Title

Using device diversity to protect data against batch-correlated disk failures

ISBN

9781595935526

Authors

Pâris, Jehan-François

Long, Darrell DE

Publication Date

2006-10-30

DOI

10.1145/1179559.1179568

Peer reviewed

Using Device Diversity to Protect Data against Batch-Correlated Disk Failures

Jehan-François Pâris*
Department of Computer Science
University of Houston
Houston, TX 77204-3010
+1 713-743-3341
paris@cs.uh.edu

Darrell D. E. Long*
Department of Computer Science
University of California
Santa Cruz, CA 95064
+1 831-459-2616
darrell@cs.ucsc.edu

ABSTRACT
Batch-correlated failures result from the manifestation of a common defect in most, if not all, disk drives belonging to the same production batch. They are much less frequent than random disk failures but can cause catastrophic data losses even in systems that rely on mirroring or erasure codes to protect their data. We propose to reduce impact of batch-correlated failures on disk arrays by storing redundant copies of the same data on disks from different batches and, possibly, different manufacturers. The technique is especially attractive for mirrored organizations as it only requires that the two disks that hold copies of the same data never belong to the same production batch. We also show that even partial diversity can greatly increase the probability that the data stored in a RAID array will survive batch-correlated failures.

Categories and Subject Descriptors
B.4.5 [**Hardware**]: INPUT/OUTPUT AND DATA COMMUNICATIONS – Reliability, Testing, and Fault-Tolerance – *Redundant design.*

General Terms
Performance, Reliability.

Keywords
Disk array reliability, disk mirroring, RAID

1. INTRODUCTION
Computer disks are now more reliable than they were twenty or thirty years ago. Their mean times to fail now span decades, which means that most disks will keep operating in a satisfactory fashion until they are retired. As a result, most small computer installations that do backups treat them as some kind of insurance policy.

The situation is quite different whenever large amounts of data must be preserved over long periods of time. First, storing terabytes, if not petabytes, of data requires a fairly large number of disks. Consider a relatively small disk farm consisting of fifty disks. Assuming a disk mean time to failure of one hundred thousand hours, this installation is likely to experience one disk failure every three months. In addition, these risks must be factored over data lifetimes that can exceed ten or even twenty years.

The best solution to guarantee the survivability of digital data over long periods of time is to introduce enough redundancy in the storage system to prevent data losses rather than trying to recover the lost data. Two techniques that can be used are mirroring and erasure codes. Mirroring maintains two or sometimes three exact copies of the data on distinct disks. Should one of these disks fail, the data will still be available on the surviving disk. An m-out-of-n code groups disks into sets of n disks that contain enough redundant data to tolerate the loss of up to $n - m$ disks. RAID level 3 and 5 organizations use $(n - 1)$-out-of-n codes [3, 7, 8] while RAID level 6 organizations use m-out-of-n codes that can protect data against two or even three disk failures [2]. Corbett *et al.* [4] have more recently proposed a provably optimum algorithm that protects data against double disk failures. Their technique stores all data unencoded, and uses only exclusive-or operations to compute parity. A major motivation for their work was the increasing occurrence of media errors (bad blocks) during the recovery of a disk failure in conventional RAID arrays.

Both mirroring and m-out-of-n codes operate on the assumption that disk failures are independent. While this assumption is generally true, correlated disk failures do happen. Some of these failures result from acts of God and other environmental factors over which we have little control. Other correlated failures are caused by installation malfunctions such as cooling failures or power surges. True correlated disk failures result from the manifestation of a common defect among disks that belong to the same fabrication batch. We call these failures *batch-correlated failures*. We propose to discuss this issue, estimate its impact on storage system reliability and propose a solution.

* Supported in part by the National Science Foundation under award CCR-0204358.

StorageSS'06, October 30, 2006, Alexandria, Virginia, USA.

2. MODELING BATCH-CORRELATED DISK FAILURES

Obtaining reliable data on disk failure modes is a very difficult task because all disk manufacturers consider these data to be proprietary. Works by Elerath and Shah [5, 6, 10, 11] are our best source of information even though their authors have refrained from publishing any hard numbers or identifying specific disk manufacturers. We can summarize their observations in the following fashion:

1. The reliability data published by disk manufacturers are obtained by exercising disks kept at temperatures above their normal operating range and extrapolating these data to ideal operational conditions that are not always realized in practice. As a result, these data tend to be very optimistic. In addition, the extrapolation factors that are used to compute the published data remain subjective.
2. Disk reliability varies within the same family of disks with the vintage, that is, the fabrication date of the disks.
3. Some batches of disks experience high infant mortality while other batches experience premature failures as their failure rates that increase over time.
4. Most disk failures are caused by fabrication defects.

Talagala [12] observed faults occurring in a 368-disk farm at the University of California, Berkeley over a period of six months and noted significant correlation between them. More recently, Baker *et al.* [1] presented a simple reliability model of storage systems that takes into account correlated faults, either when a single fault occasion others or when multiple faults result from the same defect. They mention that disks in a disk array normally come from the same manufacturing batch and conclude that this makes the array more susceptible to batch-correlated failures. They state that "the increased costs that would be incurred by giving up supply chain efficiencies of bulk purchases might make hardware diversity difficult." They also claim that continuous procurements of new disks will over time introduce some hardware diversity.

Given the dearth of reliable data on disk failures, it is difficult to justify a given model for the occurrence batch-correlated failures. Baker *et al.* [1] assume that subsequent failures have a higher probability of occurring than an initial failure and introduce a multiplicative correlation factor $\alpha < 1$ that applies to the mean time to failure once an initial failure occurs. The main advantage of their approach is its simplicity. Its main limitation is its sensitivity to scale. We can reasonably assume that batch-correlated disk failures result from the manifestation of a common defect shared by an unusual number of disks in a specific batch. Then the rate at which we would observe a sequence of batch-correlated failures should not depend on the number of the disks we are monitoring. This is not true with their model.

We propose a more realistic model. We assume that disks have two kinds of defects that manifest themselves in different fashion. Defects from the first group result in failures that happen randomly during the useful lifetime of the disks. These are the defects we are normally considering. Since they occur in an independent fashion for each individual disk, replication and erasure codes are very effective against them. Defects in the second group are much less prevalent but result in batch-correlated failures that happen in rapid succession. Their likely outcome is infrequent catastrophic failures even in systems that rely on mirroring or erasure codes to protect their data.

Consider a group of n disks all coming from the same production batch. We will consider two distinct failure processes:

1. Each disk will be subject to independent failures that will be exponentially distributed with rate λ; these independent failures are the ones that are normally considered in reliability studies.
2. The whole batch will be subject to the unpredictable manifestation of a common defect. This event will be exponentially distributed with rate $\lambda' \ll \lambda$. It will not result in the immediate failure of any disk but will accelerate disk failures and make them happen at a rate $\lambda'' \gg \lambda$.

3. EVALUATING THE IMPACT OF BATCH-CORRELATED FAILURES

Our ability to evaluate the impact of batch-correlated failures is severely limited by the lack of data. We have no reliable data about the rate at which global defects will manifest themselves in a batch of identical disks but can safely assume that they manifest themselves at a rate λ' that is much lower than the rate λ at which independent failures manifest themselves. Anecdotal evidence indicates that the rate at which individual drives fail once a global defect has occurred rarely exceeds one disk per week.

Consider, for instance, a RAID level 5 array consisting of n data disks and one spare disk that all belong to the same production batch. Since RAID level 5 organizations only protect against single disk failures, a failure of one of the $n + 1$ disks will make the array enter into a *window of vulnerability* [1] that will last until the failed disk gets replaced. Let T_R denote that time interval. If no global batch defect has manifested itself, we only have to consider normal failures. The probability that the data will survive that failure will then be

$$\exp(-n\lambda T_R). \qquad (1)$$

Assuming a failure rate of one failure every one hundred thousand hours, that is, nearly eleven and half years, and a repair time equal to one day, the probability that the data stored on an eight-disk array will survive the failure of one of its disks is 0.998 as long as no global defect has manifested itself.

Assume now that the first failure resulted from the manifestation of a global defect and that the rate λ'' at which disks will then fail is one failure per week. Replacing λ by λ'' in Eq. 1, we find out that the probability that the data stored on the array will survive the first failure is only 0.368. In other words, the data are most likely to be lost. Even with a less virulent kind of defect, say, one that would cause each disk to fail at a rate λ'' of one failure per month, the probability that the data would survive the first failure is 0.792.

Note than the same problem occurs with mirrored organizations. Consider a pair of disks such that each disk has an exact copy of the data stored on the other disk. The probability that one disk will fail after the other one has failed but before it is replaced is $\exp(-\lambda T_R)$. Assuming that disk failure rates and repair times remain the same, the probability that the data will survive the failure of one of the two disks is 0.9998 as long as no global

Table I. Probabilities that the data stored on mirrored disks will survive the first batch-correlated failure assuming it takes one day to replace a failed disk.

Replication level	*Survival rate*	
	when λ'' = one failure/week	*when λ'' = one failure/month*
Two disks from the same batch	0.867	0.967
Three disks from the same batch	0.966	0.998
Two disks from different batches	0.9998	0.9998

defect has manifested itself and disk failures can be instantly detected. Assuming that a global defect has manifested itself and results in a subsequent failure rate of one failure per week, that probability becomes 0.867, which is still an unacceptably low value for most applications.

4. PROTECTING DATA AGAINST BATCH-CORRELATED FAILURES

A possible way to reduce data losses resulting from batch-correlated failures is to accelerate the repair process, thus reducing the window of vulnerability T_R of the data. For instance, reducing the time it takes to replace a failed disk to two hours would bring to 0.920 the probability that the data would survive the first failure when λ'' is equal to one failure per week. Unfortunately, this technique suffers from two major limitations. First, the repair process requires both replacing the defective disk and storing on the new disk the data that were on the old disk. Even under the most ideal conditions, the minimum time required to perform this operation will be given by the ratio C_D/B_D of the disk capacity C_D over its bandwidth B_D. This bottleneck is not likely to disappear soon as disk capacities tend to increase at a much faster rate than disk bandwidths. Second, we cannot assume that all disk failures will be instantly detected. This problem is especially acute in archival storage system, as most data will be infrequently accessed [1, 9].

A second technique for improving data survivability consists of using storage architectures that tolerate two consecutive failures. This solution may appear especially attractive in the case of RAID arrays as it only requires the addition of a second check disk. The probability that the data will survive the initial failure then becomes the probability that at most one additional failure will occur while the disk that failed first is replaced. Since failures are distributed according to a Poisson law, this probability is

$$\begin{aligned}&\Pr[\text{no failure}] + \Pr[\text{one failure}]\\ &= (1+n\lambda''T_R)\exp(-n\lambda''T_R). \qquad (2)\end{aligned}$$

Assuming a repair time of one day and a failure rate λ'' of one failure per week, we find out that the probability that the data will survive the first batch-correlated failure is now 0.683, that is almost twice the previous value. Should the defect cause each disk to fail at the rate of one failure per month, this survival probability would be equal to 0.970, which is quite acceptable for most applications since batch-correlated failures are likely to occur much less frequently than independent failures. Unfortunately, this technique is not as cost-effective as we would want it to be. Adding a second check disk would greatly complicate the task of the RAID controller and thus its cost. The technique does not fare better with mirrored organizations as tolerating two consecutive failures would require maintaining three copies of all stored data, thus increasing by 50 percent the cost of the disks.

We propose a third solution that does not require any additional hardware. It consists of introducing diversity into disk arrays by building them with disks from different production batches, possibly from different manufacturers.

This approach is particularly attractive for mirrored organizations. Regardless of the size of the installation, we only have to ensure that the two disks that hold copies of the same data never belong to the same production batch. This would guarantee that the survival of the data will never be affected by the appearance of a common defect in these two disks. The sole remaining impact of these batch-correlated failures will be a rapid succession of failures among all the disks holding one copy of our data. While the disks holding the other copy would remain unaffected, this sudden succession of failures could overwhelm our repair process and possibly create other problems. As Table I shows, replicating data on two disks that come from two different batches always achieves a much better level of data protection than replicating data on three disks coming from the same batch.

The same approach is somewhat more difficult to implement in RAID arrays as it requires all disks in a RAID stripe to come from different production batches. This would be difficult to achieve for RAID stripes counting more than four disks. In these cases, we may have to content ourselves with introducing partial rather than total diversity in our RAID array, say, including in each stripe disks coming from two different batches.

Returning to our previous example of a RAID level 5 array consisting of n data disks and one spare disk, assume for the sake of simplicity that n is odd. Let us further assume that the array consists of two sets of $(n + 1)/2$ disks coming from different batches, say batches A and B. We can safely assume that the rate λ', at which batch defects will manifest themselves will be low enough to make the simultaneous manifestations of correlated failures in two or more different batches a very unlikely occurrence. Then the probability that the data will survive the first batch-correlated disk failure will be the product of the probabilities of no additional failures in the two sets before the failing disk can be replaced. That probability is

$$\exp\left(-\frac{n-1}{2}\lambda''T_R\right)$$

for the set that is affected by the default and

Table II. Probabilities that the data stored on a RAID level 5 array consisting of eight disks will survive the first batch-correlated failure assuming it takes one day to replace a failed disk.

Storage organization	*Survival rate*	
	when λ'' = one failure/week	*when λ'' = one failure/month*
All eight disks come from same batch	0.368	0.792
Same with an additional check disk	0.683	0.970
Disks come from two different batches	0.651	0.905
Disks come from four different batches	0.867	0.988
Each disk comes from a different batch	0.998	0.998

$$\exp(-\frac{n+1}{2}\lambda T_R)$$

for the other set. Thus the probability that the data will survive the first batch-correlated failure is

$$\exp(-\frac{n-1}{2}\lambda'' T_R)\exp(-\frac{n+1}{2}\lambda T_R) \quad (3)$$
$$\cong \exp(-\frac{n-1}{2}\lambda'' T_R)$$

as long as $\lambda'' \gg \lambda$.

Assuming a repair time of one day and a failure rate λ'' of one failure per week, we find out that the probability that the data will survive the first batch-correlated failure is 0.651 for a RAID level 5 array consisting of eight disks. This is nearly twice the survival probability of data stored on a homogenous RAID level 5 array (0.368) and almost the same survival probability of data stored on a homogeneous RAID level 6 array with seven data disks and two check disks (0.683).

We would obtain even better results if we could incorporate in our RAID level 5 array disks from four different batches. Then the probability that the data would survive the first batch-correlated failure would be 0.867. As a result, incorporating disks from four different batches would increase by 136 percent the probability that the data will survive the first batch-correlated failure (0.867) when we compare it to a RAID level 5 organization where all eight disks come from the same batch (0.368). What is even more surprising is that this simple technique protects data much better than the addition of a second check disk: the probability that the data will survive the first batch-correlated failure is now 27 percent more than that of a RAID level 6 organization whose nine disks come from the same batch.

Table II summarizes our results. As we can see, ensuring that all eight disks come from separate batches is by far the best solution as it completely eliminates the risk of batch-correlated disk failures. When this is not feasible, incorporating disks from four different batches protects data better than adding an extra check disk. Even incorporating disks from only two distinct batches has significant beneficial effects. It protects data nearly as well as adding an extra check disk when λ'' is one failure per month but not as well when the batch-correlated failure rate λ'' is one failure per week as the probability of two or more additional disk failures among the eight surviving disks is then less than the probability of one additional disk failure in the three remaining disks in the defective batch.

While these results emphasize the benefits of introducing as much disk diversity as possible in our RAID arrays, they should also convince us not to overestimate the beneficial effect of the accidental diversity obtained by replacing failed disks from the original batch by new disks coming from a different batch [1]. Returning to our example and assuming again a normal disk failure rate of one failure each one hundred thousand hours, we find out that disks will fail and be replaced at a rate of 0.7 disks per year in a disk array consisting of eight disks. We would thus have to wait more than five years to ensure that our array has less than four disks belonging to the same batch.

5. CONCLUSION

Large storage systems usually include some measure of redundancy in order to protect data against disk failures. This approach assumes that disk failures are independent events, which is only true to some extent. We have proposed a simple technique to reduce the number and impact of correlated failures caused by the sudden manifestation of a common batch defect. It consists of introducing diversity into disk arrays by building them using disks from different batches and different manufacturers. The technique is especially attractive for mirrored organizations as it only requires that the two disks that hold copies of the same data never belong to the same production batch. We have also shown that even partial diversity can greatly increase the probability that the data stored in a RAID array will survive batch-correlated failures.

We can safely conclude that incorporating disks from as many batches as possible in our storage systems constitutes the most cost-effective way to protect data against batch-correlated disk failures even if we consider the additional cost of purchasing more frequently smaller lots of disks, thus forfeiting bulk discounts.

6. REFERENCES

[1] M. Baker, M. Shah, D. S. H. Rosenthal, M. Roussopoulos, P. Maniatis, T. J. Giuli, and P. Bungale. A Fresh Look at the Reliability of Long-Term Storage. In *Proc. First EuroSys Conference* (EuroSys 2006), Leuven, Belgium, Apr. 2006.

[2] W. Burkhard and J. Menon. Disk Array Storage System Reliability. In *Proceedings of the 23rd Annual International Symposium on Fault-Tolerant Computing* (FTCS-23), Toulouse, France, June 1993, 432–441.

[3] P. M. Chen, E. K. Lee, G. A. Gibson, R. Katz, and D. Patterson, RAID, High-performance, reliable secondary storage. *ACM Computing Surveys*, *26, 2* (1994), 145–185.

[4] P. Corbett, B. English, A. Goel, T. Grcanac, S. Kleiman, J. Leong, and S. Sankar. Row-Diagonal Parity for Double Disk Failure Correction. In *Proceedings of the 3rd USENIX Conference on File and Storage Technologies* (FAST '04), San Francisco, CA, Mar.–Apr. 2004, 1–14

[5] J. G. Elerath. Specifying Reliability in the Disk Drive Industry: No More MTBF's. In *Proceedings of the 46th Annual Reliability and Maintainability Symposium* (RAMS 2000), Jan. 2000, 194–199.

[6] J. G. Elerath and S. Shah. Server Class Disk Drives: How Reliable Are They? In *Proceedings of the 50th Annual Reliability & Maintainability Symposium* (RAMS 2004), Jan. 2004, 151–156.

[7] D. Patterson, G. Gibson, R. Katz. A Case for Redundant Arrays of Inexpensive Disks (RAID). In *Proceedings of the ACM Conference on Management of Data* (SIGMOD '88), Chicago, IL, June 1988, 109–116.

[8] T. J. E. Schwarz, S.J. and W. A. Burkhard. RAID Organization and Performance. In *Proceedings of the 12th International Conference on Distributed Computing Systems*, Yokohama, Japan, June 1992, 318–325.

[9] T. J. E. Schwarz, S.J., Q. Xin, E. Miller, D. D E. Long, A. Hospodor, and S. Ng. Disk Scrubbing In Large Archival Storage Systems. In *Proceedings of the 12th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems* (MASCOTS'04), Volendam, The Netherlands, Oct. 2004, 409–418.

[10] S. Shah and J. G. Elerath. Disk Drive Vintage and Its Effect on Reliability. In *Proceedings of the 50th Annual Reliability & Maintainability Symposium* (RAMS 2004), Jan. 2004, 163–165.

[11] S. Shah and J. G. Elerath. Reliability Analysis of Disk Drive Failure Mechanisms. In *Proceedings of the 51st Annual Reliability & Maintainability Symposium* (RAMS 2005), Jan. 2005, 226–231.

[12] N. Talagala. *Characterizing Large Storage Systems: Error Behavior and Performance Benchmarks*. PhD thesis, Dept. of EECS, University of California, Berkeley, CA, Oct. 1999.