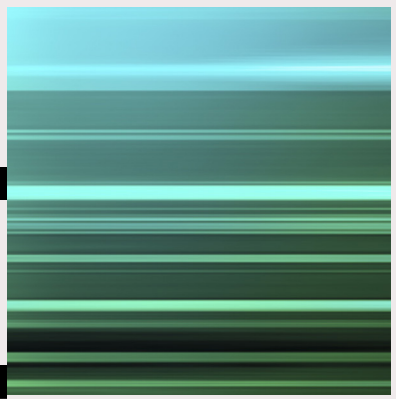

Single Failure Tolerance (1FT): Myth and Fact

Datrium™

385 Moffett Park Dr. Sunnyvale, CA 94089 844-478-8349 www.Datrium.com

Technical Report

About the Authors

Lakshmi N. Bairavasundaram is a Member of Technical Staff at Datrium, Inc. Lakshmi has worked in the file and storage systems domain for the last 14 years, including a Ph.D. from the University of Wisconsin-Madison on the topic of disk failures, and working at NetApp and Datrium. He has received 4 best-paper awards for his conference publications, and his paper on Latent Sector Error characteristics received the ACM SIGMETRICS 10-year Test-of-Time Award.

Zhe Wang is a Member of Technical Staff at Datrium, Inc. Zhe has worked on the storage pool for 4 years at Datrium. Prior to that, he worked on content-based search systems at Princeton University where he received his Ph.D. His paper on efficient indexing for similarity search received the VLDB 10-year Test-of- Time Award.

R. Hugo Patterson is the CTO, VP of Engineering, and Co-Founder at Datrium, Inc. Prior to Datrium, Hugo was an EMC Fellow serving as CTO of the EMC Backup Recovery Systems Division, and the Chief Architect and CTO of Data Domain (acquired by EMC in 2009), where he built the first deduplication storage system product. Prior to that he was the engineering lead at NetApp, developing SnapVault, the first snap-and- replicate disk-based backup product. Hugo has a Ph.D. from Carnegie Mellon. His paper on snap-and- replicate received the USENIX FAST 10-year Test-of- Time Award.

Introduction

Using 1FT would imply a 0.49% chance of data-loss in one year (and 1.95% chance of loss in system lifetime of 4 years) under typical assumptions.

Storage system reliability starts with properly managing the inevitable storage device failures. Surprisingly, despite the community's rigorous study of the topic over decades, some popular enterprise systems remain under-configured and are unable to address the significant threat posed by Latent Sector Errors (LSEs). Back in 1988, the RAID paper by Patterson et al. [1] laid out the fundamental technologies for building a system that can tolerate a whole disk failure. Subsequent rigorous studies collected a wealth of data on disk-drive failures in the field from places like supercomputing centers [2], Google [3], NetApp [4,5,6], Nutanix [7], and Backblaze [8], and considered the implications on storage system design. The key lessons from these studies are that disk failures may be more common that drive manufacturers' expected annual failure rates (AFRs) [2,3] and that Latent Sector Errors can have a profound impact on reliability and need to be considered when designing a system [4,6].

This brief delves into the myth and the fact for one system design choice — Single Failure Tolerance (1FT)— in the context of data from the above-mentioned studies. Our analysis shows that using 1FT with once-a-month disk scrubbing would imply an extremely high data-loss probability of **0.49% in one year** or 1.95% over the 4-year life of a system. Less frequent scrubbing increases the data-loss probability to as high as 18.6% in the fourth year of a system's life.

The Myth: 1FT Provides Sufficient Reliability

1FT, or Single Failure Tolerance, is a product configuration commonly quoted by hyperconvergence vendors that can handle one disk failure. In this configuration, data is typically written to disks in chunks and each chunk is stored twice (on two different disks in the pool). When a drive fails, its chunks are rebuilt from the copies spread across the pool.[1] 2FT (Double Failure Tolerance) stores three copies of the chunk on three different disks and data can be rebuilt from any of the copies.

Decades ago, reliability analysis of RAID systems centered on whether data could be rebuilt after a disk failure before a second disk failed. With hot spares, rebuilds could start immediately and rebuilds could complete quickly. Declustered geometries and distributed spares further sped up rebuild so it seemed sufficient to handle one disk failure. But, as disk capacities have soared and the industry has shifted to larger-capacity drive classes [9], a new problem emerged: a Latent Sector Error (see below) could make the content needed for rebuilding data unreadable.

The consequence is that data reconstruction fails which results in at least some data loss, though not the loss of a whole disk. The chances

[1] This setup is similar to RAID-1, except that the copies are spread across multiple nodes. One could also have an erasure-coded variant that is similar to RAID-5, but it tolerates just a single disk failure as well.

of this happening are frighteningly high. Today, LSEs are a bigger threat to reliability than the chance of a second disk failure during rebuild. The industry shift to RAID-6, or double-fault tolerant geometries, in the early 2000's was a response to this threat [10]. Double-fault tolerance became standard in enterprise storage systems not so much to protect against two total disk failures, but to protect against one disk failure and one LSE in a RAID stripe.

Does it really matter if a few chunks aren't reconstructed properly? The answer is yes, absolutely. A recent study examined the impact of undetected or uncorrected errors on distributed systems like Cassandra, and found that they are vulnerable to query errors and failures [11]. It is a myth that one mirror is good enough for an enterprise storage system, no matter how fast you can rebuild a drive in a scaled-out manner.

Latent Sector Errors (LSEs)

Sectors on disk drives can become unreadable through a variety of causes including: media imperfections, stray particles, issues when the content was written, etc [4]. These errors are referred to as latent sector errors or simply bad sectors. The errors are called "latent" because they are hidden to the system until an attempt is made to read the data.

One of the authors of this brief co-authored an extensive study of LSEs at NetApp[2] [4] that found the following:

1] **8.5% of disk drives[3] developed LSEs within two years of use.**

2] **The percentage of disk drives with LSEs also increased as the capacity of the drives increased. If the percentage of drives with errors was bad for drive capacities ten years ago, and the trend holds, the percentages will be even worse now.**

3] **About 70% of the drives with at least one LSE have more than one LSE. That is, if a disk develops LSEs, it is very likely that many sectors are affected.**

The significance of hitting LSEs during drive rebuild is not a recent discovery; it was identified as one of the primary driving factors necessitating two-drive failure protection more than a decade ago [10].

[2] The study won the 10-year Test-of-Time Award from ACM SIGMETRICS, with many follow-on studies with additional results, and new techniques for handling disk failures and latent sector errors.

[3] These numbers are for "nearline" or enterprise-quality SATA disk drives, the kind used widely in the industry now.

The Fact: Data-Loss Probability is Too High with 1FT

We use the following data points in calculating data loss probability:

1] The storage system has 30 4TB drives in a single 1FT pool, and data is replicated in chunks across the entire pool. Rebuilds scale perfectly linearly in the pool. When the system is performing a rebuild, the data being rebuilt is read from the remaining 29 drives and written into spare capacity on each of the surviving drives.

2] The chance of a disk drive failing in a year was 0.76% in the Nutanix dataset [6], 2-4% in the supercomputing dataset [2], about 2% in the Backblaze dataset [8] and as much as 8.6% in the third year of use in the Google dataset [3]. We pick 2% as a reasonable "mid-point" for enterprise-grade drives.[4]

3] The probability of a disk drive developing a latent sector error in 2 years is 8.5%. We estimated from the data in [4] that a drive with LSEs had an average of about 14 LSEs. While LSEs tend to be clustered, more than 50% of them are more than 1 MB apart in logical address [4][5]; so, we assume that they could affect 7 different chunks of data.

4] One way to reduce the chances of encountering a latent sector error during rebuild is constantly "scrubbing" the disk drives, that is, reading and verifying the content and rebuilding lost sectors[6]. The reduction in data-loss probability depends on the scrubbing frequency. Scrubbing at very high frequency without affecting customer workloads is nearly impossible due to increasing drive capacities and the 24/7 nature of many workloads. We assume that the disks are scrubbed once a month. A full scrub of the 30-drive pool once a month would consume 45 MB/s disk bandwidth 24x7[7]. Some HCI storage systems implement scrubbing [12][8].

[4] Given the diverse datasets and in the interest of brevity, we use some approximations in picking error probabilities. Drive-failure probabilities are dependent on factors such as the specific drive models, batches of drives, the age of the drives, etc. Picking 2% has the benefit of better matching the studies using enterprise-grade drives [2,6].

[5] The 50% estimate is conservative. The study [4] simply reflects the smallest distance (not the typical distances between errors that we need for the math here); that is, if there are two clusters of two LSEs each, we lean towards counting it as one chunk affected instead of two.

[6] The other way to find an LSE is when reading data to serve user requests; however, a significant part of the data is only rarely read by workloads and scrubbing is usually the main method for detecting LSEs [4].

[7] Scrubbing once a week would consume 198 MB/s 24x7, which would significantly affect regular workloads.

[8] The Nutanix Bible [12] states that scrubbing is performed when the disk utilization is not high, and does not indicate a scrubbing frequency.

Based on the assumptions, we arrive at these results (see Appendix for the detailed math):

1] The probability of at least one drive in the pool failing in one year is 1-(1-0.02)30 = 45%.

2] In the absence of disk scrubbing, the probability that we will encounter at least one LSE during rebuild is about 41%. In total, there is a 45% * 41% = 18.6% chance of data loss in the third year after LSEs have accumulated over time.

3] With a full scrub of every sector (i.e. 120 TB of data) once a month, the probability that we will encounter at least one LSE during rebuild is about 1.1%. In total, there is a 45% * 1.1% = 0.49% chance of data loss in a year or a 1.95% chance of data loss over the 4-year lifetime of the system with 1FT, even with scrubbing. That number is astonishingly high for an enterprise product.

The math here is conservative: the LSE data used in the above math is for drive sizes that were an order of magnitude smaller than they are now and LSE rates are known to increase with drive size increase; the error rates also increase significantly as the number of drives in a single 1FT pool increases; and errors are known to have correlations (e.g., we are not including the case where the system is rebuilding a large number of drives at the same time when the server hosting the drives goes down). Furthermore, it is entirely possible that many storage systems do not detect corrupt data during rebuild because of inadequate end-to-end and referential data integrity checks [13]. Lack of such checks would hide data loss/corruption issues.

Storing three copies of data with 2FT (examples include RAID 6—arrays, RF3—Nutanix, FTT2—VMware VSAN) or using erasure-coding techniques that tolerate two failures improves reliability significantly. To lose data in a system that tolerates two drive failures, there needs to be either three simultaneous drive failures, two drive failures and an LSE, or one drive failure and LSEs in both redundant copies of the same chunk. All of these are extremely improbable events and the chance of data loss is reduced by many orders of magnitude [7].

In conclusion, IT departments should be very wary of solutions using 1FT or similar offerings that handle only one drive failure, not because a second drive might fail before reconstruction completes, but because there is a good chance of discovering a Latent Sector Error (LSE) during reconstruction. Such an LSE can cause reconstruction of a portion of the data to fail resulting in data loss or corruption. Systems need to be able to handle one drive failure plus one LSE. If they can't, the probability of data loss is simply too high: 1.95% chance of data loss over system lifetime even with once-a-month scrubbing of all data. Infrastructure purchasers should insist on systems configured with 2FT (or use erasure coding that supports 2-drive failure tolerance). Our findings are in

line with the qualitative examination of the topic by industry analyst Gartner; their report on HCI underscores the same data-loss concerns that we discuss in this brief, stating "Data loss can only be prevented by using higher protection levels." [14] Anything less is rolling the dice. **Two nines of durability is just not good enough.**

References

[1] Patterson et al., "A case for redundant arrays of inexpensive disks (RAID)", Proceedings of SIGMOD '88.

[2] Schroeder and Gibson, "Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?", Proceedings of FAST'07.

[3] Pinheiro et al., "Failure Trends in a Large Disk Drive Population", Proceedings of FAST'07.

[4] Bairavasundaram et al., "An Analysis of Latent Sector Errors in Disk Drives", Proceedings of SIGMETRICS'07.

[5] Bairavasundaram et al., "An Analysis of Data Corruption in the Storage Stack", Proceedings of FAST'08.

[6] Schroeder et al., "Understanding Latent Sector Errors and How to Protect Against Them," Proceedings of FAST'10.

[7] Cano et al., "Characterizing Private Clouds: A Large-Scale Empirical Analysis of Enterprise Clusters", Proceedings of SoCC'16.

[8] Andy Klein, "Hard Drive Stats for Q2 2017", http://www.backblaze.com/blog/hard-drive-failure-stats-q2-2017/, Last accessed: September 12, 2017.

[9] Anderson et al., "More than an interface — SCSI vs. ATA", Proceedings of FAST'03.

[10] Corbett et al., "Row-Diagonal Parity for Double Disk Failure Correction", Proceedings of FAST'04.

[11] Ganesan et al., "Redundancy Does Not Imply Fault Tolerance: Analysis of Distributed Storage Reactions to Single Errors and Corruptions," Proceedings of FAST'17.

[12] Poitras, "The Nutanix Bible", http://nutanixbible.com/, Last accessed: August 25, 2017.

[13] Krioukov et al., "Parity Lost and Parity Regained", Proceedings of FAST'08.

[14] Jerry Rozeman, "Key Differences Between Nutanix, SimpliVity and VxRail HCIS Appliances, Part 2: Data Protection, Capacity Optimization and Failure Analysis", Gartner Report, ID: G00334429, Published: 26 July 2017

Appendix: Detailed Math

Symbol	Description	Value
N	Number of Drives	30
F	Single disk failure in one year probability	0.02
O	Probability of at least one disk in the pool failing in one year	= 1 - Probability of no disk failure = $1 - (1-F)^N$ = 45.45%
L	Disk having LSEs in two years probability	0.085
T	The time period for LSE probability in days	720
A	Avg. number of error chunks in an LSE disk	7
B	Scrub interval in days (on average it'll be half this number of days since a sector was last checked)	30
C	Probability that an LSE will be encountered when rebuilding from LSE disk	= 1 - Probability of no errors = 1 - None of the A errors was hit when reading 1 / N fraction of a drive[9]. = $1 - (1 - 1 / N)^A$ = 21.1%
E	Probability that an LSE will be encountered from one disk when rebuilding[10].	= ((L * C) * (B/2)) / T = 0.037%
D	**Probability of data loss in one year**	= O * Probability of hitting LSE during rebuild = O * (1 - Probability of no LSEs during rebuild) = $O * (1 - (1 - E)^{(N-1)})$ = 45.4% * 1.08% = **0.49%**

[9] It is 1 / N and not 1 / (N-1) since 1 drive worth of capacity is needed as spare space for rebuilding.

[10] In this formula, we treat L as if it is evenly spread throughout the two years for average results; the first year would be better than our numbers indicate and the second year would be worse since LSE rates increase with age.