disks that may have experienced too many recovered errors. The storage layer handles not-ready-condition errors by retrying the operation a few times. If these efforts fail, the data is reconstructed by the RAID layer from parity.

3.3 Proactive Error Detection

Our storage system periodically *scrubs* all disks as a proactive measure to detect latent sector errors and corruption errors. Two types of scrubs are performed – media scrubs and data scrubs.

Media scrubs use a SCSI VERIFY command to validate a disk sector's integrity. This command performs an ECC check of the sector's content from within the disk without transferring data to the storage layer. On failure, the command returns a latent sector error. The storage layer performs media scrubs continuously in the background, with the rate of scrub adjusted so as not to impact foreground performance. Media scrubs typically complete within 2 weeks.

A data scrub is primarily used to detect data corruption. This scrub issues read operations for each disk sector, computes a checksum over its data, compares the checksum to the on-disk 8-byte checksum, and reconstructs the sector from other disks in the RAID group if the checksum comparison fails. Latent sector errors discovered by data scrubs appear as read errors.

4. ANALYSIS METHODOLOGY

4.1 Data Collection

Our storage system has a built-in, low-overhead mechanism to log important system events back to a central repository. These messages can be enabled for a variety of system events including disk errors. These logs allow customized support based on observed events. Not all customers enable logging, although a large percentage do. Those that do, sometimes do so only after some period of initial use. We studied disk failure data in our database for a period of 32 months starting in January 2004. This left us with a substantial sample of 1.53 million disk drives of 14 disk families and 30 distinct models for our study.

4.2 Limitations

Before disks are shipped, they undergo rigorous testing both in-house and by the disk vendor. This testing eliminates disks that would have shown up in our data as highly error prone. Latent sector errors found during testing are not reflected in this study. Sectors with detected errors are automatically re-mapped through low-level formatting before they are shipped.

For a variety of reasons, disks may be removed from the system. Our study includes those disks up to the point of their removal from the system. Therefore, we may not observe errors from otherwise error prone disks after some period of time.

Nearline disks automatically perform sector reassignment for latent sector errors; see Section 3.2. Thus, latent sector errors encountered during writes for this class of disks are not propagated beyond the disk and nearline error rates do not reflect these write errors.

Our study of error rate observations reflects lower bounds. Other disk error studies, of which we are aware, suffer from these same limitations.

4.3 Sample Selection

We constrained our sample to disks for which we had a complete history in the logs. This provides us with the full sample of 1.53 million disks.

To derive statistically significant results, we often further constrain the sample set depending on the analyses being performed. For example, we sometimes use shorter time periods for our analyses so as to maximize the number of models we can study; clearly not all disk families and models have been in the field for the same duration. The disk model samples we consider may have one of the following constraints:

1. Model has at least 1000 disks in the field for time period being considered
2. Model has at least 1000 disks in the field and at least 50 error disks for time being considered

In addition to these general constraints, our samples may be conditioned on exhibiting some number of errors or of being a certain age. We also disregard the very few "outlier" disks (0.2% of error disks) with more than 1000 errors to avoid the skew caused by these numbers.

We usually present data for individual disk models, however, we sometimes report averages (mean values) for nearline disks and enterprise class disks. Since the sample size for different disk models per disk class varies considerably, we weigh the average by the sample size of each disk model in the respective class.

5. RESULTS

This section presents the results of our analysis of latent sector errors. First, we present some basic data on latent sector errors collected over 32 months from disk drives in the field. Second, we analyze the impact of various factors that affect the occurrence of latent sector errors. Third, we analyze properties and characteristics of latent sector errors. Fourth, we examine the distribution of latent sector errors across different request types. Finally, we discuss correlations between latent sector errors and other disk errors.

5.1 Conventions

We denote each disk drive model as ⟨*family-type*⟩. For anonymization purposes, *family* is a single letter representing the disk family (e.g., Seagate Cheetah 10k.7) and *type* is a single number representing the disk's particular capacity. Although capacities are anonymized to a single number, relative sizes within a family are ordered by the number representing the capacity. For example, n-2 is larger than n-1, and n-3 is larger than both n-1 and n-2. The anonymized capacities do not allow comparisons across disk families. Typically, disks in the same family only differ in the number of platters and/or read/write heads [19]. Disk families from A to E (upper case letters) are nearline disk families, while families from f to o (lower case letters) are enterprise class disk families. Thus, the number of disks in our study, N, can be expressed as

$$N = \sum_{F=\mathrm{A}}^{\mathrm{o}} N_F$$

where N_F is the number of disks in the sample of a particular disk family F. In our case, $F = \{\mathrm{A}, \ldots, \mathrm{E}, \mathrm{f}, \ldots, \mathrm{o}\}$. Note that N_F is the sum of the number of disks for each model in the family. We use the term N_M to denote the number of disks in the sample of a disk model.

We present most data as the probability of developing x latent sector errors for a particular sample of disk drives and use the shorthand notation $P(X_T \geq L)$ for describing the probability of a disk developing at least L latent sector errors within T months since the disk's first use in the field. We use $E(X_T)$ to refer to the mean number of latent sector errors developed within T months since first use. More precisely, we could express our probability as $P(X \geq L | t \leq T \wedge N_M \geq y)$, of errors that occur within time T for a single disk in our sample belonging to disk model M. The disk has been in the field for at least T months and has at least y units in the field for that time period. However, this notation would make it quite cumbersome to read the text and hence we use our shorthand notation.

5.2 General Observations

In our entire sample of 1.53 million, we find 53,820 (3.45%) disks develop one or more latent sector errors over the period of 32 months. For error disks (disks with at least one error), the median number of errors per disk is three. However, the mode is one error (30% of the error disks). Only 0.2% of error disks had more than 1000 errors per disk. Ignoring these "outlier" disks, the mean number of errors per error disk is 19.7.

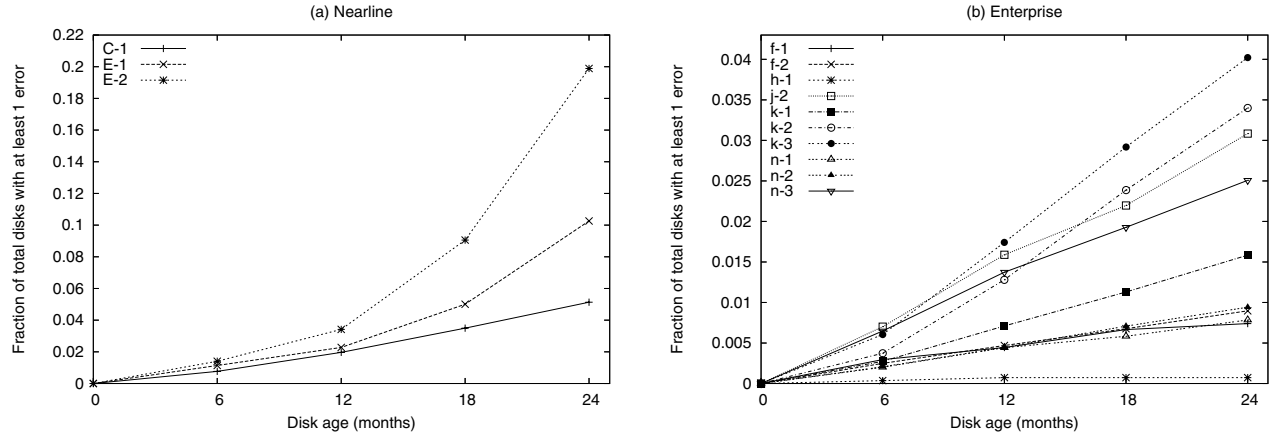

Figure 1: Impact of disk age. The probability that a disk develops latent sector errors as it ages is shown. Note that the probability is cumulative. Also, note that the y-axis scale is different for the two graphs.

Observation 1. *Enterprise class disks are less likely to develop latent sector errors than nearline disks.*

Overall, we found that nearline disks and enterprise class disks exhibit different behavior with respect to latent sector errors; about 8.5% of all nearline disks are affected by latent sector errors while only 1.9% of all enterprise class disks are affected. Therefore, most of our subsequent analyses break down results by disk class.

Looking at disks of the same age, we find that 3.15% of nearline disks and 1.46% of enterprise class disks develop at least one latent sector error within twelve months of their ship date. This sample includes 200,408 nearline disks (56% of all nearline disks in our study) across 6 disk models and 715,033 enterprise class disks (61% of all enterprise class disks in our study) across 23 disk models. Using our notation, these numbers can be represented as $P(X_{12} \geq 1)$. We present more detail about error rates as a function of time in Sections 5.3.1 and 5.4.3.

5.3 Factors

We now explore the impact of two factors on latent sector errors: the age of a disk drive, and its size.

5.3.1 Disk Drive Age

We study how the age of the disk drives affects (a) the fraction of disks that develop latent sector errors, and (b) the fraction of sectors that develop errors.

Figure 1 presents the fraction of disks that develop at least one latent sector error within a given period of time since each disk's first use. As described earlier, we include only disk models with at least 1000 units in the field ($N_M \geq 1000$) for the entire 24 month period of this study. Using our notation, we can express the graph as $P(X_t \geq 1)$ where, t={6, 12, 18, 24} months. The same sample of disks is used for all time periods. The sample includes 68,380 nearline disks across three disk models and 264,939 enterprise class disks across ten disk models.

As observed in the previous subsection (Observation 1), we see that nearline disks are more likely to develop latent sector errors. For example, almost 20% of 'E-2' disks experience latent sector errors within 24 months of their shipping. On the other hand, only 4% of 'k-3' disks, the enterprise class disk model with the highest error rate, experience latent sector errors in the same time period.

Observation 2. *The fraction of disks with latent sector errors varies significantly across manufacturers and disk models.*

We see from Figure 1 that the fraction of disks with errors at the end of 24 months could vary from 5% to 20% for nearline disks. Enterprise class disks also exhibit a significant variation.

Observation 3. *Over twenty four months, the fraction of nearline disks developing latent sector errors grows far more rapidly than the fraction of enterprise class disks with errors.*

In the case of enterprise class disks, we observe that the percentage of disks that have latent sector errors increases almost linearly with time. Thus, the probability that an enterprise class disk will develop a latent sector error in a given six month window is nearly the same within the first 24 months of use. On the other hand, this percentage for nearline disks increases super-linearly with increasing disk age. For example, the percentage of 'E-1' disks that develop latent sector errors in the time period between 18 and 24 months after shipping is 5.25%, while it is only 2.72% between 12 and 18 months after shipping. More generally, $(P(X_{t+6} \geq 1) - P(X_t \geq 1)) > (P(X_t \geq 1) - P(X_{t-6} \geq 1))$, where $t \leq 24$.

Observation 4. *Annual sector error rates vary greatly across disk models but on average are considerably worse during the second year for nearline disks.*

Figure 2 shows the annual sector error rates (ASERs) computed for the disk models, as well as the cumulative nearline and enterprise class error rates. The error rates are for the first and second year of disk use. The sample covers all drives in the field for 24 months (the same sample as in Figure 1). The figure can be represented as $E(X_t - X_{t-12})/(sectors\ per\ disk)$ for $t = \{12, 24\}$ months. Note that the figure does not show error bars since most disks have 0 errors. For nearline drives the sector error rates for the second year increase considerably over the first year.

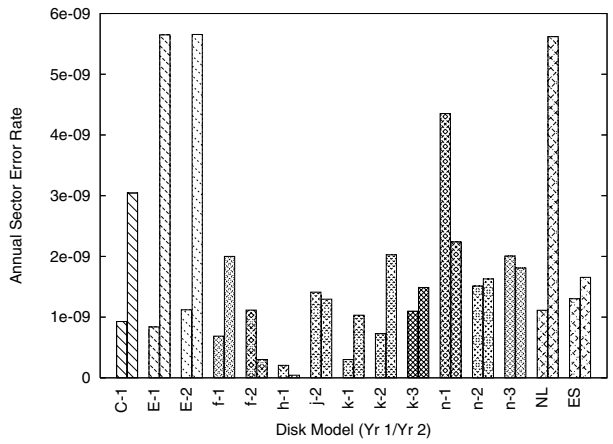

Figure 2: Annual sector error rates (ASERs). For each disk model that has been in the field for at least two years, the first bar represents Year 1 and the second represents Year 2. The NL and ES bars represent weighted averages for nearline and enterprise class drives respectively.

However, this is not the case for the enterprise class drives. About half of the enterprise class models show this trend, while half do not.

5.3.2 Disk Drive Size

Figure 3(a) shows the fraction of disks with latent sector errors across the various disk families. For each disk family, the graph groups the data by disk model (disk capacity). We restrict the disk families in the graph to those for which there are at least 1000 disks in the field with an age of at least 18 months for *each* disk size. This age maximizes the number of disk models we can study. Figure 3(a) can be represented as $P(X_{18} \geq 1)$ for different disk models.

Observation 5. *We observe that as disk size increases, the fraction of disks with latent sector errors increases across all disk models.*

We observe the same trend even for those families that did not satisfy the 1000-disk requirement with the only exception being disk family 'l'.

As disk capacity rapidly increases, storage systems will need to deal with a larger percentage of drives that develop latent sector errors. However, since many factors contribute to latent sector errors (see Section 2.2), we cannot draw any specific conclusion beyond the trend we observe in our data.

Observation 6. *The amount of probable data loss due to latent sector errors per Gigabyte does not increase or decrease consistently as disk size increases.*

Figure 3(b) presents the average number of latent sector errors per Gigabyte. It can be represented as $E(X_{18})/Capacity$. Interestingly, unlike Figure 3(a), the data does not show a consistent increase or decrease across disk size for the same disk family. Thus, we see that a higher fraction of disks with errors does not imply a greater amount of probable data loss.

5.4 Characteristics

The studies in this subsection focus on the properties and characteristics of latent sector errors.

5.4.1 Errors per Error Disk

Figure 4 shows the percentage of error disks that experience a given number of latent sector errors within a 18 month period after the ship date. We only include disk models that satisfy both the 1000 disk and 50 error disk limits. Thus, we can represent the figure using our notation as the conditional probability $P(X_{18} \leq x | X_{18} \geq 1)$ for x={1, 2, 3, 4, 5, 10, 20, 50}.

Observation 7. *A large fraction of disks with latent sector errors develop fewer than 50 errors.*

The data shows that, on average, 37% of nearline error disks and 39% of enterprise class error disks have only one error; i.e., they do not develop any additional latent sector errors after the first one. Furthermore, over 80% of error disks have fewer than 50 errors.

Since disk drives typically have thousands of spare sectors and since failed sectors can be recovered from elsewhere (e.g. from RAID), it is possible to re-map bad sectors and continue operation for a large fraction of error disks.

Observation 8. *Enterprise class and nearline disks are equally likely to develop more than one error once they develop their first error. This is in contrast to the very different probabilities of enterprise class and nearline disks developing their first error.*

While enterprise class disks seem to be more resilient to latent sector errors in general, enterprise class disks and nearline disks show similar behavior once they exhibit at least one latent sector error; compare the Nearline and Enterprise lines in Figure 4(a) and Figure 4(b), respectively. Surprisingly, some enterprise class disk models are worse than nearline disks – a larger percentage of enterprise class error disks develop many more errors than nearline error disks. However, one should note that the actual number of latent sector errors for nearline disks could be somewhat higher (as described in Section 4.2).

Observation 9. *Latent sector errors are not independent of each other. A disk with latent sector errors is more likely to develop additional latent sector errors than a disk without a latent sector error.*

We find that the occurrence of a latent sector error depends on previous occurrences of latent sector errors on the same disk. In particular, we find that the conditional probability of developing at least 1 additional error in x amount of time given that the disk has at least 1 error, $P(X_{t+x} \geq 2 | X_t \geq 1)$ is greater than the non-conditional probability of developing at least 1 error in x amount of time ($P(X_{t+x} \geq 1)$ - $P(X_t \geq 1)$). For example, $P(X_{18} \geq 2 | X_{12} \geq 1) = 0.671$, which is much greater than ($P(X_{18} \geq 1)$ - $P(X_{12} \geq 1)$) = 0.018.

5.4.2 Address Space Locality

The spatial locality of errors is often considered in the design of various existing file systems. For example, the original Fast File System (FFS) creates redundant copies of the superblock, spatially distributed, to protect against the loss of a disk head or multiple media errors on the same track or cylinder [9]. However, a recent study of file system robustness [14] found that IBM's Journaling File System (JFS) stores superblock copies close to each other in the logical address space, possibly exposing it to loss of both

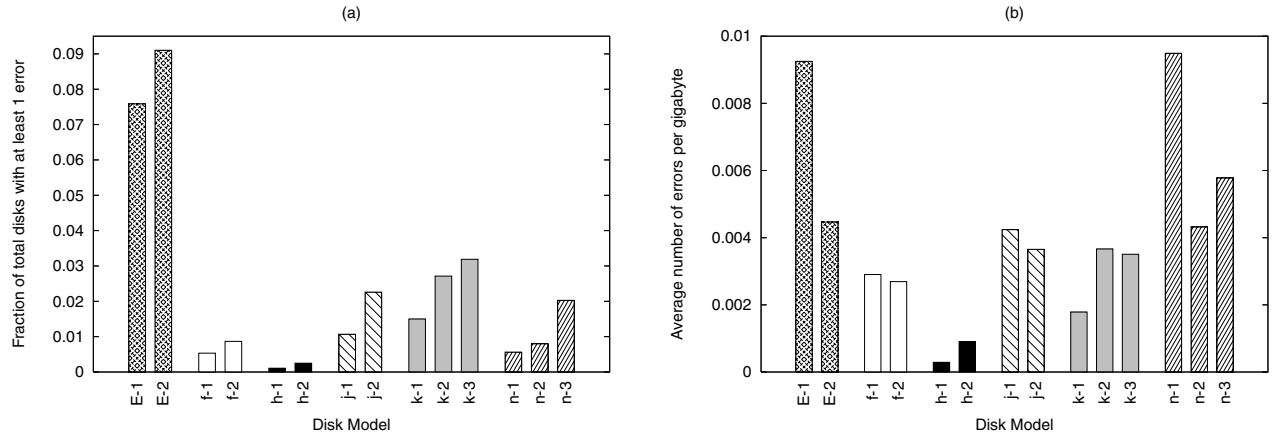

Figure 3: The impact of disk size. (a) Fraction of disks with at least one latent sector error within 18 months of shipping to the field. (b) Average number of latent sector errors per GB observed within 18 months of shipping to the field.

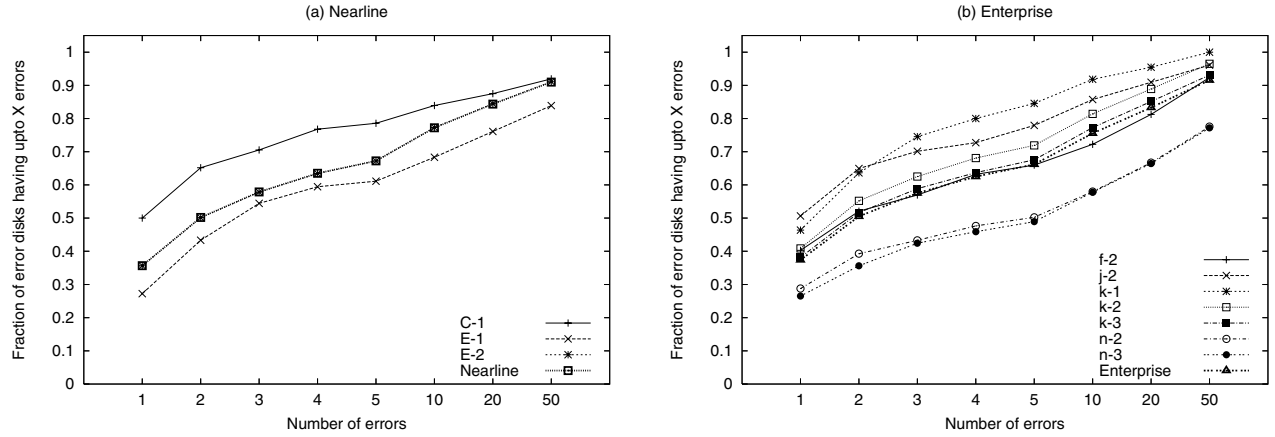

Figure 4: Errors per error disk. The fraction of error disks as a function of the number of latent sector errors that develop within a 18 month period after the ship date for (a) nearline disk models and (b) enterprise class disk models.

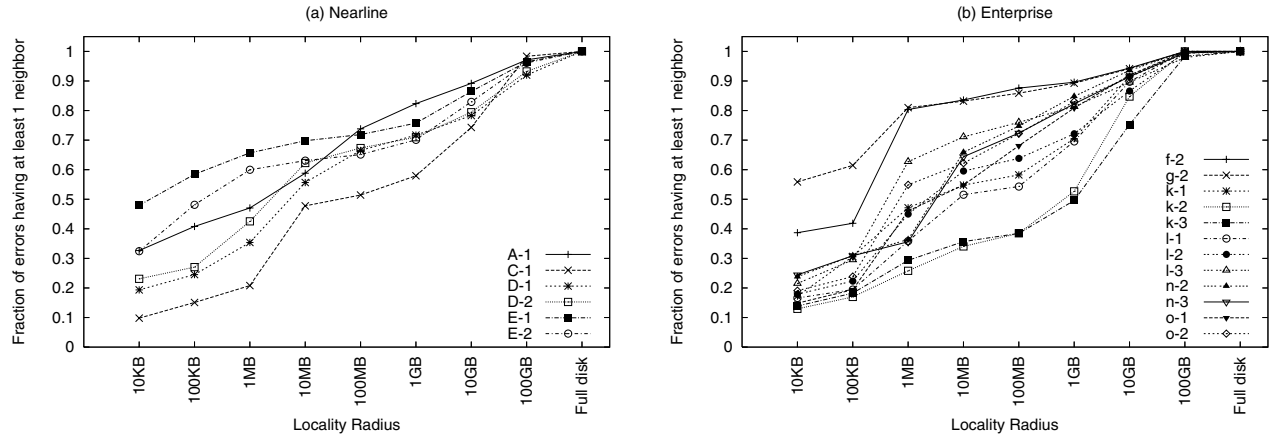

Figure 5: Address space locality. The graphs show the probability of another error within a given radius (address range) of one latent sector error.

copies. Likewise, ReiserFS places its log across a contiguous set of logical blocks [15]. Multiple latent sector errors in the log area may render the file system unusable.

Today, disk drives use a block-based interface (e.g., SCSI or ATA) that obfuscates physical block locations through complicated mapping schemes [16]. This limits file systems to use logical block locality unless more detailed information can be derived [13]. Since file system designers often make assumptions about spatial locality at the logical block level, we explore whether latent sector errors exhibit spatial locality at the logical level, referred to as *address space locality*.

Figure 5 presents the fraction of latent sector errors that have at least one other latent sector error occurring within a given radius, for disks with at least 2 errors and at most 10 errors. An upper bound of 10 errors is used in order to avoid skew introduced by disks with a large number of errors; note, the median number of errors for error disks is 3. Since address space locality is time-invariant as long as the number of errors is bounded, our sample includes all disks irrespective of their time in the field. We only include disk models that have at least 1000 total disks and 50 error disks with between 2 and 10 errors for the entire 32 months. We can express the data in our notation as $P(X_t^r \geq 1 | 2 \leq X_t \leq 10)$ with no specific restriction on time $(0 < t < 32)$, where X^r is the number of other latent sector errors in the interval $\langle a - r, a + r \rangle$ centered around sector a; sector a contains a latent sector error.

Observation 10. *There is significant locality in the occurrence of latent sector errors across logical sector addresses.*

Figure 5 shows that for most disk models the probability of other latent sector errors within a 10 MB radius of an existing error is 0.5. In fact, the probability is more than 0.6 for many models. Additionally, for many disk models, the probability increases significantly between radii of 100 KB and 1 MB. This suggests a coarse correlation between the logical and physical block space. However, we note that the observed address space locality is not perfect and may not be as correlated as system designers believe. Finally, we note that the probability varies considerably across disk models.

Figure 6 presents the mean value of X^r (X^r is the same as above) for different disk models. This figure provides an insight into how errors typically cluster together. For most models, the average number of other errors within a 10 MB radius of a latent sector error is more than 1; for some models it is as high as 2.5. When Figures 5 and 6 are compared, we see that a higher probability of a spatially local error does not necessarily imply a higher average number of spatially local errors. For example, for a 10 MB radius, 'g-2' has a higher probability of a spatially local error than 'l-3', but 'l-3' has more spatially local errors than 'g-2' on average.

5.4.3 Temporal Behavior

Another interesting characteristic of latent sector errors is their temporal behavior. We study temporal behavior in two ways: *temporal locality* and *decay*. Temporal locality is a study of how "bursty" latent sector errors are. This is useful for setting various time-based thresholds used to determine when a disk should be failed. We study temporal locality by measuring the inter-arrival rate of errors. Decay is a study of the time taken to develop e additional latent sector errors since the first latent sector error.

Figure 7 shows the percentage of latent sector errors that arrive within x minutes of the preceding error. The arrival-rates are binned by minute. We only include disk models that satisfy both the 1000-disk and 50 error disk limits. The figure can be represented as $P(X_{t+x} \geq k + 1 | X_t = k)$ for $0 < k \leq 1000$, and $0 < t < 32$ and $1 \leq x \leq 1e + 06$.

Observation 11. *All disk models exhibit high temporal locality of latent sector errors.*

Depending upon the model, between 40%-80% of errors arrive within one minute of the previous error. As can be seen, the arrival-rate distributions have very long tails. The observed locality implies that the errors are detected close in time (even though they may have developed long before they were detected). However, in our system due to scrubbing, there is typically only a short lag time between the occurrence and the discovery of an error. Thus, errors that develop at different times (e.g., a month apart) are likely to be detected at different times. It is likely that the observed temporal locality implies actual temporal locality.

Figure 8 presents the fraction of disks that develop at least e additional errors within a given time period since the discovery of the first error, for nearline and enterprise class disk classes. We use disks that developed the first error at least 6 months before the end of the study. Both nearline and enterprise class disk classes had at least 10,000 eligible units. The figure can be represented as $P(X_{t+x} \geq e+1 | X_t = 1)$ for $x = \{1, 2, 3, 4, 5, 6\}, e = \{1, 5, 10, 25, 50\}, 0 < t < 26$.

Observation 12. *Disks that develop errors beyond the first error see most of the additional errors within one month after the first error.*

First, we see that for 54.8% of nearline error disks and 62.0% of enterprise class error disks, at least one *additional* error is developed within one month of the first ever error. Second, there is a significant probability (nearline: 0.05, enterprise class: 0.10) that a disk with one error will develop 50 additional errors within one month of the first error. Third, we observe that the fraction of disks with one error that develop at least e more errors does not increase significantly with disk age for most values of e. Most of the additional errors develop within 1 month of the first error. Interestingly, this behavior is even more pronounced for enterprise class disks than for nearline disks. Finally, comparing the numbers across the two graphs, we observe that surprisingly enterprise class disks in general have a higher fraction of disks with one error that develop additional errors within a given period of time, the only exception being for $e = 1$.

5.5 Request Type Analysis

We now turn our attention to the manner in which latent sector errors are discovered by our system. Ideally, a storage system would pro-actively detect errors (e.g., through periodic scrubbing) before a user-initiated request. Sector errors detected early can be recovered from RAID-style data reconstruction and re-mapped to a new sector. Proactive detection of latent sector errors reduces the likelihood of "double-failures" in a RAID system [4].

Figure 9 presents the percentage of latent sector errors that are discovered by read, write and verify operations. In our system, read and write operations are issued in order to satisfy user or file system requests. Verify operations are issued by the media scrubber; see Section 3.3. We restrict

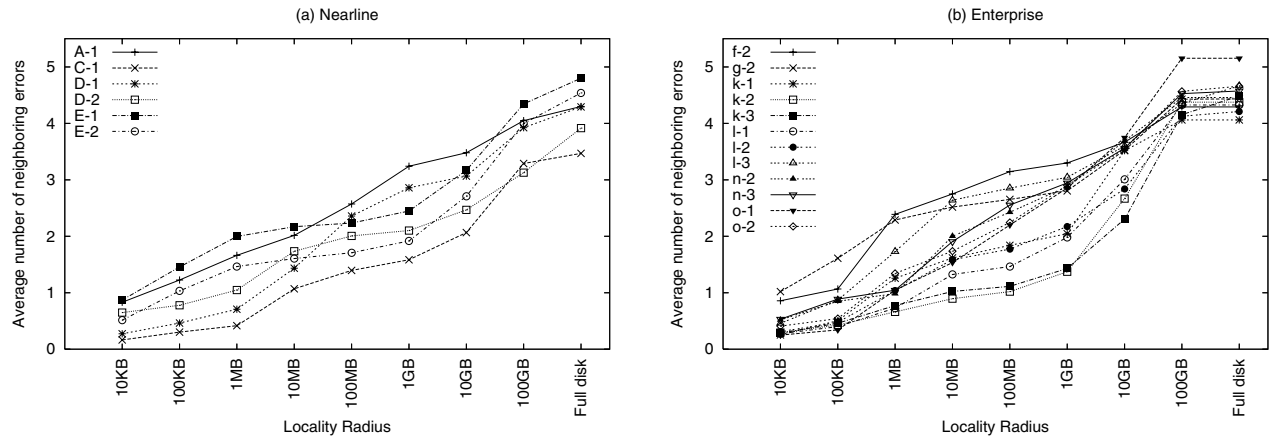

Figure 6: Count of spatially-local errors. The figure presents the mean number of other latent sector errors within a given radius (neighbors) of an existing error. The data uses only disks with 2 to 10 latent sector errors, thus limiting the maximum value possible to 9.

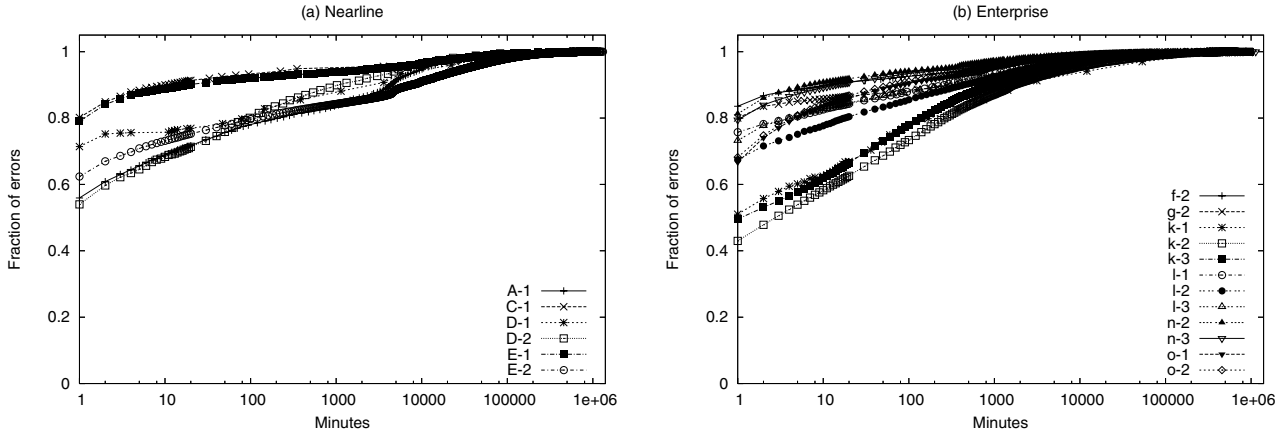

Figure 7: Inter-arrival rate. The graphs show the cumulative distribution of the inter-arrival rates of latent sector errors per minute. The fraction of errors per model is plotted against time. The arrival rates are binned by minute.

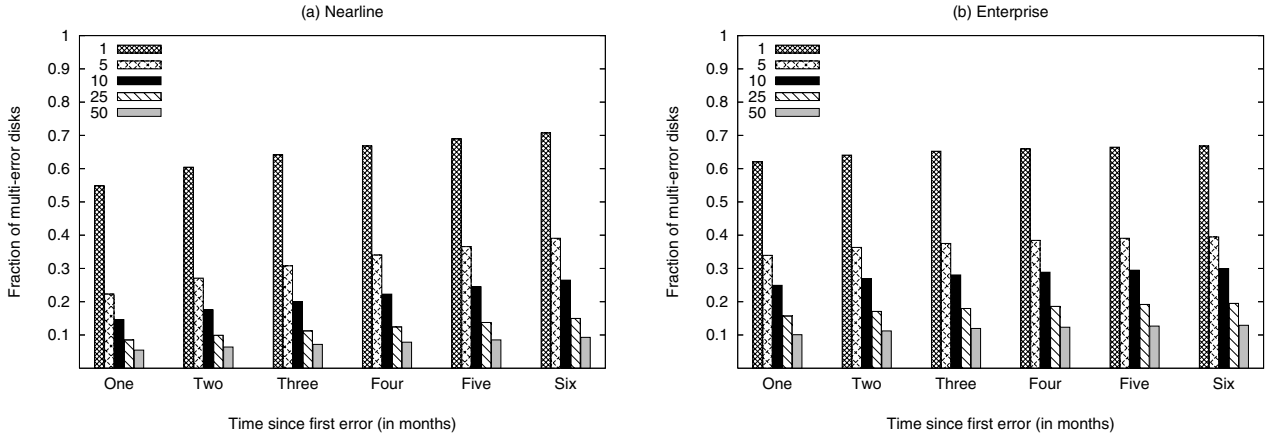

Figure 8: Temporal decay. The probability of experiencing at least 1, 5, 10, 25, and 50 additional latent sector errors within a given time period since the occurrence of the first error.

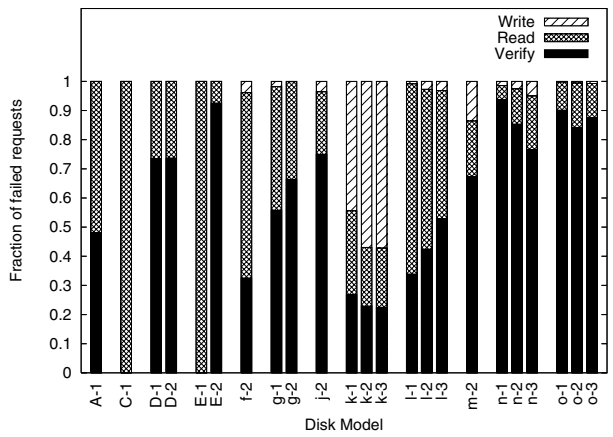

Figure 9: Request type analysis. The distribution of requests that fail due to latent sector errors across the request types read, write and verify.

models to those with at least 1000 disks in the field with at least 50 error disks in the entire 32 month study period.

OBSERVATION 13. *Disk scrubbing detects a large percentage of observed latent sector errors.*

The data shows that for many disk models, a high percentage of requests that experience a latent sector error are verify operations. On average, 86.6% of all latent sector errors in nearline disks and 61.5% of latent sector errors in enterprise class disks are discovered by verify operations, while reads discover 13.4% of errors in nearline disks and 19.1% of errors in enterprise class disks, and writes discover 0% of errors in nearline disks and 19.3% of errors in enterprise class disks. This demonstrates that the method in which our systems perform media scrubbing is useful for discovering errors. Note, since nearline disks automatically and transparently perform sector reassignment, disk writes in these systems do not report latent sector errors (see Section 4.2).

While verify operations discover a widely varying proportion of latent sector errors across disk models, on average 77.4% of all errors are detected by verify requests across all disk models. We speculate that the differences we observe are in part due to the different workloads our systems which different disk models experience.

5.6 Correlations

We now explore whether disks that exhibit latent sector errors also exhibit other kinds of errors. Specifically, we consider recovered errors and not-ready-condition errors.

5.6.1 Recovered Errors

As discussed in Section 2.2, recovered errors are errors that a disk drive encounters when accessing sectors and is able to recover from them through a combination of retries and error-correcting codes (ECC). Latent sector errors occur when such disk drive-level recovery fails. Our error logs contain recovered errors returned by enterprise class disks. We found that 52971 enterprise class disks exhibited at least one recovered errors (4.5% of enterprise class disks) over the period of 32 months ($P(Y_t \geq 1) = 0.045$, where Y is the number of recovered errors returned by a disk).

OBSERVATION 14. *There is a high correlation between latent sector errors and recovered errors for enterprise class disks.*

Interestingly, despite the fact that we observed latent sector errors in less than 2% of enterprise class disks ($P(X_t \geq 1) < 0.02$), the conditional probability of getting a latent sector error given that it experienced a recovered error is 13 times higher ($P(X_t \geq 1|Y_t \geq 1) = 0.26$). This suggests that the two kinds of errors are not independent.

Switching the variables X and Y, $P(Y_t \geq 1|X_t \geq 1)$ of observing recovered errors on a disk, given that it has a latent sector error is 0.63; i.e., 63% of enterprise class disks affected by latent sector errors also produced recovered errors. This probability varies from 0.20 to 0.78 across models.

5.6.2 Not-Ready-Condition Errors

As discussed in Section 2, a not-ready-condition error is an error during which the disk is not available to respond to requests. We found that 13% of nearline disks and 1% of enterprise class disks encountered not-ready-condition errors. Thus, with no specific restriction on time ($0 < t < 32$), $P(Z_t \geq 1) = 0.13$ for nearline disks, where Z is the number of not-ready-condition errors returned by a disk.

OBSERVATION 15. *There is a high correlation between latent sector errors and not-ready-condition errors for nearline disks.*

The conditional probability, $P(X_t \geq 1|Z_t \geq 1)$, of observing a latent sector error, given that the disk had a not-ready-condition error, is 0.38. This value is much higher than the probability of a latent sector error for a nearline disk ($P(X_t \geq 1) = 0.085$). Thus, it is highly likely that the two kinds of errors are not independent. We did not see a similar correlation in the case of enterprise class disks where $P(X_t \geq 1|Z_t \geq 1) = 0.014$ and $P(X_t \geq 1) = 0.019$.

6. TRENDS AND APPLICATIONS

The previous section presented our data according to various metrics and conditions. In this section, we interpret our data to extrapolate trends and methods for more robust design of disk-based storage systems.

6.1 Error Distribution

The probability of a failure of a unit is the basis for reliability prediction. In our context, we define a failure to be the occurrence of a latent sector error; when such an error occurs, the given request fails. However, this error is not fatal. Thanks to other mechanisms (e.g., spare sector re-mapping or reconstruction of the missing data from the parity in a RAID group), the system can in most cases recover and continue normal operation.

Figure 10 shows the probability of a latent sector error occurring on a given day since being shipped to the field. We use the same sample study as in Section 5.4.1. The light-gray line shows the probability at each day, the solid line is a 6th degree polynomial interpolation.

We notice that initially only a very latent sector errors are developed. Then, a rapid increase occurs for a period of up to about 50 days. For the enterprise class disks, this initial period is followed by a steady rate, whereas the rate for the nearline disks accelerates.

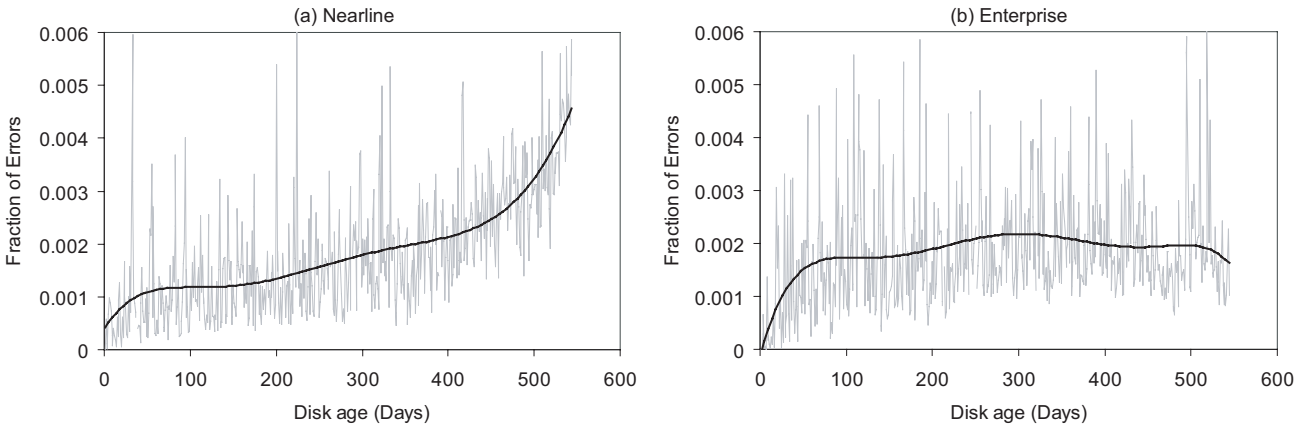

Figure 10: The distribution of observed latent sector errors per day. The graphs show the fraction of errors that were observed on a given day out of the total number of errors observed over a 18 month period. The solid line is a sixth degree polynomial interpolation to make the trend more visible.

6.2 Error Detection

Given that only 3.45% of disks in our study ever developed a latent sector error, we believe that detecting errors through a low priority background scrubbing process is sufficient. Indeed, our data shows that over 60% of all latent sector errors are discovered by the media scrubbing process, which scans the entire surface of the media at least once every two weeks. This convinces us that media scrubbing is effective in discovering problems that may result in data loss.

6.3 RAID Repair

As seen in Figure 4, the probability of a latent sector error arriving within one minute of another is over 0.4 and in many cases over 0.6. Hence, we suggest the following data structure and bi-modal process for disk reconstruction. For clarity, we assume a single-erasure-tolerant code (e.g., RAID-5) with r disks and base our conclusions on the observations from the data in Figures 1 and 7.

For each disk drive, a storage system keeps its age, the latent sector error count, and the time of the last encountered error. Once a disk fails, the disk repair process consults these statistics. If the remaining $r-1$ disks (in the RAID group) are less than one year old and have not encountered any previous errors, the repair can progress at a normal pace that minimizes the impact on the foreground workload. This can be done without compromising the MTTDL due to a complete failure of another disk within the expected repair time. If the disks are over 1 year old or if at least one of the $r-1$ experienced a failure within the last 1000 minutes, the repair process should proceed at an accelerated pace.

Our definition of normal and accelerated pace is subjective. It depends on, among others, the storage system's workload and the expected service level objectives (SLO).

7. PREVIOUS WORK

There have been very few large scale studies of disk faults. A recent study by Schroeder and Gibson [17] analyzed data involving the failure of about 70,000 disks over a period of five years and found that the failure rate increases over time. Our rate of observed latent sector errors is similar to their reported disk replacement rates. They also found that error rates are not constant with disk age. There are three key differences between their study and ours: the size of the disk pool, the length of the study, and the focus on the analysis of latent sector errors rather than on disk replacements.

Another recent study by Pinheiro et al. [12] analyzed SMART data associated with more than 100,000 disks taken over a nine month period. Using this data they determined correlations between environmental and usage factors and failures. Similar to us, they also found that the annualized failure rate (AFR) is significantly higher for drives older than one year. They also found a high-correlation between the first error and a later drive failure.

Elerath and Shah performed reliability analyses on field failure data resulting from a large pool (hundreds of thousands) of enterprise class disk drives [6, 19, 20]. They found that i) disk drive reliability is highly dependent on the model and ii) different drive failure mechanisms dominate at different points within the drive's lifespan [19, 20]. They found that a variety of factors, some completely independent of drive model (e.g., environmental factors such as heat and duty cycle), greatly impact actual disk drive reliability [6]. The data we have collected gives no indication as to how the latent sector errors occurred, thus we cannot examine correlations between these errors and environmental conditions.

Baker et al. examined the consequences of long-term digital storage [4]. They developed a reliability model that incorporates latent faults, correlated faults, and the detection time of latent faults. They found that it is critical to detect latent faults as soon as possible so that repair is fast, cheap, and reliable. The results of our study help quantify how aggressive repair ought to be to prevent data loss.

Gray and van Ingen performed a large number of read-write cycles on a small set of SATA disk drives attached to a small number of machines [7]. Over the period reported on, they observed thirty uncorrectable read errors propagated outside of the disk. Of these, only three were visible to the user. In their experience, disk errors were not the dominant source of system outages. They also found that when an uncorrectable read error occurs, there are typically many more uncorrectable read errors that follow. Similarly, Talagala found that SCSI disk drives were among the most reliable system components [22]. While SCSI disk drives had

1.9% failure rate over an 18 month period, IDE disk drives had a failure rate of 25% over the same period.

8. CONCLUSIONS

The presented analysis of latent sector errors reveals many interesting aspects of such errors especially when comparing the trends between nearline and enterprise class disks. Although not surprising, we did not expect to see such high degree of temporal locality between successive latent sector error occurrences. Likewise, we did not expect to see that the vast majority of disks developed relatively few errors during the period of 32 months. Even these few errors can cause significant data loss if not detected proactively. However, disk scrubbing appears to be beneficial and can help reduce the mean time to data loss. We believe the trends and observations from this data are significant and will help us build more reliable systems in the future. Future work could study the dependence on operating environment, workload, etc., the correlation with complete disk failures, and the probable data loss under different system configurations.

9. ACKNOWLEDGMENTS

We want to thank Jon Elerath and Sandeep Shah for their valuable insights on disk failures and analysis. Data was gathered with the help of the Autosupport team, including Aziz Htite and Larry Lancaster. Ramon del Rosario helped us verify and understand this data. Rajesh Sundaram and Mayank Saxena helped us understand how software handles different disk errors. David Ford, Steve Kleiman, Brian Pawlowski and members of the Advanced Development Group provided useful comments. Finally, we also thank Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau, Bianca Schroeder and anonymous reviewers for their insighful comments.

10. REFERENCES

[1] B. Allen. Monitoring hard disks with S.M.A.R.T. *Linux Journal*, **2004**(117):9.

[2] D. Anderson, J. Dykes, and E. Riedel. More than an interface: SCSI vs. ATA. USENIX Conference on File and Storage Technologies, p. 245–257, Apr. 2003.

[3] E. Bachmat and J. Schindler. Analysis of methods for scheduling low priority disk drive tasks. ACM SIGMETRICS Int'l. Conference on Measurement and Modeling of Computer Systems. Jun. 2002.

[4] M. Baker, M. Shah, D. S. H. Rosenthal, M. Roussopoulos, P. Maniatis, T. Giuli, and P. Bungale. A fresh look at the reliability of long-term digital storage. EuroSys2006, Apr. 2006.

[5] P. Corbett, B. English, A. Goel, T. Grcanac, S. Kleiman, J. Leong, and S. Sankar. Row-diagonal parity for double disk failure correction. USENIX Conference on File and Storage Technologies, p. 1–14, Apr. 2004.

[6] J. G. Elerath and S. Shah. Server class disk drives: how reliable are they. IEEE Reliability and Maintainability Symposium, p. 151–156, Jan. 2004.

[7] J. Gray and C. van Ingen. *Empirical measurements of disk failure rates and error rates*. MSR-TR-2005-166. Microsoft Research, Dec. 2005.

[8] J. L. Hafner, V. Deenadhayalan, K. Rao, and J. A. Tomlin. Matrix methods for lost data reconstruction in erasure codes. USENIX Conference on File and Storage Technologies, p. 15–30, Dec. 13–16, 2005.

[9] M. K. McKusick, W. N. Joy, S. J. Leffler, and R. S. Fabry. A fast file system for UNIX. *ACM Transactions on Computer Systems*, **2**(3):181–197, August 1984.

[10] Network Appliance Inc. *Introduction to Data ONTAP 7G*. TR 3356, Network Appliance Inc. Oct. 2005.

[11] D. Patterson, G. Gibson, and R. Katz. A case for redundant arrays of inexpensive disks (RAID). ACM SIGMOD Conference on the Management of Data (SIGMOD '88), p. 109–116, Jun. 1988.

[12] E. Pinheiro, W. D. Weber, and L. A. Barroso. Failure Trends in a Large Disk Drive Population. USENIX Conference on File and Storage Technologies, Feb. 13–16, 2007.

[13] F. I. Popovici, A. C. Arpaci-Dusseau, and R. H. Arpaci-Dusseau. Robust, portable I/O scheduling with the disk mimic. USENIX Annual Technical Conference, p. 297–310, Jun. 2003.

[14] V. Prabhakaran, L. N. Bairavasundaram, N. Agrawal, H. S. Gunawi, A. C. Arpaci-Dusseau, and R. H. Arpaci-Dusseau. IRON file systems. ACM Symposium on Operating Systems Principles, p. 206–220, Oct. 2005.

[15] H. Reiser. ReiserFS. http://www.namesys.com/.

[16] S. W. Schlosser, J. Schindler, S. Papadomanolakis, M. Shao, A. Ailamaki, C. Faloutsos, and G. R. Ganger. On multidimensional data and modern disks. USENIX Conference on File and Storage Technologies, p. 225–238, Dec. 2005.

[17] B. Schroeder and G. A. Gibson. Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you? USENIX Conference on File and Storage Technologies, Feb. 13–16, 2007.

[18] T. J. Schwarz, Q. Xin, E. L. Miller, D. D. Long, A. Hospodor, and S. Ng. Disk scrubbing in large archival storage systems. IEEE Int'l. Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, p. 409–418, Oct. 2004.

[19] S. Shah and J. G. Elerath. Disk drive vintage and its effect on reliability. IEEE Reliability and Maintainability Symposium, p. 163–167, Jan. 2004.

[20] S. Shah and J. G. Elerath. Reliability analyses of disk drive failure mechanisms. IEEE Reliability and Maintainability Symposium, p. 226–231, Jan. 2005.

[21] M. Sivathanu, V. Prabhakaran, A. C. Arpaci-Dusseau, and R. H. Arpaci-Dusseau. Improving storage system availability with D-GRAID. USENIX Conference on File and Storage Technologies, p. 15–30, Mar. 2004.

[22] N. Talagala. *Characterizing large storage systems: error behavior and performance benchmarks*. PhD thesis, published as Technical report UCB/CSD–99–1066. EECS Department, University of California, Berkeley, 1999.

[23] *Information Technology: SCSI primary commands (SPC-2)*. T10 Revision 5, Project 1236-D. Sept. 1998.