RAID 5: theory & reality

Summary: *RAID 5 pain is on message boards and support forums all over the net. Failed rebuilds, lost data, unhappy bosses. Why isn't RAID 5 as reliable as it is supposed to be?*

By Robin Harris for Storage Bits | June 28, 2010 -- 07:32 GMT (00:32 PDT)

In theory, RAID 5 protects your data. In reality, RAID 5 is often a painful failure. Why? Mean time-to-data-loss (MTTDL) is a fraud: actual rates of double-disk failures are 2 to 1500 times higher than MTTDL predicts.

What's behind MTTDL's failure? In A Highly Accurate Method for Assessing Reliability of RAID (http://media.netapp.com/documents/rp-0046.pdf) researchers Jon G. Elerath of NetApp (http://www.netapp.com/us/) - a major storage vendor - and Prof. Michael Pecht of the University of Maryland, compared RAID theory against actual field data. They found that MTTDL calculations inaccurate for 3 reasons:

- Errors in statistical theory of repairable systems.
- Incomplete consideration of failure modes.
- Inaccurate time-to-failure distributions.

By repairing MTTDL's theoretical basis, adding real-world failure data and using Monte Carlo simulations they found that today's MTTDL estimates are wildly optimistic. Which means your data is a lot less safe with RAID 5 than you know.

Repairable systems The typical MTTDL assumption is that once repaired - i.e. a disk is replaced with a new one - the RAID system is as good as new. But this isn't true: at best, the system is only slightly better than it was right before the failure.

One component is new - but the rest are as old and worn as they were before the failure - so the system is not "like new." The system is no less likely to fail after the repair than it was before.

The problem is that in RAID arrays repairs take time: the disk fails; a hot spare or a new disk is added; and the data rebuild starts - a process that can take hours or days - while the other components continue to age.

Net net: MTTDL calculations use the wrong failure distributions and incorrectly correlate component and system failures.

Failure modes MTTDL typically considers only catastrophic disk failures. But I've noted [see Why RAID 5 stops working in 2009 (http://www.zdnet.com/blog/storage/why-raid-5-stops-working-in-2009/162) , RAIDfail: Don't use RAID 5 on small arrays (http://www.zdnet.com/blog/storage/raidfail-dont-use-raid-5-on-small-arrays/483) and Why disks can't read - or write (http://www.zdnet.com/Why disks can't read - or write)] disks have latent errors as well. A catastrophic failure + latent error is a dual-disk failure, something RAID 5 can't handle.

Anatomy of a RAID failure There are 4 transition events in a RAID 5 failure:

- **Time to operational failure.** Drive failure distributions are not constant. Sub-populations of drives may have specific failure modes, like infant mortality, that MTTDL models do not account for.
- **Time to restore.** Minimum restore times are functions of several variables, including HDD capacity, HDD data rate, data bus bandwidth, number of HDDs on the bus and the on-going I/O load on the array. A 2 TB drive might take 40 hours or more to restore.
- **Time to latent defect.** Latent defect rates vary with usage, age and drive technology.
- **Time to scrub.** Scrubbing is a background process meant to find and repair latent errors. Busy systems have less time to scrub which increases the chance of a latent error hosing a RAID 5 rebuild. Scrub strategy has a major impact on latent error rates.

Using field-validated distributions for these 4 transition events and Monte Carlo simulations, the researchers concluded:

> The model results show that including time-dependent failure rates and restoration rates along with latent defects yields estimates of [dual-disk failures] that are as much as 4,000 times greater than the MTTDL-based estimates.

Which is why RAID 5 has caused so much trouble to so many people over the last 20 years.

The Storage Bits take As a practical matter, don't rely on 4 SATA drive RAID 5 to protect your data. The chance of a latent error and a hosed rebuild are too great - much greater than the product's engineers probably believe.

If you must use a RAID array - and I don't recommend it for home users unless you have system admin skills - make sure it protects against 2 disk failures (RAID 6 or equivalent). That means a 5 drive array at a minimum.

But there is a larger pattern here. Disk drives have a higher failure rate than vendors spec (http://storagemojo.com/2007/02/20/everything-you-know-about-disks-is-wrong/) . DRAM also has a much higher error rate (http://www.zdnet.com/blog/storage/dram-error-rates-nightmare-on-dimm-street/638) than commonly believed. And file systems (http://www.zdnet.com/blog/storage/how-microsoft-puts-your-data-at-risk/169) are also flakier than they should be.

Weird, huh? Every critical element of data integrity turns out to be much worse than commonly thought.

This isn't a conspiracy so much as a natural vendor reluctance to give out bad news about their products. That's why we need independent observers to check out product claims.

But the bigger issue with storage is that the Universe hates your data (http://www.zdnet.com/blog/storage/the-universe-hates-your-data/975?tag=mantle_skin;content) . If there's a failure mode hidden somewhere, the Universe will find it and use it.

Long term data integrity on a massive scale will require a re-tooling of vendor development and test. Detroit did that with statistical process control over the last 30 years and massively improved quality as a result.

The current piecemeal approach to mending subsystems needs to give way to a complete end-to-end systems design for data integrity. But the reality of still-rapidly-evolving storage technologies probably puts that effort at least 2 decades away.

In the interim remember that your data needs protection. Let's be careful out there.

Comments welcome, of course. I've done work for NetApp and admire the good works of co-founder Dave Hitz.

Topics: Hardware, Storage

About Robin Harris

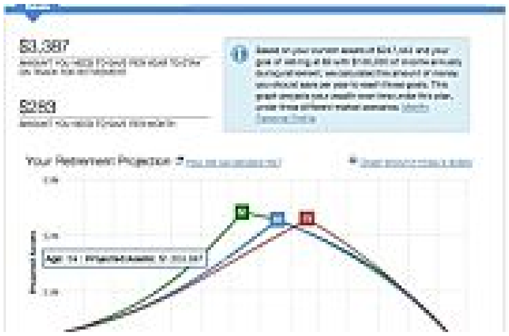
Robin Harris is Chief Analyst at TechnoQWAN LLC, based in Sedona, Arizona. He has over 30 years in the IT industry, including DEC and Sun, and degrees from Yale and the Wharton School.

You May Also Like

The most fuel-efficient vehicle in

10 Slowest-Selling Cars of April

3 reasons all-flash storage might be



$3,387

$263

Your Retirement Projection

Talkback

RAID isn't about data safety

Backups are for protecting your data. RAID is for reducing downtime.

Of course when it comes to RAID 5 I've seen more than my share double drive failures requiring a restore from backup.

voska1

28 June, 2010 08:58

Reply *Vote*

Absolutely agree

@voska1

RAID is the operational side of things. Backup is the safety side. It is a matter of convenience that RAID does provide safety in most cases, but that doesn't negate the need for backups.

Any basic setup should have a RAID to keep uptime high, should have shadow copies enabled onto a separate drive for high availability of individual files, and a comprehensive backup (preferably full nightly).

All those these things *together* mitigate the transient memory error, or drive defect, or even filesystem error.

Sometimes Robin talks about these things as if it's one system running alone on which a company's data depends.

croberts

28 June, 2010 10:22

Reply *Vote*

Agree and..

@voska1
Users/IT/buggy software are far more likely to create problems that any form of RAID won't help solve. A script for clearing a public share had its permissions changed and it deleted 1.5TB of important data. We had a backup though and go it all back.

DevGuy_z

28 June, 2010 13:17

Reply *Vote*

RE: RAID 5: theory & reality

I have been working in the same shop for 23 years. We currently have about 20 + servers. We replace our servers every

5 years. The servers are always on 24/7. In all that time we have had only one double disk failure. I believe raid 5 is great.

Zamo100

28 June, 2010 10:30

Reply Vote

Almost the exact same here...

except the companies I've worked with replace their servers more like every 7 years. In that time I've seen a couple dozen single disk failures and one double disk failure. That double disk failure happened in the mid-1990s on a Motorola server and was blamed on a Motorol tech who was called into replace a single drive and a SCSI controller. The term I remember being bantered around at the time was "over-termination" which immediately burned out the new controller and took down two drives. Being a sofware developer and not a hardware tech, I didn't pay much attention as for me it just meant the afternoon off.

jasonp@...

28 June, 2010 11:33

Reply Vote

RE: RAID 5: theory & reality

@GeneZam & others:

RAID 5 is more reliable with enterprise disks because the disks are a) lower capacity and b) have a BER of 1 in 10^-15. Combined they make enterprise disks much less likely to encounter RAID 5 problems than the SATA drives that most ZDnet readers use for home.

But don't kid yourselves. As enterprise drives get bigger and you use more SATA drives you'll bump into this. And when you do . . . .

Robin

R Harris

28 June, 2010 12:25

Reply Vote

Better have a backup.

@R Harris RAID is for reducing downtime.

DevGuy_z

28 June, 2010 13:18

Reply Vote

RE: RAID 5: theory & reality

Gee NetApp was behind this study?

Yes RAID5 Sucks, you need WAFL that only NetApp can give you to be truly safe. It won't be cheap though. ;)

I'm with GeneZan.. I've been in enterprise server and storage for over a decade and I can tell you without a doubt that the cost/benefit ratio of RAID6 just isn't there. Unless you are are using HUGE numbers of disks in your array, or for volumes actually storing backup data.

civikminded
28 June, 2010 11:34

Reply *Vote*

RE: RAID 5: theory & reality

@civikminded Cynicism is unwarranted. The article was published in the IEEE Transactions on Computers. Professor Pecht is a Fellow of the IEEE. NetApp, through Dr. Elerath, provided massive amounts of detailed disk failure info not available from anyone but a vendor.

This is Good Stuff.

Robin

R Harris
28 June, 2010 17:15

Reply *Vote*

This is like saying ...

... I don't need car insurance, because I've been driving for 10 years and have never been in a car accident.

Fault tolerance is for <i>unexpectable</i> events. Failure rates don't really follow a bell curve distribution. You spend the little bit of extra cash on all of your critical systems in the unlikely event (maybe 2 or 3 times in your entire career) you lose two disks at once. Given the low, low price of disk (even ridiculously expensive EMC disk) versus downtime, in the long-run it's a good bet.

RationalGuy
30 June, 2010 15:46

Reply *Vote*

Raid is an uptime enhancer.. not a backup tool.

Besides double disk failures, scsi memory failures, backplane failures, motherboard failures.. its all possible and the older the server, the more likely a failure.

We try to replace servers every 5 years. This past year, i was getting ready to replace a server, its backplane died before i could do that. Luckily all its information was on a SAN and i was able to redirect the shares to another server.

In all the years i have been working on servers, not one double disk failure has occurred. I have however had single disks go offline(die) or backplanes just stop working.

Obviously the newer the hardware the less likely the failure, but its not 100%. I recieved a new SAN, two months after its install, i had a disk failure. Luckily, like most SANs, there is an online hot spare and the system rebuilt overnight while still delivering data.

We are still backing up to tape but as i add more SAN's we are moving to snap shots.

We currently have 30 virtual servers and 25 physical servers and i have maintenanced over 200+ servers via side jobs/etc

Been_Done_Before

28 June, 2010 12:14

Reply *Vote*

Risk from user error is probably worse. Rely on RAID for reducing downtime.

Backup is your best protection. Most of the problems we experience have been user/IT/buggy software related. Once a bad script who's permission was changed deleted 1.5 TB of data. RAID 6 wouldn't have helped at all. Our backup helped a lot.

If you want to protect your data, do backups. If you want fail-over do RAID.

DevGuy_z

28 June, 2010 13:14

Reply *Vote*

Disk drives are not designed for reliability

Drive vendors will always tell you that data integrity and always ensuring that you can get your data back are their main concerns. And it is a huge issue with them. But the reality is that drive capacity and the higher media transfer rates it generates are still the things that drive sales, and the drive manufacturers know it. As a result, the technology is always driven to the bleeding edge, with data reliability taking a hit as a result.

zackers

28 June, 2010 15:48

Reply *Vote*

RE: RAID 5: theory & reality

@zackers

deanders

28 June, 2010 17:55

Reply Vote

That's why I trust RAID 1 0 on my desktops....

@zackers
Yes it costs more in drives but drives are cheap, and for a reason...they are not as dependable as they once were. Don't get me wrong, I use Enterprise Nearline drives and watch any reallocation activity. First sign of such I send the drive back.

But I also run a backup exec to a single drive and on my primary system I have carbonite which saved my bottom on one machine that only had RAID1 and I assumed the backup was good.....wrong! but I got all my data back, didn't lose one email, and a fresh load of the OS.
Raid 5 is convenient in that it gives you plenty of space but it is not robust enough for me. Although I say that and we have hundreds of servers running RAID5 and most have had little issue.

dunn@...

29 June, 2010 13:05

Reply Vote

RE: RAID 5: theory & reality

@dunn@... Whether drives are cheap or not depends on how much you have to store. At any time, a state-of-the-art drive has always been roughly the same dollar price for a couple of decades now.

I'm simply saying that if drive manufacturers backed off on the specs they could produce a much more reliable drive. For example, the bit cells on the media could be made larger, more reliable magnetic coatings could be used which have less storage density, servo info could be redundantly recorded, the entire drive could be made more vibration resistant, etc. Using today's technology a very reliable 500GB drive could be made for the same price or less. Certainly such a drive would get around the URE problem that is ending the effectiveness of RAID5. But nobody wants it.

Instead, it seems the industry is moving towards tiered storage including backups, allowing it to deal with the higher error rates of individual drives. And maybe that's the wiser way to go, given RAID5's other limitations such as rebuild times.

zackers

29 June, 2010 23:31

Reply Vote

RE: RAID 5: theory & reality

I have worked in shops that use RAID 5 on most servers that require high uptimes. The servers I currently manage only have about 30-50 GB used on about 140GB available in the array. The configuration I use, where possible, is a 3-drive RAID 5 with a hot spare on HP Proliant servers. Daily checking of the servers is easy with a quick scan of the drive lights. With the hot spare, a drive failure kicks-in the hot spare and the array reuilds automatically. This greatly reduces the vulnerability window.

Years ago, I had a RAID controller fail without warning and hose the server. I was able to stage a temporary server from the nightly backups from two nights previous (the server crashed before the current night's backup). Some critical files backed throughout the day to a backup server made it possible to have all key functions restored before the staff came in the next morning.

No matter what type of RAID you use, a failure like the one I experienced (considered rare, unless it happens to you) will likely cause loss of some or all data on the server.

Lesson learned: Use RAID to increase percentage of uptime, use backups to lessen stress and decrease downtime when uptime stops!

rich3page

28 June, 2010 20:22

Reply *Vote*

RE: RAID 5: theory & reality

@rich3page Yup, your story is a reminder that storage is more than just the disk drives. There are all sorts of ways a storage box can fail besides just the drives inside it. You could have bought a storage box with full redundancy (at about 2x the price), but tiered backup has other advantages that a redundant box does not.

zackers

29 June, 2010 23:18

Reply *Vote*

So what's good and affordable for SOHO users?

We currently use our local hard drives, with an Acer four disk 2TB RAID5 system as the backup. I like it as a backup as it is more reliable, but we wouldn't want to it as the working storage as it's write performance is not good, even with Gigabit connections.

Certainly for audio recording use, local high-speed drives, like our SSDs or even RAID0, are the only option for the working storage.

Patanjali

29 June, 2010 00:48

Reply *Vote*

RE: RAID 5: theory & reality

@Patanjali ...A 3-drive RAID5 (yes, only three WD2001FASS drives) has read rates of 255MB/s and writes at 160MB/s. More than enough for uncompressed HD editing and I certainly wouldn't call it "slow". (Same three drives in RAID0 - 360MB/s read/write rates.)

Alex Gerulaitis

29 June, 2010 11:42

Reply *Vote*

1 2 Next »