容许多个磁盘故障的RAID编码方法研究

刘 军[1],刘 璟[2]
(1 天津财经大学 理工学院,天津 300222;2 南开大学 信息技术科学学院,天津 300071)

摘 要:随着磁盘阵列规模的增大,同时发生多个磁盘故障的概率将大大增加,单容错编码难以满足应用对高可靠性存储的需求.分析了主要的双容错RAID编码方法及其特点,对各种双容错编码方法的冗余性能进行了比较.给出了一种基于循环置换矩阵构建的能容许三个磁盘故障的MDS交换群阵列码,其编码和解码效率较高,是大规模RAID存储系统的应用方向.

关键词:RAID编码;MDS阵列码;多磁盘故障
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1000-7180(2011)08-0009-03

Research on RAID Coding Schemes for Tolerating Multiple Disk Failures

LIU Jun[1], LIU Jing[2]
(1 College of Technology, Tianjin University of Finance and Economics, Tianjin 300222, China;
2 College of Information Technical Science, Nankai University, Tianjin 300071, China)

Abstract: As the Redundant Arrays of Inexpensive Disks (RAID) scale up, multiple disk failures are likely to take place at the same time. The single erasure code is hard to meet the requirements of high reliable storage. We reviews the coding schemes and features of various MDS array codes for tolerating up to double disk failures and compares their redundancy performances. A class of MDS Abelian group array codes for tolerating up to three disk failures is presented based on circular permutation matrices. Since the encoding and decoding are very efficient, it becomes the trend of application in large-scale RAID storage systems.

Key words: RAID coding; MDS array codes; multiple disk failures

1 引言

随着存储系统容量的增大,构成磁盘阵列的磁盘数目越来越多,考虑到磁盘的物理特性,一旦RAID系统中出现一个磁盘故障,另一个盘极可能在最近发生故障[1].这样,多个磁盘故障很容易同时发生,特别是在那些对可靠性要求极高的大规模海量信息存储系统.单容错RAID编码将会面临磁盘阵列规模增大、系统崩溃、不可恢复位故障、磁盘故障的相关性等问题,这使得多容错RAID编码方法研究成为大规模磁盘阵列存储系统领域的重要方向.

2 双容错RAID编码方法

关于容许两个磁盘故障的RAID编码方法,研究人员已经做了大量的工作,典型的编码有Reed-Solomon[2]、EVEN-ODD[3]、DH1和DH2[4]、RM2[5]、X-codes[6]、B-codes[7]等,这些都属于MDS类编码(Maximal Distance Separable codes).

MDS编码磁盘阵列的优点是磁盘冗余度最优,无论磁盘阵列规模大小,只需f个磁盘保存校验数据即可使磁盘阵列具有容许f个磁盘故障的能力.但是其校验计算和解码过程较为复杂,需要使用专用硬件来辅助编码和解码计算,才能获得较好的

收稿日期:2011-05-16;**修回日期**:2011-06-23

性能.

EVEN-ODD 编码[3]是最早实现了仅运用简单奇偶校验,并且只需两个校验盘的理论最优冗余存储构造双容错 RAID 结构的方法,其编码算法规范,但是编码和解码算法仍然比较复杂,主对角线的存在极大地影响了性能. 由于对角线校验组的规模较大,当主对角线上的数据单元改变时,全部对角线校验单元都要更新.

DH1 和 DH2 编码[4]技术通过双重奇偶校验来实现双容错. 同 EVEN-ODD 类似,它们使用的也是水平校验和对角线校验,但是 DH1 和 DH2 编码的对角线校验比 EVEN-ODD 编码简单,并且校验信息不再存储在固定的校验磁盘,而是均匀散布于整个磁盘阵列. 所以,不会因频繁小写操作而导致校验磁盘的瓶颈问题. 在 DH1 编码中,其水平校验信息存储于矩阵的对角线方向,而 DH2 与 DH1 略有不同,水平校验单元并不均匀散布,而是用一个额外的校验磁盘存储,对角线校验单元仍存储于最后一行,矩阵右下角位置多出一个空闲块. DH1 和 DH2 方法比 EVEN-ODD 方法更简单,编码、解码性能更好,校验冗余存储容量均为理论最优值,即两个磁盘的冗余存储,但是要求阵列磁盘数为素数.

RM2 编码[5]是从一个新的角度构造双容错 RAID 阵列,它将双容错数据布局构造问题转化为构造冗余矩阵 RM(Redundancy Matrix)的问题. RM2 方法中各单元的校验计算简单,但是其冗余存储容量并不是最优.

周杰等人[8]提出的基于完全图的双容错编码方法,将校验条纹和图相联系,通过分析双容错 RAID 数据布局的特征,将双容错布局的构造问题转化为简单图的划分问题,每个子图代表一个磁盘,子图的边和顶点分别表示分布到磁盘的校验单元和数据单元. 在算法实现中,包括构造算法和检验算法两部分. 构造算法用于生成可能的分组,检验算法用于检验可能分组是否满足双容错约束条件. 但是,随着顶点数的增大,搜索空间呈级数增长,进行数据布局计算的时间复杂度较高. 从冗余存储容量方面看,该方法同其它双容错编码方案比较,其冗余性能最优.

3 三容错 RAID 编码方法

从理论上讲,Reed-Solomon 编码可以扩展应用到三容错 RAID 体系结构,但是其编码和解码涉及有限域上的操作,速度很慢. 为此,Feng 等人[9]提出了一种仅有异或运算的二进制线性编码方法,其基本思想描述如下:

设 $p=m+1$ 是一个素数,$\boldsymbol{I}_m$ 是 $m\times m$ 单位矩阵(Identify Matrix),$\boldsymbol{O}_m$ 是 $m\times m$ 零矩阵,则初等循环矩阵(Elemental Cyclic Matrix) 定义为

$$\boldsymbol{E}_{m+1}=\begin{bmatrix}\vec{0} & \vec{1}\\ \boldsymbol{I}_m & \vec{0}^{\mathrm{T}}\end{bmatrix} \tag{1}$$

式中,$\vec{1}$ 是 $1\times m$ 阶 1 向量,$\vec{0}$ 是 $m\times 1$ 阶 0 向量.

容易验证:$\{\boldsymbol{I}_{m+1},\boldsymbol{E}_{m+1},\boldsymbol{E}_{m+1}^2,\cdots,\boldsymbol{E}_{m+1}^m\}$ 形成一个群(group),可以在有限域 GF(2) 上进行矩阵乘法运算.

设

$$\tilde{\boldsymbol{I}}_{m+1}=\begin{bmatrix}\boldsymbol{I}_m\\ 00\cdots0\end{bmatrix}_{(m+1)\times m} \tag{2}$$

在不引起混淆的情况下,以下我们不妨用 $\boldsymbol{I},\boldsymbol{E},\tilde{\boldsymbol{I}}$ 分别代替 $\boldsymbol{I}_{m+1},\boldsymbol{E}_{m+1},\tilde{\boldsymbol{I}}_{m+1}$.

现定义如下 $r(m+1)\times(n+1)m$ 阶二进制矩阵(其中 $r<n<m$):

$$\tilde{\boldsymbol{H}}=\begin{bmatrix}\tilde{\boldsymbol{I}} & \tilde{\boldsymbol{I}} & \tilde{\boldsymbol{I}} & \tilde{\boldsymbol{I}} & \cdots & \tilde{\boldsymbol{I}}\\ \tilde{\boldsymbol{I}} & \boldsymbol{E}\tilde{\boldsymbol{I}} & \boldsymbol{E}^2\tilde{\boldsymbol{I}} & \boldsymbol{E}^3\tilde{\boldsymbol{I}} & \cdots & \boldsymbol{E}^n\tilde{\boldsymbol{I}}\\ \tilde{\boldsymbol{I}} & \boldsymbol{E}^2\tilde{\boldsymbol{I}} & \boldsymbol{E}^4\tilde{\boldsymbol{I}} & \boldsymbol{E}^6\tilde{\boldsymbol{I}} & \cdots & \boldsymbol{E}^{2n}\tilde{\boldsymbol{I}}\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ \tilde{\boldsymbol{I}} & \boldsymbol{E}^{r-1}\tilde{\boldsymbol{I}} & \boldsymbol{E}^{2(r-1)}\tilde{\boldsymbol{I}} & \boldsymbol{E}^{3(r-1)}\tilde{\boldsymbol{I}} & \cdots & \boldsymbol{E}^{n(r-1)}\tilde{\boldsymbol{I}}\end{bmatrix} \tag{3}$$

式(3) 可以看成一个 $r\times(n+1)$ 阶块矩阵,其中每个列块包含 m 列,每个行块包含 $(m+1)$ 行.

可以证明,任意 r 个列块是一个满秩子阵,即任意 r 个列块中的列都线性无关. 因此,可以推导出与 $\tilde{\boldsymbol{H}}$ 等价的矩阵 $\boldsymbol{H}$ 作为校验矩阵:

$$\boldsymbol{H}=\begin{bmatrix}\boldsymbol{I}_m & \boldsymbol{I}_m & \boldsymbol{I}_m & \boldsymbol{I}_m & \cdots & \boldsymbol{I}_m\\ \boldsymbol{I}_m & \tilde{\boldsymbol{I}}^{\mathrm{T}}\boldsymbol{E}\tilde{\boldsymbol{I}} & \tilde{\boldsymbol{I}}^{\mathrm{T}}\boldsymbol{E}^2\tilde{\boldsymbol{I}} & \tilde{\boldsymbol{I}}^{\mathrm{T}}\boldsymbol{E}^3\tilde{\boldsymbol{I}} & \cdots & \tilde{\boldsymbol{I}}^{\mathrm{T}}\boldsymbol{E}^n\tilde{\boldsymbol{I}}\\ \boldsymbol{I}_m & \tilde{\boldsymbol{I}}^{\mathrm{T}}\boldsymbol{E}^2\tilde{\boldsymbol{I}} & \tilde{\boldsymbol{I}}^{\mathrm{T}}\boldsymbol{E}^4\tilde{\boldsymbol{I}} & \tilde{\boldsymbol{I}}^{\mathrm{T}}\boldsymbol{E}^6\tilde{\boldsymbol{I}} & \cdots & \tilde{\boldsymbol{I}}^{\mathrm{T}}\boldsymbol{E}^{2n}\tilde{\boldsymbol{I}}\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ \boldsymbol{I}_m & \tilde{\boldsymbol{I}}^{\mathrm{T}}\boldsymbol{E}^{r-1}\tilde{\boldsymbol{I}} & \tilde{\boldsymbol{I}}^{\mathrm{T}}\boldsymbol{E}^{2(r-1)}\tilde{\boldsymbol{I}} & \tilde{\boldsymbol{I}}^{\mathrm{T}}\boldsymbol{E}^{3(r-1)}\tilde{\boldsymbol{I}} & \cdots & \tilde{\boldsymbol{I}}^{\mathrm{T}}\boldsymbol{E}^{n(r-1)}\tilde{\boldsymbol{I}}\end{bmatrix} \tag{4}$$

基于循环置换矩阵(Circular Permutation Matrices,CPM),可以构建扩展型 Reed-Solomon 类编码. 假设 $(G,\oplus)$ 是一个交换群(Abelian Group),0 是单位元素,$b\in\{0,1\}$,$g\in G$,定义:

$$b\times g=g\times b=\begin{cases}0, b=0\\ g, b=1\end{cases} \tag{5}$$

设 $\boldsymbol{v}=(v_0,v_1,\cdots,v_{n-1})$ 是 G 上的向量,$\boldsymbol{b}=(b_0,b_1,\cdots,b_{n-1})$ 是 GF(2) 上的向量,定义:

$$b\times v=(b_0\times v_0)\oplus(b_1\times v_1)\oplus\cdots\oplus(b_{n-1}\times v_{n-1}) \tag{6}$$

定义 C 为交换群上的线性码，满足：

$$C=\{c=(c_0,c_1,c_2,\cdots,c_{n+3},)\mid H^*c^{\mathrm{T}}=\vec{0}^{\mathrm{T}}\} \tag{7}$$

其中，$c_i=(c_{i1},c_{i2},\cdots,c_{im})c_{ij}\in G$，并且

$$H^*=\begin{bmatrix} I_m & O_m & O_m & I_m & I_m & I_m & I_m & \cdots & I_m \\ O_m & I_m & O_m & I_m & \tilde{I}^{\mathrm{T}}E\tilde{I} & \tilde{I}^{\mathrm{T}}E^2\tilde{I} & \tilde{I}^{\mathrm{T}}E^3\tilde{I} & \cdots & \tilde{I}^{\mathrm{T}}E^n\tilde{I} \\ O_m & O_m & I_m & I_m & \tilde{I}^{\mathrm{T}}E^2\tilde{I} & \tilde{I}^{\mathrm{T}}E^4\tilde{I} & \tilde{I}^{\mathrm{T}}E^6\tilde{I} & \cdots & \tilde{I}^{\mathrm{T}}E^{2n}\tilde{I} \end{bmatrix} \tag{8}$$

对于式(7)所示的线性码，前 $r=3$ 个向量(c_0，c_1，c_2)是校验向量，其他 $n+1$ 个向量(即 c_j，$3\leqslant j\leqslant n+3$)是信息向量.

在许多应用中，交换群 G 可以是具有位异或(bit-XOR)运算的计算机字. 当 $r=2$ 时，该线性码就是 EVEN-ODD 码.

当 $r=3$ 时，式(8)为校验矩阵，编码过程表示为式(9).

$$\begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}=\tilde{H}^*\begin{bmatrix} c_3 \\ c_4 \\ \cdots \\ c_{n+3} \end{bmatrix} \tag{9}$$

具有三容错功能的 MDS 交换群阵列编码，其编码复杂度为 $3mn$ 次异或运算，解码复杂度为 $3mn+9(m+1)$次异或运算，同一般的 Reed-Solomon 编码相比，效率改进很多.

4　结束语

由于单容错 RAID 系统面临的问题越来越严重，为进一步提高大规模存储系统的可靠性，本文分析了现有的双容错 RAID 编码方法，并对各种编码方法的冗余性能进行了对比，便于在实际应用中进行权衡. 基于循环置换矩阵构建的可以容许三个磁盘故障的 MDS 交换群阵列码，代表了多容错 RAID 编码的最新研究进展，虽然其编码和解码效率比较高，但是要应用到实际 RAID 存储系统中还有许多工作要做. 未来工作应集中在多容错 RAID 编码方案的性能优化上，提高容错编码效率，使其更加实用.

参考文献：

[1] Chen P M, Lee E K, Gibson G A, et al. RAID: high—performance, reliable secondary storage[J]. ACM Computing Surveys, 1994, 26(2): 145—185.

[2] Blahut M. Algebraic codes for data transmission[M] Paris, France: Cambridge University Press, 2003.

[3] Blaum M, Brady J, Bruck J, et al. EVENODD: an efficient scheme for tolerating double disk failures in RAID architectures[J]. IEEE Transactions on Computers, 1995, 44(2): 192—202.

[4] Lee N K, Yang S B, Lee K W. Efficient parity placement schemes for tolerating up to two disk failures in disk arrays[J]. Journal of Systems Architecture, 2000, 46(15): 1383—1402

[5] Park C I. Efficient placement of parity and data to tolerate two disk failures in disk array systems. [J]. IEEE Transactions on Parallel and Distributed Systems, 1995, 6(11): 1177—1184.

[6] Xu L, Bruck J. X—code: MDS array codes with optimal encoding[J]. IEEE Transactions on Information Theory, 1999, 45(1): 272—276.

[7] Xu L, Bohossian V, Bruck J, et al. Low—density MDS codes and factors of complete graphs[J]. IEEE Transactions on Information Theory, Sept. 1999, 45(6): 1817—1826.

[8] 周杰，王刚，刘晓光，等. 容许两个盘故障的磁盘阵列数据布局与图分解的条件和存在性研究[J]. 计算机学报，2003，26(10)：1379—1386.

[9] Feng G L, Deng R H, Bao F, et al. New efficient MDS array codes for RAID part 1: reed—solomon—like codes for tolerating three disk failures[J]. IEEE Transactions on Computers, 2005, 54(9): 1071—1080.

作者简介：

刘　军　男，(1963—)，博士，教授. 研究方向为网络存储、并行与分布式系统.

刘　璟　男，(1942—)，教授，博士生导师. 研究方向为并行与分布式系统、海量存储、并行 VLSL 算法.