RIK FARROW

musings

rik@usenix.org

NOT EVERYONE WILL BE AS FASCI- nated by hardware as I am. But some people certainly are, so I will feed those who want to know more about the physical parts of the computers that, in the end, feed those of us who work with them.

I got to attend the tutorial given by Dave Anderson and Willis Whittington, both longtimers at Seagate Technologies. Anderson and Whittington shared what they could, that is, that which is not proprietary and secret, before the 5th USENIX Conference on File and Storage Technologies began in San Jose in February 2007. Later on, I will share a different perspective with you, as researchers presented two papers about disk failure rates that conflict with what drive vendors report. But for now, let me present a digest of what I learned.

First and most obvious, Anderson and Whittington speak from the manufacturer's perspective. Before you go mentally disregarding everything they say, you need to realize that they live in the world where shiny new drives get made, drives that people expect will have high capacities, high I/O rates, and low error rates and will last at least five years while costing as little as possible. This list of expectations is self-conflicting to start with. And, from a vendor's perspective, testing how long drives will survive is actually impossible, outside of field tests, at which point, said drives will be obsolete by several years.

Drive vendors see their market split into many categories, certainly more than I had considered: enterprise, near-line enterprise, home PC, notebook, and consumer devices. As a computer user and researcher, I find myself focused on just three of these: the enterprise drives, SCSI, FC, and SAS; near-line, SATA and FC; and PC, SATA. The old PC standard, ATA, is now called PATA, for Parallel ATA, and is expected to disappear, with the exception of replacement drives, very soon.

The consumer market for drives is the most volatile, with price and capacity being the driving factors. Vendors view the enterprise market differently, with reliability and high I/O rate being critical. Note that these categories were not created by drive vendors but are driven by the demands of the two biggest purchasers of hard drives: the makers of high-end servers and storage systems. If your entire business revolves around providing fast and reliable storage systems (EMC and Network Appliance as examples), then the behavior of the millions of drives you use each year influences buyer perception of your own servers.

Anderson and Whittington didn't spend much time explaining where the demand for enterprise drives comes from. I just wanted to make that clear myself, as enterprise drives are designed and manufactured to suit the needs of special classes of users. Those differences do show up as they talk about the hardware, so knowing the reason for enterprise drives helps to make sense of why enterprise drives have lower capacities, have higher I/O rates, and cost more.

Speaking of drive capacities, the areal density—the product of bits per inch times tracks per inch—is the key factor. Bits per inch (BPI) means the number of bits that can be written and later reread per linear inch. You might first think that BPI just has to do with the magnetic coating on a surface, but you would be wrong. The magnetic coating is just one of six layers, starting with the substratum, which can be aluminum or glass (used in notebook drives for its greater rigidity), and ending with a lubrication layer. The number of magnetic grains is a limiting factor for BPI, and one that gets attacked by creating smaller grains, all of nearly the same size. In the future, the magnetic coating will likely be composed of self-organized particles in the 6.3 +/- 0.3 nm range. The current particle sizes range from 8 to 15 nm.

Head technology represents another limiting factor for BPI. The head flies over the surface of a disk at about the distance of a wavelength of visible light (about 0.5 micrometers) and as fast as 118 miles per hour (in 15k rpm drives). Heads must be fabricated to exacting standards to read and write the tiny magnetic regions on narrow tracks. In the most recent advance in head technology, called perpendicular recording, the write field penetrates the media at a right angle, instead of along the surface of the media (longitudinal recording). The read portion of the head now uses GMR, Giant Magnetoresistive effect, to sense the magnetic orientation of bits. At current bit densities, 80–100 grains make up one bit.

The number of tracks that can fit within an inch is governed by head positioning, runout, and rotational vibration (RV). Head positioning is the easiest to grasp, but it requires great precision when there can be over 90,000 tracks per inch. Runout describes the shape of a track, which is never quite round. So following a track long enough to read a sector not only means seeking to the correct track but also following the track, because it does not describe a circle.

As if this feat weren't difficult enough, RV indicates the amount of vibration, created by other hard drives as they seek, by fans, and by other sources of vibration. Consider that if you have a single hard drive, each time it seeks, its case (and thus its mounting hardware) must resist the angular momentum created by swiveling the head. Now, put a bunch of hard drives into one unit, then stack many of those units up in a rack, and imagine all of the shaking going on, all in the same approximately horizontal plane within which the drives are rotating their heads. Anderson described tests of drive cabinets where one-third of the cabinets tested allowed an unacceptable level of RV for any type of drive. The more rigid the drive mounts, where metal is good and plastic bad, the better the cabinet.

This is also one of the areas where enterprise drives differ from other drives. Enterprise drives have two accelerometers, each sensing movement about the drive axis, and one drive CPU (enterprise drives have two) works to compensate for RV, keeping the head on track. All drives have positioning information created when the drives are formatted, and this information is used to keep the heads aligned with the track too. But enterprise drives can recover from more RV (21 rad/s^2) than SATA drives

(just 6 rad/s^2). If the drive fails to follow a track, there will be a read error, and the drive CPU will reattempt the read. The goal is for enterprise drives to have higher I/O rates in environments with lots of activity by avoiding having to reread sectors.

Reading the data from sectors also differs between enterprise and other drives. All drives include Error Correction Code (ECC) that allows the drive to recover from bit errors while reading data. Enterprise drives also include Error Detection Code (and this is not the CRC that's included when data is sent to the drive) and an additional IOECC that makes it possible for enterprise drives to recover from more bit errors than other drives. Enterprise drives also use additional sync marks within the data field, instead of just at the beginning of the data field, as in other drives. All of these techniques subtract from the amount of space left for data in exchange for more reliability.

I've already mentioned that future development of disk drives will require a smaller grain size and more even distribution, to increase the areal density. Another future technique will be the development of write heads that include a small laser that heats the grains before the perpendicular write head passes over them. With this Heat Assisted Magnetic Recording (HAMR), consumer drives are expected to reach capacities of 8 TB by 2013, and enterprise drives 2.4 TB.

And what about increasing disk rpms? Today, only enterprise drives spin at 15,000 rpm, with consumer drives running at 7,200 rpm and notebook drives at 5,400 rpm. Increasing the rpms increases the IO transfer rate, as more bits pass under the head in each second the faster the disks spin. But the power required to rotate a disk increases as the cube of the rpm. The roadmap does have consumer drives reaching 10,000 rpm and enterprise drives staying at 15,000 rpm. Because areal density will be increasing as well as rpms, consumer drives should reach a maximum transfer rate of 5 GB/s by 2013, but enterprise drives will actually be slower, at 4 GB/s (because their areal density is lower).

Enterprise read seek times are already about half that of consumer drives (which have seek times of 8 ms, compared to 3.7 ms for enterprise drives), and this ratio will remain about the same, with only modest improvements in read seek speed (seek times of 6.5 ms to 2.8 ms projected for 2013).

There was, of course, much more in this half-day class, and if you ever get a chance to listen to either Anderson or Whittington speak, I'd recommend that you be there if you find yourself fascinated by the details of modern disk drives.

The Competition

Not only do disk vendors compete with each other, they now find themselves competing with memory vendors as well. Various forms of flash memory have improved in speed, capacity, and number of rewrites while offering lower cost, and you can already buy flash "drives" for laptops, as well as the now ubiquitous USB memory sticks, at prices that compare well with hard drives ($10/GB for flash versus $1/GB or less for disks). As an interesting side note, IBM developed the first form of rotating magnetic memory, which cost $10,000/MB (in 1956 US dollars; perhaps $70,000/MB in today's dollars). That makes my first hard drive, at $60/MB, or $2000 for a 34-MB drive, not seem quite so outrageous.

The competition to disk vendors that I really want to address is not other hardware vendors, but researchers. During the first session at FAST '07,

two groups presented papers in which they examined hard-drive replacement rates based on field data. Disk vendors perform accelerated aging tests on the drives they build by subjecting the drives to high temperatures and high utilization (continuous IO with lots of seeks) in an attempt to tease out how long a particular drive type will last. The vendors publish the Annual Failure Rate, AFR, based on these tests. In these two papers, the researchers report actual failure rates several times higher than the ones suggested by vendors.

Bianca Schroeder (with Garth Gibson, winner of the Best Paper Award) collected information about disk failures from several High Performance Computing (HPC) centers as well as a couple of Internet service providers. Although the information collected from each of the sources differed in many ways, she statistically analyzed the data to pry out a number of interesting observations. For example, the expected rate of failures for disk drives is supposed to resemble a curve like a bathtub, with high failure rates at the beginning of drive life as well as toward its end. In Schroeder's analysis, the failure rate, which she termed ARR for Annual Replacement Rate, was highest in the third and fourth years, placing a big hump where there is supposed to be a comfortable dip in the replacement graph.

Schroeder's paper lists many observations, and I suggest you read her paper for all of them. I do want to mention another point that you need to be aware of: the possibility of a drive failing while a RAID system is in the process of rebuilding the replacement drive. The standard (and vendor) view of this process is based on the Unrecoverable Error Rate (UER), something that Anderson and Whittington discussed in their tutorial. Enterprise drives have a lower UER, 10^{-16}, compared to SATA drives, 10^{-14}. To rebuild one disk in a RAID 5 array composed of 5 500-GB SATA drives, 2^{13} bits must be read successfully, one-fifth the value of the UER for SATA drives. In other words, the odds of encountering a second error while rebuilding this RAID 5 array are 1 in 5. For people who are counting on RAID for reliable access to data, a 20% chance of failure is much too high.

Although this potential for failure already appears high, Schroeder shows that it fits poorly with observed data. First, just consider this quote from Schroeder and Gibson:

> The failure probability of disks depends for example on many factors, such as environmental factors, like temperature, that are shared by all disks in the system. When the temperature in a machine room is far outside nominal values, all disks in the room experience a higher than normal probability of failure.

I think we can agree that this makes good, intuitive sense, partially shredding the notion of relying on UER for calculating risk without looking at real data. Then Schroeder goes on to test for autocorrelation: the notion that disk failures appear to be related in time. If one just considers UER, failures should be completely random and unrelated. Schroeder shows that, in practice, disk failures appear related, exhibiting a decreasing hazard rate over time. A decreasing hazard rate implies that a subsequent disk failure is likely to occur sooner, rather than later. So the likelihood of a second disk failure while rebuilding a RAID 5 array appears much higher than a simple UER suggests.

I have felt uncomfortable when I hear or read about people relying on RAID systems with no backups. All I had to rely on was the UER, which seemed dangerous enough when applied to large arrays. But Schroeder's work makes relying on RAID without a backup appear more like expecting

lightning not to strike in the same place twice, even if the spot in question is the antenna on top of a very tall building. RAID arrays are composed of drives all within the same environment, likely the exact same type of drive manufactured perhaps within the same batch.

Schroeder also observes that she did not find any difference in the failure rates for SATA and SCSI drives in her data. One of the points in building, or buying, enterprise drives is to gain a higher level of reliability, but the data in this case do not back up that goal. I find this point very interesting, as both disk and file server manufacturers appear to believe in enterprise drives, and I suspect they have reasons that go beyond the higher profit margins in enterprise drives.

Google Drives

Schroeder and Gibson weren't the only people looking at drive failures at FAST '07. Eduardo Pinheiro, Wolf-Dietrich Weber, and Luiz André Barroso of Google Inc. wrote "Failure Trends in a Large Disk Drive Population," which examines another very large data set about disk drive failures. Google is known to collect huge amounts of data on vast distributed storage systems. Just how large is actually a secret, and Pinheiro et al. don't tell us exactly how many drives, but they do tell us that they are looking at "more than one hundred thousand," a decent-sized data set.

This paper shares one of the problems faced by Schroeder and Gibson: The exact failure time of drives is often unknown. Pinheiro et al. use the time of replacement of drives as their failure metric, and the reason for failure was not included in their data. The focus of this paper is quite different in that they examined SMART (Self-Monitoring, Analysis and Reporting Technology) data collected from disk drives to see if this data could be used to predict drive failures.

SMART has always seemed like a good idea to me. Modern disk drives are embedded systems, and having the drive expose some of the data it collects makes perfect sense. But I can't say that I have routinely run the SMART data collection tools, as I've experienced plenty of disk failures, usually at the most inappropriate times and without any useful warnings. Pinheiro reports, sadly enough, similar findings.

Like Schroeder, Pinheiro reports an AFR that is not at all bathtub-shaped, with the same big hump in the middle. The Google data actually shows peaks earlier than Schroeder's data, at years two and three. Google data also includes a utilization metric, missing from the other paper. The expected result would be that heavily utilized drives, particularly the SATA drives favored by Google, would fail more frequently. In fact, their data show no difference between heavily and lightly utilized drives, except in the first three months of use and during the fifth year. The authors suggest that the spike in failures in heavily utilized drives represents

> the survival-of-the-fittest theory. It is possible that the failure modes that are associated with higher utilization are more prominent early in the drive's lifetime. If that is the case, the drives that survive the infant mortality phase are the least susceptible to that failure mode, and the resulting population is more robust with respect to variations in utilization levels.

This finding might also account for what disk vendors discover when they stress-test new drives. However, Pinheiro et al. did not perceive any significant difference in the rate of drive failures related to higher temperatures.

Pinheiro et al. looked at four SMART data variables to see whether they can predict drive failure. Scan errors (when the drive detects an error while performing reads during background testing) do indicate that these drives are 39 times more likely to fail within 60 days than drives with no scan errors. Reallocation errors (when a drive remaps a sector because of repeated soft read errors or a hard read error) also indicate that a drive may be likely to fail sooner than drives with no errors (being 14 times more likely to fail within 60 days).

Pinheiro et al. examine seven other SMART parameters, then attempt to create predictive models for drive failures. Since some SMART data appears highly correlated with drive failures, they hoped they could create a predictive failure model. Unfortunately, using SMART data, with and without temperature values, still left 36% of all replaced drives with no failure signals at all. The authors conclude that SMART data is useful for provisioning, as it can predict the aggregate reliability of large disk populations, but it cannot suggest when an individual drive is about to die. Too bad.

Intelligent Drives

Disk drives have been getting "smarter" for many years. I did ask Dave Anderson whether programmers should make any assumptions about the relationship between logical disk layout and the physical disk. Anderson told me that we should forget any notion of "cylinders"; as for disk layout, he implied that when writing a collection of logically sequential blocks a disk would attempt to write those blocks so that they could be read again quickly. In other words, what happens inside the physical disk may be quite different from what we expect. I asked people involved with Linux and BSD filesystem design whether they knew about this, and both said they were quite aware that disks, not the filesystem designer, have control over where data gets written. I was a bit amazed, even though I had heard stories about this. Guess I am just a bit out of date.

Given that disk drives have gotten a lot smarter, perhaps it makes sense to share more responsibility for file systems. I believe that day is coming, and you will see research in other FAST papers that considers object data storage, making the disk aware of file metadata, and other newfangled notions. Take a look at the FAST summaries included in this issue for more new ideas about file systems.

I also suggest you read the filesystem articles included in this issue of *;login:*. Kirk McKusick leads off with an excellent survey of UNIX filesystem design since 1980, a must for anyone who wants to understand modern file systems. Pawel Jakub Dawidek writes a related but much more focused article about porting ZFS to FreeBSD. You can learn a lot more about ZFS as well as modern OS support for new filesystem designs by reading Dawidek's article. To provide a bit of balance, the lead authors of ext4, the newest version of the Linux ext file system lineage, explain motivations behind creating a new filesystem type, as well as the advantages they have seen in performance and capabilities in this new design. I had expected to have an article on XFS as well, but that will have to wait for another time.

We have two articles about security this month. Dan Geer has been studying security metrics for years now, and he has created a talk that examines the future of security based on current trends. If you want to have a feel for current threats and get a better idea of the security threats you can expect to be facing over the next several years, I invite you to study Geer's observations. Also in the security section, Vassilis Prevelakis demonstrates

how you can use VMware and VMs to simulate both local and routed networks for security classes.

In the Legal section, Dan Appelman concludes his two-part series on spam, blogs, U.S. law, and the system administrator. Appelman provides advice that can be followed by diligent system administrators, whether or not they work in the United States. Alexander Muentz follows Appelman, with a different way of looking at search warrants, subpoenas, and other forms of legal demands. Muentz compares these demands to a DoS attack and suggests both how to prepare for potential demands and how to handle them.

Two regular columnists opted not to submit columns for this issue, but David Blank-Edelman did decide that we needed more entertainment when learning Perl and the Acme module. Before the book review section, packed as usual, Robert Ferrell exercises his development skills with his very own filesystem design.

I've already mentioned that we have FAST summaries, but we also have the summaries from the Linux Storage & Filesystem Workshop. As you might imagine, the workshop and FAST sparked my imagination to create this longer than usual Musings—some things just fire me up.

A while back, I wrote in "Musings" that I didn't yet feel as though I was living in the future. I was referring to the future I had seen in images when I was growing up, with satellite dishes everywhere and flying cars. Since the time I wrote that column, I've acquired a microwave dish on my roof, solar panels, and a hybrid car and can claim to feel vague stirrings of the future around me. But I still run insecure operating systems, have disk drives I can't trust (but am willing to back up), and carry both a cell phone and a laptop when I travel. My own vision of the future includes more than just all-electric vehicles: it also includes a computing device I carry with me everywhere that provides secure storage, networking, and identification. Server systems, too, need to be more reliable, more secure, and easier to manage. We still have a long way to go, with lots of interesting work ahead for enterprising computer scientists.