利用图的完全 1-因子分解构造双容错数据布局

王　刚[1]，董沙莎[1]，刘晓光[1]，林　胜[1]，刘　璟[1]

（1.　南开大学信息技术科学学院计算机系, 天津市 300071）

摘　要：　本文介绍了一种 full-2 码的虚拟顶点简单图表示法，简化了双容错数据布局判定定理，最优冗余数据布局定理和双容错数据布局的构造。本文还提出了一种基于完全二部图（对应二维奇偶校验码）的完全 1-因子分解的双容错数据布局构造方法，可构造高扩展性双容错数据布局 BG-HEDP。与 B-CODE 等同类双容错数据布局相比，BG-HEDP 同样具有更新代价最优、高可靠性和低编码/解码复杂度的优点，冗余率接近最优，而扩展性更好。

关键词：　磁盘阵列；双容错编码；数据布局；完全二部图；完全 1-因子分解

中图分类号：TP 302.8；TP 333.3　　文献标识码：A　　文章编号：0372-2112(　　　　　)

Construct double-erasure-correcting Data Layout Using P1F

WANG Gang[1], DONG Sha-sha[1], LIU Xiao-guang[1], LIN Sheng[1], LIU Jing[1]

(1.Dept. of Computer, College of Information Technology Science, Nankai University, Tianjin300071, China)

Abstract:　In this paper, we present a “virtual node” simple graph representation for full-2 code (corresponds to complete graph), this representation simplifies the double-erasure-correcting data layout judgment theorem, the optimal redundancy data layout theorem and the construction of B-CODE. We also present a data layout construction method based on P1F of complete bipartite graph (corresponds to 2d parity code) in this paper, this method can produce highly extensible double-erasure-correcting data layouts (BG-HEDP). Compared with other data layouts, such as B-CODE, BG-HEDP also has optimal update penalty, high reliability and low encoding/decoding complexity, its redundancy is very close to optimal value, while it is superior in extensibility to others.

Key words:　RAID; 2-erasure-codes; data layout; complete bipartite graph; perfect 1-factorization

1 引言

磁盘阵列技术[1]是近二十年来存储领域最重要的成果之一。近年来，新型应用模式对存储系统的可靠性要求越来越高，而存储系统的一些发展趋势难以跟上这种需求，这对 RAID 技术提出了新的挑战。因此，近年来，国内外学术界、工业界在双磁盘故障容错编码和数据布局方面的研究逐渐增多。实际上，上世纪 90 年代初 Hellerstein 等人就对双容错编码进行过深入的研究[2]，多年来也不断出现新的双容错编码和数据布局[3~6]，但对双容错数据布局问题缺乏系统的研究。我们提出了一类双容错线性码的简单图表示法，以此为基础提出了双容错数据布局判定定理[7]，并在双容错数据布局的存在性方面取得了一些成果。本文介绍了我们在双容错数据布局构造方法方面的一些新的成果，第二节介绍了背景知识和前人工作，第三节讨论了基于完全图和完全二部图的 P1F 构造双容错数据布局的方法，第四节进行了总结和展望。

2 相关研究

在上世纪 90 年代初就已经出现了可恢复两个磁盘故障的 RAID6 结构，但其缺点是编码/解码复杂度很高。Hellerstein 等人提出了二维奇偶校验码、full-2 码等一系列多容错线性码[2]，其基本思想是校验分组，数据单元参与多个校验组，同组数据单元简单 XOR 运算即得到校验单元，大大提高了编码/解码性能。Hellerstein 等人还提出了针对这类线性码的校验矩阵表示方法，并提出了双容错编码评价指标：可靠性、校验开销、更新代价、校验组大小（影响重构性能）和扩展能力，这 5 个指标已经成为国内外相关研究工作评价双容错编码的主要依据。

收稿日期:0000-00-00; 修回日期:0000-00-00

基金项目:国家自然科学基金(No.906120001); 天津市科技发展计划重点项目(No.043800311, 04315111-14)

线性码编码/解码复杂度低，可靠性、更新代价方面都达到了最优，但缺点也是显而易见的——校验开销很差，特别是阵列规模较小时。而当前存储领域的一些发展趋势，使小规模磁盘阵列也需要双容错能力来保证较好的可靠性，无形中这个缺点就更为严重了。实际上，我们完全可以将双容错编码长度为 k 的校验条纹，布局在磁盘数 n<k 的阵列中，仍然保持双容错能力。这样做可能会失去两个以上磁盘故障的恢复能力，但已足够保证较好的可靠性，而冗余率上会有较大改进。EVENODD、DH1、RDP 和 DH2、RM2、B-CODE 等双容错编码/数据布局就是基于这种思路。其中，前三种编码的冗余率都达到了理论最优值 2/N，但数据单元（校验单元）可能参与超过两个（一个）校验组，造成更新代价非最优，另外三种编码则严格满足更新代价最优，我们分别称之为一类线性码和二类线性码。显然，二类线性码/布局如果冗余率达到最优，应该是各方面都较好的双容错编码/布局方案。

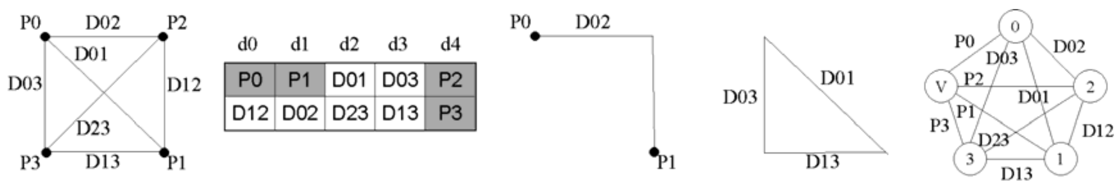


图1a. 简单图表示　　图1b. 数据布局　　图1c. 闭合校验单元子集　　图1d. 闭合数据单元子集　　图1e. 虚拟顶点表示法

图 1 full-2 码实例

利用二类线性码的特性，我们提出了用简单图的顶点表示校验单元（校验组），边表示数据单元的描述法，可以很直观地表示二类线性码[7]，图 1 给出了 10 磁盘的 full-2 码的简单图表示及其一个 6 磁盘的数据布局。我们还在此基础上提出了数据布局双容错判定定理[7]。如图 1c 所示包含两端顶点的路和图 1d 所示的圈，即为两类不可恢复的故障条纹单元集合（闭合校验单元子集和闭合数据单元子集）所对应的子图构型。实际上，这两个闭路分别对应图 1b 数据布局中两个不可恢复双磁盘故障：磁盘 0 和 1、磁盘 2 和 3。若数据布局对应的图划分中，任意两个分组的并都不包含这两类闭路，则布局具有双容错能力。

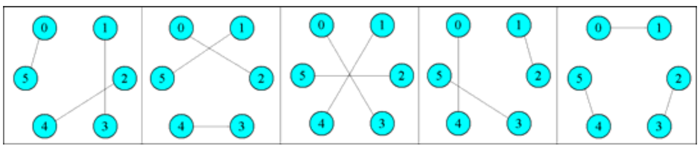


图 2 完全图 K_6 的一个 P1F

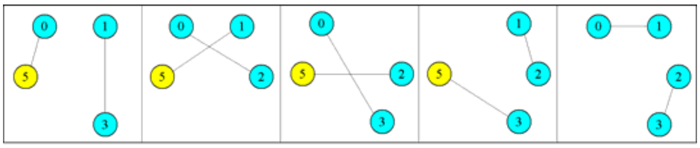
3 用图的完全 1－因子分解构造双容错布局

3．1 用完全图的 P1F 构造双容错布局

双容错判定定理给出了双容错数据布局问题研究的重要理论基础，但并未解决布局构造问题。我们对 full-2 码（对应完全图）的最优冗余双容错布局的存在性进行了研究，证明了 n 为奇数和偶数的不同情况下，最优冗余双容错布局的磁盘数下界，及其应满足的构型。我们还设计了一种利用完全图 K_{2n+2} 的完全 1-因子分解构造 2n 个校验组的 full-2 码（完全图 K_{2n}）的最优冗余布局的方法。所谓完全 1-因子分解（perfect 1-factorization，P1F），是指 G=(V, E)的一个子图集合{F_0, F_1, …, F_{k-1}}，所有 F_i 均为 1-因子（1-正则生成子图），任意子图 F_i 的边集均不相交，所有子图 F_i 的并为 G，且任意两个子图 F_i 和 F_j 的并均构成汉密尔顿回路。这种构造方法与 B-CODE 编码方法是等价的[5]，但我们的描述更为简洁、直观。生成的布局具有二类线性码的所有优点，且冗余率最优，适用范围较之 EVENODD 等编码更广：对于 2n-1 和 n 为素数的情况，存在线性复杂度的 K_{2n} 的 P1F 的公式化构造方法，对其他很多偶数，也已找到了 K_{2n} 的 P1F[8]，而且有一个著名的猜想——对所有 K_{2n} 均存在 P1F。图 2 给出了 K6 的一个 P1F。

从前文论述容易看出，利用简单图表示法研究双容错编码/数据布局问题，路、圈、分解等概念都与图论中的一般表述方式有所差异。可以简单加以改进：顶点只表示校验组，不再表示校验单元，增加一“虚拟顶点”v，校验组 w 的校验单元用边(v, w)表示，数据单元表示方式不变。图 1e 给出了图 1a full-2 码的虚拟顶点

表示。这样，两类对应不可恢复磁盘故障的闭路就均转化为圈，其中闭合校验单元子集对应包含虚拟顶点的圈，闭合数据单元子集对应不包含虚拟顶点的圈。利用虚拟顶点表示法，双容错数据布局判定定理和最优冗余数据布局定理均可得到简化。利用完全图的 P1F 构造最优冗余双容错布局的方法也简化为一个步骤：给出如图 2 所示 K_6（K_{2n+2}）的一个 P1F，其顶点 0－3（v_0－v_{2n-1}）为校验顶点，5（v_{2n+1}）为虚拟顶点，4（v_{2n}）为辅助顶点；对每个 1-因子，删除辅助顶点 4（v_{2n}）的邻边；所得分组即构成 K_4（K_{2n}）的一个最优冗余双容错布局，如图 3 所示。改进后的方法无需再考虑顶点，只需考虑边的划分即可，所得布局双容错能力的证明也变得非常简单：在删除辅助顶点的邻边后，任意两个分组的并均删除了相邻的两条边，而图的顶点减少了一个，因此任意两个分组的并均构成 K_{2n+1} 的一个长度为 2n 的路，未形成圈，故对应布局具有双容错能力。



d0	d1	d2	d3	d4
P0	P1	P2	P3	D01
D13	D02	D03	D12	D23

图 3 基于图 2 的 P1F 得到的完全图 K_4 的一个最优冗余双容错布局

3．2 用完全二部图的 P1F 构造双容错布局

B-CODE 在可靠性、校验开销、更新代价等方面都达到了很好的效果，但由于校验单元的散布，扩展能力较差。实际上，构造最优冗余布局，不一定基于完全图，非完全图同样可以。而且就目前存储技术的发展趋势看，容量应该是最容易解决的问题，最优冗余目标是否必须也要打上一个问号。以合理的冗余率，在其他方面获得更好的效果，可能比单纯追求最优冗余更好。图论领域中研究最为充分的非完全图，容易想到二部图。我们发现，利用完全二部图的 P1F，可构造出最优冗余或者接近最优冗余的双容错数据布局（此节采用简单图表示法，而非虚拟顶点表示法）。有意思的是，完全二部图对应的就是二维奇偶校验码。

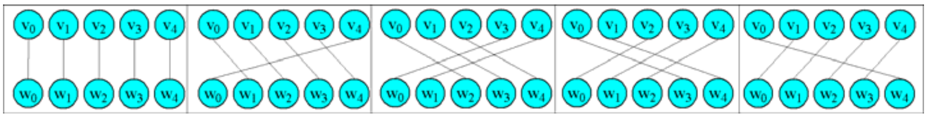


图 4 完全二部图 $K_{5,5}$ 的一个 P1F

与完全图的 P1F 问题一样，完全二部图的 P1F 问题的研究也有几十年的历史了。已知，如果完全图 K_{n+1} 存在一个 P1F，则可由此构造出完全二部图 $K_{n,n}$ 的一个 P1F[9]。这表明，对>2 的素数 p，$K_{p,p}$ 和 $K_{2p-1,\ 2p-1}$ 均存在 P1F，对<50 的奇数 p，可保证 $K_{p,p}$ 的 P1F 的存在。值得注意的是，上述命题的逆命题并不成立，这就是说，存在 P1F 的完全二部图的数目“不少于”完全图。而且，看起来寻找完全二部图的 P1F 比寻找完全图的 P1F 要“容易”得多，例如 K_{10} 只存在唯一的 P1F，而 $K_{9,9}$ 则存在 37 个不同构的 P1F[9]。对于完全图，文献[8]中给出的 P1F 公式化构造方法不是那么容易想到的。而 p 为素数的情况下，可以很容易地得到完全二部图 $K_{p,p}$ =(V, W, E)的一种 P1F {F_0, F_1, …, F_{p-1}}构造方法：令 F_i（0≤i≤p-1）包含所有边(v_j, $w_{(j+i)\%p}$)（0≤j≤p-1）即可，很容易证明这种图划分满足 P1F 的性质。图 4 给出了这种方法构造的 $K_{5,5}$ 的 P1F。

由 $K_{n,n}$ 的 P1F，将每个分组中 v_0、v_1、w_0 和 w_1 的邻边删除，再对分组 1～分组 n-1，交替选择 w_0、w_1 和 v_0、v_1 的邻接顶点（校验单元）加入分组，即构成 $K_{n-2,n-2}$ 的一个双容错数据布局，我们称之为二部图最优冗余布局（BG-ORDP）。其双容错特性的证明利用双容错数据布局判定定理很容易得到。显然，BG-ORDP 的构型和性能指标与 B-CODE 相似，但构造上更容易些，更多可选布局（更多 P1F）也为性能优化提供了更大的可能性。需要注意的是，这里的“更为容易”，是指 BG-ORDP 的适用范围更广，对给定参数，满足条件布局（P1F）的数量也更多，更容易找到，而非最优冗余布局的磁盘数范围更宽。在磁盘数上，两种布局的受限程度是一样强的，文献[7]中的定理 5 可以进一步加强为：

定理 1 对于顶点数为 n 的 d-正则图，最优冗余布局唯一可能的磁盘数是 d+2

证明：由文献[7]中定理 4 可知，为保证双容错，磁盘数 N≥d+2

而布局的冗余率 $r=n/(n+n*d/2)=2/(d+2)$，为使冗余率 r 达到最优即 2/N，显然唯一的可能就是 N=d+2。这与我们得到的最优冗余布局存在性的一些结论也是完全吻合的。

BG-ORDP 在扩展性方面也与 B-CODE 类似，是比较差的。为达到较好的扩展性，需如 EVENODD 等编码一样将校验单元布局到独立的磁盘。循此思路，我们得到了一种基于完全二部图的 P1F 的高扩展能力双容错数据布局（BG-HEDP）的构造方法。

算法 1 BG-HEDP 布局的构造算法

输入：完全二部图 $K_{n,n}=(V, W, E)$的一个 P1F $J=\{F_0, F_1, \cdots, F_{n-1}\}$

输出：$K_{n-1,n}$ 对应的双容错编码的 BG-HEDP 布局

方法：

1) 将每个 F_i（$0\leqslant i\leqslant n-1$）中顶点 v_{n-1} 的邻边删除，得到边分组 $J'=\{F_0', F_1', \cdots, F_{n-1}'\}$

2) 添加两个顶点分组 $P=V-\{v_{n-1}\}$和 $Q=W$，得到分组 $J''=\{F_0', F_1', \cdots, F_{n-1}', P, Q\}$，即为所求数据布局

图 5 给出了由图 4 中 P1F 构造而得的 BG-HEDP 布局。

d0	d1	d2	d3	d4	d5	d6
D00	D01	D02	D03	D04	v_0	w_0
D11	D12	D13	D14	D10	v_1	w_1
D22	D23	D24	D20	D21	v_2	w_2
D33	D34	D30	D31	D32	v_3	w_3
						w_4

图 5 由图 4 中 P1F 构造的 BG-HEDP 布局

定理 2 BG-HEDP 布局具有双容错能力

证明：只需证明任意两个分组的并均不包含不可恢复闭路结构即可，有三种情况：

1) P 和 Q，由于不包含边，显然不会包含闭路

2) P、Q 之一和任一边分组 F_i；由于 P 或 Q 中顶点均不相邻，F_i'中边也不相邻，显然两种闭路结构均不可能包含

3) 任意两个边分组 F_i'和 F_j；F_i 和 F_j 的并形成汉密尔顿圈，删除 v_{n-1} 的两条邻边后，F_i'和 Fj'的并构成 $K_{n-1,n}$ 的长度为 2n-2 的路，未形成圈，证毕。

显然，BG-HEDP 布局的冗余率未达到理论最优值 2/N，但两者的差距非常微小：当磁盘数 N=9 时，两者相差 6%；而 N=19 时，差距已缩小为不到 3%。注意到，对于二类线性码，若将校验单元布局于独立磁盘，是无法得到最优冗余布局的。容易证明，如果这类布局冗余率达到最优值 2/N，则平均每个数据磁盘包含的数据单元数$\geqslant n/2$（n 为校验单元数），因此至少有两个数据磁盘包含的数据单元数$\geqslant n$，这两个磁盘对应子图的并显然会形成圈，即布局无双容错能力，产生矛盾！因此，BG-HEDP 布局的冗余率已经非常接近此类布局的极限。BG-HEDP 布局的另一个问题是布局的不均衡，类似 RAID5/RAID6 采取循环重复方式即可解决此问题，既能利用所有磁盘空间，又能均匀散布校验单元（校验负载）。

3．3 扩展性的分析

所谓“扩展性”，一是指扩展布局构造算法的适用范围：即，对不存在 P1F 的情况，如何改变算法，同样能得到性能指标很好的算法；二是指文献[2]中提出的磁盘阵列扩展评价指标：即，向磁盘阵列增加新的磁盘后，如何以尽量少的数据迁移和校验计算，重组为更大规模的阵列（布局）。在这两个方面，BG-HEDP 与 BG-ORDP 和 B-CODE 相比，都具有明显优势。

对于给定磁盘数 N，若 K_{N+1}（$K_{N-2,N-2}$）不存在 P1F，构造 B-CODE 和 BG-ORDP 的最简单有效的方法是：寻找 n>N+1（N-2），且 K_n（$K_{n,n}$）存在 P1F 的最小的 n，由 K_n（$K_{n,n}$）的 P1F 构造双容错布局，将所得布局中若干分组（磁盘）删除，即可得到给定磁盘数 N 的双容错布局。但这样得到的 B-CODE 和 BG-ORDP，存在三个比较严重的问题：

1) 当删除的分组包含校验单元时，其他分组中与该校验单元属于同一校验组的数据单元也要删除，这样得到的布局构型会非常不规整，磁盘“高度”不一，而且可能存在多种“高度”。

2) 冗余率上升严重。如，11 个磁盘的 B-CODE 删除全数据分组和任意 5 个混合分组得到的 5 磁盘布

局，冗余率由 18.2%变为 55.6%。而 5 磁盘的标准 B-CODE 的冗余率为 40%。

3) 得到布局的校验组长度大小不一，差异很大，这一方面会引起读写性能上的严重问题，另一方面，也为布局的具体实现带来了巨大的困难。

BG-HEDP 若采取相同的方法，在这几方面却完全不存在问题：

1) 只需删除全数据分组，布局结构的计算非常简单，得到的布局构型也非常规则，实际上与初始布局是完全相似的——只有第二校验盘高度大 1，其他磁盘高度一致。

2) 冗余率与初始布局相比是变差了，但与同规模标准 BG-HEDP 相比，反而更优，也明显优于类似删除方式得到的 B-CODE 布局。如 11 磁盘的 BG-HEDP 删除 6 个数据分组得到 5 磁盘布局，冗余率由 19.1%变为 41.5%，而标准 5 磁盘 BG-HEDP 的冗余率为 45.5%。这是非常有趣的一个性质，我们似乎不应该构建标准 BG-HEDP，而应采取这种“降格”的构造方式，反而可得到性能更好的布局。

3) 得到布局的校验组长度基本一致，只有两种可能值，且只相差 1。

对于磁盘阵列扩展问题，很明显，通过将 K_n 对应的标准布局转化为 K_{n+m} 对应的标准布局，来融合新的磁盘，会导致大量的数据迁移（应该是全部数据）和校验计算。较好的方法是采用“降格”方式构造初始阵列，增加新的磁盘时，进行“降格”的逆操作。采用这种方法，对 BG-HEDP，只需将全 0 的新磁盘作为数据磁盘加入即可，避免了大量数据迁移和校验计算。但这对 B-CODE 和 BG-ORDP 是不可行的：一是因为添加的新磁盘中可能包含校验单元，不可能完全消除校验计算；二是如上所述，“降格”布局性能很差；更为严重的是，若严格执行“降格”的逆操作，布局的“高度”会增加，还是需要大量的数据移动，如果不增加高度，避免数据移动，则会导致扩展后阵列的性能一直很差。

4 结论

本文对利用图的完全 1-因子分解构造双容错数据布局问题进行了研究。对 full-2 编码的简单图表示，给出了一种更为简洁的虚拟顶点表示法，以此为基础，简化了 B-CODE 编码的构造算法。本文还对基于完全二部图的二进制线性码（二维码）的双容错数据布局构造问题进行了研究，提出了最优冗余的 BG-ORDP 布局构造方法和高扩展性的 BG-HEDP 布局构造方法，分析表明，BG-HEDP 的冗余率十分接近最优值，而在扩展性上有着非常明显的优势。

我们目前的一些工作，对于双容错数据布局问题的研究仅仅是刚起步，只得到了一些理论上的结果，尚有很多工作有待开展，才能达到应用于实践的程度。例如，对一类线性码，如何用图很好地描述，如何将对应的数据布局等问题转化为图论问题，都有待研究。再如，如前所述，容量通常不是最缺乏的，因此可以考虑非最优冗余布局的研究，其性能上可能有一些优势。再有，对于非公式化布局构造方法的研究，即通过搜索、优化等方法构造双容错布局，一方面可扩展 BG-HEDP 等布局的实用范围，另一方面，还可促进图论领域 P1F 等问题的研究，是理论意义和实践价值都很高的工作。另外，国内外相关研究工作，基本是以 Hellerstein 等人提出的 5 个理论指标评价编码和布局方案，读写性能分析和优化的工作少之又少，而这对双容错布局的实践应用又是十分重要的，在这方面展开工作，进行理论研究和仿真实验，应该是非常有意义的。

致谢 感谢南开大学科学计算所和南开大学创新基金对本文工作的支持

参考文献:

[1] Patterson D A, Gibson G A, Katz R H. A case for redundant arrays of inexpensive disks (RAID)[A]. ACM International Conference on Management of Data[C]. 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[J],

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] 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.

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

[9] Darryn Bryant, Barbara M Maenhaut, Ian M Wanless. A family of perfect factorisations of complete bipartite graphs[J]. Journal of Combinatorial Theory, Series A, 2002, 98(2): 328–342.

作者简介:

王　刚 男, 1974 年 3 月出生于天津市. 现为南开大学信息技术科学学院副教授. 从事海量存储、并行计算方面的研究工作. Email: wgzwp@163.com

董沙莎 女, 1978 年 4 月出生于湖北枝江. 现为南开大学信息技术科学学院博士生. 从事海量存储方面的研究工作. Email: shashadong@icbc.com.cn