Why RAID 5 stops working in 2009

Summary: *The storage version of Y2k? No, it's a function of capacity growth and RAID 5's limitations.*

By Robin Harris for Storage Bits | July 18, 2007 -- 06:18 GMT (23:18 PDT)

The storage version of Y2k? No, it's a function of capacity growth and RAID 5's limitations. If you are thinking about SATA RAID for home or business use, or using RAID today, you need to know why.

RAID 5 protects against a single disk failure. You can recover all your data if a single disk breaks. The problem: once a disk breaks, there is another increasingly common failure lurking. And in 2009 it is highly certain it will find you.

Disks fail While disks are incredibly reliable devices, they do fail. Our best data - from CMU and Google - finds that over 3% of drives fail each year in the first three years of drive life, and then failure rates start rising fast.

With 7 brand new disks, you have ~20% chance of seeing a disk failure each year. Factor in the rising failure rate with age and over 4 years you are almost certain to see a disk failure during the life of those disks.

But you're protected by RAID 5, right? Not in 2009.

Reads fail SATA drives are commonly specified with an unrecoverable read error rate (URE) of 10^14. Which means that once every 100,000,000,000,000 bits, the disk will very politely tell you that, so sorry, but I really, truly can't read that sector back to you.

One hundred trillion bits is about 12 terabytes. Sound like a lot? Not in 2009.

Disk capacities double Disk drive capacities double every 18-24 months. We have 1 TB drives now, and in 2009 we'll have 2 TB drives.

With a 7 drive RAID 5 disk failure, you'll have 6 remaining 2 TB drives. As the RAID controller is busily reading through those 6 disks to reconstruct the data from the failed drive, it is almost certain it will see an URE.

So the read fails. And when *that* happens, you are one unhappy camper. The message "we can't read this RAID volume" travels up the chain of command until an error message is presented on the screen. 12 TB of your carefully protected - you thought! - data is gone. Oh, you didn't back it up to tape? Bummer!

So now what? The obvious answer, and the one that storage marketers have begun trumpeting, is RAID 6, which protects your data against 2 failures. Which is all well and good, until you consider this: as drives increase in size, any drive failure will *always* be accompanied by a read error. So RAID 6 will give you no more protection than RAID 5 does now, *but you'll pay more anyway* for extra disk capacity and slower write performance.

Gee, paying more for less! I can hardly wait!

The Storage Bits take Users of enterprise storage arrays have less to worry about: your tiny costly disks have less capacity and thus a smaller chance of encountering an URE. And your spec'd URE rate of 10^15 also helps.

There are some other fixes out there as well, some fairly obvious and some, I'm certain, waiting for someone much brighter than me to invent. But even today a 7 drive RAID 5 with 1 TB disks has a 50% chance of a rebuild failure. RAID 5 is reaching the end of its useful life.

Update: I've clearly tapped into a rich vein of RAID folklore. Just to be clear I'm talking about a failed drive (i.e. all sectors are gone) plus an URE on another sector during a rebuild. With 12 TB of capacity in the remaining RAID 5 stripe and an URE rate of 10^14, you are highly likely to encounter a URE. Almost certain, if the drive vendors are right.

As well-informed commenter Liam Newcombe notes:

> The key point that seems to be missed in many of the comments is that when a disk fails in a RAID 5 array and it has to rebuild there is a significant chance of a non-recoverable read error during the rebuild (BER / UER). As there is no longer any redundancy the RAID array cannot rebuild, this is not dependent on whether you are running Windows or Linux, hardware or software RAID 5, it is simple mathematics. An honest RAID controller will log this and generally abort, allowing you to restore undamaged data from backup onto a fresh array.

Thus my comment about hoping you have a backup.

Mr. Newcombe, just as I was beginning to like him, then took me to task for stating that "RAID 6 will give you no more protection than RAID 5 does now". What I had hoped to communicate is this: in a few years - if not 2009 then not long after - all SATA RAID failures will consist of a disk failure + URE.

RAID 6 will protect you against this quite nicely, just as RAID 5 protects against a single disk failure today. In the future, though, you will *require* RAID 6 to protect against single disk failures + the inevitable URE and so, effectively, 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. RAID 5 won't work at all, and, instead, RAID 6 will replace RAID 5.

Originally the developers of RAID suggested RAID 6 as a means of protecting against 2 disk failures. As we now know, a single disk failure means a second disk failure is much more likely - see the CMU pdf Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You? (http://www.cs.cmu.edu/~bianca/fast07.pdf) for details - or check out my synopsis in Everything You Know About Disks Is Wrong (http://storagemojo.com/?p=383) . RAID 5 protection is a little dodgy today due to this effect and RAID 6 - in a few years - won't be able to help.

Finally, I recalculated the AFR for 7 drives using the 3.1% AFR from the CMU paper, using the formula suggested by a couple of readers - 1-96.9 ^# of disks - and got 19.8%. So I changed the ~23% number to ~20%.

Comments welcome, of course. And I got home despite a blow out on the Scottsdale's 101N in 110 degree heat. I thought of it as a Bikram Tire Changing Asana.

Topic: Hardware

About Robin Harris

Robin Harris is Chief Analyst at TechnoQWAN LLC, based in Sedona, Arizona. He has over 30 years in the IT industry, including DEC and Sun, and degrees from Yale and the Wharton School.

You May Also Like

10 Slowest-Selling Cars of April
(Wall St. Cheat Sheet)

How Wealthy People Use Credit Cards To Their Advantage
(NextAdvisor Daily)

3 reasons all-flash storage might be here to stay
(Tech Page One)

Professor Plastic: A Desktop 3-D Printer in Every Home?
(Plastics Make It Possible)

US Cellular 4G LTE Router cuts the cable cord
(SlashGear)

These Bombs Caused Casualties: 15 Career-Killing Movies
(Styleblazer)

Talkback

Been doing this for 15 years

and never have I seen this unless the drive is about ready to take a dump.

I've been in the server arena for quite some time and our file servers (which as of late have approached 1-2TB) have never encountered any errors like this rebuilding the arrays UNLESS the drive is also dying.

Again, much ado about nothing.

ITGuy04
18 July, 2007 06:39

Reply *1 Vote*

You've never had a RAID 5 rebuild fail?

I'm just curious.

Robin

R Harris
19 July, 2007 09:50

Reply *Vote*

Umm... Not at this time and Hopefully Never

Count my blessings, I have never had a RAID 5 setup fail on rebuild. I have had a mirror drop both disks because of quota on AIX, but RAID 5 has been the best of friends to me.

nucrash
19 July, 2007 10:29

Reply *Vote*

You've never had a RAID 5 rebuild fail

Yes it happens ALL the time. The thing is just because your rebuild fails does not mean you loose your data. You can start another rebuild that may or may not work and you can copy your data off the array. No data loss at all.

During the rebuild procedure the drive being rebuilt is populated with the stripes that it needs to complete the raid set. Not much really changes in the other drives until the operation is almost complete.

Even if another drive pops out of the array during the rebuild operation (which can happen and is common) you can force that drive back online with the raid controller software and be exactly back where you were previously.

One thing you DON"T want to do is "guess" on the drive with the stalest metadata and leave it plugged into a backplane while forcing another drive online. This will can cause the controller to start a rebuild on the other failed drive overwriting perfectly valid data and thus destroying your striping.

If your really paranoid pull a raid log when you first create your array or if you make any changes to it. In a worst case scenerio even if metadata is corrupted on the drives to the point that the controller cant read it you can manually put the stripe order into the controller and it will rewrite the configuation.

The MOST IMPORTANT thing to do is to keep your controller and drive firmware (yes hard drives have firmware) at the latest level and to periodically review your raid log for errors. If you catch issues early they are MUCH easier to deal with than multiple amber lights at 3am with a call into the support center of your hardware vendor.

RARE_AT_BEST
21 October, 2008 19:32

Reply 1 Vote

raid 5 failed rebuild

i too have experienced failed raid 5 rebuilds back in the day when the hardware controller was a CMD5000 that was controlling 23gb full height seagate drives in a 14 drive scsi diff configuration... yeah, failed rebuilds would occurred, and sometimes the failure would know out the entire raid, eventually forcing you to reset the whole system and restart the rebuild process over again. but these drives would fail when too many bad blocks would be accumulated in the bad black sector database, which could be cleared with a low level format.

i think the author is sensationalizing the drive failure rate, and not being clear enough that enterprise level drives do not have the same types of failures as consumer grade equipment.

take into consideration that most enterprise level raids and sans are utilizing some sort of scsi derived standard, i.e. sas, FC, ultra, iscsi, etc. scsi by it's very nature(or actually by it's design) is very conservative. if a drive starts to exhibit errors/problems, the drive will often "prefail" before actual failure. this provides a chance to recover data(i.e. rebuild raid). and ide/ata drive on the other hand simply fails, and typically failures are not as easily recoverable. there's a reason why scsi drives are more expensive than sata drives, this is one of the reasons.

additionally, the consumer versus professional markets have completely different goals: consumer markets are about biggest bang for the buck, i.e. 1.5 TB drives for under $200. professional markets are about reliability above all other matters, and it's not unusual for a 320GB SAS drive to cost $800+...

2TB drives? sure their around the corner. but don't let the size intimidate you... regardless of interface technology(sata or scsi) the proper procedure for raid maintenance is to 1) always have a cold spare and 2) always replace failed drives with a new drive, and return the old drive for wrrenty maintenance.

capsteve

21 October, 2008 22:26

Reply Vote

RE: Why RAID 5 stops working in 2009

@R Harris I am curious too, RAID 5 had been an headache for me, I gave alot of my time try to find a resolution for this, but it has always been a failure.
<a href="http://www.paperprofs.co.uk/writing-types/dissertation/">Dissertation Writing</a> | <a href="http://www.paperprofs.co.uk/writing-types/admission-essays/"> Admission Essay Writing</a> | <a href="http://www.paperprofs.co.uk/writing-types/essay/">Essay Writing</a>

lorisinclair

4 September, 2011 23:42

Reply Vote

Your storage system is not large enough

The reason you do not see this problem is that you do not have enough disk drives to be statistically significant. 1-2TB is a couple of disk drives.

In a typical enterprise data center, there are hundreds to thousands of disk drives. All of the data centers I have worked with that have a decent number of disk drives do, in fact, see this problem. Hence, this is much ado about something.

I do think that the problem is exacerbated by the behavior of the RAID controller when it encounters a failure on a disk drive. When a RAID controller is running along and gets an uncorrectable read error on a single disk drive in a RAID set, many times it will simply shut that drive down and begin a rebuild operation on the hot spare. Now, enter the problem of the probability of a second read failure on one of the remaining drives in the RAID set. That second failure will cause the RAID controller to quite possibly give up.

IMHO this is far too aggressive. Some of the newer, more intelligent RAID controllers will take the first offending drive offline but not disable it entirely. Instead, the drive is examined for the root cause of the problem and either repaired and put back into service, or it is used in conjunction with the other remaining drives to perform a more robust rebuild operation. This assumes, of course, that the drive is accessible. If the drive is dead then you are back to the problem of a data error on a second drive causing problems in the rebuild. Even so, I think that the rebuild should complete as it would normally and report enough information back to the host through sense data that the data management people can determine the extent of the problem in terms of which files and/or metadata is affected and so on.

This stuff is not easy and I agree that the higher capacity drives are increasing the exposure to data errors. I think that we need to be engineering data storage systems that assume data errors are a normal event rather than an anomaly and deal with them more appropriately than we have been.

storagelunatic

22 October, 2008 20:50

Reply Vote

RE: Why RAID 5 stops working in 2009

@storagelunatic This can be the reason too.
<a href="http://www.paperprofs.co.uk/writing-types/research-papers/">Research Paper Writing</a> |
<a href="http://www.paperprofs.co.uk/writing-types/coursework/">Coursework Writing</a>

lorisinclair
4 September, 2011 23:43

Reply 1 Vote

Not only large drives!

For smaller drives, there is lower probability of a URE during rebuild, but the probability is not zero. It's very embarassing to explain to a client that his "infallible" RAID system has indeed collapsed.

It's absurd to have massive amounts of stored data dependent on a 100% recovery. The only way to avoid increasing failures of this type as storage needs expand is to design storage architectures so that a few lost bits do not translate into global failures.

BTW: from the ratings of the original article, it appears to me that IT-ers have quite a case of denial. Thanks for rocking the boat!

w_c_mead
13 April, 2009 08:35

Reply Vote

Your whole analysis is based on a faulty assumption.

Just because you have an unrecoverable read error does not mean your RAID array is corrupted. This will most probably result in a corrupted file. In most cases the offending sector will be added to the bad sector list maintained by the drive and taken out of use. Also just because drives double in size every year doesn't mean your data does. The only relevent size is that of the data. An error on an unused portion of the drive isn't a problem.

ShadeTree
18 July, 2007 06:44

Reply Vote

Well you know the saying

"Data will always expand to the capacity of your drive space."

I know it happens to me all the time. I buy a drive and it gets full so I buy one 3 times bigger and it seems to get full almost instantly. How's that? I think the reason for me is when I have lots of space I delete less so it quickly fills up.

In the corporate world the same is true but it's an even bigger problem. When drive space is short systems get purged. If you know you have tons of room people decide not to purge. I know people who want to keep their trash permanently and get really upset when the system deletes the contents of the trash. I have no idea why people keep important documents in the trash.

voska
18 July, 2007 06:51

Reply Vote

An unrecoverable read error on rebuild...

can destroy the whole RAID in RAID5.

bjbrock
18 July, 2007 14:30

Reply Vote

And if there is a fire in your PC all your data can be destroyed.

The odds as the author stated are about one in a trillion.

ShadeTree
19 July, 2007 05:51

Reply Vote

There is an old saying

Which in the last 60 years I have found to be so true" If you can imagine it happening sooner or later it will"
Mike Hereid Sr

Michael L Hereid Sr
19 July, 2007 08:19

Reply Vote

data storage does increase at the rate of Moore's law

Drives double in size, and so does our usage of them. Why is that? I guess Parkinson's law can be extrapolated to say the data expands to fit the space allotted.

noglider@...
19 July, 2007 09:06

Reply Vote

How does that work?

How does the RAID controller know if a block is in use?

The file system writes a 1 GB file. The RAID controller writes the blocks. Then the file systems deletes the file, i.e. alters the directory to mark those blocks as free.

Are you saying the RAID controller reads the directory file, understands the allocation, and then marks those blocks as free?

I'm not getting it.

Robin

R Harris
19 July, 2007 09:34

Reply Vote

RAID controllers, sectors, and lists

With the intelligent SCSI RAID systems that I've used, the controller (or the drive itself) can mark a particular sector (or set of sectors) as bad in a list, and replace that sector by using a spare taken from another list. If the replacement is done in the drive, the RAID controller need not do the substitution, but either way, the controller should be able to map the bad area out of the drive transparently to the system software, and only need to re-build the data from the bad area and write it to the replacement block. The controller allocated the spares when formatting the RAID in the first place. This is done using the "logical block number" and a hash table, so it's pretty quick. There is performance loss when reading a replaced sector, and when a drive runs out of spares, it is time to replace it.

I'm not absolutely sure the RAID I was working with was a RAID5, though.

filker0
19 July, 2007 14:47

Reply Vote

How does that work?

The RAID controller has no idea about which sectors have user data or not. RAID controllers and disk drives are not aware of the structure or content of the data that they store. As far as the computer system is concerned, RAID controllers and disk drives do two things and only two things - read sectors and write sectors. The RAID controller and disk drives have no concept of files or directories or inodes - just sectors.

Furthermore, current generation RAID controllers do not keep lists of bad sectors on the disk drives they manage. The disk drives themselves do all of their own bad sector management. Newer RAID controllers can monitor disk drive health using SMART data to varying extents and take action if there are indications that something is wrong - like a sudden increase in correctable read errors or seek errors. This is all very specific to a RAID controller vendor and the disk drives used in the RAID.

Finally, the disk drives do not necessarily have bad sector maps like they used to. Bad sectors are managed on the track on which the occur using "sector slipping" and "alternate sector assignment" techniques. This results in little or no loss of performance except in the case where sectors are reassigned to an alternate sector on a different track in which case asses to the bad sector results in two or three "implicit" seek operations. Fortunately, off-track sector reassignments are seldom used.

Hope this clears things up a bit.

storagelunatic
22 October, 2008 20:33

Reply Vote

Isnt that what disk mainenance is for

there are ways to monitor disk health.

pcguy777
19 July, 2007 12:04

Reply Vote

Given that Optical Disks have jumped

From DVD to the new HD-DVD/ Bluray. Capacity's for backup mediums, are increasing. Currently I use 20 DVD-DL for one backup job. It costs about $1.50 for one DVD-DL.
Mind you that is personal use. The only RAID I use is 0 to increase performance, on my SATA drives.

Species8472
20 July, 2007 00:07

Reply Vote

1 2 3 4 5 6 7 8 9 Next »