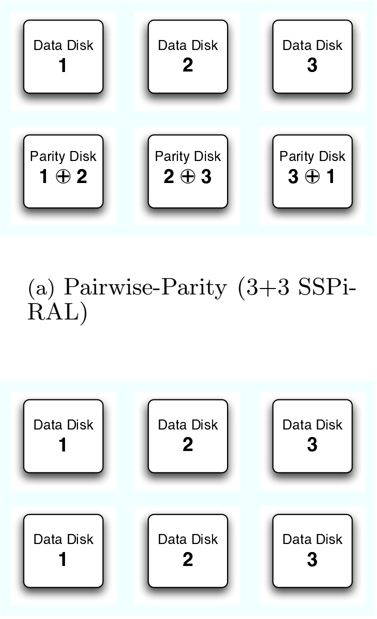
Outshining Mirrors: MTTDL of Fixed-Order SSPiRAL Layouts

Ahmed Amer[†] Jehan-François Pâris[‡] Thomas Schwarz[§]
Vincent Ciotola[†] James Larkby-Lahet[†]
[†] *University of Pittsburgh*
[‡] *University of Houston*
[§] *Santa Clara University*

1 Introduction

Complementary trends in hardware and applications are driving an increase in demand for data volume and bandwidth, resulting in an increased risk of data loss and a growing need for improved storage reliability. There is a growing need to survive the failure of multiple storage devices in larger storage arrays, as well as the need to survive the loss of multiple nodes in clustered storage. Redundant storage schemes are the obvious solution, and such applications commonly employ one of two strategies: a combination of replication and parity applied efficiently across an array of devices, or a failure-recovery scheme based on erasure coding. Computational efficiency is important when implementing redundancy schemes for disks, and so parity is particularly appealing due to its ease of computation. There are also combinations of the two approaches, but typically parity schemes tolerate only a small number of component failures, while erasure codes tend to be expensive to implement. Excellent parity-based erasure codes and layout schemes have been devised [11, 6], but prior art has focused primarily on aiming to survive a specific number of device failures. We present an argument for an efficient parity-based scheme that compares favorably to erasure codes in terms of reliability.



(a) Pairwise-Parity (3+3 SSPiRAL)

(b) 3 pairs of mirrored disks

Figure 1: *Pairwise parity vs. equivalent RAID array.*

2 SSPiRAL Description

SSPiRAL (Survivable Storage using Parity in Redundant Array Layouts) [3] is a redundant data layout scheme based solely on efficient parity computations, offering high reliability and maintainability. Every SSPiRAL layout is defined by three parameters: the degree of the system, the *x-order*, and the total number of nodes available. The degree of a SSPiRAL layout is the number of unique data nodes, while the x-order is the number of nodes that contribute to constructing a parity node. A SSPiRAL arrangement of degree 3 and x-order 2 would use no more than two nodes to build a parity node, and would need a set of six nodes to build a complete layout. Figure 1(a) shows a SSPiRAL layout of degree three and x-order two. Such a layout uses the same number of devices as a mirrored array of three striped disks, as shown in Figure 1(b) (we focus on six-disk arrangements in this abstract, but will expand our analysis to larger arrays in the final paper).

These nodes can be individual devices, servers, or storage arrays. SSPiRAL arrangements thereby distinguish between data and parity devices. As long as no devices have failed, the parity updates are efficient to compute, and SSPiRAL has performance comparable to purely striped RAID layouts such as RAID-0 arrays or striped storage clusters such as the original SWIFT distributed storage system [8]. In the example layout of Figure 1(a), data can be writ-

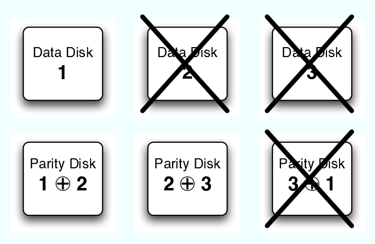


Figure 2: *SSPiRAL data layout and the loss of three nodes.*

ten across all three data blocks in parallel, increasing bandwidth, and parity nodes can almost always be calculated without requiring a read from an otherwise busy disk.

An interesting strength of a SSPiRAL layout can be demonstrated through Figure 2, which shows the loss of three of our six devices. In spite of this loss, it is possible to recover all lost data nodes. While a mirrored array *c*an survive the loss of three nodes, there are instances where it cannot survive the loss of two nodes (*e.g.*, it cannot survive the loss of any matched pair of mirrored nodes). There is *no* combination of two node losses that will cause the SSPiRAL layout in Figure 2 to lose data.

3 Reliability Analysis

In this section we evaluate the mean time to data loss (MTTDL) of a SSPiRAL disk array consisting of three data disks and three redundant disks and compare it with the respective MTTDLs of (a) a 3-out-of-6 disk array using an erasure code and (b) an array consisting of three pairs of mirrored disks. All three disk arrays consist of three data disks and three parity disks.

Our system model consists of a disk array with independent failure modes for each disk. When a disk fails, a repair process is immediately initiated for that disk. Should several disks fail, the repair process will be performed in parallel on those disks. We assume that disk failures are independent events exponentially distributed with rate λ, and that repairs are exponentially distributed with rate μ.

3.1 3+3 SSPiRAL array

Building an accurate state-transition diagram for a 3+3 SSPiRAL disk array is a task that exceeds the limitations of this paper as we have to distinguish between failures of data disks and failures of parity disks and consider the relations between each data disk and the two parity disks it shares with the two other data disks. Instead, we present here a simplified model.

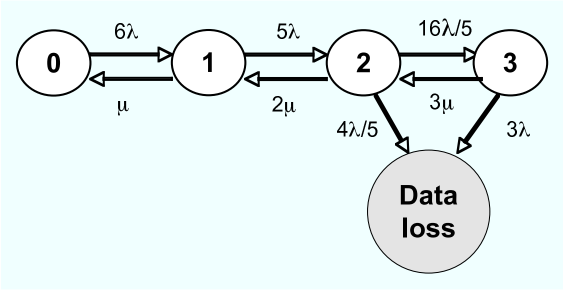


Figure 3: *3+3 disk SSPiRAL array.*

Observe first that the rate at which an array that has already two failed disks will experience a third disk failure is 4λ. Out of a total of 20 possible outcomes of this failure, only four will cause a data loss. These outcomes are

1. The failure of one data disk and its two parity disks

2. The failure of all three data disks

As a result, we will assume that the rate at which an array that has already two failed disks will incur a disk failure resulting in a data loss will be $4/20 \times 4\lambda = 4\lambda/5$ and the rate at which the same array will incur a disk failure resulting that will not affect the data will be is $16/20 \times 4\lambda = 16\lambda/5$

Figure 3 displays the simplified state transition probability diagram for a 3+3 SSPiRAL array. State $\langle 0 \rangle$ represents the normal state of the array when its six disks are all operational. A failure of any of these disks would bring the array to state $\langle 1 \rangle$. A failure of a second disk would bring the array into state $\langle 2 \rangle$. A failure of a third disk could either result in a data loss or bring the array to state $\langle 3 \rangle$. Any fourth disk failure will result in a data loss.

Repair transitions bring back the array from state $\langle 3 \rangle$ to state $\langle 2 \rangle$, then from state $\langle 2 \rangle$ to state $\langle 1 \rangle$ and, finally, from state $\langle 1 \rangle$ to state $\langle 0 \rangle$. Their rates are equal to the number of failed disks times the disk repair rate μ.

The Kolmogorov system of differential equations describing the behavior of the array is

$$
\begin{aligned}
\frac{dp_0(t)}{dt} &= -6\lambda p_0(t) + \mu p_1(t) \\
\frac{dp_1(t)}{dt} &= -(5\lambda + \mu) p_1(t) + 6\lambda p_0(t) + 2\mu p_2(t) \\
\frac{dp_2(t)}{dt} &= -(4\lambda + 2\mu) p_2(t) + 5\lambda p_1(t) + 3\mu p_3(t)
\end{aligned}
$$

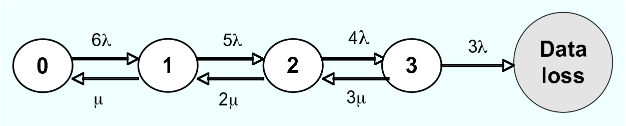


Figure 4: *3-out-of-6 array.*

$$\frac{dp_3(t)}{dt} = -(3\lambda + 3\mu)p_3(t) + \frac{16}{5}\lambda p_2(t)$$

where $p_i(t)$ is the probability that the system is in state $\langle i \rangle$ with the initial conditions $p_0(0) = 1$ and $p_i(0) = 0$ for $i \neq 0$.

The Laplace transforms of these equations are

$$\begin{aligned} sp_0^*(s) &= -6\lambda p_0^*(s) + \mu p_1^*(s) + 1 \\ sp_1^*(s) &= -(5\lambda + \mu)p_1^*(s) + 6\lambda p_0^*(s) + 2\mu p_2^*(s) \\ sp_2^*(s) &= -(4\lambda + 2\mu)p_2^*(s) + 5\lambda p_1^*(s) + 3\mu p_3^*(s) \\ sp_3(s) &= -(3\lambda + 3\mu)p_3^*(s) + \frac{16}{5}\lambda p_2^*(s) \end{aligned}$$

Observing that the mean time to data loss (MTTDL) of the array is given by

$$MTTDL = \sum_i p_i^*(0),$$

we solve the system of Laplace transforms for $s = 0$ and use this result to obtain the MTTDL of the array:

$$MTTDL = \frac{265\lambda^3 + 137\mu\lambda^2 + 37\mu^2\lambda + 5\mu^3}{60\lambda^3(5\lambda + \mu)}$$

3.2 3-out-of-6 array

Figure 4 displays the state transition probability diagram for a 3-out-of-6 disk array, that is, a disk array tolerating up to three simultaneous disk failures without data loss. State $\langle 0 \rangle$ represents the normal state of the array when its six disks are all operational. A failure of any of these disks would bring the array to state $\langle 1 \rangle$. A failure of a second disk would bring the array into state $\langle 2 \rangle$ and a failure of a third disk would always bring the array to state $\langle 3 \rangle$. A failure of fourth disk would result in a data loss. Repair transitions are identical to these of a 3+3 SSPiRAL array.

The Kolmogorov system of differential equations describing the behavior of the array is

$$\begin{aligned} \frac{dp_0(t)}{dt} &= -6\lambda p_0(t) + \mu p_1(t) \\ \frac{dp_1(t)}{dt} &= -(5\lambda + \mu)p_1(t) + 6\lambda p_0(t) + 2\mu p_2(t) \\ \frac{dp_2(t)}{dt} &= -(4\lambda + 2\mu)p_2(t) + 5\lambda p_1(t) + 3\mu p_3(t) \end{aligned}$$

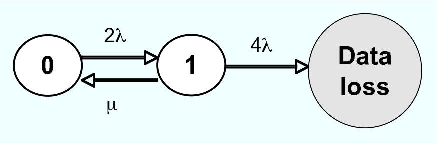


Figure 5: *Single pair of mirrored disks.*

$$\frac{dp_3(t)}{dt} = -(3\lambda + 3\mu)p_3(t) + 4\lambda p_2(t)$$

with the initial conditions $p_0(0) = 1$ and $p_i(0) = 0$ for $i \neq 0$.

Using the same techniques as in the previous case, we obtain the MTTDL of the array:

$$MTTDL = \frac{57\lambda^3 + 23\mu\lambda^2 + 7\mu^2\lambda + \mu^3}{60\lambda^4}$$

3.3 Three pairs of mirrored disks

Figure 5 displays the state transition probability diagram for a single pair of mirrored disks. State $\langle 0 \rangle$ represents the normal state of the array when its two disks are both operational. A failure of either of these disks would bring the array to state $\langle 1 \rangle$ and a failure of a second disk would result in a data loss. The sole repair transition is from state $\langle 1 \rangle$ to state $\langle 0 \rangle$

The two differential equations describing the behavior of the array are

$$\begin{aligned} \frac{dp_0(t)}{dt} &= -2\lambda p_0(t) + \mu p_1(t) \\ \frac{dp_1(t)}{dt} &= -(\lambda + \mu)p_1(t) + 2\lambda p_0(t) \end{aligned}$$

with the initial conditions $p_0(0) = 1$ and $p_1(0) = 0$.

Using the same techniques as in the two previous cases, we obtain the MTTDL of the mirrored pair:

$$MTTDL_{pair} = \frac{3\lambda + \mu}{2\lambda^2}$$

The MTTDL of an array consisting of three pairs of mirrored disks is then:

$$MTTDL = \frac{3\lambda + \mu}{6\lambda^2}$$

3.4 Results

Figure 6 displays on a logarithmic scale the MTTDLs provided by the three disk arrays. We assumed that the disk failure rate λ was one failure every one hundred thousand hours, that is, slightly less than one failure every eleven years. Disk repair times are expressed in days and MTTDLs expressed in years. As

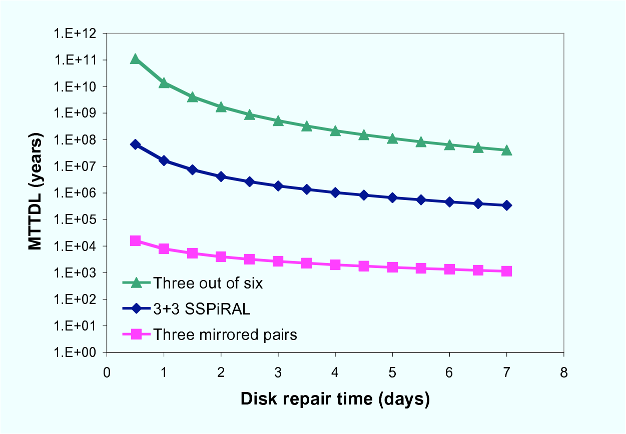


Figure 6: *Single pair of mirrored disks.*

we can see, the SSPiRAL disk array provides much better MTTDLs than the array consisting of three pairs of mirrored disks, but it falls below the MTTDL of the full 3-out-of-6 array.

4 Related Work & Conclusion

Like most of the original RAID layouts [5, 9], SSPiRAL is based solely on parity computations, and like more recent efforts [1, 7, 2, 4] SSPiRAL aims to survive the failure of multiple disks, and to achieve this goal efficiently. SSPiRAL diverges from prior efforts in its definition of efficiency. Unlike row-diagonal parity [4], SSPiRAL does not pursue the goal of optimizing capacity usage, and yet maintains the goals of optimal computational overhead and ease of management and extensibility. SSPiRAL replaces the goal of surviving a *specific* number of disk failures with the goal of surviving the most disk failures possible within the given resource constraints. The basic SSPiRAL layout discussed above can be described as an application of Systematic codes [10] across distinct storage devices. Similarly, such basic SSPiRAL layouts, in their limiting of the number of data sources, are similar to the fixed *in-degree* and *out-degree* parameters in Weaver codes [6] and the earlier $\hat{\text{B}}$ layouts [11]

The analytical results we present in this abstract demonstrate how a basic SSPiRAL array defined across six disks, and using simple pairwise parity, achieves an MTTDL superior to the mirroring of pairs of disks. This SSPiRAL layout offers lower MTTDLs than a complete three-out-of-six erasure code, but depends solely on the simplest pairwise parity computations, and still manages to offer a higher MTTDL than any scheme capable of surviving the loss of two data disks[1] (as it can survive many three-disk failures).

References

[1] G. A. Alvarez, W. A. Burkhard, and F. Cristian, "Tolerating multiple failures in RAID architectures with optimal storage and uniform declustering," in *Proceedings of the 24th ISCA*, (Denver, CO), pp. 62–72, ACM, 1997.

[2] M. Blaum, J. Brady, J. Bruck, and J. Menon, "Evenodd: An efficient scheme for tolerating double disk failures in raid architectures," *IEEE Trans. Comput.*, vol. 44, no. 2, pp. 192–202, 1995.

[3] V. Ciotola, J. Larkby-Lahet, and A. Amer, "SSPiRAL layouts: Practical extreme reliability," Tech. Rep. TR-07-149, Department of Computer Science, University of Pittsburgh, 2007. To be presented at the Usenix Annual Technical Conference 2007 poster session.

[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 FAST)*, (Berkeley, CA, USA), pp. 1–14, USENIX Association, 2004.

[5] G. A. Gibson, *Redundant Disk Arrays: Reliable, Parallel Secondary Storage.* PhD thesis, University of California at Berkeley, 1990.

[6] J. L. Hafner, "Weaver codes: Highly fault tolerant erasure codes for storage systems," in *Proceedings of FAST*, (San Francisco, CA), dec 2005.

[7] K. Hwang, H. Jin, and R. Ho, "RAID-x: A new distributed disk array for I/O-centric cluster computing," in *Proceedings of the 9th IEEE HPDC Symposium*, pp. 279–286, 2000.

[8] D. D. E. Long, B. R. Montague, and L.-F. Cabrera, "Swift/RAID: A distributed RAID system," *Computing Systems*, vol. 7, no. 3, pp. 333–359, 1994.

[9] D. A. Patterson, G. Gibson, and R. H. Katz, "A case for redundant arrays of inexpensive disks (RAID)," in *Proceedings of SIGMOD*, pp. 109–116, ACM, 1988.

[10] J. S. Plank and M. G. Thomason, "A practical analysis of low-density parity-check erasure codes for wide-area storage applications," in *Proceedings of DSN*, (Florence, Italy), June 2004.

[11] B. T. Theodorides and W. A. Burkhard, "$\hat{\text{B}}$: Disk array data layout tolerating multiple failures," in *Proceedings of MASCOTS)*, (Monterey, CA), pp. 21–32, IEEE, 2006.

[1] These results are not discussed in this abstract due to space constraints.