Low-Redundancy Two-Dimensional RAID Arrays

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

Ahmed Amer
Dept. of Computer Engineering
Santa Clara University
Santa Clara, CA
a.amer@acm.org

Thomas J. E. Schwarz, S. J.
Depto. de Informática y Ciencias de la Computación
Universidad Católica del Uruguay
11600 Montevideo, Uruguay
tschwarz@ucu.edu.uy

***Abstract*—Most extant two-dimensional RAID arrays tend to use a square matrix organization such that each row and each column of the array forms a parity stripe. These arrays require $2k$ parity disks to protect the contents of k^2 data disks against all double disk failures and most triple disk failures. We investigate the reliability of a lesser known organization that allows k parity disks to protect $k(k-1)/2$ data disks against all double and most triple failures. We found that this organization can provide the same or better mean times to data loss (MTTDLs) than organizations requiring more parity disks to protect the same number of data disks.**

Keywords-archival storage systems; RAID arrays

I. INTRODUCTION

One of the major issues in storage technology is finding cost-effective solutions for the long-term storage of archival data. The topic is growing in importance as more organizations now maintain larger amounts of data online. Archival storage systems differ from other storage systems in two important ways. First, the data they contain remain largely unmodified once they are stored. As a result, write rates are a much less important issue than in conventional storage systems. Second, these data have to remain available over time periods that can span decades.

The best way to increase the survivability of data is through the use of redundancy. Two well-known examples of this approach are mirroring and m-out-of-n codes. Mirroring maintains two replicas of the stored data while m-out-of-n codes store data on n distinct disks along with enough redundant information to allow access to the data in the event $n - m$ of these disks fail. The best-known organizations using these codes are RAID level 5, which uses an $(n - 1)$-out-of-n code, and RAID level 6, which uses $(n - 2)$-out-of-n codes.

Two-dimensional RAID arrays are disk organizations wherein each disk holding data belongs to two independent RAID arrays. Fig. 1 represents one such organization consisting of nine data disks and six parity disks: each data disk belongs to a first RAID array comprising all disks in the same row and a second RAID array comprising all disks in the same column. This organization is especially attractive when its constituting elements are RAID level 4 arrays with all their parity blocks located on a single disk. As seen in Fig. 2, the sole triple failures that can result in a data loss are the failure of one data disk and its two parity blocks. As a result, two-dimensional RAID arrays are best suited for archival storage systems where reducing failure risks is paramount and the update rate limitations of RAID level 4 arrays are not an issue.

Some of us had previously investigated the so-called "square" two-dimensional arrays with k^2 data disks and $2k$ parity disks [PSL07]. We turn this time our attention to arrays

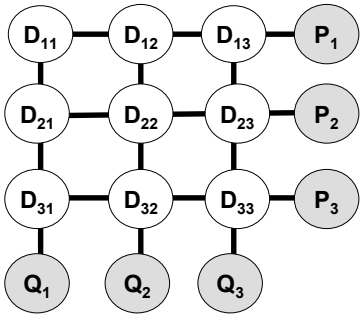


Fig. 1 – A two dimensional RAID array with nine data disks and six parity disks.

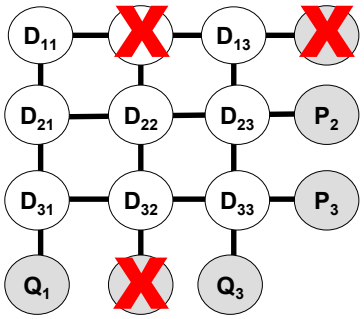


Fig. 2 – The same array experiencing the simultaneous failures of one arbitrary data disk, the parity disk in the same row and the parity disk in the same column.

requiring significantly fewer parity disks to ensure that all data disks belong to two distinct parity stripes.

We propose to show that that these array organizations can provide the same or better mean times to data loss (MTTDLs) than organizations requiring more parity disks to protect the same number of data disks against all double failures and most triple failures.

The remainder of this paper is organized as follows. Section 2 introduces our method and Section 3 discusses its impact on the mean time to data loss of the archived data. Section 4 compares it with two other redundant organizations while Section 5 surveys previous work on storage system reliability. Finally, section 6 has our conclusions.

II. OUR PROPOSAL

Square-shaped two-dimensional RAID organizations like the one depicted in Fig. 1 protect against all double disk failures because they satisfy the two following conditions:

1. Each data disk belongs to two distinct parity stripes.

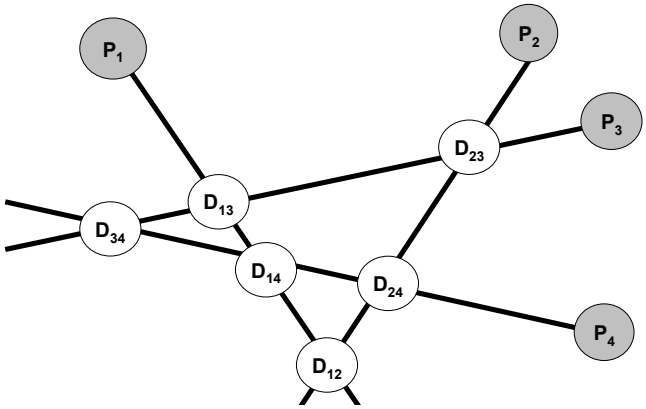


Fig. 3 – A two-dimensional RAID organization using four parity disks to protect the contents of six data disks against all double disk failures.

2. Two distinct parity stripes have at most one common data disk.

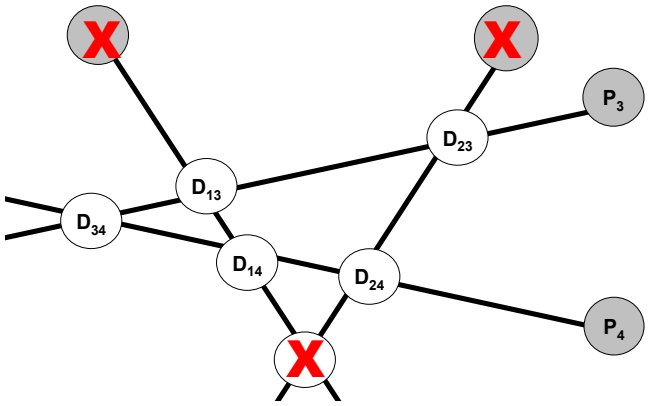
Disk array organizations that require fewer parity disks to protect against all double disk failures have been known for some time [HG+94] and dismissed as a mere curiosity. These organizations satisfy the two above conditions while only requiring k parity disks to protect $k(k-1)/2$ data disks. For instance, Fig. 3 represents an organization with six data disks and four parity disks. The figure is much easier to understand if we observe that parity stripes and their parity disk are represented by straight lines and the data disks by their intersections. The problem of finding the maximum number $m(k)$ of data disks that k parity disks can protect against all double disk failures becomes equivalent to the problem of finding the maximum number intersections that can be formed by k straight lines. We will call such disk arrays *maximal* disk arrays.

Assume that we already have a maximal disk array with $m(k)$ data disks and k parity disks. For $k = 4$, we have $m(4) = 6$. Adding one more parity disk, will allow us to form an additional parity stripe and place on this stripe k new data disks, all located at the intersection of the new parity stripe with one of the extant k parity stripes. We thus have the recurrence $m(k+1) = m(k) + k$ whose solution is:

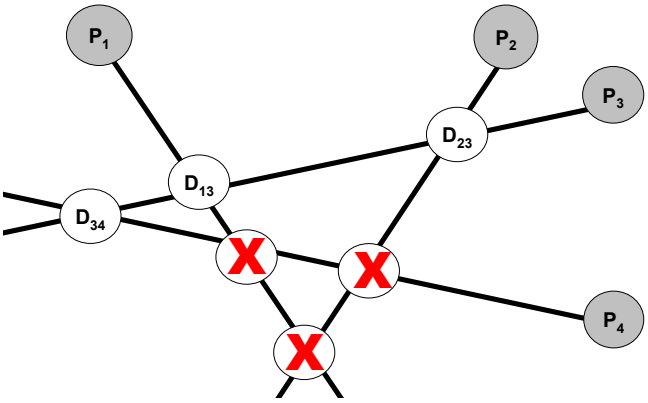
$$m(k) = k(k-1)/2 \qquad (1)$$

Returning to Fig. 1, we see that a square-shaped two-dimensional RAID array with an even number k of parity disks could protect $k^2/4$ data disks against all double disk failures. For all values of $k > 5$, that is at least 40 percent less than the number of data disks in a maximal two-dimensional array with the same number of party disks.

In addition to protecting the contents of their data disks against all double disk failures, square-shaped two-dimensional RAID arrays prevented data losses in the presence of all triple disk failures that do not involve the failure of a data disk and its two parity disks. As we can see on Fig. 4, maximal two-dimensional RAID arrays are vulnerable to a second type of triple disks failures, namely the failure of three data disks forming a triangle such that each of the three data disks has one parity stripe in common with each of the two other data disks. Let us now evaluate the impact of this additional failure mode on the reliability of these arrays.



(a) A failure of one data disk and its two parity disks.



(b) A failure of three data disks forming a triangle.

Fig. 4 – Characterizing the triple disk failures that will result in a data loss in a maximal two-dimensional disk array with four parity disks.

III. 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 simpler option is to focus on the mean time to data loss (MTTDL) of the storage system. This is the approach we will take here.

Our system model consists of a disk array with independent failure modes for each disk. Whenever 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.

Building an accurate state-transition diagram for a two-dimensional disk array is a daunting task because 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 here a simplified model based on the following approximations:

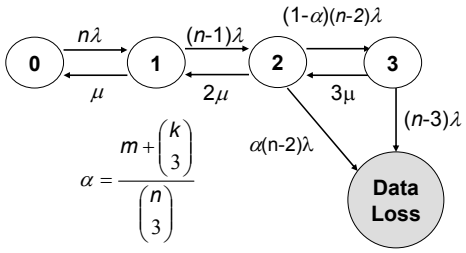


Fig. 5 – Simplified state transition probability diagram for a minimal two-dimensional array consisting of m data disks and k parity disks for a total of n disks.

1. Whenever the disk repair rate μ is much higher than the disk failure rate λ, each individual disk will be operational most of the time. Hence the probability that an array has four failed disks will be almost negligible when compared to the probability that the array has three failed disks. We can thus obtain a good upper bound of the array failure rate by assuming that the array fails whenever it has three failed disks in any of the critical configurations discussed in Section 2 or at least four failed disks regardless of their configuration. In other words, we will ignore the fact that the array can survive some, but not all, simultaneous failures of four or more disks.
2. Since disk failures are independent events exponentially distributed with rate λ, the rate at which an array that has already two failed disks will experience a third disk failure is:

$$(m+k-2)\lambda\,. \qquad (2)$$

where m is the number of data disks and k the number of parity disks.

There are $\binom{m+k}{3}$ possible configurations with 3 failed disks out of $m+k$ and $m+\binom{k}{3}$ of them correspond to the two types of fatal failures described in Fig. 4. Hence we will assume that the rate at which an array that has already two failed disks will experience a data loss will be $\alpha(m+k-2)\lambda$ with

$$\alpha = \frac{m+\binom{k}{3}}{\binom{m+k}{3}} \qquad (3)$$

where m is the number of data disks and k the number of parity disks. Conversely, the rate at which the same array will survive the loss will be $(1-\alpha)(m+k-2)\lambda$.

Fig. 5 displays the simplified state transition probability diagram for a two-dimensional array consisting of m data disks and k parity disks for a total of n disks. State <0> represents the normal state of the array when all its disks are operational. A failure of any of these n disks would bring the array to state <1>. A failure of a second disk would bring it into state <2>. A failure of a third disk will result in a data loss with probability α or bring the array to state <3> with probability $1-\alpha$.

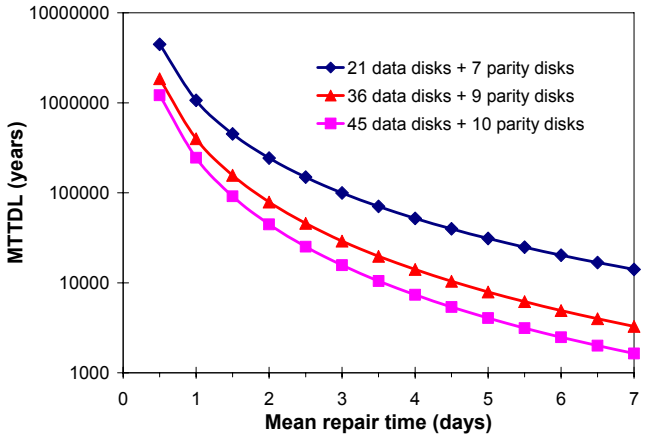


Fig. 6 – MTTDLs achieved by three maximal two dimensional RAID arrays.

As we stated earlier, we assume that any fourth disk failure will result in a data loss.

Repair transitions bring back the array 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:

$$\begin{aligned}
\frac{dp_0(t)}{dt} &= -n\lambda p_0(t) + \mu p_1(t) \\
\frac{dp_1(t)}{dt} &= -((n-1)\lambda+\mu)p_1(t) + n\lambda p_0(t) + 2\mu p_2(t) \\
\frac{dp_2(t)}{dt} &= -((n-2)\lambda+2\mu)p_2(t) + (n-1)\lambda p_1(t) + 3\mu p_3(t) \\
\frac{dp_3(t)}{dt} &= -((n-3)\lambda+3\mu)p_3(t) + (1-\alpha)(n-2)\lambda p_2(t)
\end{aligned} \qquad (4)$$

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 $i \neq 0$ and:

$$\alpha = \frac{m+\binom{k}{3}}{\binom{n}{3}}\,. \qquad (5)$$

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

$$MTTDL = \sum_i p_i^*(0)\,, \qquad (6)$$

where the $p_i^*(0)$ are the values at origin of the Laplace transforms of the $p_i(t)$ in Eq. 4, we obtain the MTTDL of the array by computing the Laplace transforms of the above system and solving it for $s = 0$. The expression we obtain is a quotient of two polynomials that are too large to be displayed. For $k = 9$ and $m = 36$, it simplifies into

$$\frac{601909\lambda^3 + 21309\lambda^2\mu + 651\lambda\mu^2 + 11\mu^3}{1980\lambda^3(3311\lambda+2\mu)}\,. \qquad (7)$$

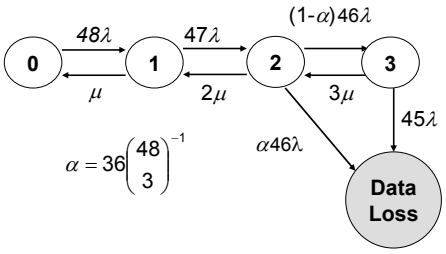


Fig. 7 – Simplified state transition probability diagram for a square-shaped two-dimensional array consisting of 36 data disks and 12 parity disks for a total of 48 disks.

Fig. 6 displays on a logarithmic scale the MTTDLs achieved by three reasonably-sized maximal two-dimensional RAID arrays for disk repair times varying between half a day and seven days. We assumed that the disk failure rate λ was one hundred thousand hours, that is, slightly less than one failure every eleven years. This failure rate is at the high end of the failure rates observed by Pinheiro et al. [PWB07] as well as Schroeder and Gibson [SG07]. MTTDLs are expressed in years and repair times in days.

As we can see, the MTTDLs of the three arrays are fairly sensitive to the average repair times. This should be expected in disk arrays counting more than twenty disks each because these arrays have a bigger aggregate disk failure rate than smaller arrays.

IV. Comparison with other disk organizations

In this section we compare the MTTDLs afforded our maximal two-dimensional RAID organization with those afforded by (a) a square-shaped two-dimensional RAID organization and (b) a more decentralized organization consisting of several independent RAID level 6 arrays.

A difficulty in comparing different two-dimensional organizations is finding organizations with the same number of data disks. For this reason, we decided to compare the MTTDLs of

1. A square-shaped two dimensional RAID array with 36 data disks and 12 parity disks, and
2. A maximal two-dimensional organization with 36 data disks and 9 parity disks.

Fig. 7 displays the simplified state transition probability diagram for a square-shaped two dimensional RAID array with 36 data disks and 12 parity disks for a total of 48 disks. Here the sole triple failures that may result in a data loss are those of a data disk and its two parity disks. Since there are $\binom{48}{3}$ possible configurations with 3 failed disks out of 48 but only 36 of them result in a data loss, the fraction a of triple failures that resulted in a data loss was:

$$\alpha = 36\binom{48}{3}^{-1} \qquad (8)$$

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

Fig. 8 displays on a logarithmic scale the MTTDLs achieved by the two array organizations. As before, we assumed that the disk failure rate λ was one failure every one hundred thousand hours and considered disk repair times between half a day and one week.

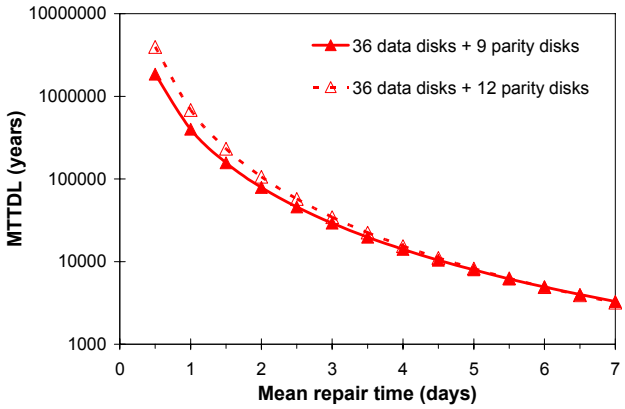


Fig. 8 – MTTDLs achieved by the two array organizations.

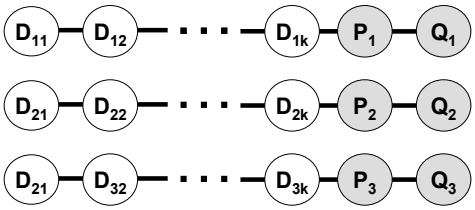


Fig. 9– A second alternative organization consisting of independent RAID level 6 arrays. (In reality the parity blocks would be equally distributed among all disks of each stripe).

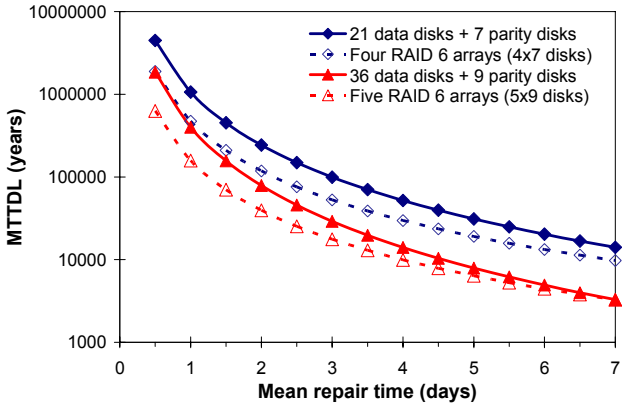


Fig. 10 – Comparing the MTTDLs achieved by our maximal two-dimensional arrays with those achieved by sets of RAID level 6 arrays with the same numbers of data disks and parity disks.

As we can see, the MTTDLs achieved by the two array organizations are fairly comparable even though the square-shaped array organization has three additional parity disks. In contrast, the MTTDLs achieved the two organizations are strongly affected by the mean disk repair times. This suggests that the most efficient organization in terms of both cost and reliability might be our maximal organization with nine parity disks supplemented by one or two spare disks in order to speed up disk repair times.

Our second benchmark was a more decentralized organization consisting of several independent RAID level 6 arrays. Fig. 9 depicts such an organization. Recalling that the MTTDL of a RAID level 6 with n disks is:

$$MTTDL(n) == \frac{(3n^2 - 6n + 2)\lambda^2 + (3n - 2)\lambda\mu + 2\mu^2}{n(n-1)(n-2)\lambda^3}, \quad (9)$$

and observing that the overall failure rate of a set of *m* identical independent disk arrays is *m* times the failure rate of one of its *m* components, we obtain the MTTDL of an organization comprising *m* RAID level 6 arrays with *n* disks each:

$$MTTDL(m;n) = (MTTDL(n)/m. \quad (10)$$

Fig. 10 compares the MTTDLs achieved by our maximal two-dimensional arrays with those achieved by sets of RAID level 6 arrays with the same numbers of disks. The comparison advantages the RAID level 6 organizations as we are comparing

1. A maximal two-dimensional organization with 21 data disks and 7 parity disks with a set of RAID level 6 arrays comprising 20 data disks and 8 parity disks.
2. A maximal two-dimensional organization with 36 data disks and 9 parity disks with a set of RAID level 6 arrays comprising 35 data disks and 10 parity disks.

Despite having one more data disk and one less parity disk than their counterparts, the two minimal two-dimensional arrays still achieve higher MTTDLs at all repair times.

V. Previous Work

Erasure coding for disk storage first appeared in RAID organizations as $(n-1)$-out-of-n codes [PGK88, SG+89, G90, SB92, CL+94]. RAID level 6 organizations use $(n-2)$-out-of-n codes to protect data against double disk failures [BM93]. EvenOdd and Row-Diagonal Parity are also parity-based schemes capable of surviving two-device failures [BB+95, CE+04, GX+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, H05, H06]. 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's LT codes, and the subsequent Tornado and Raptor variants [S06, LM+97]. HoVer and the more general GRID used parity-based layouts based on stripes arranged in two or more dimensions. These layouts largely competed based on their space efficiency [XB99, TX07], or their ability to survive more than two device failures [H05].

Redundant layouts such as Bˆ, Weaver codes [H05] 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.

VI. Conclusion

We have presented a two-dimensional RAID organization protecting data disks against all double failures and most triple failures at a lower cost than extant two-dimensional disk organizations: while these organizations require k parity disks to protect $k^2/4$ data disks, our organization can protect $k(k-1)/2$ data disks with the same number of parity disks.

In addition, we found out that our new disk organization achieved better MTTDLs than a set of RAID level 6 stripes having one additional parity disk and one less data disk.

More work is still needed to evaluate the impact of irrecoverable read errors on the reliability of these organizations.

References

[BB+95] M. Blaum, J. Brady, J. Bruck, and J. Menon, EvenOdd: An efficient scheme for tolerating double disk failures in RAID architectures, *IEEE Trans. on Computers* 44(2):192–202, 1995.

[BM93] W. A. Burkhard and J. Menon. Disk array storage system reliability. Proc. *23rd FTC symp.*, June 1993, pp. 432-441.

[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, *Proc. 3rd USENIX FAST Conf.*, 2004, pp. 1–14.

[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.

[G90] G. A. Gibson, Redundant disk arrays: Reliable, parallel secondary storage, *Ph.D. Thesis*, U. C., Berkeley, 1990.

[GX+08] W. Gang, L. Xiaoguang, L. Sheng, X. Guangjun, and L. Jing, Generalizing RDP codes using the combinatorial method, Proc. *7th IEEE Int. Symp. on Network Computing and Applications*, pp. 93–100, July 2008.

[H05] J. L. Hafner, Weaver codes: Highly fault tolerant erasure codes for storage systems, *Proc. 4th USENIX FAST Conf.*, Dec. 2005.

[H06] J. L. Hafner, HoVer erasure codes for disk arrays, *Proc. DSN Conf.*, June 2006, pp. 217–226.

[HG+94] L. Hellerstein, G. A. Gibson, R. M. Karp, R. H. Katz, D. A. Patterson, Coding Techniques for Handling Failure in Large Disk Arrays. *Algorithmica*, 12(3-4): 182– 208, June 1994

[LM+97] M. Luby, M. Mitzenmacher, M.A. Shokrollahi, D.A. Spielman, and V. Stemann, Practical loss-resilient codes, *Proc. 29th ACM STOC Symp.,* May 1997, pp. 150–159.

[LSZ09] M. Li, J. Shu, and W. Zheng, GRID codes: Stripe-based erasure codes with high fault tolerance for storage systems, *ACM Trans. on Storage,* 4(4):1–22, Jan. 2009.

[PGK88] D. A. Patterson, G. A. Gibson and R. Katz. A case for redundant arrays of inexpensive disks (RAID). *Proc. SIGMOD 1988 Conf.*, June 1988, pp. 109–116.

[PSL07] J.-F. Pâris, T. J. Schwarz and D. D. E. Long. Self-adaptive archival storage systems, *Proc. 26th IPCCC Conf.*, Apr. 2007, pp. 246-253.

[PWB07] E. Pinheiro, W.-D. Weber and L. A. Barroso, Failure trends in a large disk drive population, Proc. *5th USENIX Conference on File and Storage Technologies*, Feb. 2007, pp. 17–28.

[S06] A. Shokrollahi, Raptor codes, *IEEE/ACM Trans. on Networking* 52(6):2551–2567, June 2006.

[SB92] T. J. E. Schwarz and W. A. Burkhard. RAID organization and performance. *Proc. 12thICDCS Conf.*, June 1992, pp. 318–325.

[SG+89] M. Schulze, G. Gibson, R. Katz, R. and D. A. Patterson. How reliable is a RAID? In *Proc. Spring 1989 COMPCON Conf.*, March 1989, pp. 118–123.

[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? *Proc. 5th USENIX FAST Conf.,* Feb. 2007, pp. 1–16.

[TB06] B.T. Theodorides and W.A. Burkhard, Bˆ: Disk array data layout tolerating multiple failures, *Proc. 14th MASCOTS Symp.*, Sep. 2006, pp. 21–32.

[TX07] A. Thomasian and J. Xu, Cost analysis of the X-code double parity array, *Proc. 24th IEEE MSST Conf.*, pp. 269–274, Sep. 2007.

[XB07] L. Xu and J. Bruck, X-code: MDS array codes with optimal encoding, *IEEE Trans. Information Theory* 45(1):272–276, 1999.