Using storage class memories to increase the reliability of two-dimensional RAID arrays

Technical Report UCSC-SSRC-09-04
April 2009

Jehan-François Pâris
paris@cs.uh.edu

Ahmed Amer
amer@cs.pitt.edu

Darrell D.E. Long
darrell@cs.ucsc.edu

Storage Systems Research Center
Baskin School of Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
http://www.ssrc.ucsc.edu/

Using storage class memories to increase the reliability of two-dimensional RAID arrays

Jehan-François Pâris
Dept. of Computer Science
University of Houston
Houston, TX 77204-3010
paris@cs.uh.edu

Ahmed Amer[1]
Dept. of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
amer@cs.pitt.edu

Darrell D. E. Long[2]
Dept. of Computer Science
University of California
Santa Cruz, CA 95064
darrell@cs.ucsc.edu

Abstract

Two-dimensional RAID arrays maintain separate row and column parities for all their disks. Depending on their organization, they can tolerate between two and three concurrent disk failures without losing any data. We propose to enhance the robustness of these arrays by replacing a small fraction of these drives with storage class memory devices, and demonstrate how such a pairing is several times more reliable than relying on conventional disks alone, or simply augmenting popular redundant layouts. Depending on the ratio of the failure rates of these two devices, the substitution can double or even triple the mean time to data loss (MTTDL) of each array.

1 Introduction

RAID arrays are widely used to prevent data losses in medium to large storage systems [PGK 88, CL+94]. Their main advantages include low space overhead and excellent read throughput. Their main limitation is that they can only tolerate single disk failures.

Assuming a disk failure rate of one failure per 100,000 hours and a disk repair time of 24 hours with an eight-disk RAID, the probability of a second disk failure occurring before the failed disk can be replaced is 1.7×10^{-3}. While this is a risk that the owner of a medium-size size storage system is probably willing to take, the same is not true in storage systems composed of thousands of disks. Consider for instance an installation with one thousand disks. Assuming the same disk failure rate as before, we can predict that this installation will experience an average of 88 disk failures per year and thus has a full 15 percent probability of suffering irrecoverable data loss over a year. An even more alarming fact is the non-negligible risk of encountering an unreadable block on an otherwise operational disk while we attempt to reconstruct the contents of a failed disk, which would also result in irrecoverable data loss.

[1] Supported in part by the National Science Foundation under Award Number 0720578.
[2] Supported in part by the Petascale Data Storage Institute under Department of Energy Award DE-FC02-06ER25768.

The solution to these problems requires using redundant disk organizations that can tolerate more than one disk failure. RAID level 6 is the best known example of such organizations [SB92, BM93]. While RAID level 5 specifies one parity block per error correction group, RAID level 6 adds a second parity block thus allowing it tolerate two simultaneous disk failures.

We investigate a different approach, specifically, two-dimensional RAID arrays that organize their disks into a square matrix and maintain separate parity information for each row and each column of the matrix [PSL 07]. Depending on their layout, these arrays can tolerate two or even three simultaneous failures without experiencing any data loss. Previous investigations indicated that some two-dimensional RAID organizations could significantly increase their resilience by reorganizing themselves in the presence of one or more disk failures. We investigate the feasibility and extent of improving the reliability of such disk arrays by replacing a small fraction of their disk drives with solid-state devices in order to take advantage of the lower failure rates of these devices. Our analysis indicates that such a partial substitution can double or even triple the mean time to data loss (MTTDL) of each array depending on the ratio of the failure rates of these two devices.

In addition, we compare the reliabilities thus achieved by two-dimensional RAID arrays with those achieved by comparable RAID level 6 and mirrored organizations to conclude that two-dimensional RAID arrays are more reliable than these popular organizations. The remainder of the paper is organized as follows: Section 2 discusses storage class memories and their performance compared to conventional disks; Section 3 introduces our proposal; Section 4 evaluates their impact on array reliability; Section 5 reviews previous work; and Section 6 presents our conclusions.

2 Storage class memories

Storage class memories (SCMs) constitute a new class of non-volatile storage systems that are both cheaper than volatile main memory, and much faster than conventional disks. Unlike magnetic disk and MEMS [CG+00] technologies, SCMs have no moving

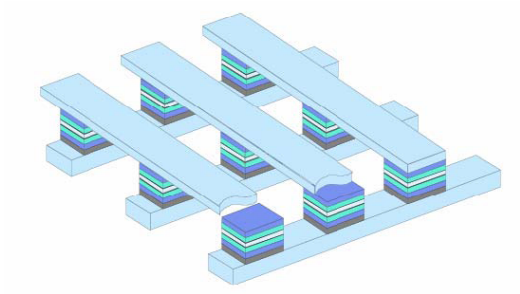

Fig. 1. General organization of a phase-change memory [N07].

Table 1. Expected specifications of PCM devices.

Parameter	*Expected Value (2012)*
Access time	100 ns
Data Rate	200–1000 MB/s
Write Endurance	10^9 write cycles
Read Endurance	no upper limit
Capacity	16 GB
Capacity growth	> 40% per year
Mean Time to Failure	10–50 million hours
Ratio of random to sequential access times	1
Active Power	100 mW
Standby Power	1 mW
Shock and Vibration resistance	> 15 g
Cost	< $2/GB
Cost reduction rate	40 percent/year

parts. In addition, they do not suffer from the potential write-speed limitations as flash memory.

We will focus here on phase-change memories (PCMs) as an exemplar of this new class of storage devices. While it is not yet clear which type of SCMs will eventually succeed on the marketplace, most of our conclusions are likely to hold for any type of SCMs.

PCMs contain no moving parts and use cross bar-type chip structures to access data. As seen in Fig.1, bits are stored at the intersection of each row and each column of the cross-bar structure. Various techniques can be used to encode this data. The most promising approach relies on the physical properties of chalcogenide materials. At room temperature, these materials can exist in two stable states, namely an amorphous state exhibiting a high resistivity and a crystalline state characterized by a much lower resistivity. Quickly heating the material above its melting temperature and then letting it quickly cool will leave the material in an amorphous state, characterized by a high resistivity. Similarly, heating the material above its crystallization temperature and then letting it cool at a relatively slower rate will leave it in a crystalline, more conductive, state.

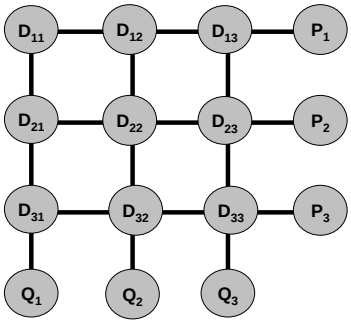

Fig. 2. A two-dimensional RAID array with nine data disks and six parity disks.

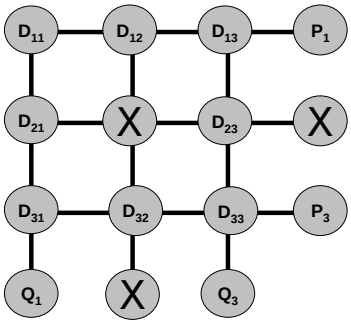

Fig. 3. One of the nine possible triple failures that will result in a data loss.

Table 1 displays the most important parameters of the first generation of SCMs. As we can see, they are almost as fast as volatile main memory and nearly as cheap as magnetic disks. In addition, they have a much better write endurance and better mean times to failures than flash memories.

3 Two-dimensional RAID arrays

Consider the two dimensional RAID array of Fig.2. It consists of nine data disks and six parity disks. Parity disks P_1, P_2 and P_3 contain the exclusive or (XOR) of the contents of the data disks in their respective rows while parity disks Q_1, Q_2 and Q_3 contain the XOR of the contents of the data disks in their respective columns. This organization offers the main advantage of ensuring that the data will survive the failure of an arbitrary pair of disks, and most failures of three disks. As seen in Fig.2, the only triple failures that result in data loss are those that include one arbitrary data disk and the two parity disks from its row and column.

The main drawback of this organization is its potentially poor write bandwidth. Assume that we want to update data stored on data disk D_{ij}. We will have to propagate the update to the respective row and column parity disks of disk D_{ij}, that is, disks P_i and Q_j. As a result, we will be unable to fully parallelize updates to data stored on two data disks that happen to be on the same row or the same column.

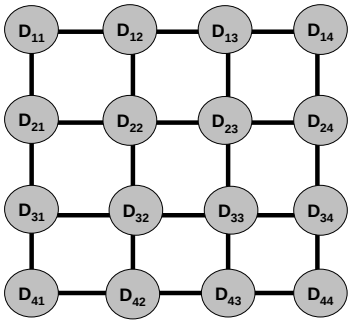

Fig. 4. A fully declustered two-dimensional RAID array with sixteen disks.

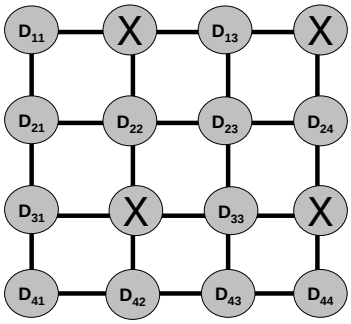

Fig. 5. One of the possible quadruple failures that will result in data loss.

One solution to this problem is *declustering*. To achieve it, we will add an additional disk to the array and distribute row parity blocks among the $n + 1$ disks of their respective rows row and column parity blocks among the $n + 1$ disks of their respective columns. As Fig.4 shows, the result of this process will be the two-dimensional equivalent of a RAID level 5 organization.

Observe that row parity blocks are now stored in each column. They now participate in the computation of the column parity blocks of the columns on which they reside. As long as a column i only contains a single failed disk, we will be able to reconstitute any row parity block located on the failed disk without involving any disk outside of column j. Similarly, column parity blocks will participate in the computation of the row parity blocks of the rows on which they reside. As long as a row i only contains a single failed disk, we will be able reconstitute any row parity block located on the failed disk without involving any disk outside of row i. As a result, triple failures similar to that displayed in Fig.3 will not result in data loss as we will always be able to reconstitute enough row or column parity blocks. The array remains vulnerable to quadruple failures involving four disks located on the same two rows and the same two columns. As seen on Fig.5, data loss occurs when four failed devices form a rectangle.

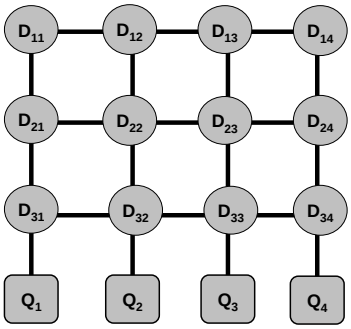

Fig. 6. A partially declustered two-dimensional RAID array with twelve disks and four SCM devices.

Even higher reliabilities could be achieved by replacing some—or all—of the array disks by more reliable devices. SCMs are an ideal candidate because of their expected lower failure rates compared to disk technology. Since they do not suffer from disk's random access performance penalty, they will also have a much higher effective I/O bandwidth.

Fig.6 describes one possible way of using both magnetic disks and SCM devices in a two-dimensional array. It consists of twelve disks and four SCM devices. The twelve disks are organized into three RAID level 5 arrays corresponding to the three top rows of the array and the four SCM devices contain the column parities of the disks. Since the SCM devices are expected to have a much higher effective I/O bandwidth than disks, they should be better able to handle parity update workloads from the three disks in their respective columns.

Observe that the SCM devices maintain the column parities of both the data blocks and the parity blocks stored on the twelve disks. As long as a column i only contains a single failed disk, we will thus be able to reconstitute any row parity block located on the failed disk without involving any disk outside of column j. This guarantees that the array will tolerate all triple device failures.

4 Reliability Analysis

Estimating the reliability of a storage system means estimating the probability $R(t)$ that the system will operate correctly over the time interval $[0, t)$ given that it operated correctly at time $t = 0$. Computing that function requires solving a system of linear differential equations, a task that becomes quickly unmanageable as the complexity of the system grows. A more practical option is to focus on the mean time to data loss (MTTDL) of the storage system, which is the approach we will take here.

Our system model consists of an array with independent failure modes for each device. When a device fails, a repair process is immediately initiated for that device. Should several devices fail, the repair process will be performed in parallel on those devices.

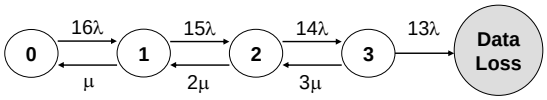

Fig. 7. Simplified state diagram for a fully declustered two-dimensional RAID array with 16 disks.

We assume that disk failures and SCM device failures are independent events exponentially distributed with respective rates λ and λ'. We further assume that that all repairs are exponentially distributed with rate μ.

We will first consider a fully declustered two-dimensional RAID array consisting 16 disks before the case of a partially declustered array with twelve disks and four SCM devices.

Building an accurate state-transition diagram for either two-dimensional disk array is a daunting task as we have to distinguish between failures of data disks and failures of parity disks, as well as between failures of disks located on the same or on different parity stripes. Instead, we present a simplified model that assumes that all quadruple failures result in a data loss.

Fig.7 displays the simplified state diagram for a fully declustered two-dimensional RAID array with 16 disks. State <0> represents the normal state of the array when its 16 disks are all operational. A failure of any of these disks would bring the array to state <1>. A failure of a second disk would bring the array into state <2> and a failure of a third disk would bring the array to state <3>. As we stated earlier, we assume that any additional failure occurring while the array already has three failed disks will result in a data loss.

Repair transitions bring the array back from state <3> to state <2>, then from state <2> to state <1> and, finally, from state <1> to state <0>. 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

$$\frac{dp_0(t)}{dt} = -16\lambda p_0(t) + \mu p_1(t)$$

$$\frac{dp_1(t)}{dt} = -(15\lambda + \mu)p_1(t) + 16\lambda p_0(t) + 2\mu p_2(t))$$

$$\frac{dp_2(t)}{dt} = -(14\lambda + 2\mu)p_2(t) + 15\lambda p_1(t) + 3\mu p_3(t)$$

$$\frac{dp_3(t)}{dt} = -(13\lambda + 3\mu)p_3(t) + 14\lambda p_2(t)$$

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

The Laplace transforms of these equations are

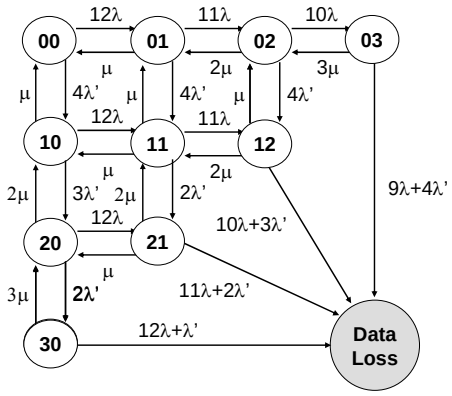

Fig. 8. Simplified state diagram for a partially declustered two-dimensional RAID array with twelve disks and four SCM devices.

$$sp_0^*(s) - 1 = -16\lambda p_0^*(s) + \mu p_1^*(s)6$$

$$sp_1^*(s) = -(15\lambda + \mu)p_1^*(s) + 16\lambda p_0^*(s) + 2\mu p_2^*(s)$$

$$sp_2^*(s) = -(14\lambda + \mu)p_2^*(s) + 15\lambda p_1^*(s) + 3\mu p_3^*(s)$$

$$sp_3^*(s) = -(13\lambda + \mu)p_3^*(s) + 14\lambda p_2^*(s)$$

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 compute the MTTDL and obtain.

$$MTTDL = \frac{6061\lambda^3 + 659\lambda^2\mu + 61\lambda\mu^2 + 3\mu^3}{21840\lambda^4},$$

Let us now consider the case where 4 of the 16 disks were replaced by SCM devices with a lower failure rate λ' than magnetic disks but the same repair rate μ. As Fig.8 shows, the state transition diagram of the new array has now ten states instead of four as we have to distinguish between failures of one of the 12 disks and failures of one of the four SCM devices. State <0, 0> represents the normal state of the array when its 16 devices are all operational. Successive failures of one of the 12 disks would first bring the array to state <0, 1> then to state <0, 2> and finally to state <0, 3>. As before, we assume that a fourth disk failure would result in a data loss. In the same way, successive failures of one of the four SCM devices would first bring the array to state <1, 0> then to state <2, 0> and finally to state <3, 0>. The three remaining states correspond to situations where the array has lost i SCM devices and j disks with $i + j \leq 3$. Repair transitions are fairly regular.

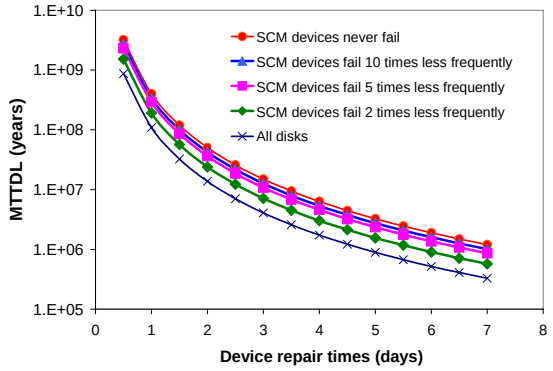

Fig. 9. Impact of SCM reliability on the Mean Time To Data Loss of the array.

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

$$\frac{dp_{00}(t)}{dt} = -(12\lambda + 4\lambda')p_{00}(t) + \mu(p_{01}(t) + p_{10}(t))$$

$$\frac{dp_{01}(t)}{dt} = -(11\lambda + 4\lambda' + \mu)p_{01}(t) + 12\lambda p_0(t) + 2\mu p_{02}(t) + \mu p_{11}(t)$$

$$\frac{dp_{10}(t)}{dt} = -(12\lambda + 3\lambda' + \mu)p_{10}(t) + 4\lambda' p_0(t) + 2\mu p_{20}(t) + \mu p_{11}(t)$$

$$\frac{dp_{02}(t)}{dt} = -(10\lambda + 4\lambda' + 2\mu)p_{02}(t) + 11\lambda p_1(t) + 3\mu p_{03}(t) + \mu p_{12}(t)$$

$$\frac{dp_{11}(t)}{dt} = -(11\lambda + 3\lambda' + 2\mu)p_{11}(t) + 12\lambda p_{10} + 4\lambda' p_{01}(t) + 2\mu p_{12}(t) + 2\mu p_{21}(t)$$

$$\frac{dp_{20}(t)}{dt} = -(12\lambda + 2\lambda' + 2\mu)p_{01}(t) + 3\lambda' p_{10}(t) + 3\mu p_{30}(t) + \mu p_{21}(t)$$

$$\frac{dp_{03}(t)}{dt} = -(9\lambda + 4\lambda' + 3\mu)p_{03}(t) + 11\lambda p_1(t)$$

$$\frac{dp_{12}(t)}{dt} = -(10\lambda + 3\lambda' + 3\mu)p_{03}(t) + 11\lambda p_{11}(t) + 4\lambda' p_{02}(t)$$

$$\frac{dp_{21}(t)}{dt} = -(11\lambda + 2\lambda' + 3\mu)p_{21}(t) + 12\lambda p_{20} + 3\lambda' p_{11}(t)$$

$$\frac{dp_{30}(t)}{dt} = -(12\lambda + \lambda' + 3\mu)p_{21}(t) + 2\lambda' p_{20}(t)$$

where $p_{ij}(t)$ is the probability that the system is in state $<i, j>$ with the initial conditions $p_{00}(0) = 1$ and $p_{ij}(0) = 0$.

Using the same techniques as in the previous example, we obtain the MTTDL of the new array as a quotient of two polynomials too large to be displayed.

When the failure rate λ' of the SCM devices goes to zero, we have

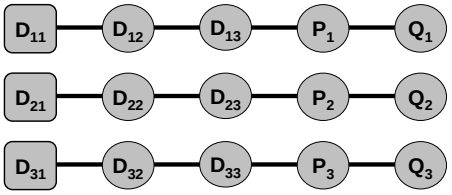

Fig. 10. An alternative organization with nine data devices (including 3 SCMs) and six check devices forming three RAID level 6 arrays.

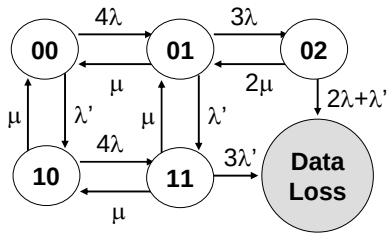

Fig. 11. State transition probability diagram for a stripe consisting of three data disks and two check disks.

$$\lim_{\lambda' \to 0} MTTDL = \frac{763\lambda^3 + 117\lambda^2\mu + 15\lambda\mu^2 + \mu^3}{1980\lambda^4}.$$

Fig.9 displays on a logarithmic scale the MTTDLs achieved by our two-dimensional array. 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. These values correspond to the high end of the failure rates observed by Pinheiro *et al.* [PWB07] and Schroeder and Gibson [PG07]. Disk repair times are expressed in days and MTTDLs expressed in years. As we can see, replacing only four disks with SCM devices can double or even triple the array MTTDL. In addition, we observe that the beneficial effects of the substitution become significant as soon as the SCM device failure rate is less than one half of the disk failure rate and reaches its maximum as soon as it becomes less than one tenth of that rate.

5 Two Alternate Organizations

In this section, we compare the reliability of our two-dimensional RAID arrays, with and without SCM devices, with those of two other organizations.

5.1 *A RAID level 6 organization*

Another way of organizing a two-dimensional RAID array with (n^2+2n) devices is to partition them into n RAID level 6 stripes each consisting of n data disks and two parity disks (we may prefer to now call these two disks *check disks*). This organization is illustrated in Fig.10. As we can see it has $((n+1)^2-1)$ devices, that is, one less device than its two-dimensional counterpart. It would protect data against the failure of up to two devices in any of its n stripes. We evaluate its MTTDL and compare it to those

obtained by our two-dimensional array. Fig.11 shows the state transition probability diagram for a single RAID level 6 stripe consisting of four disks and one SCM device. State <0, 0> represents the normal state of the array when its five devices are all operational. Successive failures of one of the four disks would first bring the array to state <0, 1> then to state <0, 2> while a third disk failure would result in a data loss. In the same way, a failure of one of the SCM device would bring the array to state <1, 0>. State <1, 1> represents the state of the array after it has lost one SCM device and one disk. Repair transitions are fairly regular.

The system of differential equations describing the behavior of each RAID level 6 stripe is

$$\begin{aligned}
\frac{dp_{00}(t)}{dt} &= -(4\lambda+\lambda')p_{00}(t)+\mu(p_{01}(t)+p_{10}(t)) \\
\frac{dp_{01}(t)}{dt} &= -(3\lambda+\lambda'+\mu)p_{01}(t)+12\lambda p_0(t)+ \\
&\quad 2\mu p_{02}(t)+\mu p_{11}(t) \\
\frac{dp_{10}(t)}{dt} &= -(4\lambda+\mu)p_{10}(t)+\lambda' p_0(t)+\mu p_{11}(t) \\
\frac{dp_{02}(t)}{dt} &= -(2\lambda+\lambda'+2\mu)p_{02}(t)+11\lambda p_{01}(t) \\
\frac{dp_{11}(t)}{dt} &= -(3\lambda+\lambda'+2\mu)p_{11}(t)+4\lambda p_{10}+\lambda' p_{01}(t)
\end{aligned}$$

Applying the same techniques as in the previous section, we obtain the MTTDL of each stripe. Since our array configuration consists of three stripes, the MTTDL of the whole array is one third that value.

Fig.12 displays, on a logarithmic scale, the MTTDLs achieved by the new array configuration and compares them with the MTTDLs achieved by a two-dimensional array with same storage capacity. As in Fig.9, the disk failure rate λ is assumed to be equal to one failure every one hundred thousand hours and the disk repair times vary between half a day and seven days. As we can see, this alternative organization achieves MTTDLs that are much lower than those achieved by the two-dimensional arrays. We can explain this discrepancy by considering that the RAID level 6 configuration is vulnerable to the failure of any three devices within a single RAID stripe while the two-dimensional organizations can tolerate all triple failures without incurring data loss.

5.2 *A mirrored organization*

Another possible way to protect data stored on n^2 data disks is to mirror them on n^2 additional devices. For the sake of our analysis, we will assume that these n^2 devices are SCM devices. Let us observe first that this new solution would require $n^2-(n+1)$ additional SCM devices to organizations such as that of Fig.6.

Applying the same techniques as before, we find that the MTTDL for a mirrored pair consisting of a disk and an SCM device is

$$MTTDL_{pair}=\frac{\lambda^2+\lambda\lambda'+\lambda'^2+\mu(2\lambda+2\lambda+\mu)}{\lambda\lambda'(\lambda+\lambda+2\mu)},$$

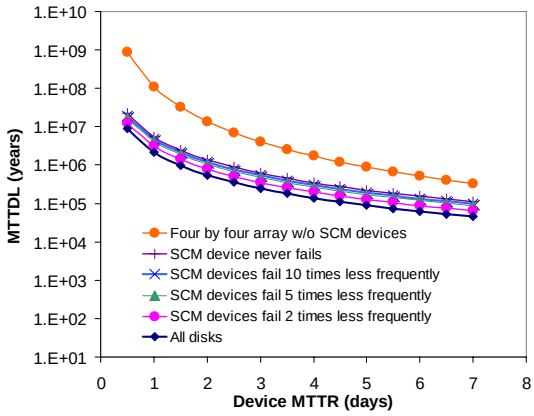

Fig. 12. Impact of SCM reliability on the Mean Time To Data Loss of a set of four RAID level 6 arrays.

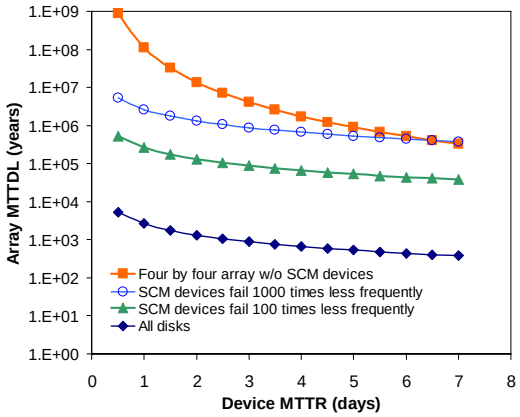

Fig. 13. Impact of SCM reliability on the Mean Time To Data Loss of a set of nine mirrored pairs each consisting of a disk and an SCM device.

where λ and λ' are the respective failure rates of disk and SCM devices and μ is the repair rates of these two devices.

Given that a mirrored organization capable of storing the same amount of data as our two-dimensional RAID array would have nine mirrored pairs, hence the MTTDL of the mirrored organization will be

$$\frac{MTTDL_{pair}}{9}.$$

Fig.13 displays on a logarithmic scale the MTTDLs achieved by such sets of mirrored pairs and compares them with the MTTDLs achieved by a two-dimensional array with same storage capacity. As in Fig.7, the disk failure rate λ is assumed to be equal to one failure every one hundred thousand hours and the disk repair times vary between half a day and seven days.

We observe that the set of nine SCM mirrors also results in much lower MTTDLs than an all-disk two-

dimensional array. It only improves upon the smaller all-disk array when the failure rate of the SCM devices is less than one thousand times that of a disk. As a result, it will require highly reliable SCMs to perform better than two-dimensional arrays, even when such arrays are not augmented with a smaller number of equivalent SCMs. This poor performance can be easily explained by observing that the mirrored organization cannot protect data against the failure of two devices in one of the mirrored pairs while the two-dimensional RAID array can tolerate the failure of *any* three devices.

The sole advantage of the mirrored organization is a simpler data reconstruction process involving a single device rather than entire rows and columns of devices. As a result, the mirrored organization will be less subject to irrecoverable read errors during the reconstruction process than the two-dimensional RAID organization.

6 Previous work

Increases in data volumes inevitably result in larger numbers of devices, which in turn result in an increased likelihood of multi-device failures, and so there has been a significant amount of work on schemes to tolerate multi-device failures. Traditional RAID schemes aimed at surviving the loss of an individual device within an array [Gib90, PGK88], and with variations of RAID-6 (and its various implementations) the goal was to survive the loss of two devices within an array [Pla08a, Pla08b, JGXJ03]. EvenOdd and Row-Diagonal Parity are also parity-based schemes capable of surviving two-device failures [BBBM95, CEG+04, GXS+08]. With these schemes the goal was to survive the requisite number of device failures while attempting to minimize the total space sacrificed for redundant storage. Other parity-based redundancy schemes included STAR, HoVer, GRID, and Bˆ, the latter of which typified the tendency of such approaches to focus on the data layout pattern, independent of the number of underlying devices [TB06, LSZ09, Haf06, HX05]. Typically, these data layouts were subsequently declustered data across homogenous, uniform, devices, and the majority could be classified as variations of low-density parity-codes as used for erasure coding in the communications domain by the Luby LT codes, and the subsequent Tornado and Raptor variants [Sho06, LMS+97, LMSS01]. Similar to the scheme we propose, HoVer and the more general GRID used parity-based layouts based on strips arranged in two or more dimensions. These layouts all assumed uniform homogenous devices, and largely competed based on their space efficiency [XB99, TX07], or their ability to survive more than two device failures [HX05, WLL+07]. Redundant layouts such as Bˆ, Weaver codes [Haf05], and our own SSPiRAL schemes [APS+07, CLLA07, ALPS08] departed from this trend, and offered redundant layouts that strictly limited the number of devices contributing to parity calculations, thereby offering a practical scheme for greater numbers of devices than typical in RAID arrays. SSPiRAL layouts were novel in their focus on individual device failures having potentially differing impact on the survivability of data. This treatment of devices as heterogeneous entities was an application of Systematic codes [PT04] across distinct storage devices, and was a departure from the typical goal of redundant storage layouts aiming to survive a predetermined number of device failures. Departing from such assumptions allows devices of varying reliability to be assigned roles with importance commensurate with their relative reliability. More recent efforts have started to investigate broader families of parity-based codes applied to heterogeneous device [GMW08], and in this work we have proposed the use, and evaluated the effect of, using devices of distinct technologies in such a manner.

7 Conclusion

Two-dimensional RAID arrays maintain separate row and column parities for all their disks. Depending on their organization, they can tolerate between two and three concurrent disk failures without losing any data. We have shown how to enhance the robustness of these arrays by replacing a small fraction of these drives by storage class memory devices that are several times more reliable than conventional disks. As we have seen, the substitution can double or even triple the mean time to data loss (MTTDL) of each array depending on the ratio of the failure rates of these two devices.

In addition, we have compared the reliabilities thus achieved by two-dimensional RAID arrays with those achieved by comparable RAID level 6 and mirrored organizations and concluded that two-dimensional RAID arrays are considerably more reliable than these two more popular organizations.

8 References

[ALPS08] A. Amer, D.D.E. Long, J.-F. Pâris, and T. Schwarz, Increased reliability with SSPiRAL data layouts, *Proc. IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)* (Baltimore, MD, USA), 2008.

[APS+07] A. Amer, J.-F. Pâris, T. Schwarz, V. Ciotola, and J. Larkby-Lahet, Out-shining Mirrors: MTTDL of Fixed-Order SSPiRAL Layouts, *Proc. International Workshop on Storage Net- work Architecture and Parallel I/Os (SNAPI)* (San Diego, CA, USA), September 2007.

[BBBM95] M. Blaum, J. Brady, J. Bruck, and J. Menon, EvenOdd: An efficient scheme for tolerating double disk failures in RAID architectures, *IEEE Trans. Computers* 44 (1995), no. 2, 192–202.

[BM93] W. A. Burkhard and J. Menon. Disk array storage system reliability. In *Proc. 23rd International Symposium on Fault-Tolerant Computing*, Toulouse, France, pp. 432–441, June 1993.

[CEG+04] P. Corbett, B. English, A. Goel, T. Grcanac, S. Kleiman, J. Leong, and S. Sankar, Row-diagonal parity for double disk failure correction, *Proc. USENIX Conference on File and Storage Technologies (FAST)* (San Francisco, CA, USA), USENIX Association, 2004, pp. 1–14.

[CG+00] L. R. Carley, G. R. Ganger, and D. F. Nagle, MEMS-based integrated-circuit mass-storage systems. *Communications of the ACM,* 43(11):73–80, Nov. 2000.

[CL+94] P. M. Chen, E. K. Lee, G. A. Gibson, R. Katz and D. A. Patterson. RAID, High-performance, reliable secondary storage, *ACM Computing Surveys* 26(2):145–185, 1994.

[CLLA07] V. Ciotola, J. Larkby-Lahet, and A. Amer, SSPiRAL layouts: Practical extreme reliability, *Tech. Report TR-07-149*, Department of Computer Science, University of Pittsburgh, 2007, Presented at the Usenix Annual Technical Conference 2007 poster session.

[E07] E-week, Intel previews potential replacement for flash memory, www.eweek.com/article2/0,1895,2021815,00.asp

[Gib90] G.A. Gibson, Redundant disk arrays: Reliable, parallel secondary storage, *Ph.D. thesis*, University of California at Berkeley, 1990.

[GMW08] K.M. Greenan, E.L. Miller, and J.J. Wylie, Reliability of XOR-based erasure codes on heterogeneous devices, *Proc. Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2008.

[GXS+08] W. Gang, L. Xiaoguang, L. Sheng, X. Guangjun, and L. Jing, Generalizing RDP codes using the combinatorial method, *NCA '08: Proc. 2008 Seventh IEEE International Symposium on Network Computing and Applications* (Washington, DC, USA), IEEE Computer Society, 2008, pp. 93–100.

[Haf05] J.L. Hafner, Weaver codes: Highly fault tolerant erasure codes for storage systems, *Proc. USENIX Conference on File and Storage Technologies (FAST)* (San Francisco, CA, USA), December 2005.

[Haf06] J.L. Hafner, HoVer erasure codes for disk arrays, *DSN '06: Proc. International Conference on Dependable Systems and Networks* (Washington, DC, USA), IEEE Computer Society, 2006, pp. 217–226.

[HX05] C. Huang and L. Xu, STAR: an efficient coding scheme for correcting triple storage node failures, *Proc. 4th conference on USENIX Conference on File and Storage Technologies (FAST)* (Berkeley, CA, USA), USENIX Association, 2005, pp. 15–15.

[JGXJ03] Z. Jie, W. Gang, L. Xiaogugang, and 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* 26 (2003), no. 10, 1379–1386.

[LMS+97] M.G. Luby, M. Mitzenmacher, M.A. Shokrollahi, D.A. Spielman, and V. Stemann, Practical loss-resilient codes, *Proc. 29th ACM Symposium on Theory of Computing (STOC)* (New York, NY, USA), ACM Press, 1997, pp. 150–159.

[LMSS01] M. Luby, M. Mitzenmacher, M.A. Shokrollahi, and D.A. Spielman, Efficient erasure correcting codes, *IEEE Trans. Information Theory* 47 (2001), no. 2, 569–584.

[LSZ09] M, Li, J, Shu, and W. Zheng, GRID codes: Strip-based erasure codes with high fault tolerance for storage systems, *Trans. Storage* 4 (2009), no. 4, 1–22.

[N07] S. Narayan, Storage class memory a disruptive technology, *Presentation at Disruptive Technologies Panel: Memory Systems of SC '07*, Reno, NV, Nov. 2007.

[PB+97] P. Pavan, R. Bez, P. Olivo, E. Zanoni,. Flash memory cells-an overview, *Proc. IEEE*, 85(8):1248–1271, Aug 1997.

[PGK88] D.A. Patterson, G. Gibson, and R.H. Katz, A case for redundant arrays of inexpensive disks (RAID). In *Proc. SIGMOD International Conference on Data Management*, Chicago, IL, pp. 109–116, June 1988.

[Pla08a] J.S. Plank, A new minimum density RAID-6 code with a word size of eight, *NCA '08: Proc. 2008 Seventh IEEE International Symposium on Network Computing and Applications* (Washington, DC, USA), IEEE Computer Society, 2008, pp. 85–92.

[Pla08b] J.S. Plank, The RAID-6 liberation codes, *Proc. 6th USENIX Conference on File and Storage Technologies (FAST)* (Berkeley, CA, USA), USENIX Association, 2008, pp. 1–14.

[PSL07] J.-F. Pâris, T. J. Schwarz and D. D. E. Long. Self-adaptive archival storage systems. In *Proc. 26th International Performance of Computers and Communication Conference*, New Orleans, LA, pp. 246–253, Apr. 2007.

[PT04] J.S. Plank and M.G. Thomason, A practical analysis of low-density parity-check erasure codes for wide-area storage applications, *Proc. 38th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)* (Florence, Italy), June 2004.

[PW07] E. Pinheiro, W.-D. Weber and L. A. Barroso, Failure trends in a large disk drive population, In *Proc. 5th USENIX Conference on File and Storage Technologies*, San Jose, CA, pp. 17–28, Feb. 2007.

[SB92] T. J. E. Schwarz and W. A. Burkhard. RAID organization and performance. In *Proc. 12th International Conference on Distributed Computing Systems*, pp. 318–325 June 1992.

[SG07] B. Schroeder and G. A. Gibson, Disk failures in the real world: what does an MTTF of 1,000,000 hours mean to you? In *Proc. 5th USENIX Conference on File and Storage Technologies*, San Jose, CA, pp. 1–16, Feb. 2007.

[SG99] M. Schulze, G. A. Gibson, R. Katz, R. and D. A. Patterson. How reliable is a RAID? In *Proc. Spring COMPCON 89 Conference*, San Francisco, CA, pp. 118–123, Mar. 1989.

[Sho06] A. Shokrollahi, Raptor codes, *IEEE/ACM Trans. Networking* 14 (2006), no. SI, 2551–2567.

[TB06] B.T. Theodorides and W.A. Burkhard, Bˆ: Disk array data layout tolerating multiple failures, *Proc. IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)* (Monterey, CA, USA), 2006, pp. 21–32.

[TX07] A. Thomasian and J. Xu, Cost analysis of the X-code double parity array, *MSST '07: Proc. 24th IEEE Conference on Mass Storage Systems and Technologies* (Washington, DC, USA), IEEE Computer Society, 2007, pp. 269–274.

[UM+03] M. Uysal, A. Merchant, G.A. Alvarez. Using MEMS-based storage in disk arrays, In *Proc. 2nd USENIX Conference on File and Storage Technologies*, San Francisco, CA, pp. 89–101, Mar.-Apr. 2003.

[WLL+07] G. Wang, X. Liu, S. Lin, G. Xie, and J. Liu, Constructing double- and triple-erasure-correcting codes with high availability using mirroring and parity approaches, *ICPADS '07: Proc. 13th International Conference on Parallel and Distributed Systems* (Washington, DC, USA), IEEE Computer Society, 2007, pp. 1–8.

[XB99] L. Xu and J. Bruck, X-code: MDS array codes with optimal encoding, *Trans. Information Theory* 45 (1999), no. 1, 272–276.