RJ10504 (ALM1303-010) March 15, 2013
Computer Science

IBM Research Report

Construction of PMDS and SD Codes Extending RAID 5

Mario Blaum
IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099
USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Construction of PMDS and SD Codes extending RAID 5

Mario Blaum
IBM Almaden Research Center
San Jose, CA 95120

March 8, 2013

Abstract

A construction of Partial Maximum Distance Separable (PMDS) and Sector-Disk (SD) codes extending RAID 5 with two extra parities is given, solving an open problem. Previous constructions relied on computer searches, while our constructions provide a theoretical solution to the problem.

Keywords: Error-correcting codes, RAID architectures, MDS codes, array codes, Reed-Solomon codes, Blaum-Roth codes, PMDS codes, SD codes.

1 Introduction

Consider an $m \times n$ array whose entries are elements in a finite field $GF(2^b)$ [4] (in general, we could consider a field $GF(p^b)$, p a prime number, but for simplicity, we constrain ourselves to binary fields). The n columns represent storage devices like SSDs, HDDs or tapes. The arrays (often called stripes also) are repeated as many times as necessary. In order to protect against a device failure, a RAID 4 or RAID 5 type of scheme, in which one of the devices is the XOR of the other ones, can be implemented. During reconstruction, the failed device is recovered sector by sector. The problem with RAID 5 is, if an additional sector is defective in addition to the one corresponding to the failed device, data loss will occur. A solution to this problem is using a second device for parity (RAID 6), allowing for recovery against two failed devices. However, this scheme may be wasteful, and moreover, it is unable to correct the situation in which in addition to the sector corresponding to the failed disk, we have two extra failed sectors in the row (we always assume that failed sectors can be identified, either by CRC or by other means, so the correcting scheme is an erasure correcting scheme). In order to overcome this problem, the so called Partial MDS (PMDS) codes [1] and Sector-Disk (SD) codes [5] were created. Very similar codes were presented in [3].

We start by giving the definition of PMDS and SD codes.

1	0	1	0	0
E	1	E	0	1
1	1	1	0	1
1	E	1	1	E

1	E	1	0	0
0	E	1	0	E
1	E	1	0	1
E	E	1	1	1

1	0	1	E	0
E	1	E	E	1
1	1	1	E	1
1	0	1	E	1

Figure 1: A 4 × 5 array with different types of failures

Definition 1.1 Let $\mathcal{C}$ be a linear $[mn, m(n-r)-s]$ code over a field such that when codewords are taken row-wise as $m \times n$ arrays, each row belongs in an $[n, n-r, r+1]$ MDS code. Then,

1. $\mathcal{C}$ is an $(r; s)$ partial-MDS (PMDS) code if, *for any* $(s_1, s_2, \ldots, s_t)$ such that each $s_j \geq 1$ and $\sum_{j=1}^{t} s_j = s$, and for any $i_1, i_2, \ldots, i_t$ such that $0 \leq i_1 < i_2 < \cdots < i_t \leq m-1$, $\mathcal{C}$ can correct up to $s_j + r$ erasures in each row i_j, $1 \leq j \leq t$, of an array in $\mathcal{C}$.

2. $\mathcal{C}$ is an $(r; s)$ sector-disk (SD) code if, for any $l_1, l_2, \ldots, l_r$ such that $0 \leq l_1 < l_2 < \cdots < l_r \leq n-1$, for any $(s_1, s_2, \ldots, s_t)$ such that each $s_j \geq 1$ and $\sum_{j=1}^{t} s_j = s$, and for any $i_1, i_2, \ldots, i_t$ such that $0 \leq i_1 < i_2 < \cdots < i_t \leq m-1$, $\mathcal{C}$ can correct up to $s_j + r$ erasures in each row i_j, $1 \leq j \leq t$, of an array in $\mathcal{C}$ provided that locations $l_1, l_2, \ldots l_r$ in each of the rows i_j have been erased.

SD codes satisfy a weaker condition than PMDS codes, but they may be sufficient in most applications. The case of $(r; 1)$ PMDS codes has been solved in [1]. In this paper, we address the case of (1;2) PMDS and SD codes. Figure 1 illustrates the difference between (1;2) PMDS and SD codes for a 4×5 array (i.e., a code of length 20): the array in the left depicts a situation that can be handled by a (1;2) PMDS but not by a (1;2) SD code; the second and the fourth rows have two erasures (denoted by E) each and there is no column containing two of these erasures. The array in the middle illustrates a situation in which the second and fourth rows have two erasures each, but the second column contains two of those erasures, which correspond to a total failure of the second device. Individual erasures in a row can always be handled by single parity (like in the first and the third rows). This situation can be handled by both (1;2) PMDS and SD codes. Finally, the array in the right shows the situation of three erasures in a row, and at most one in the remaining ones. This situation can also be handled by both (1;2) PMDS and SD codes (but not by RAID 6).

In the next section we give the construction of both (1;2) PMDS and SD codes. From now on, when we say PMDS or SD codes, we refer to (1;2) PMDS or SD codes.

2 Code Construction

Consider the field $GF(2^b)$ and let α be an element in $GF(2^b)$. The (multiplicative) order of α, denoted $\mathcal{O}(\alpha)$, is the minimum ℓ, $0 < \ell$, such that $\alpha^\ell = 1$. If α is a primitive element [4], then $\mathcal{O}(\alpha)=2^b-1$. To each element $\alpha \in GF(2^b)$, there is an associated (irreducible) minimal polynomial [4] that we denote $f_\alpha(x)$.

Let $\alpha \in GF(2^b)$ and $mn \le \mathcal{O}(\alpha)$. Consider the $(m+2)\times mn$ parity-check matrix

$$\left(\begin{array}{cccc|cccc|c|cccc} \underline{c}_0 & \underline{c}_1 & \cdots & \underline{c}_{n-1} & \underline{c}_n & \underline{c}_{n+1} & \cdots & \underline{c}_{2n-1} & \cdots & \underline{c}_{(m-1)n} & \underline{c}_{(m-1)n+1} & \cdots & \underline{c}_{mn-1} \end{array}\right) \qquad (1)$$

where $\underline{c}_i$ denotes a column of length $m+2$, and, if $\underline{e}_i$ denotes an $m\times 1$ vector whose coordinates are zero except for coordinate i, which is 1, then, for $0 \le i \le m-1$,

$$\underline{c}_{in}, \underline{c}_{in+1}, \ldots, \underline{c}_{(i+1)n-1} \quad = \quad \left(\begin{array}{ccccccc} \underline{e}_i & \underline{e}_i & \cdots & \underline{e}_i & \cdots & \underline{e}_i \\ \alpha^{in} & \alpha^{in+1} & \ldots & \alpha^{in+j} & \ldots & \alpha^{(i+1)n-1} \\ \alpha^{2in} & \alpha^{2in-1} & \ldots & \alpha^{2in-j} & \ldots & \alpha^{(2i-1)n+1} \end{array}\right) \qquad (2)$$

We denote as $\mathcal{C}^{(0)}(m,n;f_\alpha(x))$ the $[mn, m(n-1)-2]$ code over $GF(q)$ whose parity-check matrix is given by (1) and (2).

Example 2.1 Consider the finite field $GF(16)$ and let α be a primitive element, i.e., $\mathcal{O}(\alpha)=15$. Then, the parity-check matrix of $\mathcal{C}^{(0)}(3,5;f_\alpha(x))$ is given by

$$\left(\begin{array}{ccccc|ccccc|ccccc} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5 & \alpha^6 & \alpha^7 & \alpha^8 & \alpha^9 & \alpha^{10} & \alpha^{11} & \alpha^{12} & \alpha^{13} & \alpha^{14} \\ 1 & \alpha^{14} & \alpha^{13} & \alpha^{12} & \alpha^{11} & \alpha^{10} & \alpha^9 & \alpha^8 & \alpha^7 & \alpha^6 & \alpha^5 & \alpha^4 & \alpha^3 & \alpha^2 & \alpha \end{array}\right)$$

Similarly, the parity-check matrix of $\mathcal{C}^{(0)}(5,3;f_\alpha(x))$ is given by

$$\left(\begin{array}{ccc|ccc|ccc|ccc|ccc} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5 & \alpha^6 & \alpha^7 & \alpha^8 & \alpha^9 & \alpha^{10} & \alpha^{11} & \alpha^{12} & \alpha^{13} & \alpha^{14} \\ 1 & \alpha^{14} & \alpha^{13} & \alpha^6 & \alpha^5 & \alpha^4 & \alpha^{12} & \alpha^{11} & \alpha^{10} & \alpha^3 & \alpha^2 & \alpha & \alpha^9 & \alpha^8 & \alpha^7 \end{array}\right)$$

Let us point out that the construction of this type of codes is valid also over the ring of polynomials modulo $M_p(x)=1+x+\cdots+x^{p-1}$, p a prime number, as done with the Blaum-Roth (BR) codes [2]. In that case, $\mathcal{O}(\alpha)=p$, where $\alpha^{p-1}=1+\alpha+\cdots+\alpha^{p-2}$. The construction proceeds similarly, and we denote it $\mathcal{C}^{(0)}(m,n;M_p(x))$. Utilizing the ring modulo $M_p(x)$ allows for XOR operations at the encoding and the decoding without look-up tables in a finite field, which is advantageous in erasure decoding [2]. It is well known that $M_p(x)$ is irreducible if and only if 2 is primitive in $GF(p)$ [4].

Example 2.2 Consider the ring of polynomials modulo $M_{17}(x)$ and let α be an element in the ring such that $\alpha^{16} = 1 + \alpha + \cdots + \alpha^{15}$, thus, $\mathcal{O}(\alpha) = 17$ (notice, $M_{17}(x)$ is reducible). Then, the parity-check matrix of $\mathcal{C}^{(0)}(4, 4; M_{17}(x))$ is given by

$$\left(\begin{array}{cccc|cccc|cccc|cccc}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5 & \alpha^6 & \alpha^7 & \alpha^8 & \alpha^9 & \alpha^{10} & \alpha^{11} & \alpha^{12} & \alpha^{13} & \alpha^{14} & \alpha^{15} \\
1 & \alpha^{16} & \alpha^{15} & \alpha^{14} & \alpha^8 & \alpha^7 & \alpha^6 & \alpha^5 & \alpha^{16} & \alpha^{15} & \alpha^{14} & \alpha^{13} & \alpha^7 & \alpha^6 & \alpha^5 & \alpha^4
\end{array}\right)$$

We have the following theorem:

Theorem 2.1 Codes $\mathcal{C}^{(0)}(m, n; f_\alpha(x))$ and $\mathcal{C}^{(0)}(m, n; M_p(x))$ are SD codes.

Proof: According to Definition 1.1, we have to prove first that 3 erasures in the same row will be corrected. Based on the parity-check matrix of the code, this will happen if and only if, for any $0 \le i \le m-1$ and $0 \le j_0 < j_1 < j_2 \le n-1$,

$$\det\begin{pmatrix} 1 & 1 & 1 \\ \alpha^{in+j_0} & \alpha^{in+j_1} & \alpha^{in+j_2} \\ \alpha^{2in-j_0} & \alpha^{2in-j_1} & \alpha^{2in-j_2} \end{pmatrix} \neq 0$$

But the determinant of this 3×3 matrix can be easily transformed into a Vandermonde determinant on α^{j_0}, α^{j_1} and α^{j_2} times a power of α, so it is invertible in a field and also in the ring of polynomials modulo $M_p(x)$ [2].

Next we have to prove that if we have two erasures in locations i and j of row ℓ, say, $0 \le i < j \le n-1$, and two erasures in locations i' and j' of row ℓ', $0 \le i' < j' \le n-1$, $0 \le \ell < \ell' \le m-1$, such that, either $i = i'$, $i = j'$, $j' = i$ or $j = j'$, then

$$\det\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \alpha^{\ell n+i} & \alpha^{\ell n+j} & \alpha^{\ell' n+i'} & \alpha^{\ell' n+j'} \\ \alpha^{2\ell n-i} & \alpha^{2\ell n-j} & \alpha^{2\ell' n-i'} & \alpha^{2\ell' n-j'} \end{pmatrix} \neq 0$$

After some row manipulation, the inequality above holds if and only if

$$\det\begin{pmatrix} \alpha^{\ell n+i}\left(1 \oplus \alpha^{j-i}\right) & \alpha^{\ell' n+i'}\left(1 \oplus \alpha^{j'-i'}\right) \\ \alpha^{2\ell n-j}\left(1 \oplus \alpha^{j-i}\right) & \alpha^{2\ell' n-j'}\left(1 \oplus \alpha^{j'-i'}\right) \end{pmatrix} \neq 0.$$

$1 \oplus \alpha^{j-i}$ is invertible in $GF(q)$ since $1 \leq j-i < \mathcal{O}(\alpha)$, but the same is true in the polynomials modulo $M_p(x)$ [2], thus, the inequality above is satisfied if and only if

$$\det \begin{pmatrix} \alpha^{i} & \alpha^{i'} \\ \alpha^{-j} & \alpha^{(\ell'-\ell)n-j'} \end{pmatrix} \neq 0.$$

Assume that this determinant is 0. Redefining $\ell \leftarrow \ell' - \ell$, then $1 \leq \ell \leq m-1$ and we have

$$\alpha^{\ell n} = \alpha^{i'+j'-i-j}.$$

We will show that this is not possible. Assume that $i = i'$. Then,

$$\alpha^{\ell n} = \alpha^{j'-j}.$$

Assume that $j' \geq j$. Then, $\ell n = j' - j$, a contradiction since $j' - j \leq n-1$ and $n \leq \ell n < mn \leq \mathcal{O}(\alpha)$.

So, assume $j' < j$. Then, $\ell n = \mathcal{O}(\alpha) + j' - j$. But this also gives a contradiction, since $\ell n \leq mn - n \leq \mathcal{O}(\alpha) - n$, and $\mathcal{O}(\alpha) + j' - j \geq \mathcal{O}(\alpha) - n + 1$.

The cases $i = j'$, $j' = i$ and $j = j'$ are handled similarly. □

Next we show how to construct PMDS codes.

Let $\alpha \in GF(2^b)$ and $2mn \leq \mathcal{O}(\alpha)$. Consider the $(m+2) \times mn$ parity-check matrix given by (1) and, for $0 \leq i \leq m-1$,

$$\underline{c}_{in}, \underline{c}_{in+1}, \ldots, \underline{c}_{(i+1)n-1} = \begin{pmatrix} \underline{e}_i & \underline{e}_i & \ldots & \underline{e}_i & \ldots & \underline{e}_i \\ \alpha^{2in} & \alpha^{2in+1} & \ldots & \alpha^{2in+j} & \ldots & \alpha^{2(i+1)n-1} \\ \alpha^{4in} & \alpha^{4in-1} & \ldots & \alpha^{4in-j} & \ldots & \alpha^{(4i-1)n+1} \end{pmatrix} \quad (3)$$

We denote the $[mn, m(n-1)-2]$ code over $GF(q)$ whose parity-check matrix is given by (1) and (3) as $\mathcal{C}^{(1)}(m, n; f_\alpha(x))$. The same can be done with the ring of polynomials modulo $M_p(x)$, in which case we denote the code $\mathcal{C}^{(1)}(m, n; M_p(x))$.

Example 2.3 As in Example 2.2, consider the ring of polynomials modulo $M_{17}(x)$ and let α be an element in the ring such that $\mathcal{O}(\alpha) = 17$ and $\alpha^{16} = 1 + \alpha + \cdots + \alpha^{15}$. Then, the parity-check matrix of $\mathcal{C}^{(1)}(2, 4; M_{17}(x))$ is given by

$$\left(\begin{array}{cccc|cccc} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^8 & \alpha^9 & \alpha^{10} & \alpha^{11} \\ 1 & \alpha^{15} & \alpha^{14} & \alpha^{13} & \alpha^{16} & \alpha^{15} & \alpha^{14} & \alpha^{13} \end{array}\right)$$

Theorem 2.2 Codes $\mathcal{C}^{(1)}(m,n;\alpha;q)$ and $\mathcal{C}^{(1)}(m,n;M_p(x))$ are PMDS codes.

Proof: As in Theorem 2.1, we have to prove first that three erasures in the same row will always be corrected.

Based on the parity-check matrix of the code, this will happen if and only if, for any $0 \leq i \leq m-1$ and $0 \leq j_0 < j_1 < j_2 \leq n-1$,

$$\det\left(\begin{array}{ccc} 1 & 1 & 1 \\ \alpha^{2in+j_0} & \alpha^{2in+j_1} & \alpha^{2in+j_2} \\ \alpha^{4in-j_0} & \alpha^{4in-j_1} & \alpha^{4in-j_2} \end{array}\right) \neq 0$$

Again, the determinant of this 3×3 matrix can be transformed into a Vandermonde determinant on α^{j_0}, α^{j_1} and α^{j_2} times a power of α, so it is invertible in a field and also in the ring of polynomials modulo $M_p(x)$.

Next we have to prove that if we have two erasures in locations i and j of row ℓ, say, $0 \leq i < j \leq n-1$, and two erasures in locations i' and j' of row ℓ', $0 \leq i' < j' \leq n-1$, $0 \leq \ell < \ell' \leq m-1$, then

$$\det\left(\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \alpha^{2\ell n+i} & \alpha^{2\ell n+j} & \alpha^{2\ell' n+i'} & \alpha^{2\ell' n+j'} \\ \alpha^{4\ell n-i} & \alpha^{4\ell n-j} & \alpha^{4\ell' n-i'} & \alpha^{4\ell' n-j'} \end{array}\right) \neq 0$$

After some row manipulation, the inequality above holds if and only if

$$\det\left(\begin{array}{cc} \alpha^{2\ell n+i}\left(1 \oplus \alpha^{j-i}\right) & \alpha^{2\ell' n+i'}\left(1 \oplus \alpha^{j'-i'}\right) \\ \alpha^{4\ell n-j}\left(1 \oplus \alpha^{j-i}\right) & \alpha^{4\ell' n-j'}\left(1 \oplus \alpha^{j'-i'}\right) \end{array}\right) \neq 0$$

Again, $1 \oplus \alpha^{j-i}$ is invertible in $GF(q)$ and in the ring of polynomials modulo $M_p(x)$, thus, the inequality above is satisfied if and only if

$$\det\left(\begin{array}{cc} \alpha^{i} & \alpha^{i'} \\ \alpha^{-j} & \alpha^{2(\ell'-\ell)n-j'} \end{array}\right) \neq 0.$$

Assume that this determinant is 0. Redefining $\ell \leftarrow \ell' - \ell$, then $1 \leq \ell \leq m-1$ and we have

$$\alpha^{2\ell n} = \alpha^{i'+j'-i-j}.$$

But this is not possible. In effect, assume first that $i'+j' \geq i+j$. Then, since $2n \leq 2\ell n < \mathcal{O}(\alpha)$, we would have $2\ell n = i'+j'-i-j$, a contradiction since $i'+j'-i-j \leq 2(n-1)$.

So, assume $i'+j' < i+j$. Then, $2\ell n = \mathcal{O}(\alpha)+i'+j'-i-j$. This also gives a contradiction, since $2\ell n \leq 2mn - 2n \leq \mathcal{O}(\alpha) - 2n$, and $\mathcal{O}(\alpha)+i'+j'-i-j \geq \mathcal{O}(\alpha)-2n+2$. □

3 Conclusions

We have presented constructions of PMDS and SD codes extending RAID 5 with two extra parities, solving an open problem since previous constructions were based on computer search. It is an open problem to extend the results to more parities.

References

[1] M. Blaum, J. L. Hafner and S. Hetzler, "Partial-MDS Codes and their Application to RAID Type of Architectures," IBM Research Report, RJ10498, February 2012.

[2] M. Blaum and R. M. Roth, "New Array Codes for Multiple Phased Burst Correction," IEEE Trans. on Information Theory, vol. IT-39, pp. 66-77, January 1993.

[3] C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Li and S. Yekhanin, "Erasure Coding in Windows Azure Storage," 2012 USENIX Annual Technical Conference, Boston, Massachussetts, June 2012.

[4] F. J. MacWilliams and N. J. A. Sloane, "The Theory of Error-Correcting Codes," North Holland, Amsterdam, 1977.

[5] J. S. Plank, M. Blaum and J. L. Hafner, "SD Codes: Erasure Codes Designed for How Storage Systems Really Fail," FAST 13, San Jose, CA, February 2013.