Why RAID 6 stops working in 2019

Three years ago I warned that RAID 5 would stop working in 2009. Sure enough, no enterprise storage vendor now recommends RAID 5. Now it's RAID 6, which protects against 2 drive failures. But in 2019 even RAID 6 won't protect your data. Here's why.

By Robin Harris for Storage Bits | February 22, 2010 -- 06:50 GMT (22:50 PST) | Topic: Storage

Three years ago I warned that RAID 5 would stop working in 2009 (http://blogs.zdnet.com/storage/?p=162). Sure enough, no enterprise storage vendor now recommends RAID 5.

They now recommend RAID 6, which protects against two drive failures. But in 2019 even RAID 6 won't protect your data. Here's why.

The power of power functions I said that even RAID 6 would have a limited lifetime.

> . . . RAID 6 in a few years will give you no more protection than RAID 5 does today. This isn't RAID 6's fault. Instead it is due to the increasing capacity of disks and their steady URE rate.

Late last year Sun engineer, DTrace co-inventor, flash architect and ZFS developer Adam Leventhal, did the heavy lifting to analyze the expected life of RAID 6 as a viable data protection strategy. He lays it out in the Association of Computing Machinery's Queue magazine, in the article Triple-Parity RAID and Beyond (http://queue.acm.org/detail.cfm?id=1670144), which I draw from for much of this post.

The good news: Mr. Leventhal found that RAID 6 protection levels will be as good as RAID 5 was until 2019.

The bad news: Mr. Leventhal assumed that drives are more reliable than they really are. The lead time may be shorter unless drive vendors get their game on. More good news: one of them already has - and I'll tell you who that is.

The crux of the problem RAID arrays are groups of disks with special logic in the controller that stores the data with extra bits so the loss of 1 or 2 disks won't destroy the information (I'm speaking of RAID levels 5 and 6, not 0, 1 or 10). The extra bits - *parity* - enable the lost data to be reconstructed by reading all the data off the remaining disks and writing to a replacement disk.

The problem with RAID 5 is that disk drives have read errors. SATA drives are commonly specified with an unrecoverable read error rate (URE) of 10^14. Which means that once every 200,000,000 sectors, the disk will not be able to read a sector.

2 hundred million sectors is about 12 terabytes. When a drive fails in a 7 drive, 2 TB SATA disk RAID 5, you'll have 6 remaining 2 TB drives. As the RAID controller is reconstructing the data it is very likely it will see an URE. At that point the RAID reconstruction stops.

Here's the math: (1 - 1 /(2.4 x 10^10)) ^ (2.3 x 10^10) = 0.3835

You have a 62% chance of data loss due to an uncorrectable read error on a 7 drive RAID with one failed disk, assuming a 10^14 read error rate and ~23 billion sectors in 12 TB. Feeling lucky?

RAID 6 RAID 6 tackles this problem by creating enough parity data to handle 2 failures. You can lose a disk *and* have a URE and *still* reconstruct your data.

Some complain about the increased overhead of 2 parity disks. But doubling the size of RAID 5 stripe gives you dual disk protection with the same capacity. Instead of a 7 drive RAID 5 stripe with 1 parity disk, build a 14 drive stripe with 2 parity disks: no more capacity for parity and protection against 2 failures.

Digital nirvana, eh? Not so fast, my friend.

Grit in the gears Mr. Leventhal points out is that a confluence of factors are leading to a time when even dual parity will not suffice to protect enterprise data.

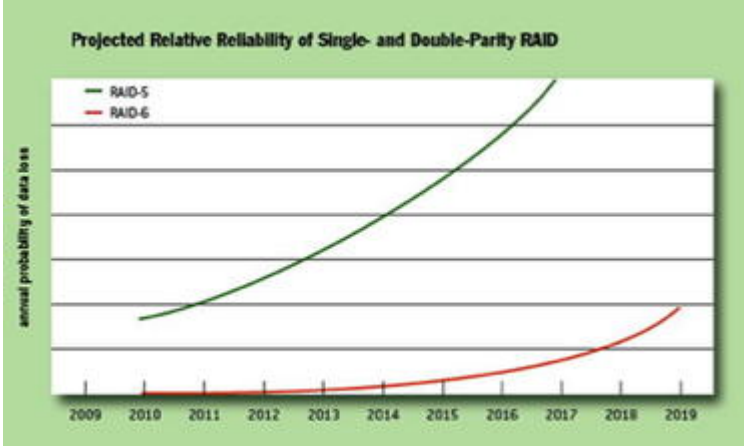
Consider:

- **Long rebuild times.** As disk capacity grows, so do rebuild times. 7200 RPM full drive writes average about 115 MB/sec - they slow down as they fill up - which means about 5 hours minimum to rebuild a failed drive. But most arrays can't afford the overhead of a top speed rebuild, so rebuild times are usually 2-5x that.
- **More latent errors.** Enterprise arrays employ background disk-scrubbing to find and correct disk errors before they bite. But as disk capapcities increase scrubbing takes longer. In a large array a disk might go for months between scrubs, meaning more errors on rebuild.
- **Disk failure correlation.** RAID proponents assumed that disk failures are independent events, but long experience has shown this is not the case: 1 drive failure means another is much more

likely.

Simplifying: bigger drives = longer rebuilds + more latent errors -> greater chance of RAID 6 failure.

Mr. Leventhal graphs the outcome:



(https://www.zdnet.com/i/story/60/88/000805/relative_reliability_r5_vs_r6.jpg)

Courtesy of the ACM

By 2019 RAID 6 will be no more reliable than RAID 5 is today.

The Storage Bits take For enterprise users this conclusion is a Big Deal. While triple parity will solve the protection problem, there are significant trade-offs.

21 drive stripes? Week long rebuilds that mean arrays are always operating in a degraded rebuild mode? Wholesale move to 2.5" drives? Functional obsolescence of billions of dollars worth of current arrays?

Home users can relax though. Home RAID is a bad idea (http://blogs.zdnet.com/storage/?p=116): you are much better off with frequent disk-to-disk backups and an online backup like CrashPlan (http://www9.crashplan.com/landing/index.html) or Backblaze (http://www.backblaze.com/).

What is scarier is that Mr. Leventhal assumes disk drive error rates of 1 in 10^16. That is true of the small, fast and costly enterprise drives, but most SATA drives are 2 orders of magnitude less: 1 in 10^14.

With one exception: Western Digital's Caviar Green, model WD20EADS, is spec'd (http://www.wdc.com/en/products/products.asp?DriveID=576) at 10^15, unlike Seagate's 2 TB ST32000542AS

(http://www.seagate.com/ww/v/index.jsp?name=st32000542as-bcuda-lp-sata-2tb-hd&vgnextoid=1f70e5daa90b0210VgnVCM1000001a48090aRCRD&locale=en-US#tTabContentSpecifications) or Hitachi's Deskstar 7K2000 (http://www.hitachigst.com/tech/techlib.nsf/techdocs/6A7E7E6848832B7786257603007AAF5E/%24file/DS7K2000_DS_final (pdf).

Comments welcome, of course. Oddly enough I haven't done any work for WD, Seagate or Hitachi, although WD's indefatigable Heather Skinner is a pleasure to work with. I did work at Sun years ago and admire what they've been doing with ZFS, flash, DTrace and more.

RELATED TOPICS: HARDWARE | REVIEWS | MOBILITY | DATA CENTERS | CLOUD