An Extension for Construction of Systematic MDS Codes With Minimum Repair Bandwidth

Liang ZHAN, Songtao LIANG

Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433, P. R. China

11210240067@fudan.edu.cn, 11110240013@fudan.edu.cn

***Abstract*— Distributed storage systems based on erasure coding usually provide redundancy to increase the reliability and the storage efficiency. One main challenge in the construction of this kind of system is the repair problem: if a node storing encoded information fails, in order to maintain the same level of reliability, we need to create encoded information at a new node.**

Among various kinds of coding schemes presented in recent years, Wu puts forward a construction of systematic (n,k) -MDS codes for $2k \le n$ that achieves the minimum repair bandwidth when repairing from $k+1$ nodes. In this paper, we optimize Wu's method to make the repairing coefficients have more wide-ranging choices.

***Keywords*—Distributed storage, erasure coding, MDS code, minimum repair bandwidth, repair**

I. INTRODUCTION

In the past decade, the demand for distributed data storage has increased significantly due to the explosive growth of the amount of data all over the world. According to a study sponsored by the information storage company EMC, the volume of world's data grows even faster than Moore' Law [2], [3]. For instance, the storage space used for photo storage in Facebook was over $20PB$ in 2011 and was keeping increasing by $60TB$ every week [4].

At the same time, node failures in the distributed storage systems become a norm rather than a rare exception gradually. Therefore, redundancy must be introduced into the system in order to make it fault-tolerant. Between the two common strategies to provide redundancy, erasure coding is preferred to straightforward replication by reason of storage efficiency. Storage services like CleverSafe and Wuala have adopted the use of erasure codes in their systems(see e.g., [5], [6]).

As to erasure coding, Maximum Distance Separable (MDS) codes can achieve the optimal storage efficiency among various erasure codes. Given two positive integers n and $n > k$, an MDS code can be used for reliability: divide a file of size B into k fragments (each of size B/k) and encode them into n fragments (of the same size) with MDS code. In this way, any set of k coded fragments out of these n which provides the minimum data desired suffice to recover the original file (see Fig.1 for an example).

Despite the optimal storage efficiency of MDS codes, it is totally unnecessary to recover the whole original file if we only need to repair one failed node. In fact, one main challenge for distributed storage systems based on erasure coding is the repair problem: if a node storing encoded information fails, in order to maintain the same level of reliability, we need to create encoded information at a new node. The naive repair strategy based on (n,k) - MDS code needs the newcomer to download k encoded fragments from a subset of other live nodes to reconstruct the whole file, which generates a total repair bandwidth of B $(k \times B/k)$. Thus, the consideration of the repair bandwidth gives rise to new design challenges.

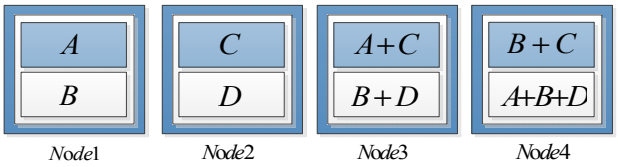


Fig. 1. A $(4,2)$ - MDS binary erasure code. Each storage node stores two fragments that are linear binary combinations of the original data fragments A, B, C, D. The total size is 4 fragments, and any 2 out of the 4 storage nodes suffice to recover the whole data.

There are mainly three kinds of repair models proposed to reduce repair bandwidth as illustrated in Fig.2: exact repair, functional repair, and exact repair of systematic parts.

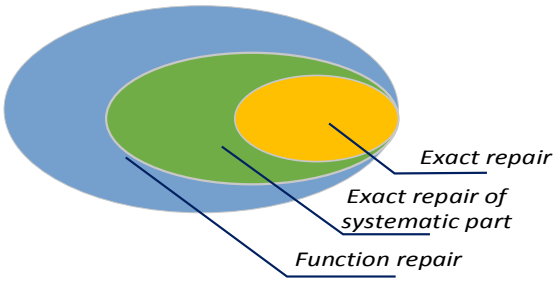


Fig. 2. Three typical repair models

In exact repair, the new node accesses some live nodes and exactly regenerates the lost fragments. In functional repair, the newly generated fragments may be different from the original ones as long as the MDS code property is still maintained after repair. The exact repair of the systematic part is a hybrid repair model in which the systematic part follows an exact repair while the nonsystematic part implements a functional

repair. Besides, the systematic property of the codes means one copy of the data exists in uncoded form. Obviously, the data retrieval process can be largely simplified with systematic codes because data can be read directly from the uncoded copy without requiring decoding in normal cases.

II. REVIEW: WU'S CONSTRUCTION

In this section, we review the construction proposed by Wu in [1]. To combine the advantages of both MDS property and the systematic property in practice, Wu presented a construction of systematic (n,k) -MDS codes for $2k \le n$ that achieves the minimum repair bandwidth when repairing from $k+1$ nodes. A hybrid exact and functional repair model was adopted in Wu's method so that the systematic symbols are exactly repaired and the nonsystematic parts follow a functional repair model.

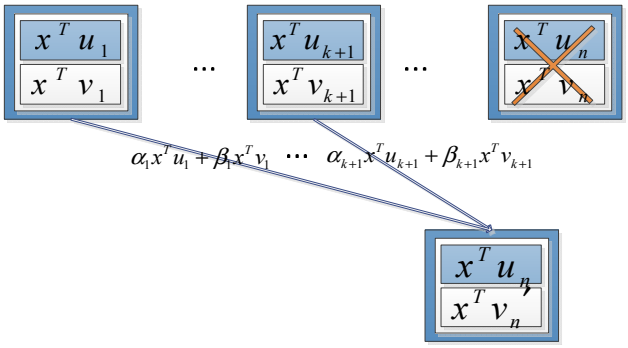


Fig. 3. Illustration of Wu's scheme

The proposed scheme of Wu's construction is illustrated in Fig.3. Suppose node n failed and is repaired by accessing the existing nodes $\{1,\ldots,k+1\}$. The new node downloads $\alpha_i x^T u_i + \beta_i x^T v_i$ from each node of $\{1,\ldots,k+1\}$. With these $k+1$ downloaded pieces, the new node computes two symbols $x^T u_n$ and $x^T v'_n$ as follows:

$$\sum_{i=1}^{k+1} (\alpha_i x^T u_i + \beta_i x^T v_i) = x^T u_n \quad (1)$$

$$\sum_{i=1}^{k+1} \rho_i (\alpha_i x^T u_i + \beta_i x^T v_i) = x^T v'_n \quad (2)$$

Given a finite field $\mathbb{F}$ whose size is greater than

$$d_0 = 2\binom{2n-1}{2k-1} \quad (3)$$

As Wu proved in [1], suppose the old code specified by $\{u_i, v_i\}$ is an $(2n,2k)$-MDS code defined over $\mathbb{F}$, when node n fails, there exists an assignment of the variables $\{\alpha_i, \beta_i, \rho_i\}$ such that (3) and (4) are satisfied and the repaired code continues to be an $(2n,2k)$-MDS code.

According to the proof in [1], the procedure of the code construction algorithm proposed by Wu presented as follows:

ALGORITHM 1. THE ALGORITHM FOR REGENERATING CODES

1) Initialize the code with any $(2n,2k)$ systematic MDS code over $\mathbb{F}$;
2) Draw a vector ξ from $\mathbb{F}^{2k}$;
3) Compute the resulting v'_n and check if it is linearly independent from the subset $\binom{2n-1}{2k-1}$ of $\{u_1,\ldots,u_n,v_1,\ldots,v_{n-1}\}$ with cardinality $2k-1$. If this property is satisfied, the process is over. If not, go back to step 2.

Let us compare Wu's scheme with other existing schemes. Without loss of generality, we only consider the case $d = k+1$ here. Generally speaking, in the network coding scheme [7] for the functional repair mode, it can't provide the systematic feature. In the interference alignment scheme for the exact repair model [8], the systematic symbols can only store in the first k nodes. As to the scheme of [9], although it can achieve the cut bound on total repair bandwidth, the MDS feature may lose as the code repairs.

III. EXTENSION TO WU'S CONSTRUCTION

In this section, we give an extension to the constructive scheme given by Wu in [1]. Let us look at a simplified example illustrated in Fig.4.

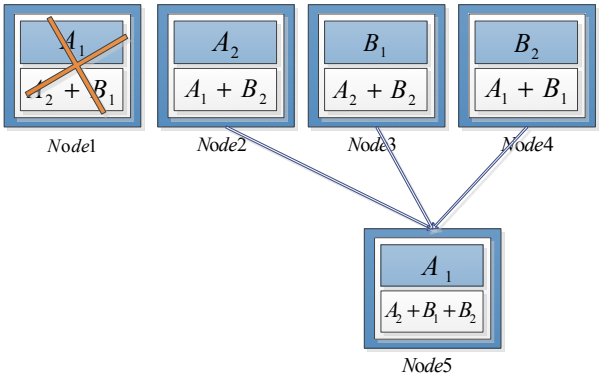


Fig. 4. A $(4,2)$ -systematic MDS code. Each storage node stores two fragments that are linear binary combinations of the original fragments A_1, A_2, B_1, B_2 of the data object.

TABLE I. PARAMETERS ASSIGNED IN TWO CONSTRUCTIONS

	$(\alpha_2,\beta_2,\omega_2,\rho_2)$	$(\alpha_3,\beta_3,\omega_3,\rho_3)$	$(\alpha_4,\beta_4,\omega_4,\rho_4)$	**Result**
Wu	$(0,1,1,?)$	$(0,0,1,?)$	$(1,0,1,?)$	False
Extend	$(0,1,1,0)$	$(1,1,0,1)$	$(1,0,1,0)$	True

The original data that consisted of fragments A_1, A_2, B_1, B_2 was encoded by a $(4,2)$ -systematic MDS code. Then, let the encoded fragments be stored in 4 nodes. Without loss of generality, suppose node1 failed and we use the Wu's scheme to conduct the repair process. The newcomer node5 should now access the existing 3 nodes to download the desired fragments. In order to satisfy the equation (1) above, we can consider $\alpha_2 = 0$ and $\beta_2 = 1$ at first as shown in Table1. Subsequently, we can figure out that $\alpha_3 = 0$, $\beta_3 = 0$, $\alpha_4 = 1$ and $\beta_4 = 0$. However, when we turn to repair the second row

of node1, we find no available vector $\{\alpha_i, \beta_i, \rho_i\}$ exist to satisfy the equation (2) any more though the encoded fragments in node3 suffice to provide the fragment required by node5.

Inspired by this case, we modify (1) and (2) by adding an extra parameter ω_i into them. So the two symbols $x^T u_n$ and $x^T v_n'$ can be computed as follows:

$$\sum_{i=1}^{k+1} \omega_i (\alpha_i x^T u_i + \beta_i x^T v_i) = x^T u_n \quad (4)$$

$$\sum_{i=1}^{k+1} \rho_i (\alpha_i x^T u_i + \beta_i x^T v_i) = x^T v'_n \quad (5)$$

Let us go back to the example illustrated in Fig. 4. We can still consider $\alpha_2 = 0$ and $\beta_2 = 1$. Since we have added a new parameter ω_i in (4), we can make ω_2 equal to 1. Subsequently, we can assign $\{\alpha_3, \beta_3, \omega_3, \alpha_4, \beta_4, \omega_4\}$ to $\{1,1,0,1,0,1\}$. Apparently, equation (4) can be satisfied in this way. Then, we can focus on the equation (5). If we assign $\{\rho_2, \rho_3, \rho_4\}$ to $\{0,1,0\}$, the second row of the node1 can be functional repaired with an assignment of $A_2 + B_1 + B_2$ properly and the repaired code continues to be a MDS code simultaneously.

With the discussion above, we further proposed a theorem extensive to Wu's in [1] as follows:

Theorem 1: Let $\mathbb{F}$ be a finite field whose size is greater than

$$d_0 = 3\binom{2n-1}{2k-1} \quad (6)$$

Suppose the old code specified by $\{u_i, v_i\}$ is an $(2n, 2k)$-MDS code defined over $\mathbb{F}$. When node n fails, there exists an assignment of the variables $\{\alpha_i, \beta_i, \omega_i, \rho_i\}$ such that (4) and (5) are satisfied and the repaired code continues to be an $(2n, 2k)$-MDS code.

Proof: We can prove theory 1 in a way similar to Wu's proof in [1]. Let us examine the equation (4) first. Introduce two vectors as follows:

$$\eta = [\alpha_1\omega_1, \beta_1\omega_1, ..., \alpha_{k+1}\omega_{k+1}, \beta_{k+1}\omega_{k+1}]^T \quad (7)$$

$$A = [u_1, v_1, ..., u_{k+1} v_{k+1}] \quad (8)$$

Let η_i denote the i-th entry of η and a_i denote the i-th entry of A. Then the equation (4) can be represented as follows:

$$A\eta = u_n \quad (9)$$

Since the $2n$ length-$2k$ vectors $\{u_i, v_i\}$ form an $(2n, 2k)$-MDS code as the premise mentioned above, any $2k$ columns of A should be linearly independent. So the $2k$ columns have a full rank of $2k$, i.e., $rank(A) = 2k$. Considering that η has $2k+2$ entries, the solutions to (9) have two degrees of freedom relatively. In order to get a unique solution to (9), we can particularly assign $\eta_1 = \alpha_1\omega_1, \eta_2 = \beta_1\omega_1$ so that any other η_i can be uniquely determined by η_1 and η_2. Subsequently, let the variables $\{\omega_2, ..., \omega_{k+1}\}$ be collectively represented by a vector γ with k entries over the finite field $\mathbb{F}$. Thereby, we can finally calculate $\{\alpha_2, \beta_2, ..., \alpha_{k+1}, \beta_{k+1}\}$ with a random assignment of γ in $\mathbb{F}^k$ relatively.

After considering (4), $k+4$ degrees of freedom are remained for us in total. Represent the variables $\{\alpha_1, \beta_1, \omega_1, \rho_1, ..., \rho_{k+1}\}$ with a vector ξ which contains $k+4$ entries in $\mathbb{F}$. From (5), v_n' can be determined as

$$v'_n = \sum_{i=1}^{k+1} \rho_i (\alpha_i u_i + \beta_i v_i) \quad (10)$$

Note that $\{\alpha_i, \beta_i\}$ are affine functions of $\alpha_1\omega_1$ and $\alpha_2\omega_2$. Thus, each entry of v_n' is a multivariate polynomial in ξ with a total degree no more than 3.

From the analysis above, it remains to prove that we can choose a vector $\xi \in \mathbb{F}^{k+4}$ so that the repaired node $\{u_1, ..., u_n, v_1, ..., v_{n-1}, v_n'\}$ continues to be an $(2n, 2k)$-MDS code. In fact, it is equivalent to prove v_n' can be made linearly independent of any $2k-1$ subset of $U = \{u_1, ..., u_n, v_1, ..., v_{n-1}\}$. For any $2k-1$ subset S of $\{1, ..., 2n-1\}$, let U_S denote the $2k \times (2k-1)$ matrix whose columns are given by the vectors in U indexed by S. Then we can get the condition below:

$$\prod_{S \subset \{1,...,2n-1\}, |S|=2k-1} \det([U_S, v'_n]) \neq 0 \quad (11)$$

From (10) and the analysis above, the left hand side of (11) is equivalent to a multivariate polynomial in ξ whose total degree is at most $3\binom{2n-1}{2k-1}$.

Claim 1: For any $S \subset \{1, ..., 2n-1\}$ with $S = 2k-1$, $\det([U_S, v'_n]) \neq 0$ for some $\xi \in \mathbb{F}^{k+4}$.

Proof of Claim: the matrix U_S can be viewed as a set of $2k-1$ column vectors. Thus there must exist a node, say i^*, in $1, ..., k+1$, satisfying either $u_{i^*} \notin U_S$ or $v_{i^*} \notin U_S$ or both.

Without loss of generality, suppose $u_{i^*} \notin U_S$ for $i^* \in \{1, ..., k+1\}$. Let $\alpha_{i^*} = 1$, $\beta_{i^*} = 0$, $\omega_{i^*} = 1$, then there exists a unique solution to (9). Further, we can assign $\rho_{i^*} = 1$ and all the other $\rho_i = 0$. Thus, $v_n' = u_{i^*}$. Note that $u_{i^*} \notin U_S$ and the old code $\{u_1, v_1, ..., u_n, v_n\}$ was an $(2n, 2k)$-MDS code, so we can get the relationship below with this choice of ξ:

$$\det([U_S, v'_n]) = \det([U_S, u_{i^*}]) \neq 0 \quad (12)$$

The case $v_{i^*} \notin U_S$ follows similarly. ■

Claim 1 is proved based on the implied condition that $\det([U_S, v'_n])$ is a nonzero multivariate polynomial in ξ, i.e., the left hand side of (11) is a nonzero multivariate polynomial in ξ. According to the Schwartz-Zippel Theorem (quoted below as Lemma 1), for a finite field whose size is greater

than d_0 , there is an assignment of $\xi \in \mathbb{F}^{k+4}$ such that (11) holds. Therefore, Theorem1 follows. ■

Lemma 1: (Schwartz-Zippel Theorem (see, E.g., [10])): Let $P \in \mathbb{F}[x_1,...,x_n]$ be a non-zero multivariate polynomial of total degree (the total degree is the maximum degree of the additive terms and the degree of a term is the sum of exponents of the variables) $d_0 \geq 0$ over a field $\mathbb{F}$. Let $\mathbb{S}$ be a finite subset of $\mathbb{F}$, i.e., $\mathbb{S} \subseteq \mathbb{F}$ and let $r_1, r_2, \ldots, r_n$ be selected independently and uniformly at random from $\mathbb{S}$. Then

$$\Pr[P(r_1, r_2, ..., r_n) = 0] \leq \frac{d_0}{|\mathbb{S}|} \quad . \tag{13}$$

As is proved in [1], the above scheme can also give a construction of systematic (n,k) -MDS codes for $2k \leq n$ that achieves the minimum repair bandwidth when repairing from $k+1$ nodes.

A. Code Construction Algorithm

According to the proof of Theorem 1 and the Schwartz-Zippel Theorem, if we uniformly and independently draw each entry of ξ from a sufficiently large finite field $\mathbb{F}$, then (11) will establish with high probability. Thus, we can initialize the code using any $(2n,2k)$ systematic MDS code over a finite field $\mathbb{F}$ at first. Then, draw a vector $\xi = \{\alpha_1, \beta_1, \omega_1, \rho_1, ..., \rho_{k+1}\}$ from $\mathbb{F}^{k+4}$ and another vector $\gamma = \{\omega_2, ..., \omega_{k+1}\}$ from $\mathbb{F}^k$. Next, we should check if the resulting v_n' maintains the $(2n,2k)$ -MDS code property. Repeat the he random drawing process until the desired property is met.

B. Performance Evaluation Compared With Wu's

From the discussion above, we can find that the proposed extension to Wu's construction can provide us more degrees of freedom when choosing the repair coefficients $\{\alpha_i, \beta_i, \omega_i, \rho_i\}$. Particularly, as the example illustrated above, Wu's construction may become inefficient under some special conditions where the extended construction can perform better yet. It implies that the proposed extensive construction can provide us more combination choices among the $k+1$ nodes connected by the new node. Considering the practical network situations and the requirements in different applications, we can benefit quite more from the more wide-ranging repair coefficients.

However, the proposed extension construction itself may require a larger finite field as shown in the Theorem 1 above compared with Wu's construction. Thus, there are both advantages and disadvantages in them. Finally, it is worth mentioning that the proposed extended construction can be easily specialized to Wu's construction with an assignment of $\{\omega_1, ..., \omega_{k+1}\} = \{1, ..., 1\}$ if necessary.

IV. CONCLUSION

In this paper, we have introduced the repair problem which exists in distributed storage systems based on erasure coding. Various repair strategies have been proposed to reduce the repair bandwidth. Particularly, we focus on a construction presented by Wu which combines the advantages of both MDS property and the systematic property in practice and achieves the minimum repair bandwidth when repairing from $k+1$ nodes. Based on Wu's construction, we have proposed an extension in this paper which optimizes Wu's method to make the repairing coefficients more wide-ranging so that more combination choices among the $k+1$ nodes connected by the new node can be provided. This extended construction can obviously make up for the deficiency of Wu's construction under some special conditions as illustrated in Fig. 4 and surely provide more benefits when considering the actual network situations. In other words, the extended construction proposed by us is more generalized compared with Wu's. Further study on detailed application conditions is required.

REFERENCES

[1] Y. Wu. (2009, Aug.). A construction of systematic MDS codes with minimum repair bandwidth. IEEE Trans. Inf. Theory. [Online]. Available: http://arxiv.org/abs/0910.2486.

[2] World's data more than doubling every two years — Driving Big Data opportunity, new IT roles. Available: http://www.emc.com/about/news/press/2011/20110628-01.htm.

[3] IDC says world's storage is breaking Moore's law, more than doubling every two years, http://enterprise.media.seagate.com/2011/06/insideit-storage/idc-says-worlds-storage-is-breaking-mooreslaw-more-than doubling-every-two-years/, 2012.

[4] D. Beaver, S. Kumar, H. C. Li, J. Sobel, and P. Vajgel, Finding a needle in Haystack: Facebook's photo storage, in Proc. 9th USENIX Conference on Operating Systems Design and Implementation (OSDI), 2010.

[5] The Coding for Distributed Storage wiki. Available: http://tinyurl.com/storagecoding.

[6] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, "A survey on network codes for distributed storage," in IEEE Proceedings, vol. 99, pp. 476 – 489, Mar. 2011.

[7] Y. Wu, "Existence and construction of capacity-achieving network codes for distributed storage," presented at the IEEE Int. Symp. Information Theory (ISIT), Seoul, Korea, Jun. 2009.

[8] Y.Wu and A. G. Dimakis, "Reducing repair traffic for erasure coding-based storage via interference alignment," presented at the IEEE Int. Symp. Information Theory (ISIT), Seoul, Korea, Jun. 2009.

[9] K. V. Rashmi, N. B. Shah, P. V. Kumar, and K. Ramchandran, Exact Regenerating Codes for Distributed Storage Jun. 2009 [Online]. Available: http://arxiv.org/abs/0906.4913.

[10] R. Motwani and P. Raghavan, *Randomized Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 1995.

[11] Haibin Kan and Hong Shen, A relation between the characteristic generators of a linear code and its dual, IEEE Transactions on Information Theory, Vol. 51, No. 3, March 2005.

[12] Haibin Kan and Hong Shen, A counterexample for the open problem on the minimal delay of orthogonal designs with maximal rates, IEEE Transactions on Information Theory, Vol. 51, No. 1, January 2005.

[13] C. Yuan and Haibin Kan, A characterization of solvability for a kind of networks, Science in China(F), Vol.55, No.4, 747-754, 2012.

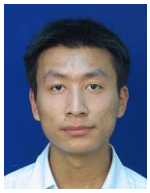
[14] Yuan Li and Haibin Kan, Complex Orthogonal Designs with Forbidden 2×2 Submatrices, IEEE Transactions on Information Theory, Vol. 58, No. 7, July 2012.

[15] C. Yuan, Haibin Kan, X. Wang, H. Imai, A construction method of matroidal networks, Science in China(F), Vol.55, No.11, 2445-2453, 2012.

[16] Haibin Kan and Hong Shen, Lower bounds of the minimal delays of complex orthogonal designs with maximal rates, IEEE Transactions on Communications, Vol. 54, No. 3, March 2006.

Liang Zhan received his bachelor's degree in Architecture from Huazhong University of Science and technology, China, in 2010. He is now pursuing a master's degree at the School of Computer Science in Fudan University. His current research interests include network coding and distributed storage.

Mr. Zhan became a member of IEEE in 2014. He is now a member of Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, P. R. China.



Songtao Liang received his BS and MS degrees from the School of Computer Science, Harbin Institute of Technology and Fudan university, China, in 2008 and 2011. He is currently pursuing PhD degree at the School of Computer in Fudan university. His research interests include network coding and distributed storage.