# 基于EEOD码的一种有效的数据分布策略

万武南，索 望，张金全，陈 运

(成都信息工程学院网络工程系 成都 610225)

【摘要】在EVENODD码的基础上，提出一种新的基于EEOD码的数据布局分布策略。该策略只需要三个额外的存储设备保存校验信息，能容许任意三个存储设备同时故障。策略中的校验方程组用图的回路表示，可将其顶点逐步消除，把译码过程归结为图回路的叠加。讨论了基于EEOD码数据分布策略的性能，与其他策略相比，其容灾能力大幅度提高，编码和译码过程只需要简单的异或运算，但空间利用率和系统吞吐量的影响非常小。

关 键 词 数据分布策略； EEOD码； EVENODD码； 可靠性
中图分类号 TP333 文献标识码 A

# An Effective Data Distribution Strategy Based on EEOD Codes

WAN Wu-nan，SUO Wang，ZHANG Jin-quan，CHEN Yun

(Department of Network Engineering, Chengdu University of Information Technology Chengdu 610225)

**Abstract** In this paper, a new data distribution strategy is described, which called the extend EVENODD codes (EEOD codes). It can recover any triple storage nodes failures with only three extra nodes for parity information. The equations are represented by circle paths of graph. Detailed EEOD code's decoding algorithms are proposed for correcting various triple node failures with successive elimination of graph vertex. It shows that the decoding complexity of the EEOD code is much lower than those of the existing comparable codes. The EEOD code is very valuable for storage systems which need higher reliability.

**Key words** data distribution strategy; EEOD codes; EVENODD codes; reliability

可靠性是分布式存储系统最重要的指标之一，也是当今信息社会对信息存储的迫切需求[1]。文献[2]提出了利用编码技术构建高可靠性的分布式存储系统，其基本思想为：一个目标文件可划分为$m$块，然后利用某编码把$m$块编码为$n$块，并分别存储在$n$个不同的存储设备上；当发生特殊状况(自然灾害、战争、意外损坏等)造成其中某些存储设备部分或者全部发生损坏时，可以通过存放在该系统其他存储设备上经过编码的冗余数据恢复损坏的数据，从而增强系统的安全性。

目前，容许两个存储设备同时故障的数据分布策略的编码有EVENODD码[3]、X码、B码[4]、S码[5]等；容许多个设备同时故障，特别是容许三个设备同时故障的数据分布策略的编码，如Blaum码[6]、WEAVER码[7]、HoVer码[8]、HDD1码和HDD2码[9]等。但是Blaum码解码方法是解多项式环上的一组线性方程，解码算法不易实现、复杂度高。HoVer码、WEAVER码不是MDS码，冗余盘数目并不是随着磁盘阵列系统总盘数的线性增长，冗余信息量太大、代价昂贵。而HDD1码和HDD2码的解码过程需要做线性方程高斯消元求解，其解码复杂度等于9。

本文在EVENODD码的基础上，提出了一种扩展EVENODD码——EEOD码，能容许任意三个存储设备同时故障，同时校验方程组用图的回路表示，译码过程可看作图回路的叠加。与其他策略相比，冗余率达到最优，编译码复杂度和更新复杂度都相对较低。

## 1 EEOD码的编码方法

### 1.1 EVENODD码

为了能够承受三个磁盘同时故障，EEOD在EVENODD码的基础上进行了扩展，其编码矩阵为$m$+3列，行为$m$–1列，其中前$m$列存放原始数据，后3列存放冗余校验数据。EEOD码的前两列冗余的构

收稿日期：2007 – 09 – 10
基金项目：现代通信国家重点实验室基金资助项目(9140C1101050705)；四川省教育厅科研基金资助项目(2006C033)
作者简介：万武南(1978 – )，女，博士，主要从事信息安全、编码理论等方面的研究；索望(1978 –)，男，讲师，主要从事信息安全、3G通信等方面的研究；张金全(1974 – )，男，讲师，主要从事信息安全等方面的研究；陈 运(1958 –)，女，教授，主要从事信息安全、3G通信等方面的研究.

造与EVENODD完全一样，后一列为增加1列冗余校验列，则三列冗余校验位构造公式为：

$$c_{u,m}=\sum_{t=0}^{m-1}c_{u,t} \qquad S_1=\sum_{t=1}^{m-1}c_{m-1-t,t}$$

$$c_{u,m+1}=S_1+\sum_{\substack{t=0\\ t\neq u-1}}^{m-1}c_{\langle u-t\rangle_m,t} \qquad S_2=\sum_{t=1}^{m-1}c_{\langle m-1-2t\rangle_m,t}$$

$$c_{u,m+2}=S_2+\sum_{\substack{t=0\\ \langle u-2t\rangle_m\neq m-1}}^{m-1}c_{\langle u-2t\rangle_m,t}$$

式中　$m$的值为大于等于2的素数；$\langle a\rangle_m=a(\mathrm{mod}\,m)$；$c_{i,j}$为第$j$列第$i$行的信息位或校验位。

### 1.2　几何描述

为了更好描述EEOD码，下面从几何角度来描述编码过程。(8,5,4)EEOD码三列校验列编码几何示意图如图1所示。图中，三校验列的每个校验位分别是沿斜率(0,1,2)经过的格点对应的信息位异或运算的值，黑点的坐标点对应的信息位为零，即$c_{4,u}=0$，$0\le u\le m-1$(实际并不存在)，称为虚拟信息位；白点表示信息位。虚线表示调节因子$S$，实线表示校验列的每个校验位(与调节因子$S$异或之前的值)。

EEOD码的每个二维阵列码字$C$的信息位可作为平面坐标上的格点，横坐标表示信息位的列号，纵坐标表示信息位的行号，且坐标轴的取值是模$m$运算，取值范围为$0,1,\cdots,m-1$。如某点坐标为(1，2)，该点表示码字$C$的$c_{1,2}$信息位，即对应的二维阵列中第2列第3行的信息位。其中，坐标轴上纵坐标为$m-1$的坐标点，其对应信息位的值全为零，是增加的虚拟坐标点。

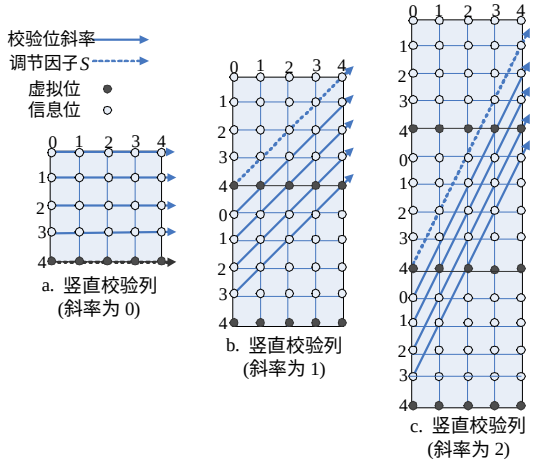


图1　(8,5,4)EEOD码的编码几何示意图

每列校验位调节因子$S_i$的值从坐标轴$(0,m-1)$格点开始，沿某固定斜率的直线经过的格点对应的信息位异或运算的值。校验列的第$i$个校验位则是从$(0，i)$格点开始，沿某固定斜率的直线经过的格点对应的信息位异或运算的值，再与该校验列调节因子$S_i$的值进行异或运算的值，EEOD码的3列校验列的斜率分别为$(0,1,2)$。

从几何的角度看，EEOD码是在EVENODD码的基础上增加了其斜率为2的校验列，因而EEOD码的3列校验列的斜率分别为$(0,1,2)$。(8,5,4)EEOD码三列校验列编码几何示意图如图1所示。图中三校验列的每个校验位分别是沿斜率(0,1,2)经过的格点对应的信息位异或运算的值。

## 2　EEOD码的译码算法

EEOD码译码主要可分为两种情况：第一种为出错列为两列的情况，这与EVENODD码纠两列删除错译码算法类似；第二种为出错列为三列的情况，若出错三列中有一列为校验列，与前一种类似。本文讨论最复杂的一种情况，出错的三列全部为信息列的译码过程。

设丢失的信息列分别为$i$、$j$、$k$三列，$0\le i<j<k\le m-1$，这种情况的详细译码过程如下：首先通过$m$、$m+1$、$m+2$三校验列计算出调节因子$S'$、$S_1$、$S_2$，即$S'=\sum_{u=0}^{m-2}c_{u,m}$，$S_1=S'+\sum_{u=0}^{m-2}c_{u,m+1}$，$S_2=S'+\sum_{u=0}^{m-2}c_{u,m+2}$。则可根据以下计算公式：$\tilde{S}_u^{(0)}=\sum_{\substack{t=0\\ t\neq i,j,k}}^{m}c_{u,t}$，

$$\tilde{S}_u^{(1)}=S_1+c_{u,m+1}+\sum_{\substack{t=0\\ t\neq i,j,k}}^{m-1}c_{\langle u-t\rangle_m,t}\quad,\quad \tilde{S}_u^{(2)}=S_2+c_{u,m+2}+\sum_{\substack{t=0\\ t\neq i,j,k}}^{m-1}c_{\langle u-2t\rangle_m,t}$$

，得到只含有$i$、$j$、$k$三列信息位的校验算子，分别为$\tilde{S}^{(0)}=\tilde{S}_0^{(0)}$，$\tilde{S}_1^{(0)},\cdots,\tilde{S}_{m-1}^{(0)}$；$\tilde{S}^{(1)}=\tilde{S}_0^{(1)}$，$\tilde{S}_1^{(1)},\cdots,\tilde{S}_{m-1}^{(1)}$；$\tilde{S}^{(2)}=\tilde{S}_0^{(2)},\tilde{S}_1^{(2)},\cdots,\tilde{S}_{m-1}^{(2)}$，$0\le u\le m-1$。

其译码思想如下：从解线性方程组的角度可知，校验算子$\tilde{S}^{(0)}$、$\tilde{S}^{(1)}$、$\tilde{S}^{(2)}$作为一组线性方程组，其中$c_{u,i}$、$c_{u,j}$、$c_{u,k}$为未知变量。译码过程则是解这组线性方程。而$\tilde{S}^{(0)}$、$\tilde{S}^{(1)}$、$\tilde{S}^{(2)}$的每个方程中至少含有两个未知变量，不能直接求解未知变量。通过一定的变换，把原方程组转化为一组只含有$j$列未知变量的循环方程组，记$\tilde{S}'_u$，$0\le u\le m-2$；并且方程组中每个方程至多只含有两个未知变量，其中第一个方程$\tilde{S}'_u$中的两个未知变量中含有信息位$c_{m-1,j}$，而$c_{m-1,j}=0$，实际只有一个未知变量。因此可以通过这组循环方程组依次求解出第$j$列的未知变量，即求

解出第$j$列的信息位。则原方程组变换为循环方程组的过程，称为“消元过程”。消元过程可依次第$i$、$k$两列的未知变量消掉，只剩下第$j$列的未知变量。

例 (8,5,4)EEOD码消元过程实例。设出错列的分别为$i$=1、$j$=2、$k$=4和$i$=0、$j$=2、$k$=4。

图2给出了通过回路叠加恢复中间列信息位的消元过程。第一种情况称为非对称出错，即$j-i\neq k-j$，如图2a所示，通过图的回路叠加，依次把第1、4列的信息位消掉，只剩下第2列信息位，并且最终得到至多只含有第2列两个未知的变量的一组循环方程组，具体过程如下：根据图2a可知，$a\to b\to c\to d\to a$称为一条回路。每条回路经过$i$、$k$列的未知变量的度都为2，经过$j$列的未知变量的度为1，只含两个未知变量的方程可由两条回路得到。

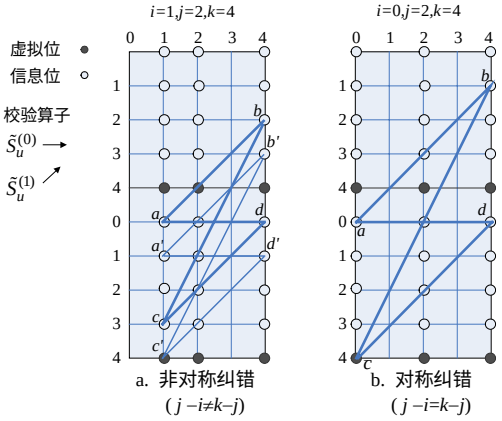


图2 (8,5,4)EEOD码的恢复中间列译码消元过程

$a\to b\to c\to d\to a$回路对应校验方程的变换为：$c_{4,2}+c_{1,2}+c_{0,2}+c_{2,2}=\tilde{S}_1^{(1)}+\tilde{S}_0^{(2)}+\tilde{S}_4^{(1)}+\tilde{S}_0^{(0)}$；$a'\to b'\to c'\to d'\to a'$回路则对应的校验方程的变换为：$c_{0,2}+c_{1,2}+c_{2,2}+c_{3,2}=\tilde{S}_2^{(1)}+\tilde{S}_1^{(2)}+S_0'^{(1)}+S_0'^{(0)}$。然后两条回路上的未知变量叠加就可以得到循环方程组的第一个方程为：

$$\tilde{S}_0'=c_{4,2}+c_{3,2}=\tilde{S}_2^{(1)}+\tilde{S}_1^{(2)}+\tilde{S}_0^{(1)}+\tilde{S}_1^{(0)}+\tilde{S}_1^{(1)}+\tilde{S}_0^{(2)}+\tilde{S}_4^{(1)}+\tilde{S}_0^{(0)}$$

依照此方法可依次得到至多只含$j$列两未知变量的循环方程组为：

$$\tilde{S}_1'=c_{3,2}+c_{2,2}=\tilde{S}_0^{(1)}+\tilde{S}_4^{(2)}+\tilde{S}_3^{(1)}+\tilde{S}_4^{(0)}+\tilde{S}_1^{(1)}+\tilde{S}_0^{(2)}+\tilde{S}_4^{(1)}+\tilde{S}_0^{(0)}$$

$$\tilde{S}_2'=c_{2,2}+c_{1,2}=\tilde{S}_4^{(1)}+\tilde{S}_3^{(2)}+\tilde{S}_2^{(1)}+\tilde{S}_3^{(0)}+\tilde{S}_0^{(1)}+\tilde{S}_4^{(2)}+\tilde{S}_3^{(1)}+\tilde{S}_4^{(0)}$$

$$\tilde{S}_3'=c_{1,2}+c_{0,2}=\tilde{S}_3^{(1)}+\tilde{S}_2^{(2)}+\tilde{S}_1^{(1)}+\tilde{S}_2^{(0)}+\tilde{S}_4^{(1)}+\tilde{S}_3^{(2)}+\tilde{S}_2^{(1)}+\tilde{S}_3^{(0)}$$

因此根据循环方程组依次把$j$列信息列恢复出来，然后可根据EVENODD码译码算法恢复出丢失的$i$、$k$两列。

若出错的三信息列$i$、$j$、$k$满足$j-i=k-j$，则称为对称出错，如图2b所示，则消元过程可简化，只需要一条回路就能得到循环方程组。在图2b中，$a\to b\to c\to d\to a$一条回路就可以得到只含有$j$列的两个未知变量的循环方程组的一个方程。

## 3 EEOD码的性能分析

在数据分布策略中，编译码复杂度是衡量数据分布策略性能的一个重要的参数。EEOD码与其他阵列码一样，编译码过程主要是异或运算，因此码的编码复杂度通可定义为每个码字编码总的异或次数与码字信息位的总比特位之比。EEOD码整个编码过程需要总异或次数为$b(m-1)^2+2(b(m-1)^2+b(m-2))$；而EEOD码总的信息位比特数为$(m-1)mb$，则EEOD的编码复杂度，即每个比特需要的异或次数为$3-(m+1)/(m-1)m$。随着$m$值的增大，EEOD码的每比特信息位需要异或的次数接近3。

根据EEOD码纠三列信息列译码算法，首先计算两列垂直校验列的共同调节因子需要$2(m-1)+m-2$次异或运算，计算$\tilde{S}^{(0)}$、$\tilde{S}^{(1)}$、$\tilde{S}^{(2)}$三列校验算子需要$3(m-3)(m-1)+2(m-1)$；然后采用消元算法恢复第$j$列的所有信息位需要异或次数为$(4l_d-1)(m-1)+(m-2)$；最后恢复出其余两列的全部信息位则需要异或次数为$4(m-1)-1$，EEOD码纠三列信息列译码过程总得译码大约次数为$(4l_d-2+3m)(m-1)-3$，因此译码复杂度为$3+\dfrac{(4l_d-2)(m-1)-3}{m(m-1)}$。

本文比较EEOD码与其Blaum码及基于XOR的复损码译码复杂度。EEOD码的译码过程$l_d$依靠具体丢失信息列的位置确定。为了便于分析，$l_d$取平均值。设码的信息列为$m$列，校验列为3列。Blaum码纠三列信息列译码总异或次数为$(3m+21)(m-1)$。基于XOR的复损码译码过程中需要异或总的次数为$krL^2$。其中，$k$、$r$分别为码的信息位和冗余位；$L$为有限域的大小GF($2^L$)；忽略有限域的操作为$r^2$。

从表1可以看出，码的长度比较短时(信息列的列数)，EEOD码的每个信息位需要的异或次数最少，译码性能最好。而RS类纠删码译码复杂度最高，并且随域的扩大而增加。随着码长的变大，EEOD码每个信息位恢复需要的异或次数逐渐接近4。Blaum码单从译码复杂度考虑优于EEOD码，但其译码算法的实现比较复杂，不易于软件实现。

表1　三类编码每比特译码所需异或次数

| 磁盘数 | Blaum码 | EEOD码 | RS码 |
|---|---|---|---|
| 3 | 10.000 | 3.670 | 9 |
| 5 | 7.200 | 3.620 | 9 |
| 7 | 6.000 | 3.680 | 9 |
| 11 | 4.909 | 3.810 | 12 |
| 13 | 4.615 | 3.846 | 12 |
| 17 | 4.235 | 3.882 | 15 |
| 19 | 4.105 | 3.894 | 15 |
| 23 | 3.913 | 3.850 | 15 |
| 29 | 3.740 | 3.931 | 15 |
| 31 | 3.678 | 3.935 | 15 |

## 4　结 束 语

在分布式存储系统中，有较高的可靠性和吞吐量、较好的I/O性能及简单的编译码算法的数据分布策略具有重要的实际意义和理论研究价值。在EVENODD码基础上，本文提出了一类新的基于EEOD码的数据分布策略,不但可以容许任意三个设备出错，而且冗余率达到最优，编译码复杂度和更新复杂度都相对较低。适用于实时性较强的连续存储，如摄影图像存储等，特别对巨型阵列优势较大。

## 参 考 文 献

[1] XIN Q, MILLER E L, SCHWARZ T J. Reliability mechanisms for very large storage systems[C]//In Proceedings of the 20th IEEE /11th NASA Goddard Conference on Mass Storage Systems and Technologies. Los Alamitos, Calif: IEEE Computer Society Press, 2003.

[2] PATTERSON D A, GIBSON G A, KATZ R H. A case for redundant arrays of inexpensive disks(RAID)[C]//In Proc. of International Conference on Management of Data (SIGMOD). Chicago IL: ACM Press, 1988.

[3] BLAUM M, BRADY J, BRUCK J, et al. EVENODD: an efficient scheme for tolerating double disk failures in RAID architectures[J]. IEEE Trans. Comput, 1995, 44(2) :192-202.

[4] XU L. Highly available distributed storage systems[D]. Pasadena, California: California Institute of Technology, 1997.

[5] KATTI R, RUAN Xiao-yu. S-code: New distance-3 MDS array codes with optimal encoding[C]//In: Proceedings of IEEE ICASSP'05. Piscataway, NJ: IEEE Signal Processing Society, 2005.

[6] BLAUM M, BRUCK J, VARDY A. MDS array codes with independent parity symbols[J]. IEEE Trans. Inform. Theory, 1996, 42(2): 529-542.

[7] HAFNER J L. Hover erasure codes for disk arrays[C]//In DSN-2006: the International Conference on Dependable Systems and Networks. Washington DC: IEEE Computer Society, 2006.

[8] HAFNER J L. Weaver codes: highly fault tolerant erasure codes for storage systems[C]//FAST-2005: 4th Usenix Conference on File and Storage Technologies. San Francisco, CA: USENIX Association Berkeley, 2005.

[9] TAU Chih-shing, WANG T I. Efficient parity placement schemes for tolerating triple disk failures in raid architectures[C]//In Proceedings of the 17th International Conference on Advanced Information Networking and Applications(AINA'03). Washington DC: IEEE Computer Society, 2003.

编　辑　黄　莘