

COMPUTERWORLD

Storage

June 21, 2007

Opinion: Real-World Disk Failure Rates Offer Surprises

Disks don't have a sweet spot of low failure rate

June 21, 2007 (LinuxWorld) At this year's Usenix File and Storage Technologies Conference in San Jose, we were treated to two papers studying failure rates in disk populations numbering over 100,000. These kinds of data sets are hard to get: First you have to have 100,000 disks, then you have to record failure-related data faithfully for years on end, and then you have to release the data in a form that doesn't get anyone sued (see "Disk drive failures 15 times what vendors say, study says").

The storage community has salivated after this kind of real-world data for years, and now we have not one, but two long-term studies of disk failure rates. The conference hall was packed during these two presentations. When the talks were done, we stumbled out into the hallway, dazed and excited by the many surprising results. Heat is erroneously correlated with failure! Failures show short- and long-term correlation! SMART (self-monitoring, analysis and reporting technology) errors do mean the drive is more likely to fail, but one-third of drives die with no warning at all! The size of the data sets, the quality of analysis and the nonintuitive results win these two papers a place on the Kernel Hacker's Bookshelf.

The first paper (and winner of Best Paper at the conference) was "Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?," by Bianca Schroeder and Garth Gibson. They reviewed failure data from a collection of 100,000 disks, over a period of up to five years. The disks were part of a variety of HPC clusters and an Internet service provider. Disk failure was defined as the disk being replaced. The date of replacement was used as the date of the failure, since determining exactly when a disk failed was not possible.

Their first major result was that the real-world annualized failure rate (average percentage of disks failing per year) was much higher than manufacturers' estimates: an average of 3% vs. the estimated 0.5% to 0.9%. Disk manufacturers obviously can't test disks for a year before shipping them, so they stress test disks in high-temperature, high-vibration, high-workload environments, and use data from previous models to estimate MTTF. Only one set of disks had a real-world failure rate less than the estimated failure rate, and one set of disks had a 13.5% annualized failure rate!

More surprisingly, they found no correlation between failure rate and disk type -- SCSI, SATA or Fibre Channel. The most reliable disk set was composed of only SATA drives, which are commonly regarded to be less reliable than SCSI or Fibre Channel.

In another surprise, they debunked the "bathtub model" of disk failure rates. In this theory, disks experience a higher "infant mortality," or initial rate of failure, then settle down for a few years of low failure rate and then begin to wear out and fail. The graph of the probability vs. time looks like a bathtub, flat in the middle and sloping up at the ends. Instead, the real-world failure rate began low and steadily increased over the years. Disks don't have a sweet spot of low failure rate.

Failures within a batch of disks were strongly correlated over both short and long time periods. If a disk had failed in a batch, then there was a significant probability of a second failure up to at least two years later.

If one disk in your batch has just died, you are more likely to have another disk failure in the same batch. Scary news for RAID arrays with disks from the same batch. A recent paper featured at the 2006 Workshop on Storage Security and Survivability, "Using Device Diversity to Protect Data against Batch-Correlated Disk Failures," by Jehan-François Pâris and Darrell D. E. Long, calculated the increase in RAID reliability from mixing batches of disks. Using more than one kind of disk increases costs, but with the combination of data from these two papers, RAID users can calculate the value of the extra reliability and make the most economical decision.

A second paper, "Failure Trends in a Large Disk Drive Population," by Eduardo Pinheiro, Wolf-Dietrich Weber and Luiz André Barroso, reported on disk failure rates at Google. They used a Google tool for recording system health parameters and many other staples of Google software (MapReduce, BigTable, etc.) to collect and analyze the data. They focused on SMART statistics -- the built-in disk drive monitoring in many modern disk drives that records statistics about scan errors and blocks relocated.

The first result agrees with the first paper: The annualized failure rate was much higher than estimated, between 1.7% and 8.6%. They next looked for correlation between failure rate and drive utilization (as estimated by the amount of data read or written to the drive). They find a much weaker correlation between higher utilization and failure rate than expected, with low utilization disks often having higher failure rates than medium utilization disks and, in the case of the three-year-old vintage of disks, higher than the high-utilization group.

Now for the most surprising result. In Google's population of cheap ATA disks, high temperature was erroneously correlated with failure! In the authors' words: "In fact, there is a clear trend showing that lower temperatures are associated with higher failure rates. Only at very high temperatures is there a slight reversal of this trend."

This correlation held true over a temperature range of 17 to 55 degrees Celsius. Only in the three-year-old disk population was there correlation between high temperatures and failure rates. My completely unsupported and untested hypothesis is that drive manufacturers stress test their drives in high-temperature environments to simulate longer wear. Perhaps they have unwittingly designed drives that work better in their high-temperature test environment at the expense of a more typical low-temperature field environment.

Finally, they looked at the SMART data gathered from the drives. Overall, any kind of SMART error correlated strongly with disk failure. A scan error occurs when the disk checks data in the background, reading the entire disk. Within eight months of the first scan error, about 30% of drives would fail completely. A reallocation error occurs when a block can't be written, and the block is reassigned to another location on disk. A reallocation error resulted in about 15% of affected drives failing within eight months. On the other hand, 36% of the drives that failed had no warning whatsoever, either from SMART errors or from exceptionally high temperatures.

For Google's purposes, the predictive power of SMART is of limited utility. Replacing every disk that had a SMART error would end up replacing good disks that will run for years to come about 70% of the time. For Google, this isn't cost-effective, since all its data is replicated several times. But for an individual user for whom losing his disk is a disaster, replacing the disk at the first sign of a SMART error makes eminent sense. I have personally had two laptop drives start spitting SMART errors in time to get my data off the disk before it died completely.

Overall, these are two exciting papers with long-awaited real-world failure data on large disk populations. We should expect to see more publications analyzing these data sets in the years to come.

Valerie Henson is a Linux file systems consultant specializing in file system check and repair.

Copyright © 2008 Computerworld Inc. All rights reserved. Reproduction in whole or in part in any form or medium without express written permission of Computerworld Inc. is prohibited. Computerworld and Computerworld.com and the respective logos are trademarks of International Data Group Inc.