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 lay-out 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.

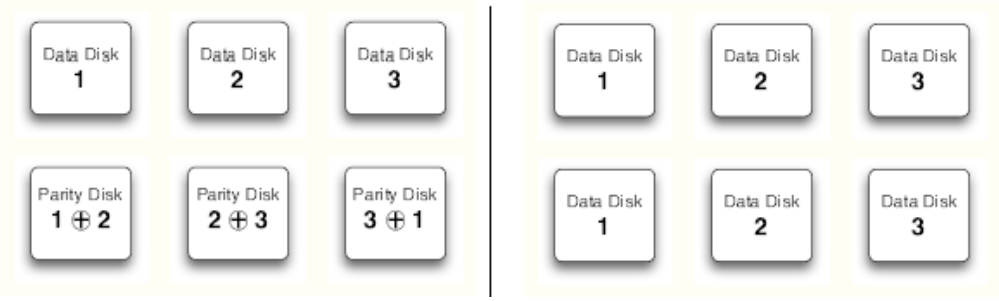

Figure 1: 3+3 SSPiRAL Layout versus 3+3 Mirrored Disk

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 egree 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). 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 written 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 can 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.

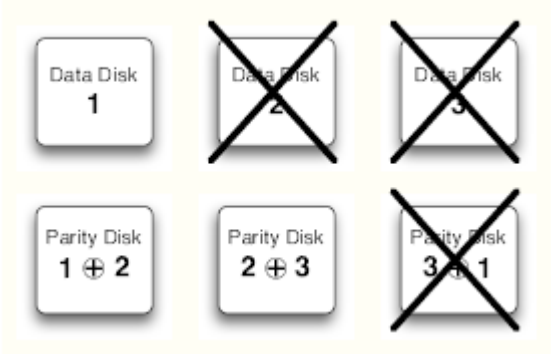

Figure 2: SSPiRAL data layout and the loss of three nodes

3 Reliability Analysis

We now evaluate the reliability of small SSpiral layouts and compare them to mirrored disks (RAID level 1) as well as RAID level 5 and RAID level 10. We model only hardware failures and we assume that disk scrubbing is used to prevent any data loss due to bit rot or any other type of data loss limited to a block. We measure system reliability in terms of data loss, using the two criteria of Mean Time To Data Loss (MTTDL) and loss expectancy at the end of the economic lifespan of the devices, which we set somewhat arbitrarily to five years.

3.1 SSPiRAL Arrays

Building a completely accurate state-transition diagram for a SSPiRAL array exceeds the limitations of this paper as we would have to distinguish between failures of data disks and failures of parity disks. These distinctions are necessary for complete accuracy since complexity of recalculating data previously stored in a lost drive differs. Instead, we abstract from these details and aggregate states as much as possible. We capture a system with i failed disks in a state S_i. We have a repair transition from State S_i to State S_{i-1}, $i \geq 1$, which is taken with rate $i \cdot \mu$, where μ is the inverse of the average repair time. Thus,

our model assumes independent repairs of any failed devices. Experiences with similar Markov models show that the repair distribution has much less influence on MTTDL than the average repair time. The repair time itself is composed of the time to detection, issue of the service call and wait for the replacement of the failed device followed by the reconstruction of the data previously stored in the failed disk. The latter component is typically several hours since for example a 1 TB disk is fully read at 10MB/sec (about 10% to 20% of the capacity of an idle device) in ~ 28 hours. It increases proportionally with the size of the disk and decreases inversely proportionally with the read/write rate. We have also failure transitions that leave State S_i with a combined rate of $(N-i)\lambda$. Partially or totally, a failure transition leads to the next State S_{i+1}, complemented by a transition to the Failure State, which is absorbing.

Our model is limited by the Markovian assumption of independent repairs and failures and by the modeling of repair in particular. Nevertheless, models of this type have been confirmed by simulation to be reasonably accurate.

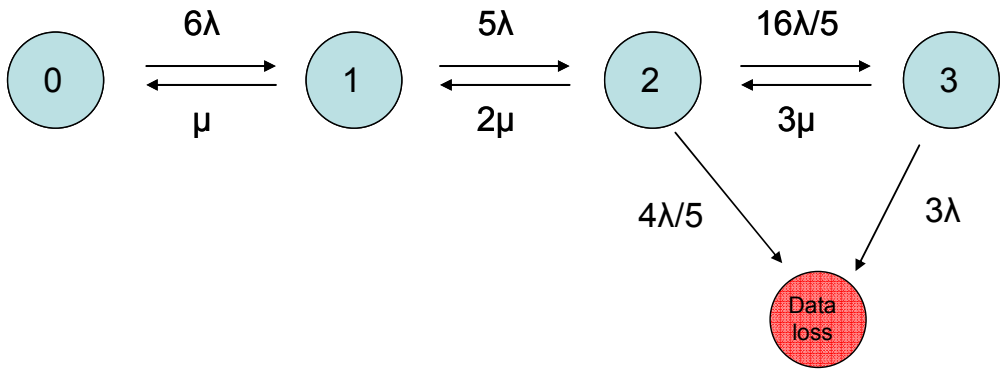

Figure 3: 3+3 SSPiRAL Markov Model

3.1.1 The 3+3 SSPiRAL Array

The 3+3 SSPiRAL array has an x-value of two and encompasses six disks. Its layout is given in Figure 2. Clearly, loss of four disks (or more) has to lead to data loss. (We are of course assuming that all data disks indeed contain data.) A case by case distinction shows that there is never data loss if any two disks have failed and that in 4 out of the $\binom{6}{3}$=20 ways in which three out of six disks can fail data loss occurs. In more detail, data loss occurs if

(1) all data disks have failed

(2) a data disks and the two parity devices containing its data have failed.

As a result, we have a state transition from S_3 to S_4 taken with rate $\frac{16}{20}\cdot 4\cdot\lambda = \frac{4}{5}\cdot\lambda$ and a transition for S_3 to the absorbing (data loss) state with rate $\frac{4}{20}\cdot 4\cdot\lambda = \frac{4}{5}\lambda$. The complete Markov model is in Figure 3. We write $p_i(t)$ for the probability that the system is in State

S_i at time t and obtain the corresponding *Kolmogorov* system of linear differential equations as

$$\frac{dp_0}{dt} = -6\lambda p_0(t) + \mu p_1(t)$$
$$\frac{dp_1}{dt} = 6\lambda p_0(t) - (5\lambda + \mu) p_1(t) + 2\mu p_2(t)$$
$$\frac{dp_2}{dt} = 5\lambda p_1(t) - (4\lambda + 2\mu) p_2(t) + 3\mu p_3(t)$$
$$\frac{dp_3}{dt} = 16/5\, \lambda p_2(t) - (3\lambda + 3\mu) p_3(t)$$

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

Table 1: Data Loss Probability with various disk MTBF 1/λ and average repair time 1/μ for the 3+3 SSPiRAL array.

1/λ	1/μ	4 year	5 year	20 years	100 years
50000	30	3.02E-06	3.78E-06	1.51E-05	7.56E-05
100000	30	3.78E-07	4.73E-07	1.89E-06	9.46E-06
1000000	30	3.78E-10	4.73E-10	1.89E-09	9.47E-09
50000	100	3.34E-05	4.17E-05	1.67E-04	8.37E-04
100000	100	4.18E-06	5.23E-06	2.10E-05	1.05E-04
1000000	100	4.19E-09	5.24E-09	2.10E-08	1.05E-07

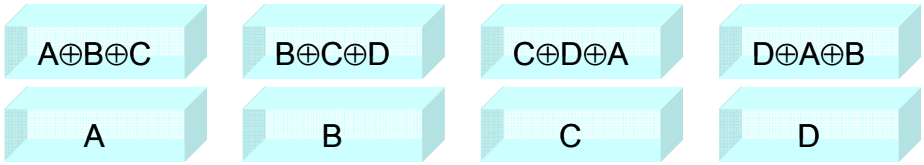

Figure 4: 4+4 SSPiRAL layout with 8 disks and $x = 3$.

3.1.2 The 4+4 SSPiRAL Array with $x = 3$

We present the layout of the 4+4 SSPiRAL array with x-value 3 in Figure 4. The modeling of the 4+4 array is quite similar to the previous one. A case-by-case enumeration shows that there is no data loss if up to three disks have failed. An information theoretical argument shows that loss of five disks needs to lead to data loss. We consider the remaining case (failure of four disks) in more detail. We make a case distinction according to the number of data disks.

No lost data disks: Data loss is then impossible.

One lost data disk: Assume that data disk A (see Figure 4) has failed. Three parity drives have also failed and one remains available. If this one is A⊕B⊕C, C⊕D⊕A, or D⊕A⊕B, then we can reconstruct the data previously in A. In the remaining case, all disks with contents reflecting A are lost and data loss is inevitable. Hence, we have data loss in 4 of the 16 cases where one data disk is lost.

Two lost data disks: First, we assume that two neighboring data disks in Figure 4 are unavailable. Let these be A and B. Two of the parity drives are also available. If B⊕C⊕D or C⊕D⊕A are among them, then we achieve directly the contents of B and A, respectively. In the remaining case, C, D, A⊕B⊕C, and D⊕A⊕B are available. Since any reconstruction has to use XORing as a primitive operation and since C, D, A⊕B⊕C, D⊕A⊕B, C⊕D are the elements of a set closed under XORing, the contents of A and B remain unavailable. Of the 4*6 subcases, four lead to data loss. Second, we assume that two non-neighbors in Figure 4 are available. Let these be A and C. If B⊕C⊕D or D⊕A⊕B are available, we obtain with B and D directly C or A respectively and hence A and C from the other parity drive contents. This leaves the case where B, D, A⊕B⊕C, and D⊕A⊕B are available. Taking XORs of these four available objects, we obtain a set that additionally contains B⊕D and A⊕C, but that is closed under further taking of pair-wise parity. Hence, of the 2*6 subcases, two lead to data loss.

Three lost data disks: Assume that disk A is available. We obtain the remaining data disks contents in three of the four cases as follows:

- Available are A, A⊕B⊕C, B⊕C⊕D, and C⊕D⊕A. Then also B = B⊕C⊕D ⊕ A⊕B⊕C ⊕ A, C = A⊕B⊕C ⊕ B⊕C⊕D ⊕ C⊕D⊕A, D = B⊕C⊕D ⊕ A⊕B⊕C ⊕ A.
- Available are A, A⊕B⊕C, B⊕C⊕D and D⊕A⊕B. Then B = B⊕C⊕D ⊕ C⊕D⊕A ⊕ A, C = A⊕B⊕C ⊕ B⊕C⊕D ⊕ D⊕A⊕B, and D = A⊕B⊕C ⊕ B⊕C⊕D ⊕ A.
- Available are A, B⊕C⊕D, C⊕D⊕A, and D⊕A⊕B. Then B = C⊕D⊕A ⊕ B⊕C⊕D ⊕ A, C = B⊕C⊕D ⊕ D⊕A⊕B ⊕ A, and D = B⊕C⊕D ⊕ C⊕D⊕A ⊕ D⊕A⊕B.

In the remaining case, A, A⊕B⊕C, C⊕D⊕A, and D⊕A⊕B are available. By taking all possible pair-wise parity, we obtain the set S = {A, A⊕B⊕C, C⊕D⊕A, D⊕A⊕B, B⊕C, C⊕D, D⊕B} which is closed under this operation. Hence, in this case the array suffers data loss. In toto, of the 4*4 = 16 subcases, 4 lead to data loss.

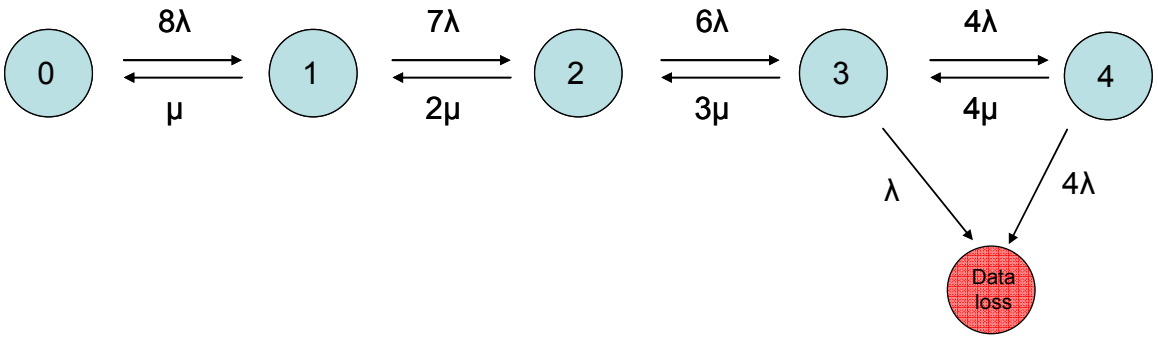

Figure 5: 4+4 SSPiRAL Markov Model

Four lost data disks: Since A = A⊕B⊕C ⊕ C⊕D⊕A ⊕ D⊕A⊕B, B = D⊕A⊕B ⊕ A⊕B⊕C ⊕ B⊕C⊕D, C = A⊕B⊕C ⊕ B⊕C⊕D ⊕ C⊕D⊕A, D = B⊕C⊕D ⊕ C⊕D⊕A ⊕ D⊕A⊕B, there is no data loss.

To summarize, out of a total of $\binom{8}{4} = 70$ ways for four out of eight disks to fail, 14 lead to data loss. We can now use our insight to calculate the Markov model given in Figure 5. There is a combined rate of 5λ of failure transitions out of State S_3. The rate of the transition from S_3 to the absorbing state is (14/70)·5λ = λ and of the transition from State S_3 to State S_4 is (56/70)·5λ = 4λ.

The resulting Kolmogorov system is

$$\frac{dp_0}{dt} = -8\lambda p_0(t) + \mu p_1(t)$$
$$\frac{dp_1}{dt} = 8\lambda p_0(t) - (7\lambda + \mu) p_1(t) + 2\mu p_2(t)$$
$$\frac{dp_2}{dt} = 7\lambda p_1(t) - (6\lambda + 2\mu) p_2(t) + 3\mu p_3(t)$$
$$\frac{dp_3}{dt} = 6\lambda p_2(t) - (5\lambda + 3\mu) p_3(t) + 4\mu p_4(t)$$
$$\frac{dp_4}{dt} = 4\lambda p_3(t) - (4\lambda + 4\mu) p_4(t)$$
$$p_0(0) = 1, p_1(0) = p_2(0) = p_3(0) = p_4(0) = 0$$

Its solution requires solving polynomial equations of degree 5. Therefore, we cannot give a general closed form solution other than for specific values of λ and μ. They nevertheless allow us to calculate the four and five year probabilities of data loss. We also calculated the reliability for much larger time spans and tabulate the results in Table 2.

Table 2: Data Loss Probability with various disk MTBF 1/λ and average repair time 1/μ for the 4+4 SSPiRAL array.

1/λ	1/μ	4 year	5 year	20 years	100 years
50000	30	8.45E-09	1.06E-08	4.23E-08	2.12E-07
100000	30	5.29E-10	6.61E-10	2.65E-09	1.32E-08
1000000	30	5.28E-14	6.63E-14	2.65E-13	1.33E-12
50000	100	3.10E-07	3.88E-07	1.56E-06	7.78E-06
100000	100	1.94E-08	2.43E-08	9.77E-08	4.89E-07
1000000	100	1.95E-12	2.44E-12	9.80E-12	4.91E-11

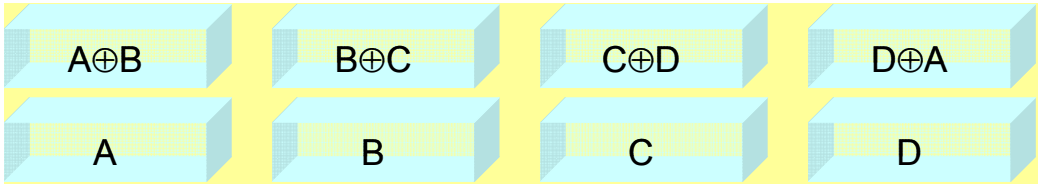

Figure 6: 4+4 SSPiRAL layout with $x = 2$.

3.1.3 The 4+4 SSPiRAL Array with $x = 2$

For direct comparison purposes, we also consider the SSPiRAL array with $x = 2$ and eight disks (Figure 6). To study its data survival, we use a technique similar to that developed by Hellerstein et al. [H+92] and later expended by [J+03] that is based on interpreting a 2-failure correcting layouts using parity calculations as a type of mathematical design called configuration (see [Gr96]). The dual is then a regular graph. In this representation, vertices are parity disks and edges are data disks. An edge is connected to a graph if the corresponding data disk contributed to the parity. The result of this representation is in Figure 7

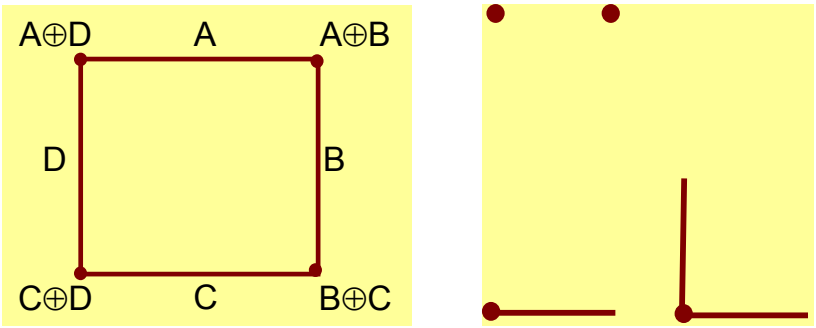

Figure 7: Graph Representation of the 4+4 SSPiRAL layout with $x = 2$ (left), two neighbor, club, and two neighbor double club patterns (right)

We now model the loss of one or more devices as a subgraph of this graph. We can reconstruct lost data by using parity calculations, which we can represent by the following graph theoretical operations. Given an edge and an adjacent vertex, we can reconstruct the other vertex. For example, given A and A⊕B, we can reconstruct B. Given two adjacent edges, we can reconstruct the vertex at the intersection between these edges. For example, given A and B, we can reconstruct A⊕B. Operations that are more complicated do not reconstruct additional data. As a consequence of this insight, a loss pattern leading to data loss must contain a cycle of adjacent edges (in our case the loss of A, B, C, and D) or a path consisting of edges and vertices that starts and ends in a vertex. For example, A⊕B, A, A⊕D is a (minimal) loss pattern with data loss. We can now calculate the probabilities that a failure leads to data loss.

We use the now standard notations for the Markov model. In addition to the absorbing, data loss state, we have states S_i representing the system when i disks have failed.

If there are none, one, or two failures, no data loss occurs. The three loss patterns with data loss consist of two edges with the connecting edge. Hence, the chances are 4/56 = 1/14. Since the total rate of failure transitions out of S_2 is 6λ, the system transitions from S_2 at a rate $3\lambda/7$ to the data loss state and at rate $39\lambda/7$ to S_3.

We now calculate the data loss probability for the fourth failure. Assume now that we are in one of the 52 cases that does not represent data loss after failure of three devices. The failure of an additional device only results in data loss if it creates either a pattern vertex-edge-vertex or a cycle edge-edge-edge-edge. 4 out of the 52 cases consist of three edges. In this case, failure of the data disk representing the other edge leads to data loss. This

happens with probability 1/5. Some of the 52 cases representing failure of three devices contain a "single club" pattern formed by a vertex, an adjoining edge, but not the other adjoining edge. We can pick the vertex in four different ways and then one of the edges. The other failed device must not be the "other" edge, hence we can pick this in four different ways. This gives us 32 possibilities. The "double club" consists of a vertex and two adjoining edges, for this, we have four possibilities. We also have the "two neighbor" pattern, consisting of two adjacent vertices with the other failure represented not by the combining edge. There are 8 cases that contain both a club and a two neighbor pattern. There are 8 cases of a two neighbor pattern that does not contain a club nor the three failure pattern. Of these, 4 are made up of three vertices.

We now calculate the probability that an additional failure leads to data loss. In the 24 cases of a single club without a two neighbors pattern, only failure of the data disk represented by the other vertex of the club's edge leads to data loss. In the 4 cases of the three vertices, three out of the five possibilities for an additional failure leads to data loss. In the 4 cases of the two neighbors without club and three vertices (e.g. pattern A⊕D, C, A⊕B, the "face", possibly rotated) , only 1 out of the 5 possibilities for further failure leads to data loss. In the 8 cases of a club with two neighbors, two out of the five possibilities for further device failure lead to data loss. In the 4 cases of the double club, two out of the five possibilities lead to further failure.

In total, the fourth failure induces data loss with probability

$$\frac{24\cdot 1+4\cdot 3+4\cdot 1+8\cdot 2+4\cdot 2}{52\cdot 5}=\frac{16}{65}.$$

Since the total rate of failure transitions out of S_3 is 5λ, the transition rate from S_3 to the absorbing state is 5λ·64/(52·5) = 16λ/13 and from S_3 to S_4 49λ/13. We give the Markov model in Figure 8.

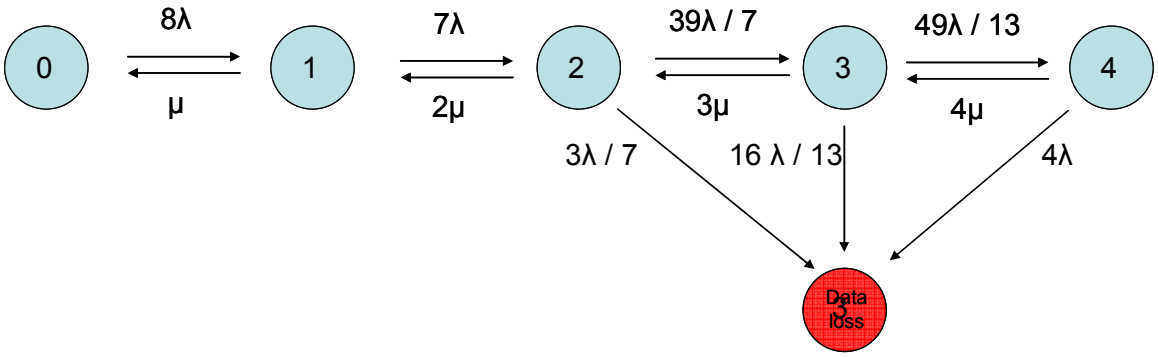

Figure 8: Markov model for the 4+4 SSPiRAL with x = 2.

The Kolmogorov system is

$$\frac{dp_0}{dt} = -8\lambda p_0(t) + \mu p_1(t)$$
$$\frac{dp_1}{dt} = 8\lambda p_0(t) - (7\lambda + \mu) p_1(t) + 2\mu p_2(t)$$
$$\frac{dp_2}{dt} = 7\lambda p_1(t) - (6\lambda + 2\mu) p_2(t) + 3\mu p_3(t)$$
$$\frac{dp_3}{dt} = \frac{39}{7}\lambda p_2(t) - (5\lambda + 3\mu) p_3(t) + 4\mu p_4(t)$$
$$\frac{dp_4}{dt} = \frac{49}{13}\lambda p_3(t) - (4\lambda + 4\mu) p_4(t)$$
$$p_0(0) = 1, p_1(0) = p_2(0) = p_3(0) = p_4(0) = 0$$

Table 3: Survival rate for the SSPiRAL array with $x = 2$ and 8 disks

1/λ	1/μ	4 year	5 year	20 years	100 years
50000	30	3.02E-06	3.78E-06	1.51E-05	7.56E-05
100000	30	3.78E-07	4.72E-07	1.89E-06	9.46E-06
1000000	30	3.78E-10	4.73E-10	1.89E-09	9.47E-09
50000	100	3.33E-05	4.17E-05	1.67E-04	8.36E-04
100000	100	4.18E-06	5.23E-06	2.10E-05	1.05E-04
1000000	100	4.19E-09	5.24E-09	2.10E-08	1.05E-07

A closed form solution involves again solving polynomial equations of degree 5 and is too unyielding to be presented here. We give the survival rates after 4 and 5 years in Table 1. Compared with the SSPiRAL array also with 8 disks but with $x = 3$, the numbers are worse by about three powers of ten. We can guess this behavior already from the Markov model.

3.2 Mirrored Layout

Mirroring is functionally the simplest way to induce redundancy. Two copies are written, but either copy can satisfy reads. We can restore its contents by simply accessing the other copy.

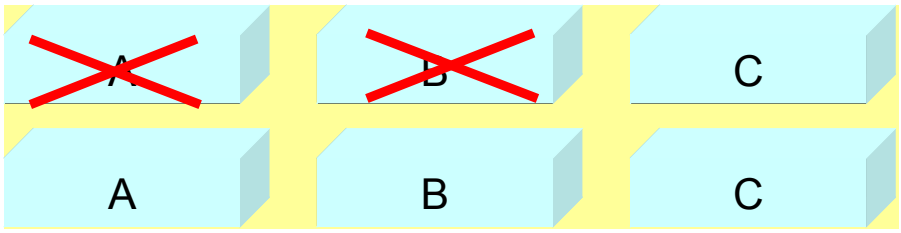

Figure 9: 3+3 Mirrored Layout after loss of two drives without data loss.

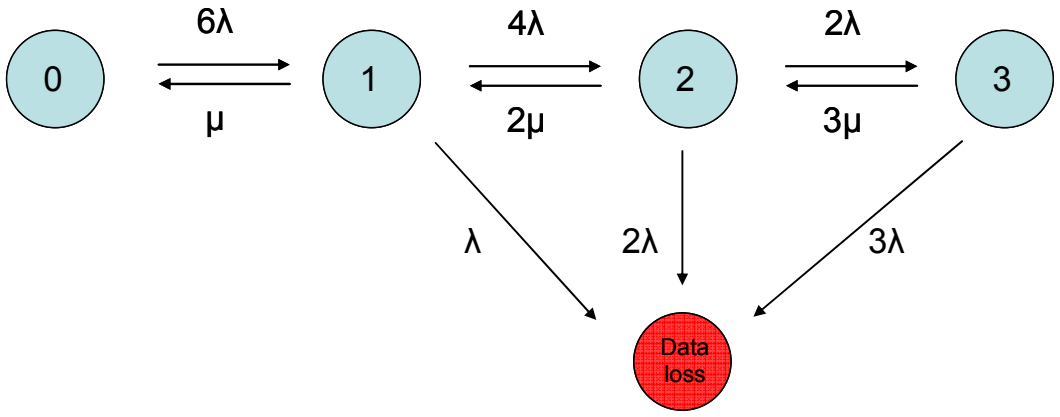

Figure 10: Markov model for the 3+3 mirrored layout

3.2.1 The 3+3 Mirrored Layout

We present the 3+3 mirrored layout in Figure 1. The array can tolerate any loss of a single drive and definitely any loss of four drives leads to data loss. However, two drives containing the same data can lead to data loss, this happens with probability 1/5 after loss of a single drive. (There are 5 drives left and loss of the one containing the same data as the already failed drive leads to data loss.) If the array has tolerated two failures without data loss, it is (modulo renaming of disks) in the situation depicted in Figure 9. The chance that an additional loss looses access to the data in A or B is 1/2. As a result, we obtain the Markov model depicted in Figure 10.

The Kolmogorov system is

$$\frac{dp_0}{dt} = -6\lambda p_0(t) + \mu p_1(t)$$
$$\frac{dp_1}{dt} = 6\lambda p_0(t) - (5\lambda + \mu) p_1(t) + 2\mu p_2(t)$$
$$\frac{dp_2}{dt} = 4\lambda p_1(t) - (4\lambda + 2\mu) p_2(t) + 3\mu p_3(t)$$
$$\frac{dp_3}{dt} = 2\lambda p_2(t) - (3\lambda + 3\mu) p_3(t)$$
$$p_0(0) = 1, p_1(0) = p_2(0) = p_3(0) = 0$$

While it is possible to give the closed form of the solution, it is really to involved to find place in this article. Nevertheless, it enables us to calculate the 4 and 5 year data loss rate of such a system. We present the results in Table 4.

Table 4: Survival rate of the mirrored array with 6 disks

1/λ	1/μ	4 year	5 year	20 years	100 years
50000	30	2.51E-03	3.14E-03	1.25E-02	6.11E-02
100000	30	6.30E-04	7.87E-04	3.15E-03	1.56E-02
1000000	30	6.31E-06	7.88E-06	3.15E-05	1.58E-04
50000	100	8.31E-03	1.04E-02	4.09E-02	1.89E-01
100000	100	2.09E-03	2.61E-03	1.04E-02	5.11E-02

1000000	100	2.10E-05	2.62E-05	1.05E-04	5.26E-04

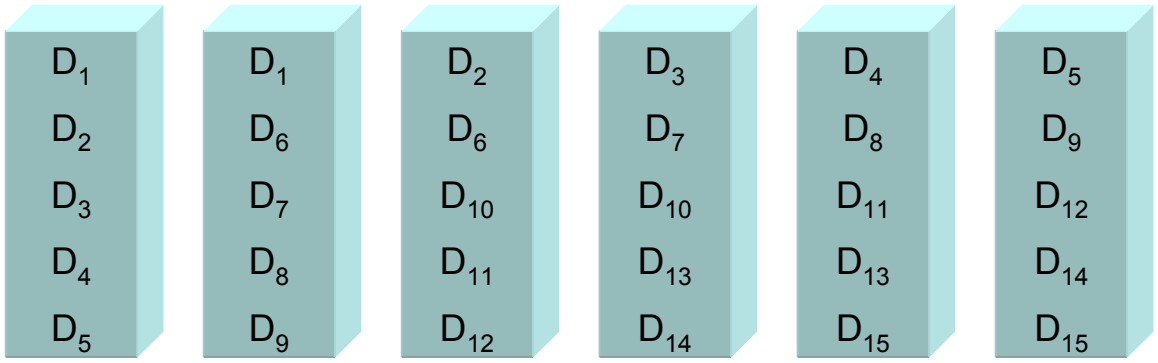

Figure 11: 3+3 Declustered Mirrored Array

3.2.2 The 3+3 Declustered Layout

Fast repair times not only contribute to reliability but also minimize the time that a disk array spends in degraded mode. *Declustering* equalizes the relationships between disks due to shared data and distributes the load increases evenly. At the same time, it increases disk performance in the degraded mode and shortens repair times. We decluster a small array with *n* mirrored disks by dividing the contents of each disk into *n*-1 blocks. Each block is stored in exactly two different disks. Figure 11 gives the data layout for a declustered mirrored array with six disks. Such an array always survives one disk failure and always suffers data loss with the failure of a second device. On the other hand, reconstruction is now distributed over all survivors of a single failure and therefore up to *n*-1 faster than a non-declustered layout.

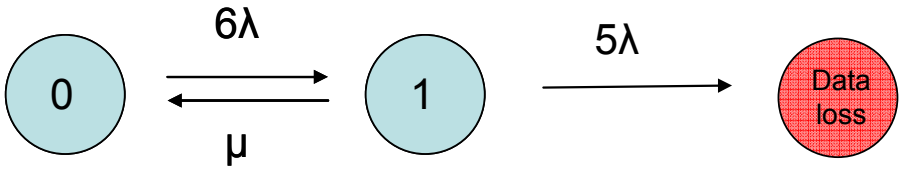

Figure 12: Markov model for the declustered mirrored array with six disks

The resulting Markov model is simpler than previous ones and given in Figure 6 for six disks. It contains only two non-failure states, State 0, in which all devices are functioning and the degraded State 1, where a single disk has failed. We have three state transitions, a failure transition with rate 6λ from State 0 to State 1, a second failure transition from State 1 to the absorbing state with rate 5λ, and a repair transition from State 1 to State 0 taken with rate μ. The resulting *Kolmogorov* system of differential equations has the form

$$\frac{dp_0}{dt} = -6\lambda p_0(t) + \mu p_1(t)$$

$$\frac{dp_1}{dt} = 6\lambda p_0(t) - (5\lambda + \mu) p_1(t)$$

with initial condition $p_0(t) = 1$, $p_1(t) = 1$.

This equation has the explicit solution

$$p_0(t) = \frac{1}{2\mathrm{D}}\left(\begin{array}{l}\left(\exp(\frac{1}{2}t(-11\lambda-\mu-D)-\exp(\frac{1}{2}t(-11\lambda-\mu+D)\right)(\lambda-\mu)\\ +\left(\exp(\frac{1}{2}t(-11\lambda-\mu-D)+\exp(\frac{1}{2}t(-11\lambda-\mu+D)\right)D\end{array}\right)$$

$$p_1(t) = \frac{6\lambda\left(\exp(\frac{1}{2}t(-11\lambda-\mu+D)-\exp(\frac{1}{2}t(-11\lambda-\mu-D)\right)}{D}$$

where

$$D = \sqrt{\lambda^2 + 22\lambda\mu + \mu^2}\ .$$

We can use this expression to calculate the probability for data loss after a certain period. We tabulate the results in Table 5. Comparing the results is difficult since declustering speeds up the process of copying data previously stored in the lost drive by a factor of 5. However, the same speed-up does of course not pertain to failure discovery and the initiation of the recovery. Therefore, we include additional repair times in our scheme. However, the numbers are slightly better than for the undeclustered, mirrored array.

Table 5: Data Loss Probability of the declustered mirrored array with six disks

1/λ	1/μ	4 year	5 year	20 years	100 years
50000	5	2.10E-03	2.62E-03	1.05E-02	5.12E-02
100000	5	5.25E-04	6.57E-04	2.62E-03	1.31E-02
1000000	5	5.26E-06	6.57E-06	2.63E-05	1.31E-04
50000	20	8.34E-03	1.04E-02	4.10E-02	1.89E-01
100000	20	2.10E-03	2.62E-03	1.04E-02	5.11E-02
1000000	20	2.10E-05	2.63E-05	1.05E-04	5.26E-04
50000	30	1.25E-02	1.55E-02	6.08E-02	2.69E-01
100000	30	3.14E-03	3.92E-03	1.56E-02	7.56E-02
1000000	30	3.15E-05	3.94E-05	1.58E-04	7.88E-04
50000	100	4.02E-02	5.01E-02	1.86E-01	6.43E-01
100000	100	1.03E-02	1.29E-02	5.07E-02	2.29E-01
1000000	100	1.05E-04	1.31E-04	5.25E-04	2.62E-03

3.2.3 The Mirrored Array with 8 disks

The derivation of the Markov model and the Kolmogorov system proceeds in strict analogy to the case of 6 disks. We give the Markov model in Figure 13 and the data loss probability during the economic lifespan of the array in Table 6.

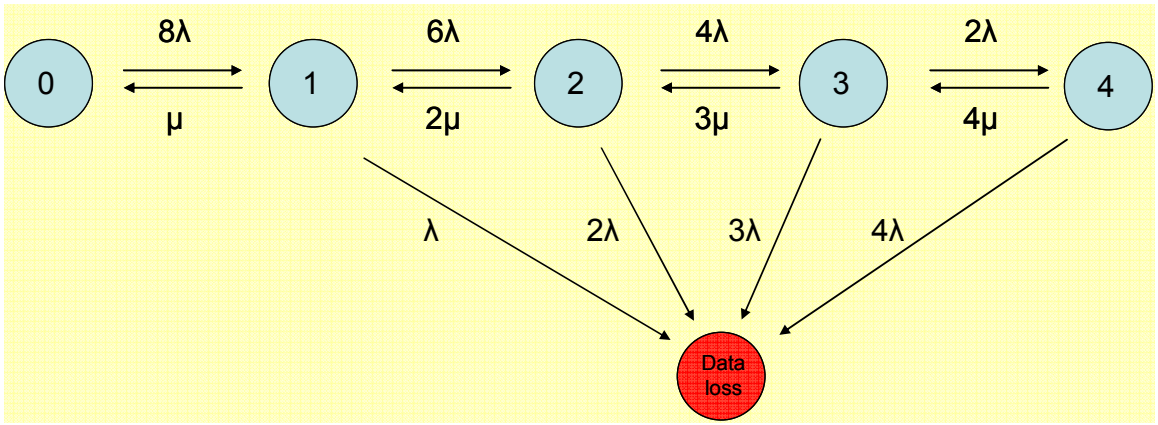

Figure 13: Markov model for the mirrored array with eight disks

Table 6: Data Loss Probability of the mirrored array with eight disks

1/λ	1/μ	4 year	5 year	20 years	100 years
50000	30	3.35E-03	4.19E-03	1.67E-02	8.06E-02
100000	30	8.40E-04	1.05E-03	4.19E-03	2.08E-02
1000000	30	8.41E-06	1.05E-05	4.21E-05	2.10E-04
50000	100	1.11E-02	1.38E-02	5.42E-02	2.43E-01
100000	100	2.78E-03	3.48E-03	1.39E-02	6.75E-02
1000000	100	2.80E-05	3.50E-05	1.40E-04	7.01E-04

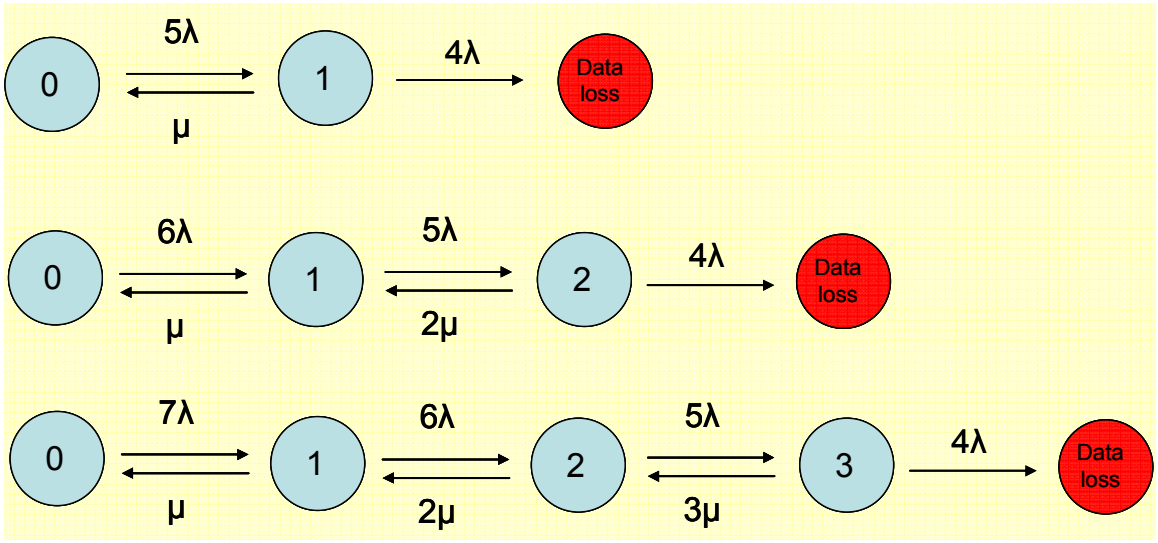

Figure 14: Markov models for the 4+*k* Array, *k* = 1, 2, 3

3.3 4+k Array

For comparison purposes, we consider a disk array with 4 data disks and one (Level 5), two (Level 6), three (Level 6') parity disks. An array made up of *m* data disks and *k* parity disks can survive exactly *k* failures and will always suffer data loss with the k+1st failure. There are a variety of ways to calculate the contents of the parity disks. Litwin *et al.* use Reed Solomon codes, but one where the contents of the first parity drive are the

normal, XOR parity. Some proposals calculate the contents of two parity drives using only XORing [BB+95, CE04]. Other proposals are applicable to much larger systems than the ones that we consider here [Ha05, TB06] or add more than k parity drives to obtain k failure tolerance [PT04]. However, the Galois field calculations necessary to calculate parity using a generalized Reed Solomon code have higher bandwidth than disk accesses and present no hindrance to this implementation. We give the Markov models in **Error! Reference source not found.**

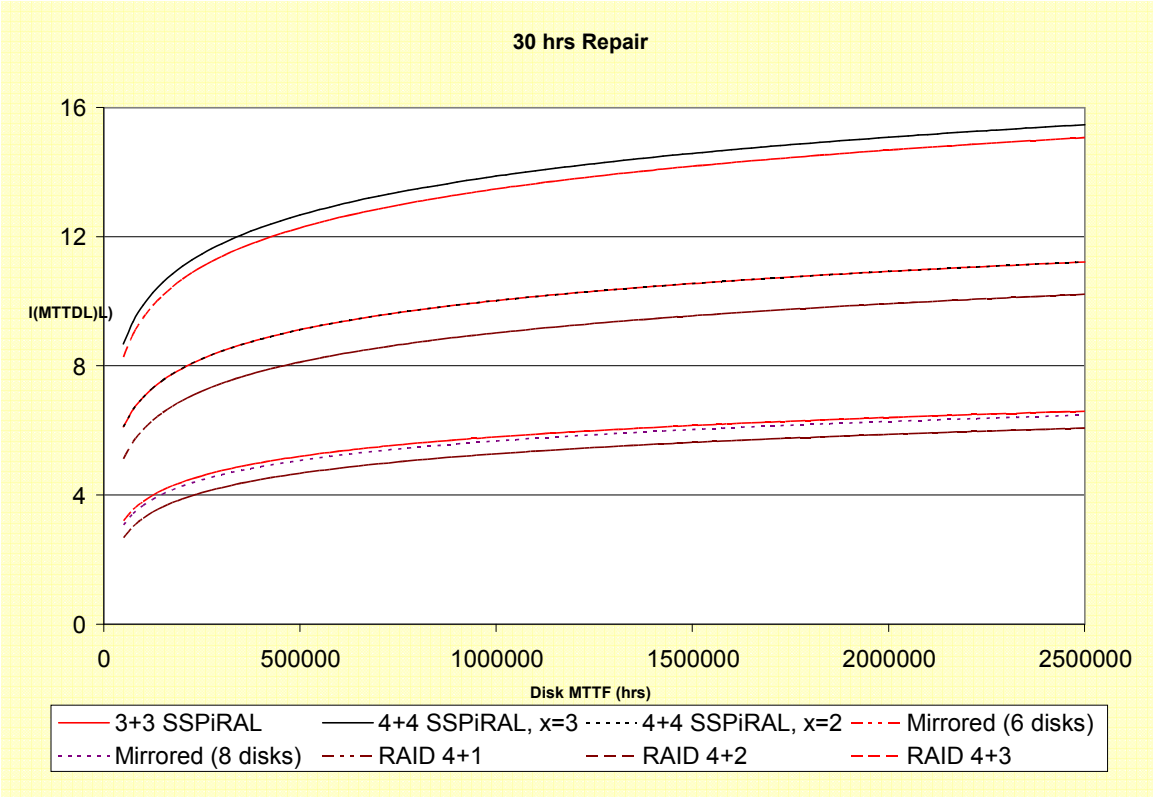

Figure 15: $\log_{10}$(MTTDL) in years for various arrays. The curves for the 3+3 SSPiRAL, x =2, and the 4+4 SSPiRAL, x=2, are almost identical. Average repair time is 30 hrs.

4 Mean Time to Data Loss Comparisons

Traditionally, the reliability of disk arrays has been measured in the mean time to data loss. Unfortunately, calculating the failure rate from the MTTDL is erroneous since the survival of a Markovian system with absorbing state is no longer exponentially distributed. In fact, using MTTDL to provision an array with specified data loss probability during its economic lifespan can lead to large overprovisioning of the array. Nevertheless, MTTDL is a single figure of reliability and we therefore compare the MTTDL for all but the declustered mirrored array, which we leave out since assigning a comparable average repair time for it is difficult and probably misleading.

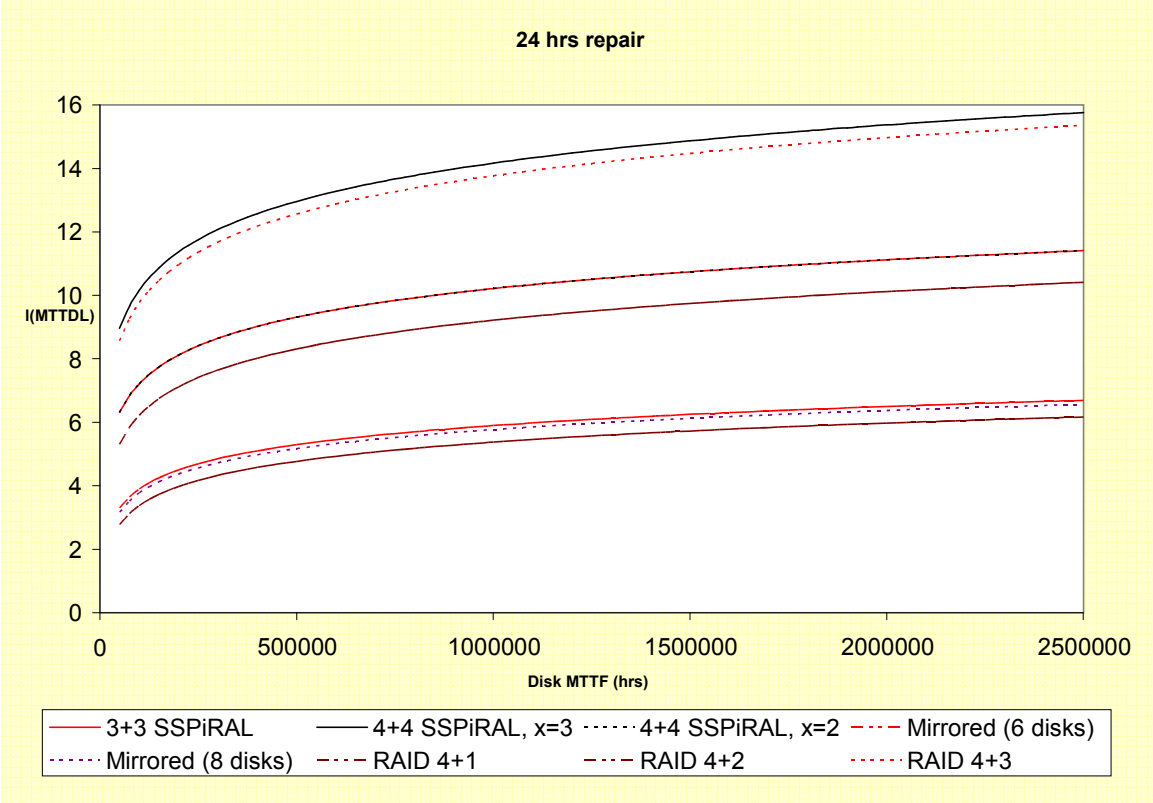

Figure 16: $\log_{10}$(MTTDL) in years for various arrays. The curves for the 3+3 SSPiRAL, x =2, and the 4+4 SSPiRAL, x=2, are almost identical. Average repair time is 24 hrs.

We give the results of our MTTDL calculations in Figure 15, Figure 16, and Figure 17. There, the x-axis gives the MTTF of the individual device, ranging from the unfortunately sometimes-realistic 50,000 hrs over the typical specification of over 1,000,000 hrs to 2,500,000 hrs, a value that can only be achieved under ideal circumstances, as far as we know. The y-axis gives the decadic logarithm of the MTTF in years.

The ranking is the same as that obtained from previous calculations. The 4+4 SSPiRAL scheme with x = 3 is the most reliable among those we considered. In fact, the 4+4 array consisting of 4 data and 4 parity disks would do considerably better. The 4+3 array comes a close second. This can easily be explained. A single data item has the same protection in both layouts, but the SSPiRAL scheme uses one disk more for parity generation. Next comes the almost identical performance of the 4+4 SSPiRAL configuration with x=2 and the 3+3 SSPiRAL also with x = 2, in this order. Then comes RAID 4+2, followed by the two mirrored layouts that we considered. The six disk configuration has marginally better reliability. The worst configuration, also no surprise, is the 4+1 array.

In general, the ranking can be glanced from the Markov models where the shortest path from the initial state to the data loss state dominates the MTTDL determination. What is

surprising is the large distance between the 4+4 SSPiRAL, x = 3 scheme and the other SSPiRAL schemes.

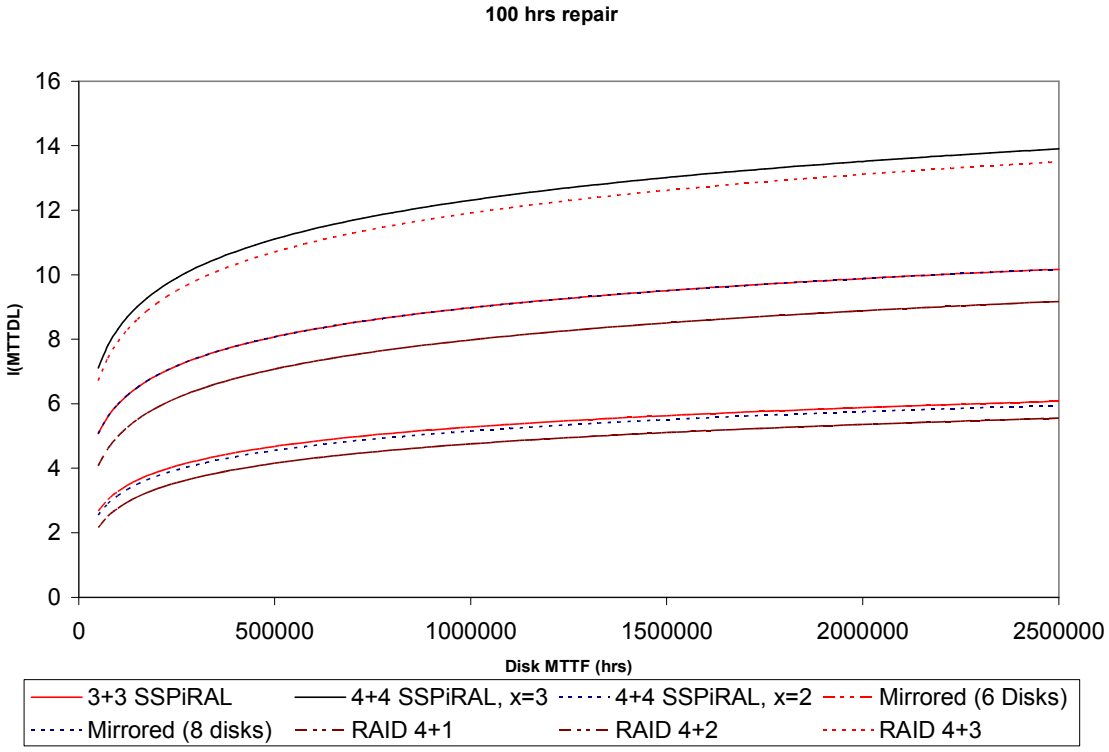

Figure 17: $\log_{10}$(MTTDL) in years for various arrays. The curves for the 3+3 SSPiRAL, x =2, and the 4+4 SSPiRAL, x=2, are almost identical. Average repair time is 100 hrs.

References

[ABC97] 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.

[BB+95] M. Blaum, J. Brady, J. Bruck, and J. Menon, "Even-odd: An efficient scheme for tolerating double disk failures in raid architectures," IEEE Transactions on Computing, vol. 44, no. 2, pp. 192–202, 1995.

[CLA07] 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.)

[CE+04] 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), pp. 1–14, USENIX, 2004.

[Gi90] G. A. Gibson, Redundant Disk Arrays: Reliable, Parallel Secondary Storage. PhD thesis, University of California at Berkeley, 1990.

[Gr96] H Gropp: Configuration. Chapter IV.6 in: The CRC Handbook of Combinatorial Designs, edited by C. Colbourn and J. Dinitz, CRC Press, 1996.

[Ha05] J. L. Hafner, "Weaver codes: Highly fault tolerant erasure codes for storage systems," in Proceedings of FAST, (San Francisco, CA), 2005.

[HG+92] L. Hellerstein, G. Gibson, R. Karp, R. Katz, and D. Patterson: Coding Techniques for Handling Failures in Large Disk Arrays. Algorithmica, Vol. 12 (2-3), p. 182-208.

[HJH00] 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.

[J+03] Z. Jie, W. Gang, L. Xiaogugang, L. Jing: The Study of Graph Decompositions and Placement of Parity and Data to Tolerate Two Failures in Disk Arrays: Conditions and Existence", Chinese Journal of Computer, vol. 26(10), p. 1379-1386, Oct. 2003.

[LMS05] W. Litwin, R. Moussa, T. Schwarz: LH*RS – A Highly-Available Scalable Distributed Data Structure, Transactions on Database Systems (TODS). Vol. 30(3). September 2005.

[LMC94] 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.

[PGK88] 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.

[PT04] 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.

[TB06] B. T. Theodorides and W. A. Burkhard, "^ B: Disk array data layout tolerating multiple failures," in Proceedings of MASCOTS), (Monterey, CA), pp. 21–32, IEEE, 2006.