Disk Failure Investigations at the Internet Archive

Thomas Schwarz[3], Mary Baker[2], Steven Bassi[3], Bruce Baumgart[1], Wayne Flagg[1], Catherine van Ingen[4], Kobus Joste[5], Mark Manasse[4], Mehul Shah[2]

[1] Internet Archive, 116 Sheridan Avenue, Presidio, San Francisco, CA 94129
[2] HP Labs, 1501 Page Mill Road, Mail Stop 1183, Palo Alto, CA 94304
[3] Department of Computer Engineering, Santa Clara University, Santa Clara, CA 95053
[4] Microsoft Research, BARC, 455 Market Street, #1690, San Francisco, CA 94105
[5] Bixdata LLC, 45 W 21st St, Suite 6D, New York, NY 10010

1. Introduction

As storage sites grow in size to thousands of disks, and as the need to predict availability and reliability increases, researchers and designers need a better quantitative understanding of the ways that disks fail or lose data. Unfortunately, these numbers are hard to come by. Disk manufacturers have some approximations to these numbers from their own (mostly accelerated) testing and through customer returns. Those data are not directly relevant to what their customers observe when deploying systems. There are several reasons for this.

- Their focus is not historical; rather, they intend to determine quality of current products (and the budget for warranty funds).
- Most quality tests are based on artificially accelerated testing that tries to predict the failure behavior of disks over their economic lifespan through tests that last at best a few months.
- Forensic (non-accelerated) quality data do not address silent data corruption (bit-rot).
- A large number of returned disks show no problem, though the customer problem that precipitated the disk return was real.

Some large disk customers, such as Network Appliance [ES03], [SE04], [SE05] have failure data; however, most customers do not systematically gather data, because IT groups are in the business of operating storage systems, not data-gathering about them. As a result, much academic and corporate research is based on anecdotes and back of the envelope calculations, e.g. in Schwarz et al. [S&al04].

The Internet Archive is making the first public-domain attempt to obtain and disseminate data about disk failures and disk data loss. The activity has several goals. Historically the first goal was to use already collected data on disk failure over several years at several installations to determine failure rates and thereby better manage the Archive data assets. Very shortly into our investigation, it became obvious that data collection should be planned. Planned measurements give better consistency and easier evaluation. Consequently, the idea of an *Archive Observatory* emerged. We are currently planning experiments to test for bit rot on a large number of drives that became available when the Internet Archive switched to better enclosures.

2. Internet Archive

The Internet Archive is a nonprofit organization based in San Francisco established to preserve Web sites by taking "regular snapshots". It has now extended its mission to preserve as much digital and digitalizable data as possible. Its data set currently grows at about 25 TB/month. Its *Wayback* machine not only allows future historians to access data that would otherwise not be archived, but has already become a daily tool for investigating such things as trademark disputes. It complements its small staff with highly dedicated and educated volunteers.

The Internet Archive stores its archival data at several sites, including sites in Alexandria, Egypt, Amsterdam, Netherlands, and several sites in San Francisco, CA. For cost reasons, it stores data on desktop ATA disks. These are now located in four-disk pizza-box-form-factor nodes that replace ones with a bulkier form factor. Of the four disks, one is the Linux OS boot disk, while the others only store data configured as a JBOD. There are 40 storage nodes in a rack and the San Francisco clusters have 36 racks. Over time, the environmental quality of sites has differed. At one point, a

main site did not have any air-conditioning, but those servers have now been transferred to a much more adequate location.

The workload of the boot disks is similar to other server applications, whereas disks that only carry data have a small write load and a usually small but occasionally intense read load. The Internet Archive stores most of its current data in ~100MB archive files (ARC files). These files are written once and very rarely accessed. However, there are extreme hotspots caused by demand for certain data. For example, after the big Indian Ocean tsunami of 2005, videos of that event were uploaded and subsequently became very popular downloads. As another example, when the music group *The Grateful Dead* changed its policy on allowing fans to offer personal recordings of concerts on their own websites, the change triggered "panic" music downloads from the Archive.

As an archive, the Internet Archive is interested in storing all of its data. Some data is more important than others and is either mirrored or triplicated over several systems. Other data, such as data from the Wayback machines is not so important since crawls at different times read and archive often the same web pages. Therefore, only one copy is stored. If this is lost, then the state of that website cannot be reconstructed, though most of the content of the lost visit is found in the ARC file of the previous or subsequent visit. To assess the integrity of its data, the Internet Archive calculates and stores the MD5 hash value of each archive file. It periodically accesses the files to recalculate the MD5 hash. In some instances, these diagnostic accesses are the majority of the accesses seen by a file. In addition, the Archive also stores SMART statistics and kernel logs.

One set of past data, mainly from American installations, covers 4973 disks. It will soon be augmented by more recent data from 2005 and 2006. In addition, there is a collection of manual log entries recording when disks were taken out of service and what data located on them was "evacuated" for reconstruction and storage. The logs also have anecdotal comments. All digitally available data is made freely available at the Internet Archive. In addition, our results are public and will be posted on the Internet Archive website [IA06].

3. Goals and Methodology

Much of our work is currently concentrated on using archival failure data to determine disk failure rates and bit rot rates as a function of environmental factors, disk age and disk vintage. Trying to make sense out of the data set taught us how hard the analysis is. Disk failure, sector failure, and bit-rot (silent corruption) rates are a quotient of occurrence counts over total service times. Both numerator and denominator are hard to determine. At the level of the user / system administrator, disk failures, file system failures, controller failures, bus failures, etc. all appear as a generic failure. Designing and hardening disk based systems, however, requires understanding the failure types in greater detail.

For example, we observe instances where a disk behaves erratically; but, when moved to another node the disk functioned well. There are at least four different possible root causes for this behavior, and each one has very different implications for the interpretation of the data. First, the original disk mounting failed, allowing the disk to vibrate. When moved, the new mounting was secure. Second, the original interface cable (or controller) for the disk was damaged. Any disk that replaced the original disk will unfortunately also fail intermittently. Third, the disk has partially failed. When moved, new data is written to the disk in different places and / or the failing locations were remapped. Fourth, the disk was indeed behaving erratically.

Gathering accurate population data is even more difficult. Disks are identified by a serial number, though there were about a dozen instances where the serial number was a fake zero identifier. While in principle disks remained in the node in which they were initially mounted, some problems were taken care off by swapping disks. Manual logs were not always accurate, understandably, since disk swaps are typically triggered by failures.

SMART makes a large set of environmental, error and other metrics available. We used the Linux smartctl utility to interpret the SMART data [SCTL]. SMART data interpretation and format can vary between manufacturers (three in the case of the Internet Archive) and some SMART counters (such as the disk power-on lifetime counter) consist of only 16 bits that rolls over after 45.5 days on a considerable part of our population. Nevertheless, SMART allows us to identify whether a disk has been in continuous use (and not taken out and later returned to service), to gather environmental data, and most importantly, to analyze the condition of a disk through SMART error codes.

To test for bit-rot, MD5 digests were used on complete files. MD5 is a widely used cryptographic hash function with a 128 bit hash value (see RFC 1321). An MD5 digest changes its value with extremely high probability if the file changes.

As with almost all real world data, our measurements have gaps or periods without data. Gaps may be caused physically, for example by powering down a system. Gaps may also be caused by losing or corrupting the actual SMART or MD5 measurement data. As such, tracking a disk over time is difficult. We developed the notion of "chains" in order to deal with this A chain is a set of measurements of a single object that are contiguous, where with "contiguous" we mean a measurements taken so close to each other that we can infer that the disk or file stayed in place. We then developed database schemes in order to capture data on MD5 and SMART chains. All statistics can then be computed over both the cumulative chain time or, occasionally, cumulative wall clock time. We can also preferentially select short or long chains for analysis. All of this work is also in preparation for the *Archive Observatory*, our instrumentation of the Internet Archive storage system.

4. Preliminary Results

4.1. Bit-rot Data

We gathered and analyzed data on MD5 that give some insight into the occurrence of bit-rot. The same data were used in an assessment of long-term reliability of digital storage [B&al06]. The state history of each archive file is represented by a series of MD5 digests. A total of 112,865,205 MD5 digests for about 4,717,158 files are available, though so far only a small fraction (1,496,572 records representing 36,169 files) has been analyzed. Not all changes of an MD5 digest indicate actual bit-rot. Often, a transient change of an MD5 digest seemed to be caused by a temporary inability to access the file. Out of 10 disks where a different MD5 value for a file was recorded, only two were ultimately attributed to likely disk failure. Sometimes, the MD5 digest switches to a new value but then returns to the original value, most likely the result of a problem with the node, controller, bus, or processor computing the MD5, but not with the disk if other files on the same drive do not experience MD5 digest changes.

4.2. SMART Data

We collected the SMART data in a single database. In a separate log that we have not yet succeeded in integrating, we have data on the load for each block device, including read and write operations per second.

Our current data do not yet allow us to characterize why disks leave the system. A disk that powers down and does not restart (i.e. a classical disk failure) and a disk that is decommissioned look the same to our database. Currently, we have 93,385 chains. 470 of these chains exhibit failure, sometimes reflecting oscillation between failure and normal state. We call this behavior *bouncing*. These data represent total measurements of 1,039,516.27 days. In 18 chains, SMART indicated a failure free state followed by a failure state for the rest of the chain. 6 of these disks left the population. A total of 36 out of 4,978 disks (0.7%) exited the population after exhibiting SMART indicated failures. If a disk powers down and then does not come up, we only see a disk exiting the population; we are not able to say whether the exit was planned or due to failure.

SMART data indicate either a healthy or failing state of a disk. We observed 136 "*oscillating*" chains that change state more than once (out of the 470 chains with failure). However, 94 of these seem to have a spurious indication of a healthy state. The remaining 42 oscillating chains represent 27 different disks out of a population of 4978. In one case only, the disk was moved to another node, and in this case, the disk continued to form an oscillating chain. 8 out of the 27 disks seemed to have left the population permanently and prematurely. All 8 formed oscillating chains, typically 2-3. In these cases, SMART gave failure warnings before the disk actually stopped working or before it was pulled from the machine.

Our next step is to correlate SMART data with data on bit-rot. We have identified a number of disks that housed files whose MD5 either changed or oscillated. Our database allows us to correlate these events and SMART reports of physical conditions.

The Internet Archive has logs that record disks being removed from nodes because of instability or failure. With our database, we are now in the position to access SMART data for these drives before their removal.

4.3. Disk Failure Rates

Based on maintenance records, the Archive does not see systematic disk infant mortality. However, it experienced in 2005 two incidents of high failure rates in two unrelated batches of disks. The manufacturer attributed this to shipping or handling mistreatment. Our Archive Observatory has now data for 2489 disks for almost the complete year 2005. Disk failures appeared at a rate of 2.00%, but the data also show a combined failure rate of 2.54% for motherboard, CPU, memory, et cet. failure. In this period, almost half the replacements were due to substantial sector failure. Replacements in the past have been as high as 6%. The lower values now observed are in line with [GI05].

5. Future Work

Obviously, much work still needs to be done before our initial target, is completed. At the same time, our efforts have shown the need to gather data in a more systematic way, i.e. the setting up of the *Archive Observatory*. We are also starting bit-rot measurements. For our first set of experiments, we will be using sixty nodes with up to 500 decommissioned drives and simulate an archive (write once, read often) environment. We will then measure the amount of bit-rot observed for sixty days. Actually, this experiment will exclude certain causes of bit-rot such as wayward writes (that overwrite data on a neighboring track). To our knowledge, this is the first large-scale attempt to observe bit-rot in the laboratory. It expands on the experiment done by Gray and v. Ingen reported in [GI05], in which many fewer errors were observed than predicted by the drive manufacturer specifications. Currently, at the beginning of April 2006, almost a PB of data has been stored and checked; no genuine incident of bit-rot has been discovered, but a number of false initializations have been seen [IA06].

Even if we completely succeed in analyzing disk failure behavior at the Internet Archive, we only obtain a single data point. We would like to invite other institutions with large disk populations to join the effort to measure actual disk failure behavior. Ultimately, we hope to increase the openness about failure rates and modes actually seen in practice.

Acknowledgments

We would like to thank the staff members of the Internet Archive, especially James Akers, Joerg Bashir, Jeff Gerard, Bill Moyer, Brad Tofel, and Eric Volpe. Special thanks go to Brewster Kahle, the digital librarian at the Internet Archive. Steven Bassi and Thomas Schwarz would like to thank Microsoft Research (Bay Area Research Center and Silicon Valley Research Center) and in particular Jim Gray and Roy Levin for generous financial support through a research gift.

References

[B&al06] M. Baker, M. Shah, D. Rosenthal, M. Roussopoulos, P. Maniatis, T. Giuli, and P. Bungale: A Fresh Look at the Reliability of Long-term Digital Storage, Proc. of Eurosys, Leuven, Belgium, 2006.

[ES03] J. Elerath and S. Shah: Disk Drive Reliability Case Study: Dependence Upon Fly-Height and Quantity of Heads, Proc., Annual Reliability and Maintainability Symposium (RAMS), 2003.

[GI05] J. Gray, C. v. Ingen: Empirical Measurements of Disk Failure Rates and Error Rates, Microsoft Research Technical Report MSR-TR-2005-166.

[IA06] Bit-rot experiment results published at http://www.us.archive.org/ao/wip.

[S&al04] T. J. E. Schwarz, Q. Xin, E. L. Miller, D. D. E. Long, A. Hospodor, S. Ng: Disk Scrubbing in Large Archival Storage Systems. 12th IEEE Intern. Symp. on Modeling, Analysis and Simulation of Computer and Telecommunications Systems (MASCOTS), Volendam, The Netherlands.

[SE05] S. Shah and J. Elerath: Disk Drive Vintage and Its Effects on Reliability. Proc., Annual Reliability and Maintainability Symposium (RAMS), 2004.

[SE05] S. Shah and J. Elerath: Reliability Analysis of Disk Drive Failure Mechanisms, Proc., Annual Reliability and Maintainability Symposium (RAMS), 2005.

[SMCTL] Smartctl web man page at at smartmontools.sourceforge.net