Nutanix’s RF2 Configuration for Flash Storage: 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

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 one drive failure. Variants of RAID have been used universally in enterprise storage products over the years. The primary factor under consideration for the choice of RAID variant is the probability of data loss. Space efficiency and performance overheads of the scheme are important additional factors.

Using RF2 on a flash-based storage system would imply a 0.4-1.1% chance of data loss in one year (and 1.7-4.3% chance of data loss in system lifetime of 4 years) under typical assumptions; customers should demand RF3

We have previously shown that the probability of data loss under single failure tolerance (Nutanix's RF2 configuration is an example) in storage systems using disk drives is simply too high [2]. In this paper, we examine the same for flash storage systems. Recent studies by Facebook [3], Google [4], Microsoft [5], and Nutanix [6] provide a wealth of data on flash-drive failures in the field. The key lessons from these studies are as follows. First, full SSD failures are common. Second, SSDs encounter Latent Sector Errors (also called uncorrectable errors) at an alarming rate — higher than one would expect from vendor specifications derived from accelerated testing. This combination can have a profound impact on reliability and needs to be considered when designing a system.

When we examine the myth and the fact for RF2 in the context of data from the above-mentioned studies, we find that using RF2 with once-a-month drive scrubbing implies an extremely high data-loss probability of 0.4-1.1% per year or 1.7-4.3% over the 4-year life of a flash-based storage system. Not scrubbing increases the data-loss probability to as high as 50% in one year. The data-loss probability range for RF2 on flash drives is comparable to that for RF2 on disk drives, with the higher end of the range more than the 0.49% per year number for disk drives [2]. Thus, irrespective of whether the storage medium is flash or disk, RF2 vs. RF3 is not really a choice: given the substantial risk of data loss with RF2, one should pick RF3 every time.

The Myth: RF2 Provides Sufficient Reliability

RF2, or Replication Factor of 2, is a product configuration quoted by Nutanix that can handle one drive failure[1]. In this configuration, data is typically written to drives in chunks and each chunk is stored twice (on two different drives in the pool). When a drive fails, its chunks are rebuilt from the copies spread across the pool. The alternative configuration, RF3, provides double failure tolerance by storing three copies of the chunk on three different drives so that data can be rebuilt from any of the copies.

[1] Other hyperconvergence vendors have similar single failure tolerance configurations.

Decades ago, reliability analysis of RAID systems centered on whether data could be rebuilt after a drive failure before a second drive 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 drive failure. But, that thinking does not account for a new problem: 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 drive's contents. The chances of this happening are frighteningly high. Today, LSEs are a bigger threat to reliability than the chance of a second drive failure during rebuild. The fundamental issue is that in order to rebuild the data lost on a drive, you would always have to read at least full drive's worth of data to rebuild it, and the chance of hitting an LSE is related to how much data is read during rebuild, not how fast a rebuild can be done. Double-fault tolerance became standard in enterprise storage systems starting in the early 2000's not so much to protect against two total drive failures, but to protect against one drive failure and one LSE in a RAID stripe [7]. The studies of the prevalence of uncorrectable errors in SSDs shows that the approach is well-worth continuing [3,4].

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 [8]. 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.

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)

Latent sector errors (LSEs) happen on disk drives due to media imperfection, stray particles, etc. Given that flash drives do not have moving parts like disk drives do, a reasonable question to ask is whether flash drives have LSEs at all. The answer is that they do in fact have LSEs. Studies by Facebook [3] and Google [4] find that a large percentage of their drives encounter "uncorrectable errors" — where a read operation gets more corrupted bits than can be fixed using drive ECC. Microsoft's study [5] focuses on full SSD failure, but notes that the uncorrectable error rate in the field is between one and three orders of magnitude higher than SSD vendor specifications[2]. The studies do not attribute uncorrectable

[2] The uncorrectable error rates are for the MLC drives in the study, but the orders of magnitude divergence from vendor specifications underscores the importance of real-world field data over specifications derived from accelerated testing.

errors to a specific failure mechanism; for example, due to the lack of correlation with correctable bit errors, the Google study attributes uncorrectable errors to “larger scale issues with the device” than on factors like retention errors in individual cells.

We use the data from the Google study [4] for our data-loss probability calculations due to the limitations of the Facebook study; some examples of the limitations are that the Facebook study includes only MLC[3] drives and that the failure data is from a snapshot in time (as opposed to studying all drives in an N-year timespan). We summarize the Google study and the data points we use below[4]:

1] The study includes data on MLC, eMLC, and SLC flash drives over a six-year total time period, including any errors experienced by the drives in their first four years of life. The SSDs use commodity flash, but have custom firmware. We assume that Google’s firmware is not substantially different than that of SSD vendors at handling flash from a reliability standpoint[5].

2] The study delves into different error types — both correctable and uncorrectable ones — and also looks into characteristics of “bad” blocks (erase blocks that are marked as “bad” when block access errors are encountered).

3] Uncorrectable errors can be encountered during user reads as well as device-internal reads (say, garbage collection). There is data loss at the drive level when an uncorrectable error is encountered; higher-level redundancy — i.e., the topic of this report — is needed to avoid actual data loss. Depending on the drive model, 20-90% of the drives encounter uncorrectable errors in four years. This rate is significantly higher than the rate of LSEs on disk drives in the NetApp study [9].

4] The conditional probability of encountering another uncorrectable error on the same drive in a month after an uncorrectable error is as high as 30% while the unconditional probability for a random month is only 2%. Since an uncorrectable error typically removes a block from further usage and erase block sizes are usually large[6], this indicates that multiple data chunks are likely to be affected on a drive with uncorrectable errors. In the study, Google found that the median number of bad blocks was 2-4 and the mean numbers ranged from about 200 to about 2000[7]. Beyond 2-4 bad blocks, there is a 50% chance that hundreds of bad blocks will follow. The paper does not indicate how soon these additional bad blocks appear (so that we may account for scrubbing frequency). To be conservative, we assume that the typical number of bad blocks encountered during rebuild on a drive with bad blocks is 2-4.

[3] MLC stands for “multi-level cell.” Enterprise storage systems usually use eMLC (enterprise MLC) or SLC (single-level cell) flash drives.

[4] We focus on the most-relevant parts of the study: drive replacement rates, uncorrectable errors, and bad blocks.

[5] The study discusses the impact of program-erase (PE) cycles, and finds that many drives with PE cycles well below the limit run into uncorrectable errors; therefore, the errors cannot simply be attributed to poor flash management by SSD firmware.

[6] The Google study does not indicate precisely how big erase blocks are in their drives. However, erase blocks are usually large enough that losing multiple blocks would likely cause multiple data chunks to be lost.

[7] According to the paper, a “bad” block and an uncorrectable error are not exactly the same; correctable errors like write errors can also lead to a block being marked as bad. That said, there is substantial overlap between the two categories, and the Google study mentions that most bad blocks are discovered during reads.

5] In the first four years in the field, 6-30% of the drives undergo repair (at least temporarily considered “failed” or unusable), and 4-10% of the drives are permanently replaced in four years. The paper does not indicate whether we should consider all of the data lost on the drives that undergo repair. To be conservative, we pick the replacement percentage. Note that the lower end of the range (4% in four years) is a close match to the flash failure rate of 0.72% per year observed by the Nutanix study [6] and the 0.2-1% observed by the Microsoft study[8] [5].

The Fact: Data-Loss Probability is Too High with RF2

The detailed math is shown in the appendix for each of the drive models used in the Google study. We use the following assumptions in our calculations:

1] We pick a pool size of 60 drives since commonly-used flash drives are about half the capacity of commonly-used disk drives, and we want to match the 30-disk-drive pool capacity in our earlier technical report [2]. The eMLC drives in the study are 2TB in size and would match those expectations. We would need many more of the SLC drives to match the capacity (up to ~250 drives), but we maintain the drive count at 60 for all drive types so that the drive count of the pool is not unreasonably high.

2] One way to reduce the chances of encountering an uncorrectable error during rebuild is constantly “scrubbing” the drives, that is, reading and verifying the content and rebuilding lost pages or blocks[9]. 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 storage system performs once-a-month scrubbing of drive contents to discover uncorrectable before they could cause data loss in a rebuild scenario[10].

Using the data and our assumptions, we arrive at the following results for one of the drive models in the Google study -- “eMLC-B”; the appendix includes the data for all the drive models.

1] The probability that at least one out of the drives in the pool will fail in one year is 43.6%.

2] With once-a-month scrubbing of the entire pool of drives, the probability that an LSE will be encountered during rebuild is 1.8%. Thus, the chances that there’ll be data loss in one year is 43.6% x 1.8% = 0.79%.

[8] The eMLC drive in Microsoft’s study was the drive at the lower end of the failure rate at 0.2%, but those drives were also younger than the MLC drives.

[9] The other way to find an uncorrectable error is when reading data to serve user requests; however, a significant part of the data is only rarely read by workloads and therefore storage systems use scrubbing as the main method for detecting the errors [9].

[10] Some hyper-convergence vendors perform scrubbing [10]; we do not know if scrubbing is performed for flash drives as well as disk drives (we would certainly recommend it), or the frequency of scrubbing.

3] The chances of data loss in typical system lifetime of four years is 1 - (1 - 0.0079)^4 = 3.14%.

4] We can perform the same calculations for the case without any scrubbing; as one would expect the data-loss probability is extremely high — 36% in one year.

Across the drive models, the probability of data loss in one year ranges from 0.4% to 1.1% with once-a-month scrubbing[11]. Again, as in the case of disk drives, this number is shockingly high! The math here is conservative: we use bad block counts far lower than the average counts seen in the Google study.

It is also important to note that it is entirely possible that many storage systems do not detect corrupt or lost data during rebuild because of inadequate end-to-end and referential data integrity checks [11]. Lack of such checks would hide data loss/corruption issues.

One interesting note is that a Nutanix system with RF2 stores two copies of customer data and three copies of Nutanix's metadata [10], so that data loss due to LSEs during rebuild would affect customer data while Nutanix's metadata itself remains protected. It seems to be a somewhat surprising design choice to expose customer data to such high risk that one doesn't want for metadata.

Storing three copies of data with RF3 (or similar offerings from other vendors) 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 [6].

Customers are sometimes steered away from RF3 or similar offerings due to higher space usage or poorer performance due "write amplification" (substantially higher amount of I/O may be needed to perform writes). These issues are due to the limitations of many of the hyper-converged solutions. These issues can be eliminated with a storage-system design that performs erasure coding of data to deliver space efficiency, and uses a log-structured file system to deliver performance efficiency [12].

Conclusion

IT departments should be very wary of solutions using RF2 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.

[11] Without scrubbing, the data loss probability ranges from 25% to 50% in one year; to calculate, substitute the scrub interval value of 30 days with 1440 days in the math in the appendix.

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: 1.7% to 4.3% chance of data loss over system lifetime even with once-a-month scrubbing of all data. This data-loss probability range for RF2 on flash drives is comparable to that for RF2 on disk drives.

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" [13]. Anything less is rolling the dice.

References

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

[2] Bairavasundaram et al. "Single Failure Tolerance (1FT): Myth and Fact", Datrium Technical Report, 2017, http://www.datrium.com/wp-content/uploads/2017/10/1FT-Myth-and-Facts.pdf. Last Accessed: December 2017.

[3] Meza et al., "A Large-Scale Study of Flash Memory Failures in the Field", Proceedings of SIGMETRICS'15.

[4] Schroeder et al., "Flash Reliability in Production: The Expected and the Unexpected", Proceedings of FAST'16.

[5] Narayanan et al., "SSD Failures in Datacenters: What? When? and Why?", Proceedings SYSTOR'16.

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

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

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

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

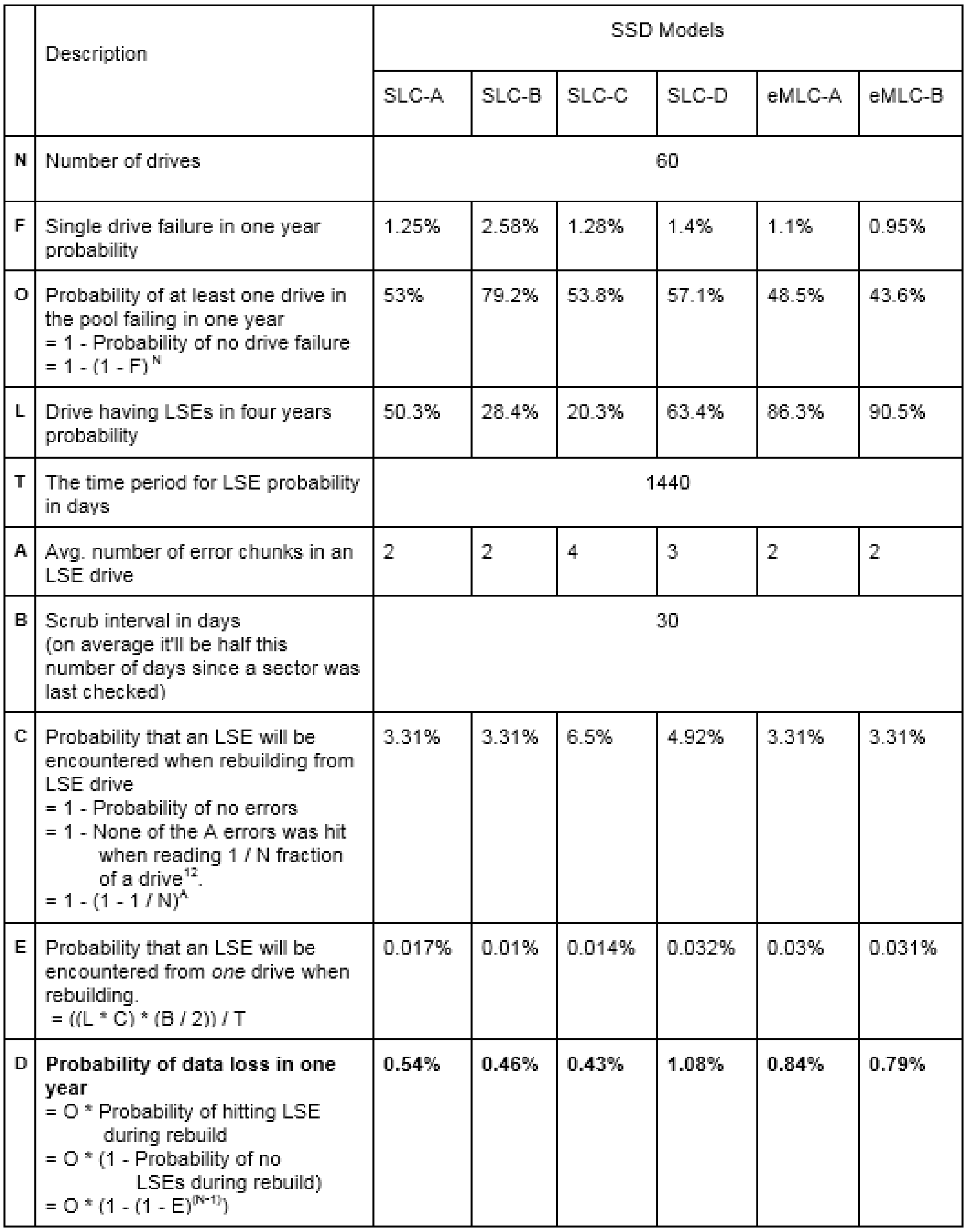
[10] Poitras, "The Nutanix Bible", http://nutanixbible.com/, Last accessed: August 25, 2017.

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

[12] Datrium Inc, "Always-On Erasure Coding", http://www.datrium.com/wp-content/uploads/2017/11/Erasure-Coding-WP.pdf. Last accessed: December 2017.

[13] 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 for Each SSD Model



	Description	SSD Models					
		SLC-A	SLC-B	SLC-C	SLC-D	eMLC-A	eMLC-B
N	Number of drives	60					
F	Single drive failure in one year probability	1.25%	2.58%	1.28%	1.4%	1.1%	0.95%
O	Probability of at least one drive in the pool failing in one year = 1 - Probability of no drive failure $= 1 - (1 - F)^N$	53%	79.2%	53.8%	57.1%	48.5%	43.6%
L	Drive having LSEs in four years probability	50.3%	28.4%	20.3%	63.4%	86.3%	90.5%
T	The time period for LSE probability in days	1440					
A	Avg. number of error chunks in an LSE drive	2	2	4	3	2	2
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 drive = 1 - Probability of no errors = 1 - None of the A errors was hit when reading 1 / N fraction of a drive[12]. $= 1 - (1 - 1 / N)^A$	3.31%	3.31%	6.5%	4.92%	3.31%	3.31%
E	Probability that an LSE will be encountered from *one* drive when rebuilding. $= ((L * C) * (B / 2)) / T$	0.017%	0.01%	0.014%	0.032%	0.03%	0.031%
D	**Probability of data loss in one year** = O * Probability of hitting LSE during rebuild = O * (1 - Probability of no LSEs during rebuild) $= O * (1 - (1 - E)^{(N-1)})$	**0.54%**	**0.46%**	**0.43%**	**1.08%**	**0.84%**	**0.79%**

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