最优冗余双容错数据布局研究*

王　刚[1]；刘晓光[1]；董沙莎[1]；刘　璟[1]

（1. 南开大学 信息技术科学学院，　天津 300071）

摘要：在磁盘阵列双容错编码简单图表示法和双容错数据布局判定定理的基础上，提出了最优冗余布局定理，证明了 full-2 码（对应完全图）的双容错数据布局的磁盘数下界，及最优冗余布局的构型。并给出了一种基于完全图的完全 1-因子分解的 full-2 码最优冗余双容错数据布局构造方法，与其他双容错编码/布局相比，该方法生成的布局具有高可靠性、更新代价最优、冗余率最优、低编码/解码复杂度等优点，且构造方法适应性更强。

关键词：计算机应用　双容错数据布局　简单图表示法　完全 1-因子分解

中图分类号：TP302

Research on Optimal Redundancy double-erasure-correcting Data Layout

WANG　Gang[1]；　LIU　Xiao-guang[1]；　DONG　Sha-sha[1]；　LIU　Jing[1]

(1. *Institute of Information Technical Science, Nankai University, Tianjin* 300071)

Abstract: On the basis of the simple graph representation of RAID double-erasure-correcting codes and the double-erasure-correcting data layout judgement theorem, the optimal redundancy data layout theorem was proposed, and the theorem proves the lower bound of the number of disks of double-erasure-correcting data layout of full-2 code (corresponds to complete graph) and the structure properties of the optimal redundancy data layouts. A construction method for optimal redundancy double-erasure-correcting data layout based on perfect 1-factorization was presented. Compared with other double-erasure-correcting codes/data layouts, the data layouts produced by this method have high reliability, optimal update penalty, optimal redundancy and low encoding/decoding complexity. Moreover this construction method has high applicability.

Key words: Computer Application　double-erasure-correcting data layout　simple graph representation　perfect 1-factorization

0　引言

上世纪 80 年代末期，Patterson 等人提出了廉价磁盘冗余阵列技术（Redundant Arrays of Inexpensive Disks，RAID）[1]，这是近二十年来存储领域最重要的成果之一，目前仍是构造大规模存储系统的关键技术之一。近几年，互联网和网格技术的飞速发展，对存储技术提出了新的挑战，特别是对存储系统可靠性、可用性的要求越来越高，因此磁盘阵列多容错编码的研究逐渐增多。

较早提出的双容错编码是 RAID6 结构，其可靠性、冗余率、更新代价都很好，但编码/解码复杂度较高。上世纪 90 年代初 Hellerstein 等人提出了二维奇偶校验码、full-2 码、full-3 码等一系列多容错线性码[2]，通过校验条纹分组，每个数据单元参与多个校验组来提供多容错能力。每个校验组等同于一个 RAID5 条纹，编码/解码复杂度远远优于 RAID6。Hellerstein 等人还提出了线性码的校验矩阵表示法[2]，用一个 0/1 矩阵表示磁盘（矩阵列）和校验组（矩阵行）的参与关系，图 1a 给出了一个 8 磁盘的二维码，图 1b 即为其校验矩阵表示。文献[2]的另一重要贡献是提出了 5 个多容错编码评价指标：可靠性、冗余率、更新代价、校验组大小和扩展性。

线性码的缺点是冗余率高，当阵列规模较小时尤为突出。而近年来应用需求的发展所带来的存储系统网络化等趋势，和硬盘技术自身发展的一些特

收稿日期：
济金项目：国家自然科学基金重大计划项目（编号：90612001），天津市科技发展计划重点项目（编号：043800311，04315111-14），南开大学创新基金，论文受到南开大学科学计算所支持
作者简介：王刚（1974-），男，副教授。研究方向：海量存储，并行计算。E-mail:wgzwp@163.com

点，使得中小规模的磁盘阵列也需要多容错能力，这就无形中放大了这一缺点。研究者于是考虑多容错数据布局方式——在一个磁盘上放置多容错编码中归属于同一个校验条纹的多个条纹单元。通过合理布置数据单元和条纹单元，仍可保证双容错能力，但冗余率却大大降低。EVENODD[3]、DH1 和 DH2[4]、RM2[5]、B-CODE[6]、RDP[7]等双容错编码/布局就是此类成果。其中，EVENODD、DH1 和 RDP 编码的冗余率都达到了理论最优的 2/N（磁盘数为 N），但校验条纹中的数据（校验）单元可能参与不止两个（一个）校验组，这显然可能影响更新代价，我们称之为一类线性码。而 DH2、RM2、B-CODE 编码则严格满足每个数据（校验）单元参与两个（一个）校验组的性质，保证了更新代价最优，我们称之为二类线性码。

图1a. 二维编码　　图1b. 校验矩阵　　图1c. 简单图表示

图 1. 二维奇偶校验码及其校验矩阵和简单图

Fig.1 2d-parity code, parity check matrix and graph representation

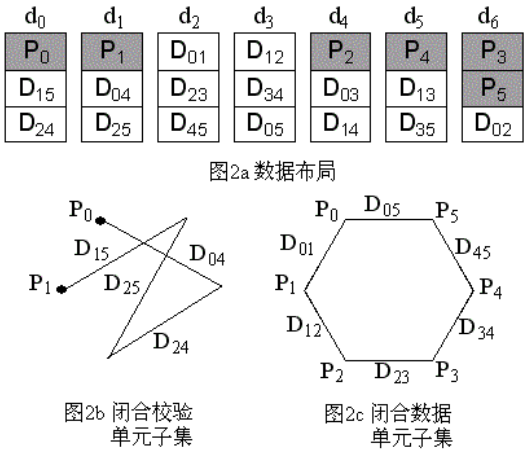

图2a 数据布局

图2b 闭合校验单元子集　　图2c 闭合数据单元子集

图 2. 数据布局不可恢复故障对应的图表示

Fig.2 graph representation of unrecoverable stripe units set

我们对双容错编码/数据布局问题进行了研究，提出了二类线性码的简单图表示法[8]，用顶点表 P_i 示校验单元 i（校验组 i），用边 D_{ij} 表示数据单元（参与校验组 i 和校验组 j），图 1c 给出了图 1a 二维码的简单图表示。图表示法与校验矩阵表示法一样严谨、简洁，而且将数据布局问题转化为图划分问题，可借助图论理论进行双容错数据布局问题的研究。我们提出了双容错数据布局判定定理[8]。如图 2，图 2b 所示两端为顶点，中间由边构成的路，和图 2c 所示完全由边构成的圈，这两种闭路结构即对应两类不可恢复磁盘故障条纹单元集合（闭合校验单元子集和闭合数据单元子集）。实际上，两个闭路实例即为图 2a 数据布局中不可恢复的两个双磁盘故障：磁盘 0 和磁盘 1、磁盘 2 和磁盘 3。双容错数据布局判定定理指出：若数据布局对应的图划分，任意两个分组的并都不包含这两类闭路，则数据布局具有双容错能力。

1　最优冗余数据布局

双容错数据布局判定定理给出了双容错数据布局研究的重要基础，但未解决构造问题。我们对 full-2 码（对应完全图）的双容错数据布局的存在性进行了研究，得到了双容错布局对应的图划分应满足的一些性质（其中“分组”意为图划分的子图）。

性质 1　一个分组至多含有图的一个顶点

证明：如果一个分组包含两个顶点 P_i 和 P_j，考虑包含边 D_{ij} 的分组，两个分组的并包含闭合校验单元子集，因此布局不能容双错，矛盾！

性质 2　如果一个分组中包含了图的一个顶点，则以该顶点为端点的边不能出现在这个分组中

证明：如果一个分组包含顶点 P_i 和边 D_{ij}，则与包含顶点 P_j 的分组的并包含闭合校验单元子集，矛盾！

性质 3　如果两边相邻，不可属于同一分组

证明：如果一个分组中包含边 D_{ij} 和 D_{jk}，则与包含边 D_{ki} 的分组的并包含闭合数据单元子集，矛盾！

性质 4　任何两个分组的并，最多包含 n-1 条边

证明：如果包含的边数≥n，子图必然包含圈，矛盾！

性质 5　如果两个分组各含一个顶点，则其并最多包含 n-2 条边

证明：如果包含 n-1 边，则要么形成圈，要么恰为汉密尔顿圈去掉一条边形成的路，而任意两个顶点均会落在这条路上，构成 1）型闭路，矛盾！

于是可得最优冗余布局定理。

定理 1　最优冗余布局定理：n 个校验组的 full-2 码（完全图 K_n），其双容错布局满足如下性质：

1) 若 n 为偶数，分组数（磁盘数）下界为 n+1（此时冗余率最低）。每组含 n/2 个条纹单元，其中有 n 个分组各由 1 个顶点（校验单元）和(n-2)/2 条边（数据单元）组成，另一个分组由 n/2 条边组成。
2) 若 n 为奇数，分组数下界为 n+2。其中，有一

个分组由 1 个顶点和(n-1)/2 条边共(n+1)/2 个条纹单元组成。其他分组均含(n-1)/2 个条纹单元，其中有两个分组由(n-1)/2 条边组成，其他分组由 1 个顶点和(n-3)/2 条边组成。

证明：为叙述简便，下文用 N_0、N_1、…表示分组包含的校验单元（顶点）数目，用 M_0、M_1、…表示数据单元（边）数目。

a)　证明 n 为偶数时，分组数下界为 n+1

由性质 1，分组数≥n，假定存在分组数为 n 的双容错布局，则有 $N_0=N_1=\ldots=N_{n-1}=1$。

则 $M_0+M_1+\ldots+M_{n-1}=\frac{n(n-1)}{2}$，均值为 $\frac{n-1}{2}$，而 n 为偶数，则至少存在一个 $M_i \geq \frac{n}{2}$。

若其他 M 均 $< \frac{n-2}{2}$，则 $M_i > n-1$，与性质 5 矛盾。

若存在 $M_j \geq \frac{n-2}{2}(j \neq i)$，则 $M_i + M_j \geq n-1$，与性质 5 矛盾。得证。

b)　证明 n 为偶数时，达到分组数下界的布局必具有 1)所述构型

由性质 1，不妨假定 $N_0=N_1=\ldots=N_{n-1}=1$，$N_n=0$，先证明 $M_n = \frac{n}{2}$

首先，由性质 5，对所有 $0 \leq i, j \leq n-1$，有 $M_i + M_j \leq n-2$。

因此 $M_0 + \cdots + M_{n-1} \leq \frac{n(n-2)}{2}$，所以 $M_n \geq \frac{n}{2}$。

假设 $\frac{n}{2} < M_n \leq n-1$，由性质 4，$M_0, \cdots, M_{n-1} \leq n-1-M_n$，则

$$M_0 + \cdots + M_{n-2} + (M_{n-1} + M_n)$$
$$< (n-1-\frac{n}{2}) \times (n-1) + n - 1 = \frac{n(n-1)}{2}$$

矛盾！因此 $M_n = \frac{n}{2}$

则 $M_0 + \cdots + M_{n-2} + M_{n-1} = \frac{n(n-2)}{2}$，类似 a)可证明 $M_0 = \cdots = M_{n-1} = \frac{n-2}{2}$。得证。

c)　证明 n 为奇数时，分组数下界为 n+2

类似 a)可证分组数>n。假设分组数为 n+1，$N_0=N_1=\ldots=N_{n-1}=1$，$N_n=0$。

由性质 5，对任意 $0 \leq i, j \leq n-1$，有 $M_i + M_j \leq n-2$，且 n 为奇数，因此至多有一个 $M_i \geq \frac{n-1}{2}$，其他 $M_j \leq \frac{n-3}{2}, j \neq i$。

因此

$$M_0 + \cdots + M_{n-2} + M_{n-1}$$
$$\leq \frac{n-3}{2}(n-2) + (n-2) = \frac{(n-1)(n-2)}{2}$$

则 $M_n \geq n-1$，与性质 4 矛盾。得证。

d)　证明 n 为奇数时，达到分组数下界的布局必然具有 2)所述的构型

不妨假定 $N_0=N_1=\ldots=N_{n-1}=1$，$N_n= N_{n+1}=0$

类似 c)，易知 $M_n + M_{n+1} \geq n-1$，由性质 4，$M_n + M_{n+1} = n-1$。

因此 $M_0 + \cdots + M_{n-2} + M_{n-1} = \frac{(n-1)(n-2)}{2}$

M_n 和 M_{n+1} 中至少有一个 $\geq \frac{n-1}{2}$，由性质 4，M_0、…、M_{n-1} 均 $\leq \frac{n-1}{2}$，且不可能均 $\leq \frac{n-3}{2}$，至少有一个为 $\frac{n-1}{2}$。

而由性质 5，只可能有一个为 $\frac{n-1}{2}$，其他均为 $\frac{n-3}{2}$，而 M_n 和 M_{n+1} 只能均为 $\frac{n-1}{2}$。得证。

定理 1 证毕。

容易看出，若 K_{2n} 存在 2n+1 个磁盘的双容错布局，其冗余率达到理论最优，而对 K_{2n-1}，即便存在 2n+1 个磁盘的双容错布局，其冗余率也非理论最优。实际上，两种布局的磁盘数都是奇数，无偶数磁盘布局方案。但容易发现，将 K_{2n+2} 的最优冗余布局的纯数据单元磁盘删除，所得布局仍具有双容错能力，冗余率也保持理论最优。因此，无论磁盘数奇偶，均利用 K_{2n} 构造最优冗余布局即可。

2　最优冗余数据布局构造方法

定理 1 仍未解决最优冗余双容错数据布局构造问题，借助图论领域完全 1-因子分解问题的研究成果，我们可设计出高效的构造方法。所谓图 G(V, E)的一个完全 1-因子分解（perfect 1-factorization，P1F），指 G 的一个划分{ F_0, F_1, …, F_{k-1}}，每个 F_i 均为 G 的一个 1-因子（1-正则生成子图），且任意两个 1-因子均不相交，而它们的并构成汉密尔顿圈。文献[9]给出了 n 或 2n-1 为素数时，构造完全图 K_{2n} 的 P1F 的线性时间复杂度算法。对其他很多偶数，也已找到 P1F。而 EVENODD 等编码，要求磁盘数与素数相关，显然本方法适应性更强。图 3 给出了 K_6 的一个 1-因子分解。

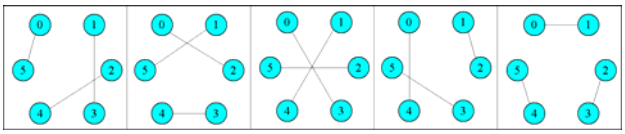

图 3. 完全图 K_6 的一个 1-因子分解

Fig.3 a P1F of K_6

算法 1　full-2 码最优冗余双容错布局构造算法

输入：完全图 K_{2n} 的一个 P1F J={F_0, F_1, …, F_{2n-2}}，F_i 中包含顶点 P_{2n-1} 的边为 $D_{2n-1\,i}$（0≤i≤2n-2）。

输出：完全图 K_{2n-2}（2n-2 个校验组的 full-2 码）对应的双容错编码的最优冗余双容错布局。

方法：

1)　对每个 F_i（0≤i≤2n-2），删除顶点 P_{2n-1} 和边 $D_{2n-1\,i}$，得到 J’ = {F_0’, F_1’, …, F_{2n-2}’}，图 4 给出了图 3 的 P1F 进行此操作后的结果。

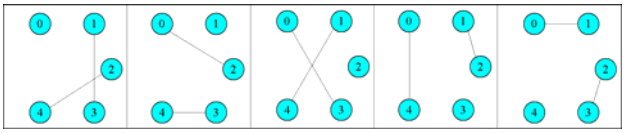

图 4. 经过步骤 1 后的结果

Fig.4 the result of step 1

2)　对每个 F_i’（0≤i≤2n-2）删除顶点 P_{2n-2} 及其邻接边（F_{2n-2}’仅删除顶点），得到 J” = {F_0”, F_1”, …, F_{2n-2}”}，图 5 给出了图 4 中结果经此步骤后的结果。J”即为所求的最优冗余布局，我们规定分组 F_i”（0≤i≤2n-3）只包含孤立顶点 P_i，不包含其他顶点。

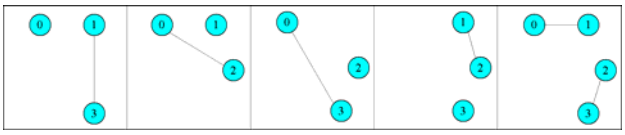

图 5. 算法 1 的执行结果

Fig.5 the final result of algorithm 1

定理 2　算法 1 输出的布局具有双容错能力：

证明：

由 P1F 定义，对任意 F_i 和 F_j（0≤i, j≤2n-2），$F_i \cup F_j$ 形成汉密尔顿圈。

F_i’∪F_j’由 $F_i \cup F_j$ 删除顶点 P_{2n-1} 和边 $D_{2n-1\,i}$、$D_{2n-1\,j}$ 而得，因此构成长度为 2n-2，$P_i \rightarrow P_j$ 的路。

对任意 F_i”和 F_j”（0≤i, j≤2n-3），F_i”∪F_j”为 F_i’∪F_j’删除边 $D_{2n-2\,k}$ 和 $D_{2n-2\,k'}$（k≠i，k’≠j）和 P_i、P_j 之外的所有顶点。考虑到磁盘阵列布局的实际意义，这里 n≥2，因此 k=j 和 k’=i 不可能同时成立，因此 F_i”∪F_j”的结构有两种情况：1）分别以 P_i、P_j 为端点的两条不相交的路（图 4 的 F_0 和 F_1）或，2）一条以 P_i 或 P_j 为端点的路和另一顶点形成的孤立顶点（图 4 的 F_0 和 F_2），均未形成不可恢复闭路。

对任意 F_i”（0≤i≤2n-3）和 F_{2n-2}”，F_i”∪F_{2n-2}”为 F_i’∪F_j’删除边 $D_{2n-2\,k}$（k≠i）和 P_i 外所有顶点，必然形成以 P_i 和 P_k 为端点的路，但只包含顶点 P_i，也未形成闭路。证毕。

算法 1 本质上与 B-CODE 编码构造方法[5]是等价的，但更为简洁、清晰，基于双容错数据布局判定定理的证明也更为简单。算法 1 生成的布局，即 B-CODE，其特性完全吻合文献[8]中结论及本文的性质 1－性质 5 和定理 1。

若对于磁盘数 2n-2 或 2n-1，尚未找到 K_{2n} 的 P1F，一种较好的布局构造方法是：取 n’为>n，且 $K_{2n'}$已找到 P1F 的最小整数，利用算法 1 构造 2n’-1 个磁盘的最优冗余双容错布局，删除若干分组（磁盘）后转换为 2n-2 个或 2n-1 个磁盘的布局。需注意的是，删除校验单元，其相关数据单元也要同时删除，由此造成的分组大小不一，可采用类似 RAID5 的循环重复方式解决。如此构造的布局，冗余率未达到理论最优，但非常接近，使得算法 1 的适用范围更广。

算法 1 生成的双容错布局，由于源于二类线性码，因此其更新代价是最优的，编码/解码复杂度低，通过数据布局的方式，校验开销也达到了理论最优值。虽然 full-2 码能够恢复的一些三个以上磁盘故障，数据布局方式无此能力，但双容错能力已经可以保证很好的可靠性。因此，以此构造的双容错磁

盘阵列，各方面指标都达到较好的效果。当然，这种布局的扩展能力较差，实际上这也是 DH、RM2 等校验分散编码方案共有的缺点，我们已经借助 P1F 得到了一种具有独立校验分组的高扩展能力的双容错数据布局的构造方法，限于篇幅，这里不再详述。

3　总结

本文利用图论方法，对双容错数据布局的存在性及其构造方法进行了研究。给出了 full-2 码最优冗余双容错数据布局存在的必要条件，并借助完全图的完全 1-因子分解问题的研究成果，对最优冗余布局的充分条件进行了研究。发现 B-CODE 编码就是我们研究的最优冗余布局方案，给出了更为简洁的构造方法和正确性证明。生成的布局在冗余率、更新代价及可靠性方面都达到最优，而构造方法的适应性优于 EVENODD 等编码。

本文和文献[8]的工作，对于双容错数据布局问题只是刚刚开始，尚有很多工作有待开展，才能达到应用于实践的程度。如目前的图描述法很多表述与图论领域的一般表述方式有差异，我们已经在研究更好的虚拟顶点表示法。还有一类线性码的表示方法，三容错、四容错编码的表示方法，非最优冗余布局构造方法，数据布局性能分析和性能优化等等，都是很有价值的工作，我们已着手对其中一些问题展开进一步的研究工作。

参　考　文　献

［1］D A Patterson, G A Gibson, R H Katz. A case for Redundant Arrays of Inexpensive Disks (RAID) ［C］// ACM International Conference on Management of Data. Chicago: ACM Press, 1988: 109-116.

［2］Lisa Hellerstein, Garth A Gibson, Richard M Karp, Randy H Katz, David A Patterson. Coding Techniques for Handling Failures in Large Disk Arrays［J］. Algorithmica, 1994, 12(2/3): 182-208.

［3］M Blaum, J Brady, J Bruck, J Menon. EVENODD: An Efficient Scheme for Tolerating Double Disk Failures in RAID Architectures[J]. IEEE Trans Computers, 1995, 44(2): 192-202.

［4］C Park. Efficient Placement of Parity and Data to Tolerate Two Disk Failures in Disk Array Systems［J］. IEEE Trans Parallel Distrib Syst, 1995, 6(11): 1177-1184.

［5］L Xu, V Bohossian, J Bruck, D G Wagner. Low-Density MDS Codes and Factors of Complete Graphs［J］. IEEE Transactions on Information Theory, 1999, 45(6): 1817-1826.

［6］Nam-Kyu Lee, Sung-Bong Yang, Kyoung-Woo Lee. Efficient Parity Placement Schemes for Tolerating up to Two Disk Failures in Disk Arrays［J］. Journal of Systems Architecture, 2000, 46(15): 1383-1402.

［7］Peter Corbett, Bob English, Atul Goel, Tomislav Grcanac, Steven Kleiman, James Leong, Sunitha Sankar. Row-Diagonal Parity for Double Disk Failure Correction［C］// Proceedings of the Third USENIX Conference on File and Storage Technologies. San Francisco: USENIX Association, 2004.

［8］周杰, 王刚, 刘晓光, 刘璟. 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.
ZHOU Jie, WANG Gang, LIU Xiao-guang, LIU Jing. The Study of Graph Decompositions and Placement of Parity and Data to Tolerate Two Failures in Disk Arrays : Conditions and Existence［J］. Chinese Journal of Computers, 2003, 26(10): 1379-1386.

［9］W D Wallis. One-Factorizations[M]. Boston: Kluwer Academic Publishers, 1997.

［10］董沙莎. 双容错 RAID 数据布局方法的研究［D］. 天津: 南开大学, 2003.
DONG Sha sha. The Research of Data

Placement Scheme for Tolerating Double Disk Failures in RAID Architectures［D］. Tianjin: Nankai University, 2003.