

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



WIPO | PCT



(10) 国际公布号
WO 2014/005279 A1

(43) 国际公布日
2014 年 1 月 9 日 (09.01.2014)

- (51) 国际专利分类号:
G06F 12/00 (2006.01) *G06F 11/14* (2006.01)
- (21) 国际申请号: PCT/CN2012/078100
- (22) 国际申请日: 2012 年 7 月 3 日 (03.07.2012)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (71) 申请人 (对除美国外的所有指定国): 北京大学深圳研究生院 (PEKING UNIVERSITY SHENZHEN GRADUATE SCHOOL) [CN/CN]; 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。
- (72) 发明人: 及
- (75) 发明人/申请人 (仅对美国): 李挥 (LI, Hui) [CN/CN]; 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。侯韩旭 (HOU, Hanxu) [CN/CN]; 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。缪伟豪 (MIU, Weihao) [CN/CN]; 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。李硕彦 (LI, Shuoyan) [CN/CN]; 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。

(CN)。朱兵 (ZHU, Bing) [CN/CN]; 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。郭涵 (GUO, Han) [CN/CN]; 中国广东省深圳市南山区西丽镇丽水路深圳大学城北大校区, Guangdong 518055 (CN)。

(74) 代理人: 深圳市科吉华峰知识产权事务所 (SHENZHEN KINDWALF INTELLECTUAL PROPERTY FIRM); 中国广东省深圳市南山区深南西路深南花园裙楼 A 区四层 402 室, Guangdong 518057 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH,

[见续页]

(54) Title: METHOD AND DEVICE FOR CONSTRUCTING DISTRIBUTED STORAGE CODE CAPABLE OF ACCURATE REGENERATION

(54) 发明名称: 一种可精确再生的分布式存储码的构建方法及装置

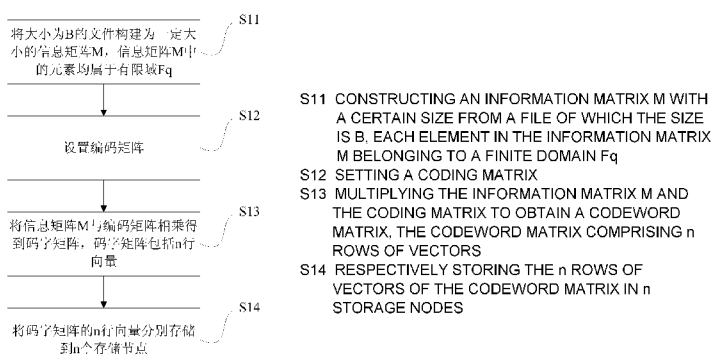


图 1 / FIG. 1

(57) Abstract: Provided is a method for constructing a distributed storage code capable of accurate regeneration, which comprises the steps of: constructing an information matrix M with a certain size from a file to be stored, each item of data in the file being a certain element in the information matrix; setting a coding matrix Ψ , the coding matrix Ψ comprising n rows of vectors, and n being the number of distributed storage nodes used for storage; multiplying the information matrix M and the coding matrix Ψ to obtain a codeword matrix C , the codeword matrix also comprising n rows of vectors; and respectively storing the n rows of vectors of the codeword matrix C in n storage nodes, all the elements in all the matrixes belonging to a finite domain F_q of which the size is q . In addition, also provided is a device for achieving the abovementioned method. The method and device for constructing a distributed storage code capable of accurate regeneration have the following beneficial effects: the realization thereof is easier, and the required finite domain is smaller.

(57) 摘要:

[见续页]



WO 2014/005279 A1



CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, **本国际公布:**

IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO,
RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,

CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

— 包括国际检索报告(条约第 21 条(3))。

提供一种可精确再生的分布式存储码的构建方法，包括如下步骤：将要存储的文件构建为其大小一定的信息矩阵 M ，其中，所述文件中的每个数据均为所述信息矩阵中的某一元素；设定编码矩阵 Ψ ，所述编码矩阵 Ψ 包括 n 行向量；其中， n 是用于存储的分布式存储节点数；将所述信息矩阵 M 和所述编码矩阵 Ψ 相乘，得到码字矩阵 C ；所述码字矩阵同样包括 n 行向量；将所述码字矩阵 C 的 n 行向量分别存储在 n 个存储节点上；其中，上述所有矩阵中的元素均属于大小为 q 的有限域 F_q 。另外还提供一种实现上述方法的装置。实施可精确再生的分布式存储码的构建方法及装置，具有以下有益效果：其实现较为容易、所需的有限域也较小。

一种可精确再生的分布式存储码的构建方法及装置

技术领域

本发明涉及网络分布式存储领域，更具体地说，涉及一种可精确再生的分布式存储码的构建方法及装置。

背景技术

本发明是在发明人先前的发明基础上一种存储数据的方法、装置及分布式网络存储系统”，(PCT/CN2012/071177)的进一步发明。

2000年诞生的网络编码（Network Coding，NC）是继C.E.Shannon发表《A Mathematical Theory of Communication》解决信道容量极限问题后的一个全新突破，它解决了网络通信中单/多源对多接收点组/广播如何达到网络容量极限的问题。传统网络通信节点上的路由交换机只完成存储转发功能。NC指出如果允许路由交换机对输入信息流进行编码再发送，将使得网络节点既实现路由功能又实现编码功能。在这种全新的体系结构下，网络性能可以达到最大流传输的理论极限。

伴随存储系统规模的发展，故障概率也显著提高，人们对存储系统也提出了更高的容错要求。现有技术主要是通过纠错码来实现分布式存储的可靠性。比之于RAID系统，常用的RAID-5产品只能提供单盘故障的修复，能够修复双盘故障的恢复的RAID-6逐渐也走入了实用。容单盘错的RAID-5系统原理来自于奇偶校验，而为达到各方面性能的优化，容双盘故障的RAID-6系统需要采用“特制”的纠错编码。

在分布式存储系统中，原始的数据文件被编码成若干小块并分别存储在各个存储节点，用户需要该数据文件时只需要下载部分存储节点的数据即可。而编码方式一般为最大距离码maximum-distance-separable (MDS)，比如RS（Reed-Solomon）码。首先，我们用有限域 F_q 中的元素个数来表示数据文件的大小， B 表示文件包含的原始个数，其中有限域的大小为 q 。使用RS编码， k 个元素的数据被编码成 n 个元素并分别存储在 n 个存储节点，用户只需要下载 n 中的任意 k 个存储节点的数据即可获得该 k 个元素的数据文件，我们称这一过程为数据重建。目前使用RS编码存储数据的分布式存储系统有RAID-6 [1], OceanStore[2] 和 Total Recall [3]。

当存储系统中有存储节点失效时，为了维持系统的可靠性和相应的容错性，我们必须修复失效节点，恢复失效节点的数据并存储在一个新的存储节点，该新节点为失

效节点的替代节点。在自修复过程中，很简单的方法是新的替代节点下载任意的现存 k 个存储节点的所有数据并重建原始数据，然后为新节点使用RS编码，精确恢复出失效节点的数据。但是这样做的缺点是：为恢复一个存储节点的数据而下载整个数据文件 B ，对于修复带宽来说是一种浪费。再生码就是解决该问题的有效方法。

在传统的RS编码中，存储在存储节点的每个数据块被看作为有限域 F_q 的一个信息符号，每个节点中的信息符号只允许在有限域 F_q 的线性操作，为修复一个失效节点的数据，修复带宽不可能小于整个数据文件的大小 B 。相反，再生码的码字是一个向量字母，因此每个数据块都可以看作是由有限域 F_q 中的 α 个符号组成的。这样有限域中的线性操作允许存储节点传输存储数据中的一部分。除了参数 α 外， d 和 β 也是再生码的参数。根据文[A. G. Dimakis, P. B. Godfrey, M. Wainwright, and K. Ramchandran, "Network coding for distributed storage systems," in Proc. 26th IEEE Int. Conf. Computer Communications (INFOCOM), Anchorage, AK, May 2007, pp. 2000–2008.]的介绍，失效节点会从现存存储节点中的 d 个节点各自下载 $\beta \leq \alpha$ 个符号，该过程为自修复过程，自修复过程中需要下载的数据总量 $d\beta$ 称为修复带宽。另外，在自修复过程中下载数据的 d 个存储节点称为帮助节点。一般情况下，再生码的平均修复带宽小于文件大小 B 。

在再生码中，功能修复是指新节点 f' 代替失效节点 f 并且 n 个存储节点能够保持数据重建和再生特性；而精确修复是指新节点 f' 代替失效节点 f 并且 f' 节点存储的数据与失效节点 f 存储的数据完全一样。相比于功能修复，精确修复有以下优势。首先，在存储系统中，服务器是知道全局的编码系数的，而功能修复则需要服务器更新修复节点的编码系数。其次，在功能修复中，因为编码系数地的变更，修复函数和数据重建的解码函数均需要重新调整。另外，对于功能修复，每次修复过程均需要增加包头以更新编码系数。而精确修复则不需要以上操作，而且精确修复码可以编码成系统码。系统再生码是指在 k 个存储节点存储了 $k\alpha$ 个没有经过任何编码操作的码字符号。对于再生码的系统码部分，精确修复系统码是一定的。

再生码的概念是由文[A. G. Dimakis, P. B. Godfrey, M. Wainwright, and K. Ramchandran, "Network coding for distributed storage systems," in Proc. 26th IEEE Int. Conf. Computer Communications (INFOCOM), Anchorage, AK, May 2007, pp. 2000–2008.]给出的，文中指出如果允许存储节点存储的数据量大于 B/k ，那么就可以减少修复带宽。

文[A. G. Dimakis, P. B. Godfrey, M. Wainwright, and K. Ramchandran, “Network coding for distributed storage systems,” in Proc. 26th IEEE Int. Conf. Computer Communications (INFOCOM), Anchorage, AK, May 2007, pp. 2000–2008]提出了失效节点功能修复的最小带宽问题, 数据重建的条件转化为网络编码的组播问题, 通过分析该网络的最小割得到了式(2)。文[A. Duminuco and E. Biersack, “A practical study of regenerating codes for peer-to-peer backup systems,” in Proc. 29th IEEE Int. Conf. Distributed Computing Systems (ICDCS), Jun. 2009, pp. 376–384]研究了再生码的计算复杂度问题, 文中考虑的是使用随机线性网络编码的功能修复。

精确再生码问题首先由文[Y. Wu and A. Dimakis, “Reducing repair traffic for erasure coding-based storage via interference alignment,” in Proc. IEEE Int. Symp. Information Theory (ISIT), Seoul, South Korea, Jul. 2009, pp. 2276–2280]给出。文中指出, 当 $k=2, d=n-1$ 时可以构造出MSR精确码, 其中编码策略是基于无线网络中干扰队列的概念。但是文中并没有精确的构造出MSR码而且构造该码需要相当大的有限域。

对于一般化参数的精确再生码, 文[K. V. Rashmi, N. B. Shah, P. V. Kumar, and K. Ramchandran, “Explicit construction of optimal exact regenerating codes for distributed storage,” in Proc. 47th Annu. Allerton Conf. Communication, Control, and Computing, Urbana-Champaign, IL, Sep. 2009, pp. 1243–1249]给出了 $d=n-1, k$ 为任意值的MBR码。在文中提出的MBR码中, 因为在失效节点的精确再生过程中没有涉及计算, 所以构造该码的复杂度很低, 有限域的大小也仅需要 n^2 。另外, 该文还构造了 $d=k+1$ 的精确MSR码。文[Y. Wu, “A construction of systematic MDS codes with minimum repair bandwidth,” IEEE Trans. Inf. Theory, 2012]综合考虑了功能修复和精确修复, 其中参数 $d=k+1, n>2k$ 。但是构造该码的复杂度很高而且需要很大的有限域。

综上所述, 在现有技术中, 虽然对精确再生码有所研究, 但普遍缺乏较为具体的实现方法、所需的有限域较大。

发明内容

本发明要解决的技术问题在于, 针对现有技术的上述缺乏具体的实现方法、所需有限域较大的缺陷, 提供一种容易实现、所需有限域较小的一种可精确再生的分布式存储码的构建方法及装置。

本发明解决其技术问题所采用的技术方案是：构造一种可精确再生的分布式存储码的构建方法，所述分布式存储码为最小修复带宽再生码或最小存储带宽再生码，所述构建方法包括如下步骤：

A)将要存储的文件构建为其大小一定的信息矩阵 M ，其中，所述文件中的每个数据均为所述信息矩阵中的某一元素；

B)设定编码矩阵 ψ ，所述编码矩阵 ψ 包括 n 行向量；其中， n 是用于存储的分布式存储节点数；

C)将所述信息矩阵 M 和所述编码矩阵 ψ 相乘，得到码字矩阵 C ；所述码字矩阵同样包括 n 行向量；

D)将所述码字矩阵 C 的 n 行向量分别存储在 n 个存储节点上；

其中，上述所有矩阵中的元素均属于大小为 q 的有限域 F_q 。

更进一步地，所述编码矩阵 ψ 的为 $n \times d$ 矩阵，所述信息矩阵 M 为 $d \times \alpha$ 矩阵；所述编码矩阵 ψ 为事先确定且独立于所述存储文件；

其中， d 是所述数据再生时的帮助节点数， α 是所述码字矩阵 C 中每个数据块包含的有限域 F_q 中的元的个数。

更进一步地，所述编码矩阵 ψ 由两个子矩阵组合或运算后组合而成；所述信息矩阵 M 由两个子矩阵组合而成，信息矩阵 M 的元素是对称的。

更进一步地，所述分布式存储码为最小修复带宽再生码；所述步骤 A) 中，所述信息矩阵 M 为 $d \times d$ 的对称矩阵， $M = \begin{bmatrix} S & T \\ T^t & 0 \end{bmatrix}$ ；所述 S 为 $k \times k$ 的矩阵，其上三角部分共有 $\binom{k+1}{2}$ 个不同元素，其下三角部分与其上三角部分对称；所述 T 为 $k \times (d-k)$ 矩阵，其具有 $k(d-k)$ 个元；所述 T^t 为所述 T 的转置矩阵；所述 $\binom{k+1}{2} + k(d-k)$ 为所述信息矩阵 M 包括的信息符号个数；

所述步骤 B) 中，其编码矩阵 $\psi = [\phi \ \Delta]$ ；其中， ϕ 为 $(n \times k)$ 矩阵， Δ 为 $(n \times (d-k))$ 矩阵；且矩阵 ϕ 和矩阵 Δ 满足：矩阵 ϕ 中的任意 d 行线性独立以及矩阵 Δ 中的任意 k 行线性独立。

更进一步地, 所述矩阵 ϕ 和矩阵 Δ 分别为柯西矩阵或范德蒙德矩阵。

更进一步地, 所述分布式存储码为最小存储带宽再生码; 所述步骤 A) 中, 其信息矩阵 $M = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$, 其中, S_1 和 S_2 均为 $(\alpha \times \alpha)$ 矩阵, S_1 和 S_2 的上三角部分分别由 $\binom{\alpha+1}{2}$ 个不同信息符号填充, 选择 S_1 和 S_2 的下三角部分使得 S_1 和 S_2 均为对称矩阵;

所述步骤 B) 中, 编码矩阵 $\psi = [\phi \ \Delta\phi]$, 其中, ϕ 为 $(n \times \alpha)$ 矩阵, Δ 为 $(n \times n)$ 的对角矩阵; 选择编码矩阵 ψ 包括的元素使得所述编码矩阵 ψ 中的任意 d 行向量线性独立、矩阵 ϕ 中的任意 α 行向量线性独立以及矩阵 Δ 中的 n 个对角元素均不同。

更进一步地, 所述编码矩阵 ψ 为范德蒙德矩阵; 所述编码矩阵 ψ 的第 i 行向量为 $\psi_i = [1 \ x_i \ \dots \ x_i^{d-1}]$ ($i=1, \dots, n$); 矩阵 $\Delta = \text{diag}\{x_1^\alpha, x_2^\alpha, \dots, x_n^\alpha\}$; 其中, 有限域 F_q 的大小不小于 $n(d-k+1)$, 其元素表示为 $x_i = g^{i-1}$, g 为有限域 F_q 的乘法群的生成元。

本发明还涉及一种实现上述可精确再生的分布式存储码的构建方法的装置, 所述分布式存储码为最小修复带宽再生码或最小存储带宽再生码; 所述构建装置包括:

信息矩阵构建单元: 用于将要存储的文件构建为其大小一定的信息矩阵 M , 其中, 所述文件中的每个数据均为所述信息矩阵中的某一元素;

编码矩阵设置单元: 用于设定编码矩阵 ψ , 所述编码矩阵 ψ 包括 n 行向量; 其中, n 是用于存储的分布式存储节点数;

码字矩阵取得单元: 用于将所述信息矩阵 M 和所述编码矩阵 ψ 相乘, 得到码字矩阵 C ; 所述码字矩阵同样包括 n 行向量;

存储分配单元: 用于将所述码字矩阵 C 的 n 行向量分别存储在 n 个存储节点上;

其中, 上述所有矩阵中的元素均属于大小为 q 的有限域 F_q 。

更进一步地, 所述分布式存储码为最小修复带宽再生码; 所述信息矩阵构建单元包括信息矩阵形成模块、矩阵 S 形成模块、矩阵 T 形成模块及矩阵 T 转置模块; 所述信息矩阵形成模块按照 $M = \begin{bmatrix} S & T \\ T^t & 0 \end{bmatrix}$ 得到其大小为 $d \times d$ 的、对称的信息矩阵 M ; 所述矩

阵 S 形成模块用于形成 S 矩阵；所述矩阵 T 形成模块用于形成 T 矩阵；所述矩阵 T 转置模块用于由所述矩阵 T 得到其转置矩阵 T' ；所述矩阵 S 为 $k \times k$ 的矩阵，其上三角部分共有 $\binom{k+1}{2}$ 个不同元素，其下三角部分与其上三角部分对称；所述矩阵 T 为 $k \times (d-k)$ 矩阵，其具有 $k(d-k)$ 个元；所述信息矩阵 M 包括的信息符号个数为 $\binom{k+1}{2} + k(d-k)$ ；其中， d 是所述数据再生时的帮助节点数， α 是所述码字矩阵 C 中每个数据块包含的有限域 F_q 中的元的个数； k 是数据重构时需要下载其码字的存储节点数；

所述编码矩阵设置单元包括编码矩阵 ψ 形成模块、矩阵 ϕ 形成模块和矩阵 Δ 形成模块；其中，所述编码矩阵 ψ 形成模块用于按照编码矩阵 $\psi = [\phi \ \Delta]$ 得到编码矩阵；所述矩阵 ϕ 形成模块用于形成矩阵 ϕ ；矩阵 Δ 形成模块用于形成矩阵 Δ ； ϕ 为 $(n \times k)$ 矩阵， Δ 为 $(n \times (d-k))$ 矩阵；且矩阵 ϕ 和矩阵 Δ 满足：矩阵 ϕ 中的任意 d 行线性独立以及矩阵 Δ 中的任意 k 行线性独立。

更进一步地，所述分布式存储码为最小存储带宽再生码；所述信息矩阵构建单元包括矩阵 S_1 形成模块、矩阵 S_2 形成模块和信息矩阵 M 形成模块；其中，所述信息矩阵 M 形成模块用于按照 $M = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$ 得到信息矩阵；所述矩阵 S_1 形成模块用于形成 S_1 矩阵；所述矩阵 S_2 形成模块用于形成 S_2 矩阵；矩阵 S_1 和 S_2 均为 $(\alpha \times \alpha)$ 矩阵，矩阵 S_1 和 S_2 的上三角部分分别由 $\binom{\alpha+1}{2}$ 个不同信息符号填充，选择矩阵 S_1 和 S_2 的下三角部分使得矩阵 S_1 和 S_2 均为对称矩阵；

所述编码矩阵设置单元包括编码矩阵形成模块、矩阵 ϕ 形成模块、对角矩阵形成模块以及矩阵元素选择模块；其中，所述编码矩阵形成模块用于按照 $\psi = [\phi \ \Lambda\phi]$ 得到编码矩阵 ψ ；所述矩阵 ϕ 形成模块用于得到其大小为 $(n \times \alpha)$ 的 ϕ 矩阵；所述对角矩阵形成模块用于得到其大小为 $(n \times n)$ 的对角矩阵 Λ ，并将其与所述矩阵 ϕ 相乘得到 $\Lambda\phi$ ；所述矩阵元

素选择模块用于选择编码矩阵 ψ 包括的元素使得所述编码矩阵 ψ 中的任意 d 行向量线性独立、所述矩阵 ϕ 中的任意 α 行向量线性独立以及所述矩阵 Λ 中的 n 个对角元素均不同。

实施本发明的可精确再生的分布式存储码的构建方法及装置，具有以下有益效果：由于采用设定的编码矩阵，并通过该编码矩阵与信息矩阵相乘得到码字矩阵，将码字矩阵不同的行向量作为编码模块分配到不同的存储节点储存，这使得其实现较为容易、所需的有限域也较小。

附图说明

图 1 是本发明可精确再生的分布式存储码的构建方法及装置第一实施例中方法的流程图；

图 2 是所述第一实施例中信息矩阵的构建流程图；

图 3 是所述第一实施例中编码矩阵的构建流程图；

图 4 是所述第一实施例中存储节点 1 数据失效后精确再生的示意图；

图 5 是所述第一实施例中装置的结构示意图；

图 6 是本发明可精确再生的分布式存储码的构建方法及装置第二实施例中存储节点 1 数据失效后精确再生的示意图；

图 7 是所述第二实施例中装置的结构示意图。

具体实施方式

下面将结合附图对本发明实施例作进一步说明。

如图 1 所示，在本发明可精确再生的分布式存储码的构建方法及装置第一实施例中，其分布式存储码的构建方法包括如下步骤：

步骤 S11 将大小为 B 的文件构建为一定大小的信息矩阵 M ，信息矩阵 M 中的元素均属于有限域 F_q ：在本步骤中，将需要存储的文件（其大小为 B ）转换为信息矩阵 M ，该信息矩阵具有设定的大小和形式，在不同的情况下，该信息矩阵的大小和格式不同；这与希望构建的码的类型有关。在第一实施例中，以构建一个 MBR 码为例，说明了其信息矩阵 M 的具体构建方法；在第二实施例中，则以一个构建一个 MSR 码为例，说明其信息矩阵的具体构建方法及大小。关于具体的步骤，稍后将分别描述。

步骤 S12 设置编码矩阵：在本步骤中，设置编码矩阵 ψ ，该编码矩阵 ψ 同样与不同类型的码有关，因此，在第一实施例和第二实施例中其构建的步骤及得到的编码矩阵 ψ

也是稍有不同。关于其具体步骤，同样在稍后分别描述。

步骤 S13 将信息矩阵 M 与编码矩阵相乘得到码字矩阵，码字矩阵包括 n 行向量：在本步骤中，将上述得到的信息矩阵和编码矩阵相乘，得到码字矩阵 C ，该码字矩阵 C 包括了 n 个行向量。

步骤 S14 将码字矩阵的 n 行向量分别存储到 n 个存储节点：在本步骤中，将，上述得到的码字矩阵 C 的 n 个行向量分别取出，并存储到不同的储存节点中。这些行向量作为存储节点存储的编码模块被存储。

在第一实施例中，希望得到并存储在各存储节点中的是 RS 再生码的一种，MBR 码。这种码是 RS 码的一个子集，其自然满足再生码的一些共性。具体来讲，在第一实施例中，于 $[n,k,d]$ 再生码的参数， k 和 d 满足 $k \leq d \leq n-1$ (1)

其中， k 是正常下载时需要由该节点下载存储编码模块内容的存储节点数， d 是修复数据需要下载的存储节点数（即失效节点需要下载的节点数），而 n 是存储该文件内容的总的存储节点数。

文[Y. Wu, A. G. Dimakis, and K. Ramchandran, “Deterministic regenerating codes for distributed storage,” in Proc. 45th Annu. Allerton Conf. Control, Computing, and Communication, Urbana-Champaign, IL, Sep. 2007.]给出了再生码的参数所必须要满足的

$$\text{条件: } B \leq \sum_{i=0}^{k-1} \min(\alpha, (d-i)\beta) \quad (2)$$

该条件也是再生码的主要结论。显然，我们想同时减小参数 α 和 β 的值。最小化 α 值就得到了最小存储的方法，最小化 β 可以得到最小修复带宽的方法。从式 (2) 可以推导出参数 α 和 β 不可能同时减小，因此在参数 α 和 β 的选择上有一个折中。该折中曲线上的两个极点码分别称为最小存储码（minimum storage regeneration，MSR）和最小修复带宽码（minimum bandwidth regeneration，MBR）。最小化 α 得到 MSR 码的参数满足

$$\alpha = \frac{B}{k}$$

$$\beta = \frac{B}{k(d-k+1)} \quad (3)$$

同样的, 最小化 β 得到MBR码的参数满足

$$\beta = \frac{2B}{k(2d-k+1)}$$

$$\alpha = \frac{2dB}{k(2d-k+1)} = d\beta \quad (4)$$

最佳的 $[n,k,d]$ 再生码的参数 (α, β, B) 满足以下两个条件:

- 1) 参数 (α, β, B) 在式 (1) 中取等号;
- 2) 减少 α 或者 β 会导致新的参数不满足式 (1)。

MSR再生码的参数 (α, β, B) 满足 (3), 而MBR码的参数 (α, β, B) 满足 (4)。

显然, MSR再生码和MBR再生码均是最佳再生码。在本发明实施例中, 由于任意的正整数 β 均可以由并行的多个 $\beta=1$ 组成, 取 $\beta=1$ 。

令 $\beta=1$, (3) 和 (4) 分别为 $\alpha = d - k + 1$ (5)

$B = k(d - k + 1)$ (6) 以及

$$\alpha = d \quad (7)$$

$$B = kd - \binom{k}{2} \quad (8)。$$

由于在本发明实施例中使用矩阵乘积框架构造精确修复码 (包括 MBR 码和 MSR 码), 在该框架下, 存储节点中的每个码字可以用码字矩阵 $C(n \times \alpha)$ 的第 i 行表示, 码字矩阵 C 的每一行均有 c 个符号, 而码字矩阵 C 是由

$$C = \psi M \quad (9)$$

求出, 其中 ψ 为 $n \times d$ 的编码矩阵, M 为 $d \times \alpha$ 的信息矩阵。编码矩阵 ψ 是事先就确定的且独立于信息符号 B 。信息矩阵 M 包含了 B 个信息符号, 其中的符号有可能是一样

的。码字矩阵 C 的第 i 行可以表示为

$$c_i^t = \psi_i^t M \quad (10),$$

其中 ψ_i^t 为编码矩阵 Ψ 的第 i 行, 字母 “t” 用来表示矩阵的转置。在本发明具体实施例中, 所有的符号均属于大小为 q 的有限域 F_q 。

数据重建是指客户端从任意的 k 个存储节点获取 $k\alpha$ 符号并解码出信息矩阵 M 。

客户端下载的 k 个存储节点用 $\{i_1, i_2, \dots, i_k\}$ 表示, 第 j 节点将信息向量 $\psi_{i,j}^t M$ 传输给客户端。这样客户端可以收到数据矩阵 $\Psi_{DC} M$, 其中 Ψ_{DC} 是矩阵 Ψ 的 k 行 $\{\psi_{i_1}, \dots, \psi_{i_k}\}$ 子矩阵。所以, 客户端就可以利用矩阵 Ψ 和 M 的特性解码出信息码字。这个过程并不涉及数据或存储节点上的数据失效。

当数据失效时, 需要再生该存储节点的失效数据。在存储节点的失效数据的再生过程中, μ_i 为长度为 α 的向量, 它为向量 ψ_i 的一部分。为了恢复失效节点 f , 代替节点 f 的新节点需要从现存存储节点中选择 d 个 $\{h_1, \dots, h_d\}$ 并各下载一个符号, 这 d 个节点称为帮助节点, 每个帮助节点传输一个符号给新节点, 该符号为其存储的 α 个符号的内部运算, 帮助节点 h_j 传输的符号为 $\psi_{h_j}^t M \mu_f$ 。因此代替节点可以获得矩阵 $\Psi_{repair} M \mu_f$, 其中 Ψ_{repair} 为矩阵 Ψ 的 d 行 $\{\psi_{h_1}, \dots, \psi_{h_d}\}$ 子矩阵。在再生过程中, 各个帮助节点只需要知道失效节点 f 的编码系数并不需要其他的编码系数。

为了能够修复失效节点的数据, 帮助节点的个数 d 必须满足第一个不等式。有限域 F_q 中的再生码的参数集合为 $\{n, k, d, \alpha, \beta, B\}$, 其中 $\{n, k, d\}$ 为主参数, 而 $\{\alpha, \beta, B\}$ 为辅参数。

在第一实施例中, 为构造的精确 MBR 码, 使参数 $\beta=1, k \leq d \leq n-1$ 。参数 $\{\alpha, B\}$

满足式 (7) (8), (8) 可以写成 $B = \binom{k+1}{2} + k(d-k)$ 。所以 $[n, k, d]$ MBR 码的参数为

($\alpha = d$, $\beta = 1$, $B = \binom{k+1}{2} + k(d-k)$)。令 S 为 $k \times k$ 的矩阵, 该矩阵的上三角部分共有 $\binom{k+1}{2}$ 个不同元素, 选择矩阵的下三角部分的元素是使矩阵 S 为对称矩阵, B 中剩下的 $k(d-k)$ 符号用来填充矩阵 $T(k \times (d-k))$ 。信息矩阵 M 为 $(d \times d)$ 的对称矩阵, 且

$$M = \begin{bmatrix} S & T \\ T^t & 0 \end{bmatrix}。 (11)$$

信息矩阵 M 的对称性会在失效节点的恢复过程中用到。请参见图 2, 在图 2 中, 示出了第一实施例中信息矩阵 M 的一个具体的构建流程, 其具体包括:

步骤 S21 分别形成矩阵 S 的上、下三角部分: 在本步骤中, 按照上述构建矩阵 S 的方法, 先构建该矩阵的上三角部分, 并使其与该矩阵的下三角部分对称, 进而得到矩阵 S 。

步骤 S22 形成矩阵 T : 在本步骤中, 将构建上述矩阵 S 后还没有用到的、原文件中包括的 $k(d-k)$ 个符号用于填充矩阵 T , 进而得到其大小为 $k \times (d-k)$ 的矩阵 T 。

步骤 S23 得到矩阵 T 的转置矩阵: 由于已经得到矩阵 T , 在本步骤中, 求得矩阵 T 的转置矩阵。

步骤 S24 组合上述矩阵 S 、矩阵 T 以及矩阵 T 的转置矩阵, 得到信息矩阵 M : 在本步骤中, 按照上面所述的信息矩阵的构成方法 (公式 11), 由上述步骤 S21-S23 中得到的各矩阵求得信息矩阵。

对于第一实施例中的编码矩阵而言, 编码矩阵 ψ 为 $(n \times d)$ 矩阵 $\psi = [\phi \Delta]$, 其中 ϕ 为 $(n \times k)$ 矩阵, Δ 为 $(n \times (d-k))$ 矩阵。矩阵 ϕ 和 Δ 满足以下两个条件: 矩阵 ϕ 中的任意 d 行线性独立; 矩阵 Δ 中的任意 k 行线性独立; 这两个条件可以通过选择柯西 (cauchy) 矩阵或者范德蒙德矩阵来满足。请参见图 3, 在图 3 中, 示出了第一实施例中编码信息矩阵 ψ 的一个具体的构建流程, 包括:

步骤 S31 形成矩阵 ϕ : 在本步骤中, 按照上述要求得到 $n \times k$ 的矩阵 ϕ ; 该矩阵

当然需要满足上面所讲的两个条件。

步骤 S32 形成矩阵 Δ ：在本步骤中，按照上述要求得到 $n \times (d-k)$ 的矩阵 Δ ；同样，该矩阵同样需要满足上面所讲的两个条件。

步骤 S33 组合上述矩阵得到编码矩阵 Ψ ：由于已经得到上述的两个矩阵，在本步骤中，按照上述公式 $\Psi = [\phi \Delta]$ 得到编码矩阵。

当得到上述的信息矩阵 M 和编码矩阵 Ψ 后，将其相乘，就得到码字矩阵 C ，将得到的码字矩阵 C 的 n 个行向量分别存储到 n 个存储节点，即完成分布式网络存储的编码过程。

对于以上给出的精确再生码（MBR 码），任意失效节点的精确再生通过从 $(n-1)$ 现存节点中任选 d 个节点并各下载一个信息符号。其过程如下：令 ψ_f^t 为失效节点 f 的编码向量，所以存储在节点 f 的 d 个信息符号为

$$\psi_f^t M \quad (12).$$

为恢复失效节点 f 存储数据的新节点连接任意 d 个帮助节点 $\{h_j | j=1, \dots, d\}$ ，而后帮助节点 h_j 计算该节点的内部乘积 $\psi_f^t M \psi_f$ 并将其传输给替代节点，在该恢复过程中

向量 $\mu_f = \psi_f$ 。最终替代节点获得了 d 个符号 $\psi_{repair} M \psi_f$ ，其中

$$\beta \psi_{repair} = \begin{bmatrix} \psi_{h_1}^t \\ \psi_{h_2}^t \\ \dots \\ \psi_{h_d}^t \end{bmatrix}.$$

在构造过程中易知矩阵 Ψ_{repair} 可逆，所以替代节点可以通过左乘矩阵 Ψ_{repair}^{-1} 获得 $M \psi_f$ 。又因为 M 是对称的，所以 $(M \psi_f)^t = \psi_f^t M$ ，(13)

而 $\psi_f^t M$ 即为存储在失效节点的数据。

当所有存储节点上（或下载所述涉及的存储节点）的数据都没有失效时，对以上给出的精确再生码，客户端可以通过下载任意的 k 个存储节点的数据即可获得所有的信息符号 B 。其过程如下：客户端下载的 k 个存储节点数据的编码矩阵为

$$\psi_{DC} = [\phi_{DC} \Delta_{DC}] \quad (14),$$

ψ_{DC} 为矩阵 Ψ 的 $(k \times d)$ 子矩阵, 也就是矩阵 Ψ 的 k 行向量组成的子矩阵。所以客户端可以获得

$$\psi_{DC} M = [\phi_{DC} S + \Delta_{DC} T^t \phi_{DC} T] \quad (15)。$$

由以上构造过程知矩阵 ψ_{DC} 是可逆的, 用矩阵 $\psi_{DC} M$ 左乘矩阵 ψ_{DC}^{-1} 可以计算出原始数据 T 以及 S 。

此外, 在第一实施例中, 一个具体的失效数据修复的例子如图 4 所示, 在图 4 中, $n=6$, $k=3$, $d=4$, 以及 $\alpha=d=4$, $B=9$ 。选择有限域的大小 $q=7$ 。矩阵 S 和 T 分别为

$$S = \begin{bmatrix} u_1 & u_2 & u_3 \\ u_2 & u_4 & u_5 \\ u_3 & u_5 & u_6 \end{bmatrix}, \quad T = \begin{bmatrix} u_7 \\ u_8 \\ u_9 \end{bmatrix} \quad (16)。$$

所以信息矩阵 M 为

$$M = \begin{bmatrix} u_1 & u_2 & u_3 & u_7 \\ u_2 & u_4 & u_5 & u_8 \\ u_3 & u_5 & u_6 & u_9 \\ u_7 & u_8 & u_9 & 0 \end{bmatrix} \quad (17)。$$

编码向量 Ψ 为 (6×4) 的范德蒙德矩阵

$$\Psi = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 1 \\ 1 & 3 & 2 & 6 \\ 1 & 4 & 2 & 1 \\ 1 & 5 & 4 & 6 \\ 1 & 6 & 1 & 6 \end{bmatrix} \quad (18)。$$

图 4 给出了该例子的码字矩阵 $C = \Psi M$ 并给出了当节点 1 失效时的精确再生过程。为了再生出节点 1 的存储信息, 帮助节点 (在该例子中为节点 2, 4, 5, 6) 各自传输

它们的内部乘积 $\psi_i^t M [1 \ 1 \ 1 \ 1]^t$ ($i=2, 4, 5, 6$)。然后替代节点即可通过乘 ψ_{repair}^{-1} 恢复出失效节点的数据, 其中

$$\psi_{repair} = \begin{bmatrix} 1 & 2 & 4 & 1 \\ 1 & 4 & 2 & 1 \\ 1 & 5 & 4 & 6 \\ 1 & 6 & 1 & 6 \end{bmatrix} \quad (19)。$$

在第一实施例中, 还涉及一种实现上述方法的装置。如图 5 所示, 该装置包括信息矩阵构建单元 51、编码矩阵设置单元 52、码字矩阵取得单元 53 以及存储分配单元 54。其中, 信息矩阵构建单元 51 用于将要存储的文件构建为其大小一定的信息矩阵 M , 其中, 该文件中的每个数据均为所述信息矩阵中的某一元素; 编码矩阵设置单元 52 用于设定编码矩阵 ψ , 编码矩阵 ψ 包括 n 行向量, n 是用于存储的分布式存储节点数; 码字矩阵取得单元 53 用于将上述信息矩阵 M 和编码矩阵 ψ 相乘, 得到码字矩阵 C ; 所述码字矩阵 C 同样包括 n 行向量; 存储分配单元 54 用于将所述码字矩阵 C 的 n 行向量分别存储在 n 个存储节点上; 在第一实施例中, 上述所有矩阵中的元素均属于大小为 q 的有限域 F_q 。

在第一实施例中, 信息矩阵构建单元 51 包括信息矩阵形成模块 511、矩阵 S 形成模块 512、矩阵 T 形成模块 513 及矩阵 T 转置模块 514; 信息矩阵形成模块 511 按照 $M = \begin{bmatrix} S & T \\ T^t & 0 \end{bmatrix}$ 得到其大小为 $d \times d$ 的、对称的信息矩阵 M ; 矩阵 S 形成模块 512 用于形成 S 矩阵(如何形成的步骤已在上面的描述中记载); 矩阵 T 形成模块 513 用于形成 T 矩阵; 而矩阵 T 转置模块 514 用于由所述矩阵 T 得到其转置矩阵 T^t ; 其中, 上述矩阵 S 为 $k \times k$ 的矩阵, 其上三角部分共有 $\binom{k+1}{2}$ 个不同元素, 其下三角部分与其上三角部分对称; 矩阵 T 为 $k \times (d-k)$ 矩阵, 其具有 $k(d-k)$ 个元; 信息矩阵 M 包括的信息符号个数为 $\binom{k+1}{2} + k(d-k)$; 其中, d 是所述数据再生时的帮助节点数, α 是所述码字矩阵 C 中每个数据块包含的有限域 F_q 中的元的个数; k 是数据重构时需要下载其码字的存储节点数;

而编码矩阵设置单元包括编码矩阵 ψ 形成模块521、矩阵 ϕ 形成模块522和矩阵 Δ 形成模块523；其中，编码矩阵 ψ 形成模块521用于按照编码矩阵 $\psi = [\phi \ \Delta]$ 得到编码矩阵；矩阵 ϕ 形成模块522用于形成矩阵 ϕ ；矩阵 Δ 形成模块523用于形成矩阵 Δ ；其中，正如上面所述， ϕ 为 $(n \times k)$ 矩阵， Δ 为 $(n \times (d-k))$ 矩阵；且矩阵 ϕ 和矩阵 Δ 满足：矩阵 ϕ 中的任意 d 行线性独立以及矩阵 Δ 中的任意 k 行线性独立。

在第二实施例中，以一个MSR码为例，说明精确修复分布式存储码的构建方法及装置。其与第一实施例有较多的相同之处，但是，由于编码的不同，使得其方法中的一些步骤或装置中的一些模块细节有一些差别。为此，将较为仔细地描述这些差别。在第二实施例中，MSR再生码的参数为 $\beta = 1$ 以及 $[n, k, d \geq 2k-2]$ ，参数 $\{\alpha, B\}$ 满足式(5)和(6)。以 $d=2k-2$ 的MSR精确再生码为例详细说明如下：

当 $d=2k-2$ 时有

$$\alpha = d - k + 1 = k - 1 \quad (20) \text{ 以及}$$

$$d = 2\alpha \quad (21),$$

$$B = k\alpha = \alpha(\alpha + 1) \quad (22)。$$

定义 $(d \times \alpha)$ 信息矩阵 M 为 $M = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$ (23)，其中 S_1 和 S_2 均是 $(\alpha \times \alpha)$ 矩阵， S_1

和 S_2 的上三角部分分别由 $\binom{\alpha+1}{2}$ 个不同信息符号填充。所以，所有的 $B = \alpha(\alpha + 1)$ 信息符号都已经填充到矩阵 S_1 和 S_2 中。然后选择矩阵 S_1 和 S_2 的下三角部分使得矩阵 S_1 和 S_2 均是对称矩阵。

在第二实施例中，编码矩阵 Ψ ($n \times d$) 为

$$\psi = [\phi \ \Lambda\phi] \quad (24)。$$

其中 ϕ 为 $(n \times \alpha)$ 矩阵， Λ 为 $(n \times n)$ 的对角矩阵。选择 Ψ 合适的元素以满足以下

条件：矩阵 Ψ 中的任意 d 行向量线性独立；矩阵 ϕ 中的任意 α 行向量线性独立；矩阵 Λ 中的 n 个对角元素均不同。

以上三个条件可以通过合适的选择矩阵 Ψ 为范德蒙德矩阵来满足。也就是令矩阵 Ψ 的第 i 行向量为 $\psi_i = [1 \ x_i \ \dots \ x_i^{d-1}]$ ($i = 1, \dots, n$)，矩阵 Λ 为 $\Lambda = \text{diag}\{x_1^\alpha, x_2^\alpha, \dots, x_n^\alpha\}$ 。为了满足第三个条件，需要选择有限域 F_q 的大小不小于 $n(d-k+1)$ ，可以用 $x_i = g^{i-1}$ 表示有限域的元素，其中 g 为有限域 F_q 的乘法群的生成元。

在第二实施例的 MSR 码构造框架中，码字矩阵 C 为 $C = \Psi M$ ，矩阵 C 中第 i 行包含有 α 个码字符号，这 α 个码字符号存储在第 i 个存储节点中。

值得一提的是，在第二实施例中，当构建信息矩阵和编码矩阵时，同样也具有与第一实施例中相似的步骤。只不过由于其编码矩阵和信息矩阵的结构或表示不同，构建时的具体步骤也稍有不同。例如，在第二实施例中，构建信息矩阵时，不是如第一实施例中一样通过矩阵 S 、矩阵 T 以及矩阵 T 的转置矩阵来得到信息矩阵，而是通过先构建矩阵 S_1 、矩阵 S_2 ，然后再得到信息矩阵；在第二实施例中构建编码矩阵时，也是先得到符合上述定义的矩阵 ϕ ，得到对角矩阵 Λ ，之后再按照上述公式 24，得到编码矩阵 ψ 。由此可知，除了具体内容不同，上述步骤还是与第一实施例中相似的。所以，在第二实施例中不再具体描述信息矩阵 M 和编码矩阵 ψ 得到的具体步骤。

在第二实施例中得到的 $[n, k, d]$ MSR 码的精确恢复过程和数据重建过程分别为描述如下。

对于第二实施例中构造出的 MSR 精确修复码，为了精确再生出失效节点的存储数据，可以从 $(n-1)$ 个现存节点中选择 $d=2k-2$ 个存储节点并各自下载一个符号即可精确再生该失效数据。令失效节点的编码向量为

$$\begin{bmatrix} \phi_f^t & \lambda_f \phi_f^t \end{bmatrix}_M = \phi_f^t S_1 + \lambda_f \phi_f^t S_2. \quad (25)$$

失效节点 f 的替代节点连接任意的 d 个帮助节点 $\{h_j | j=1, \dots, d\}$ ，然后帮助节点计算其内部的乘积 $\psi_{h_j}^t M \phi_f$ 并将其发送给替代节点。因此，在这里向量 $\mu_f = \phi_f$ 。替

代节点可以从 d 个帮助节点获得 d 个信息符号 $\psi_{repair} M \phi_f$ ，其中

$$\psi_{repair} = \begin{bmatrix} \psi_{h_1}^t \\ \psi_{h_2}^t \\ \dots \\ \psi_{h_d}^t \end{bmatrix}。$$

通过以上构造方法知 $(d \times d)$ 矩阵 ψ_{repair} 是可逆的。因此，替代节点获得了 $M \phi_f = \begin{bmatrix} S_1 \phi_f \\ S_2 \phi_f \end{bmatrix}$ 。

因为矩阵 S_1 和 S_2 是对称矩阵，所以替代节点可以通过转置得到 $\phi_f^t S_1$ 和 $\phi_f^t S_2$ 。这样，替代节点可以得到

$$\phi_f^t S_1 + \lambda_f \phi_f^t S_2, \quad (26)$$

这也就是失效节点存储的数据。

对于第二实施例中构造出的 MSR 精确修复码，客户端可以通过连接并下载任意 k 个存储节点的数据即可重建出所有 B 个信息符号。令客户端下载的数据的编码向量为

$$\psi_{DC} = [\phi_{DC} \quad \Lambda_{DC} \phi_{DC}] \quad (27)$$

矩阵 ψ_{DC} 为矩阵 ψ 的 $(k \times d)$ 子矩阵， ψ 的 k 行向量对应为客户端下载的 k 个存储节点的编码向量。因此，客户端可以得到符号

$$\psi_{DC} M = [\phi_{DC} \quad \Lambda_{DC} \phi_{DC}] \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} = [\phi_{DC} S_1 \quad \Lambda_{DC} \phi_{DC} S_2]。 \quad (28)$$

然后客户端计算

$$[\phi_{DC} S_1 \quad \Lambda_{DC} \phi_{DC} S_2] \phi_{DC}^T = \phi_{DC} S_1 \phi_{DC}^T + \Lambda_{DC} \phi_{DC} S_2 \phi_{DC}^T。 \quad (29)$$

令矩阵 P 和 Q 分别为

$$P = \phi_{DC} S_1 \phi_{DC}^T \quad (30)$$

$$Q = \phi_{DC} S_2 \phi_{DC}^T. \quad (31)$$

因为矩阵 S_1 和 S_2 是对称矩阵，所以矩阵 P 和 Q 也都是对称矩阵。所以式 (29) 可以写成矩阵

$$P + \Lambda_{DC} Q. \quad (32)$$

令该矩阵的第 (i, j) 元素为

$$P_{ij} + \lambda_i Q_{ij}, \quad (33)$$

同样的，该矩阵的第 (j, i) 元素为

$$P_{ji} + \lambda_j Q_{ji} = P_{ij} + \lambda_j Q_{ij} \quad (34),$$

式 (35) 中等号是因为矩阵 P 和 Q 的对称性。因为在构造码字的时候，我们选择 λ_i 互不相同，因此根据式 (33) 和 (34) 客户端可以解码出 P_{ij} 和 Q_{ij} 对于 $i \neq j$ 。

首先考虑矩阵 P ，客户端收到的编码矩阵 ϕ_{DC} 为

$$\phi_{DC} = \begin{bmatrix} \phi_1' \\ \phi_2' \\ \dots \\ \phi_{\alpha+1}' \end{bmatrix}. \quad (35)$$

矩阵 P 的所有非对角元素都是可解的，除了对角元素外，第 i 行向量的元素为

$$\phi_i' S_1 [\phi_1 \quad \dots \quad \phi_{i-1} \quad \phi_{i+1} \quad \dots \quad \phi_{\alpha+1}]. \quad (36)$$

然而，因为在构造码字的时候编码矩阵的所有子矩阵均是可逆的，所以客户端可以得到

$$\{\phi_i' S_1 | 1 \leq i \leq \alpha + 1\}. \quad (37)$$

选择其中的前 α 行，也就是

$$\begin{bmatrix} \phi_1^t \\ \phi_2^t \\ \dots \\ \phi_\alpha^t \end{bmatrix} S_1 \quad (38)。$$

式 (38) 左边的矩阵是可逆的, 所以客户端可以计算出 S_1 。同样的道理, 客户端也可以计算出数据 S_2 。

请参见图 6, 在图 6 中, 示出了第二实施例中一个 $n=6$, $k=3$, $d=4$, $\alpha=d-k+1=2$, $B=k\alpha=6$ 的具体例子中的数据修复过程。选择有限域的大小 $q=13$ 。信息矩阵 S_1 和 S_2 分别为

$$S_1 = \begin{bmatrix} u_1 & u_2 \\ u_3 & u_4 \end{bmatrix}, \quad S_2 = \begin{bmatrix} u_4 & u_5 \\ u_5 & u_6 \end{bmatrix} \quad (49),$$

所以信息矩阵 M 为

$$M = \begin{bmatrix} u_1 & u_2 \\ u_2 & u_3 \\ u_4 & u_5 \\ u_5 & u_6 \end{bmatrix} \quad (40)。$$

选择 ψ 为 $(6 \times 2)C = \psi_l^t M \begin{bmatrix} 1 & 1 \end{bmatrix}^t$ 范德蒙德矩阵

$$\psi = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 1 \\ 1 & 4 & 3 & 12 \\ 1 & 5 & 12 & 8 \\ 1 & 6 & 10 & 8 \end{bmatrix} \quad (41)。$$

所以 (6×2) 矩阵 ϕ 和 (6×6) 对角矩阵 Λ 分别为

$$\psi_{repair} = \begin{bmatrix} 1 & 2 & 4 & 8 \\ 1 & 4 & 3 & 12 \\ 1 & 5 & 12 & 8 \\ 1 & 6 & 10 & 8 \end{bmatrix} S_2 \psi_1 = \begin{bmatrix} u_4 + u_5 \\ u_5 + u_6 \end{bmatrix} d \geq 2k - 2, \quad (42).$$

$$\Lambda = \begin{bmatrix} 1 & & & & & \\ & 4 & & & & \\ & & 9 & & & \\ & & & 3 & & \\ & & & & 12 & \\ & & & & & 10 \end{bmatrix}$$

图 6 给出了 (6×2) 码字矩阵 $C = \psi M$ 以及节点失效后的精确修复过程。当节点 1 失效时, 为了再生出节点 1 的码字, 帮助节点 2, 4, 5, 6 分别传输其各自的内部乘积 $\psi_l^t M [1 \ 1]^t$, $l=2, 4, 5, 6$ 。替代节点用矩阵 ψ_{repair}^{-1} 乘其收到的符号即可解码出 $S_1 \psi_1$ 和 $S_2 \psi_1$, 其中

$$\psi_{repair} = \begin{bmatrix} 1 & 2 & 4 & 8 \\ 1 & 4 & 3 & 12 \\ 1 & 5 & 12 & 8 \\ 1 & 6 & 10 & 8 \end{bmatrix} \quad (43),$$

$$S_1 \psi_1 = \begin{bmatrix} u_1 + u_2 \\ u_2 + u_3 \end{bmatrix}, \quad S_2 \psi_1 = \begin{bmatrix} u_4 + u_5 \\ u_5 + u_6 \end{bmatrix}. \quad (44)$$

在第二实施例中还涉及一种实现上述方法的装置, 如图 7 所示, 该装置包括信息矩阵构建单元 71、编码矩阵设置单元 72、码字矩阵取得单元 73 以及存储分配单元 74。其中, 信息矩阵构建单元 71 用于将要存储的文件构建为其大小一定的信息矩阵 M , 其

中, 该文件中的每个数据均为所述信息矩阵中的某一元素; 编码矩阵设置单元 72 用于设定编码矩阵 ψ , 编码矩阵 ψ 包括 n 行向量, n 是用于存储的分布式存储节点数; 码字矩阵取得单元 73 用于将上述信息矩阵 M 和编码矩阵 ψ 相乘, 得到码字矩阵 C ; 所述码字矩阵 C 同样包括 n 行向量; 存储分配单元 74 用于将所述码字矩阵 C 的 n 行向量分别存储在 n 个存储节点上; 在第一实施例中, 上述所有矩阵中的元素均属于大小为 q 的有限域 F_q 。

所述信息矩阵构建单元 71 包括矩阵 S_1 形成模块 712、矩阵 S_2 形成模块 713 和信息矩阵形成模块 711; 其中, 信息矩阵形成模块 711 用于按照 $M = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$ 得到信息矩阵; 矩阵 S_1 形成模块 712 用于形成 S_1 矩阵; 矩阵 S_2 形成模块 713 用于形成 S_2 矩阵; 矩阵 S_1 和矩阵 S_2 均为 $(\alpha \times \alpha)$ 矩阵, 矩阵 S_1 和 S_2 的上三角部分分别由 $\binom{\alpha+1}{2}$ 个不同信息符号填充, 选择矩阵 S_1 和 S_2 的下三角部分使得矩阵 S_1 和 S_2 均为对称矩阵; 编码矩阵设置单元 72 包括编码矩阵形成模块 721、矩阵 ϕ 形成模块 722、对角矩阵形成模块 723 以及矩阵元素选择模块 724; 其中, 编码矩阵形成模块 721 用于按照 $\psi = [\phi \quad \Lambda\phi]$ 得到编码矩阵 ψ ; 矩阵 ϕ 形成模块 722 用于得到其大小为 $(n \times \alpha)$ 的 ϕ 矩阵; 对角矩阵形成模块 723 用于得到其大小为 $(n \times n)$ 的对角矩阵 Λ , 并将其与矩阵 ϕ 相乘得到 $\Lambda\phi$; 矩阵元素选择模块 724 用于选择编码矩阵 ψ 包括的元素使得所述编码矩阵 ψ 中的任意 d 行向量线性独立、矩阵 ϕ 中的任意 α 行向量线性独立以及矩阵 Λ 中的 n 个对角元素均不同。

以上所述实施例仅表达了本发明的几种实施方式, 其描述较为具体和详细, 但不能因此而理解为对本发明专利范围的限制。应当指出的是, 对于本领域的普通技术人员来说, 在不脱离本发明构思的前提下, 还可以做出若干变形和改进, 这些都属于本发明的保护范围。因此, 本发明的保护范围应以所附权利要求为准。

权利要求书

1、一种可精确再生的分布式存储码的构建方法，其特征在于，所述分布式存储码为最小修复带宽再生码或最小存储带宽再生码，所述构建方法包括如下步骤：

A)将要存储的文件构建为其大小一定的信息矩阵 M ，其中，所述文件中的每个数据均为所述信息矩阵中的某一元素；

B)设定编码矩阵 ψ ，所述编码矩阵 ψ 包括 n 行向量；其中， n 是用于存储的分布式存储节点数；

C)将所述信息矩阵 M 和所述编码矩阵 ψ 相乘，得到码字矩阵 C ；所述码字矩阵同样包括 n 行向量；

D)将所述码字矩阵 C 的 n 行向量分别存储在 n 个存储节点上；

其中，上述所有矩阵中的元素均属于大小为 q 的有限域 F_q 。

2、根据权利要求 1 所述的可精确再生的分布式存储码的构建方法，其特征在于，所述编码矩阵 ψ 的为 $n \times d$ 矩阵，所述信息矩阵 M 为 $d \times \alpha$ 矩阵；所述编码矩阵 ψ 为事先确定且独立于所述存储文件；

其中， d 是所述数据再生时的帮助节点数， α 是所述码字矩阵 C 中每个数据块包含的有限域 F_q 中的元的个数。

3、根据权利要求 2 所述的可精确再生的分布式存储码的构建方法，其特征在于，所述编码矩阵 ψ 由两个子矩阵组合或运算后组合而成；所述信息矩阵 M 由两个子矩阵组合而成，信息矩阵 M 的元素是对称的。

4、根据权利要求 3 所述的可精确再生的分布式存储码的构建方法，其特征在于，所述分布式存储码为最小修复带宽再生码；所述步骤 A) 中，所述信息矩阵 M 为 $d \times d$ 的对称矩阵， $M = \begin{bmatrix} S & T \\ T^t & 0 \end{bmatrix}$ ；所述 S 为 $k \times k$ 的矩阵，其上三角部分共有 $\binom{k+1}{2}$ 个不同元素，其下三角部分与其上三角部分对称；所述 T 为 $k \times (d-k)$ 矩阵，其具有 $k(d-k)$ 个元素；所述 T^t 为所述 T 的转置矩阵；所述 $\binom{k+1}{2} + k(d-k)$ 为所述信息矩阵 M 包括的信息符号个数；

所述步骤 B) 中，其编码矩阵 $\psi = [\phi \ \Delta]$ ；其中， ϕ 为 $(n \times k)$ 矩阵， Δ 为 $(n \times (d-k))$ 矩

阵；且矩阵 ϕ 和矩阵 Δ 满足：矩阵 ϕ 中的任意 d 行线性独立以及矩阵 Δ 中的任意 k 行线性独立。

5、根据权利要求 4 所述的可精确再生的分布式存储码的构建方法，其特征在于，所述矩阵 ϕ 和矩阵 Δ 分别为柯西矩阵或范德蒙德矩阵。

6、根据权利要求 3 所述的可精确再生的分布式存储码的构建方法，其特征在于，所述分布式存储码为最小存储带宽再生码；所述步骤 A) 中，其信息矩阵 $M = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$ ，

其中， S_1 和 S_2 均为 $(\alpha \times \alpha)$ 矩阵， S_1 和 S_2 的上三角部分分别由 $\binom{\alpha+1}{2}$ 个不同信息符号填充，

选择 S_1 和 S_2 的下三角部分使得 S_1 和 S_2 均为对称矩阵；

所述步骤 B) 中，编码矩阵 $\psi = [\phi \quad \Lambda\phi]$ ，其中， ϕ 为 $(n \times \alpha)$ 矩阵， Λ 为 $(n \times n)$ 的对角矩阵；选择编码矩阵 ψ 包括的元素使得所述编码矩阵 ψ 中的任意 d 行向量线性独立、矩阵 ϕ 中的任意 α 行向量线性独立以及矩阵 Λ 中的 n 个对角元素均不同。

7、根据权利要求 6 所述的可精确再生的分布式存储码的构建方法，其特征在于，所述编码矩阵 ψ 为范德蒙德矩阵；所述编码矩阵 ψ 的第 i 行向量为 $\psi_i = [1 \quad x_i \quad \dots \quad x_i^{d-1}]$ ($i=1, \dots, n$)；矩阵 $\Lambda = \text{diag}\{x_1^\alpha, x_2^\alpha, \dots, x_n^\alpha\}$ ；其中，有限域 F_q 的大小不小于 $n(d-k+1)$ ，其元素表示为 $x_i = g^{i-1}$ ， g 为有限域 F_q 的乘法群的生成元。

8、一种可精确再生的分布式存储码的构建装置，其特征在于，所述分布式存储码为最小修复带宽再生码或最小存储带宽再生码；所述构建装置包括：

信息矩阵构建单元：用于将要存储的文件构建为其大小一定的信息矩阵 M ，其中，所述文件中的每个数据均为所述信息矩阵中的某一元素；

编码矩阵设置单元：用于设定编码矩阵 ψ ，所述编码矩阵 ψ 包括 n 行向量；其中， n 是用于存储的分布式存储节点数；

码字矩阵取得单元：用于将所述信息矩阵 M 和所述编码矩阵 ψ 相乘，得到码字矩阵 C ；所述码字矩阵同样包括 n 行向量；

存储分配单元：用于将所述码字矩阵 C 的 n 行向量分别存储在 n 个存储节点

上;

其中, 上述所有矩阵中的元素均属于大小为 q 的有限域 F_q 。

9、根据权利要求 8 所述的分布式存储码的构建装置, 其特征在于, 所述分布式存储码为最小修复带宽再生码; 所述信息矩阵构建单元包括信息矩阵形成模块、矩阵 S 形成模块、矩阵 T 形成模块及矩阵 T 转置模块; 所述信息矩阵形成模块按照 $M = \begin{bmatrix} S & T \\ T^t & 0 \end{bmatrix}$ 得到其大小为 $d \times d$ 的、对称的信息矩阵 M ; 所述矩阵 S 形成模块用于形成 S 矩阵; 所述矩阵 T 形成模块用于形成 T 矩阵; 所述矩阵 T 转置模块用于由所述矩阵 T 得到其转置矩阵 T^t ; 所述矩阵 S 为 $k \times k$ 的矩阵, 其上三角部分共有 $\binom{k+1}{2}$ 个不同元素, 其下三角部分与其上三角部分对称; 所述矩阵 T 为 $k \times (d-k)$ 矩阵, 其具有 $k(d-k)$ 个元; 所述信息矩阵 M 包括的信息符号个数为 $\binom{k+1}{2} + k(d-k)$; 其中, d 是所述数据再生时的帮助节点数, α 是所述码字矩阵 C 中每个数据块包含的有限域 F_q 中的元的个数; k 是数据重构时需要下载其码字的存储节点数;

所述编码矩阵设置单元包括编码矩阵 ψ 形成模块、矩阵 ϕ 形成模块和矩阵 Δ 形成模块; 其中, 所述编码矩阵 ψ 形成模块用于按照编码矩阵 $\psi = [\phi \ \Delta]$ 得到编码矩阵; 所述矩阵 ϕ 形成模块用于形成矩阵 ϕ ; 矩阵 Δ 形成模块用于形成矩阵 Δ ; ϕ 为 $(n \times k)$ 矩阵, Δ 为 $(n \times (d-k))$ 矩阵; 且矩阵 ϕ 和矩阵 Δ 满足: 矩阵 ϕ 中的任意 d 行线性独立以及矩阵 Δ 中的任意 k 行线性独立。

10、根据权利要求 8 所述的分布式存储码的构建装置, 其特征在于, 所述分布式存储码为最小存储带宽再生码; 所述信息矩阵构建单元包括矩阵 S_1 形成模块、矩阵 S_2 形成模块和信息矩阵 M 形成模块; 其中, 所述信息矩阵 M 形成模块用于按照 $M = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$ 得到信息矩阵; 所述矩阵 S_1 形成模块用于形成 S_1 矩阵; 所述矩阵 S_2 形成模块用于形成 S_2 矩

阵；矩阵 S_1 和 S_2 均为 $(\alpha \times \alpha)$ 矩阵，矩阵 S_1 和 S_2 的上三角部分分别由 $\binom{\alpha+1}{2}$ 个不同信息符号填充，选择矩阵 S_1 和 S_2 的下三角部分使得矩阵 S_1 和 S_2 均为对称矩阵；

所述编码矩阵设置单元包括编码矩阵形成模块、矩阵 ϕ 形成模块、对角矩阵形成模块以及矩阵元素选择模块；其中，所述编码矩阵形成模块用于按照 $\psi = [\phi \quad \Lambda\phi]$ 得到编码矩阵 ψ ；所述矩阵 ϕ 形成模块用于得到其大小为 $(n \times \alpha)$ 的 ϕ 矩阵；所述对角矩阵形成模块用于得到其大小为 $(n \times n)$ 的对角矩阵 Λ ，并将其与所述矩阵 ϕ 相乘得到 $\Lambda\phi$ ；所述矩阵元素选择模块用于选择编码矩阵 ψ 包括的元素使得所述编码矩阵 ψ 中的任意 d 行向量线性独立、所述矩阵 ϕ 中的任意 α 行向量线性独立以及所述矩阵 Λ 中的 n 个对角元素均不同。

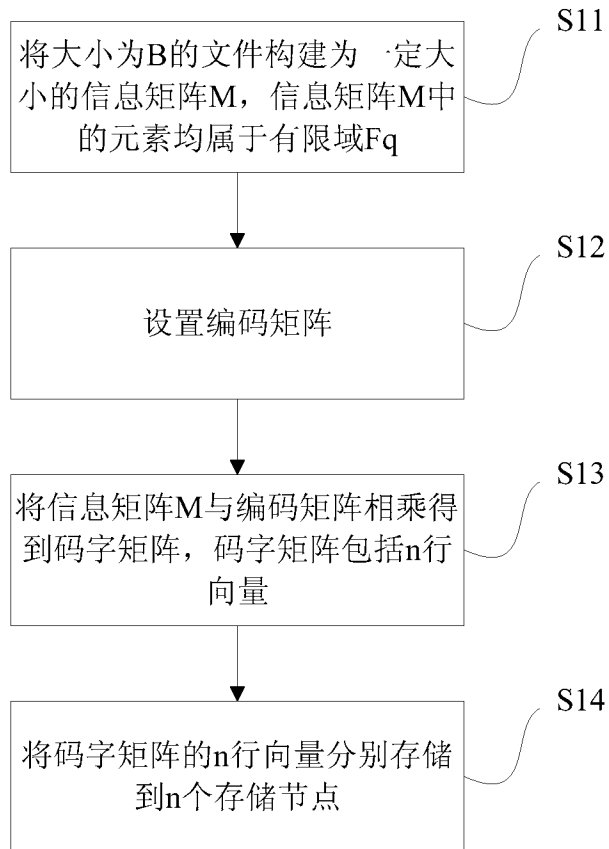


图 1

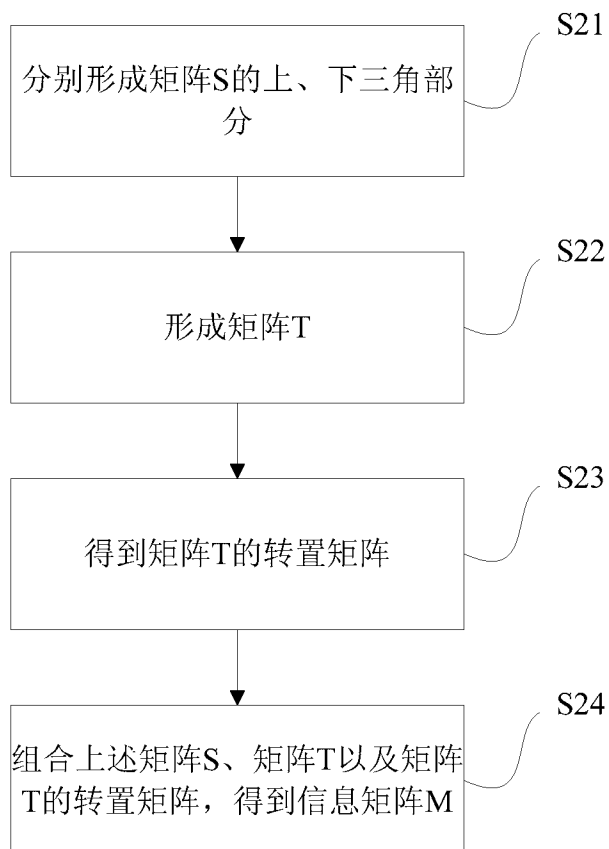


图 2

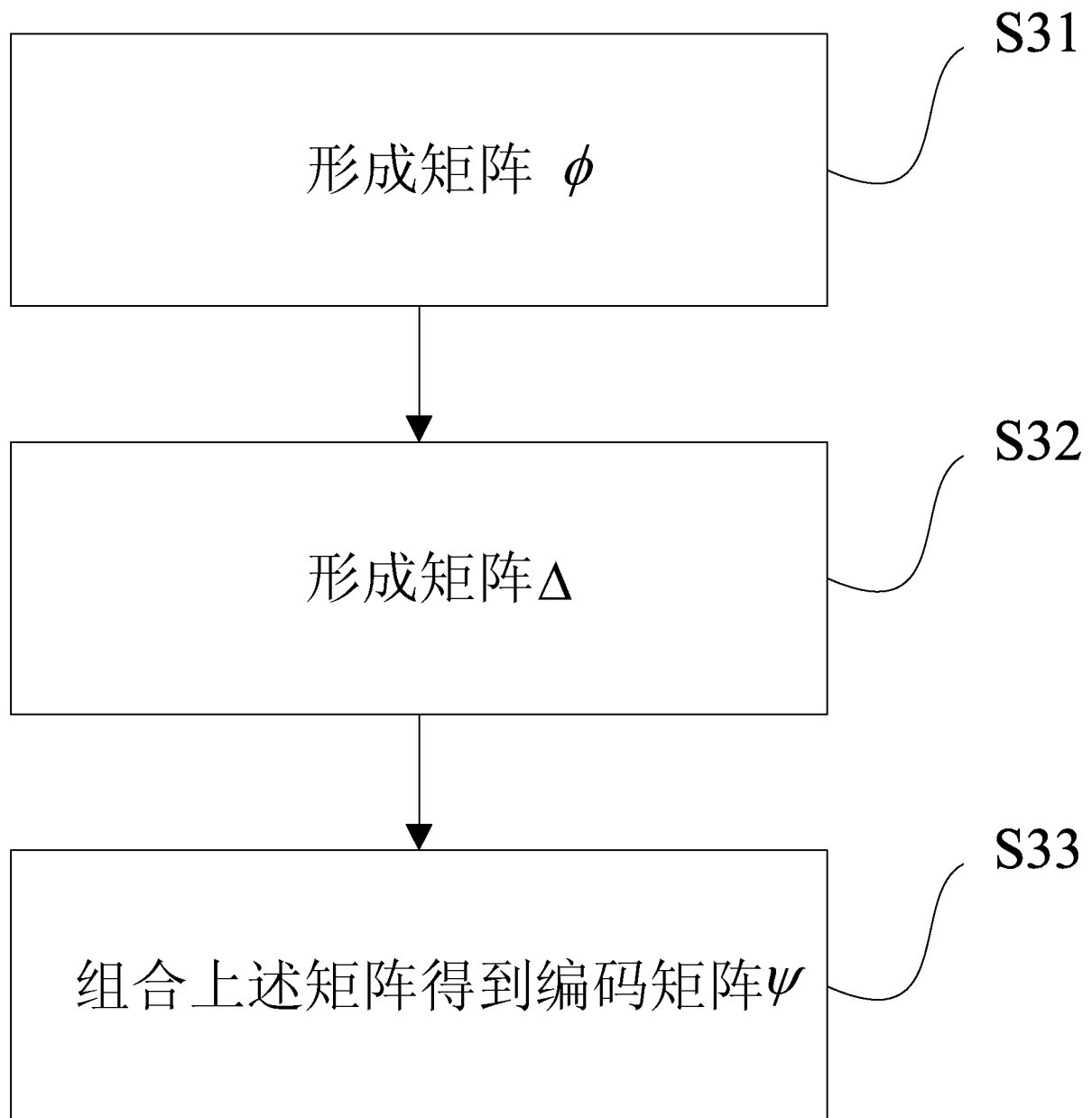


图 3

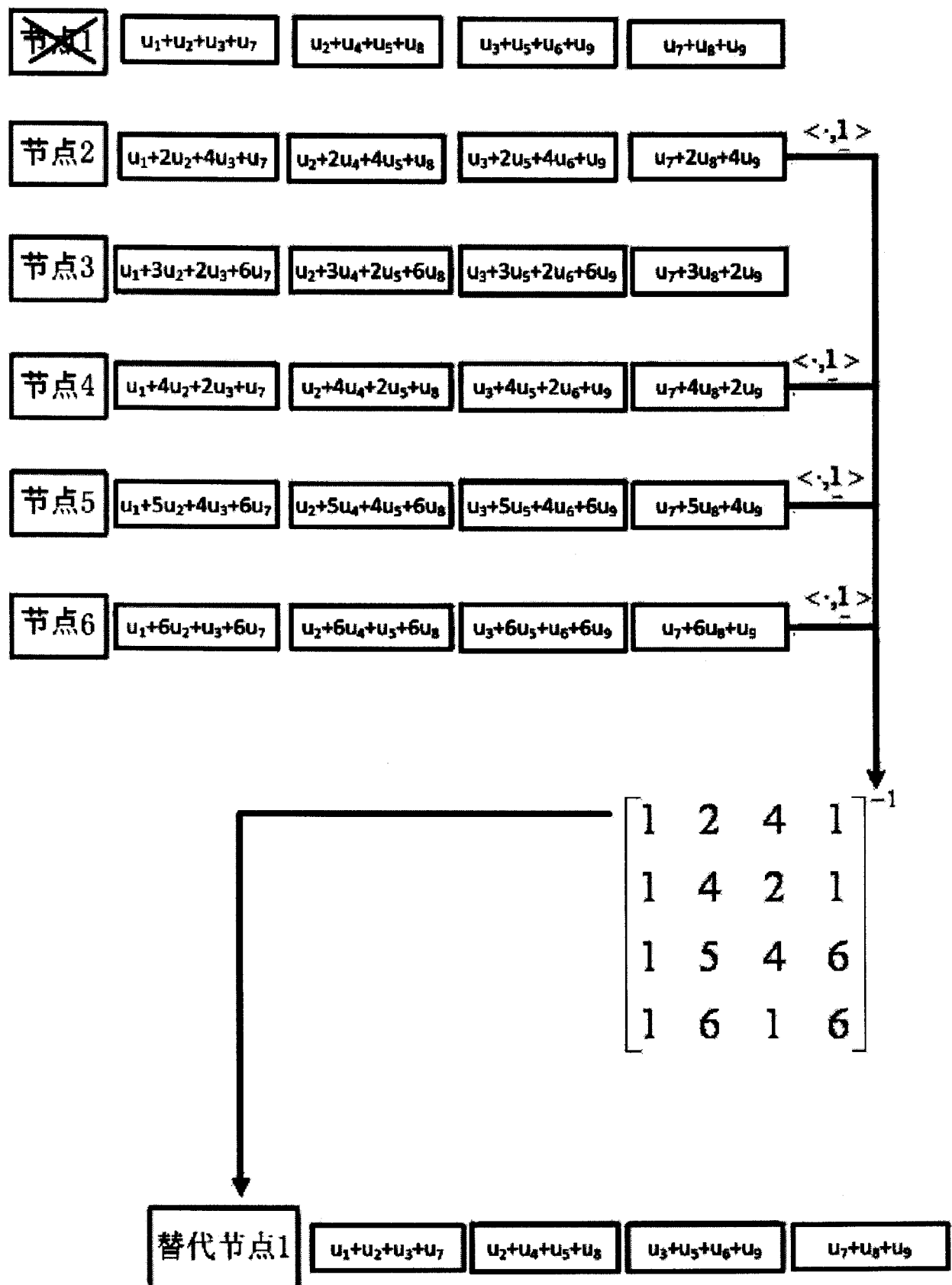


图 4

图 4

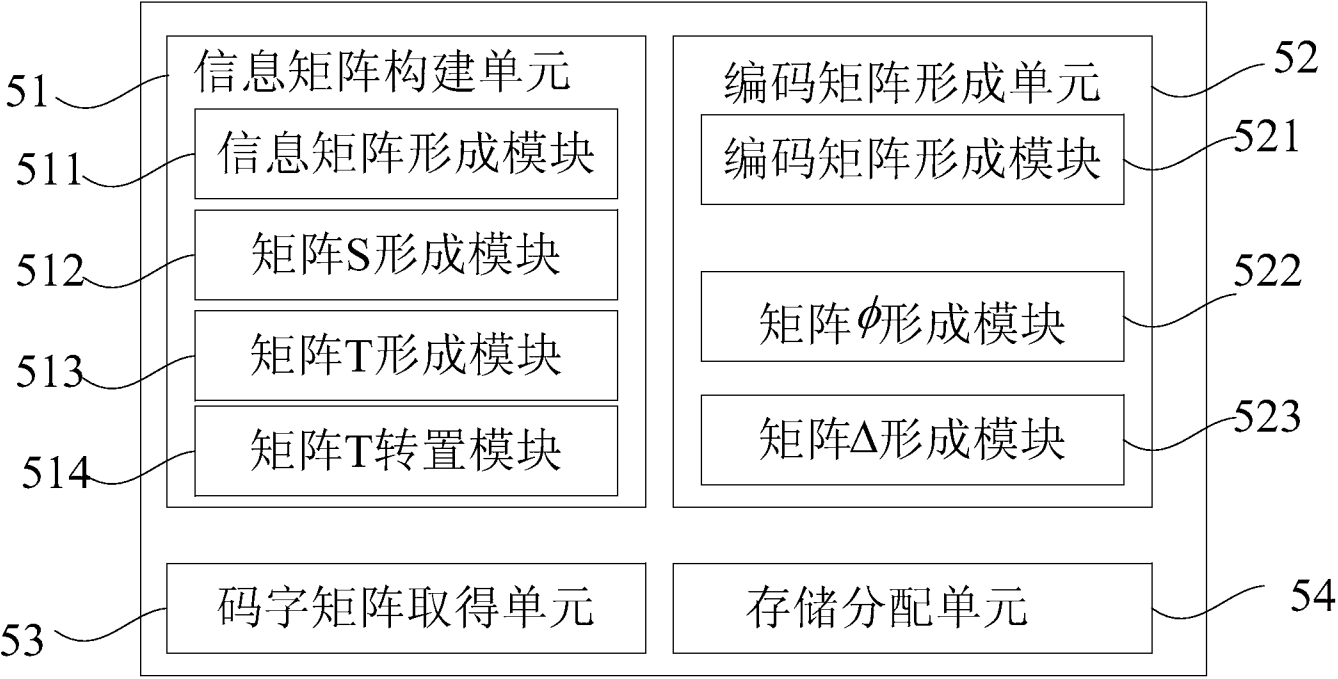


图 5

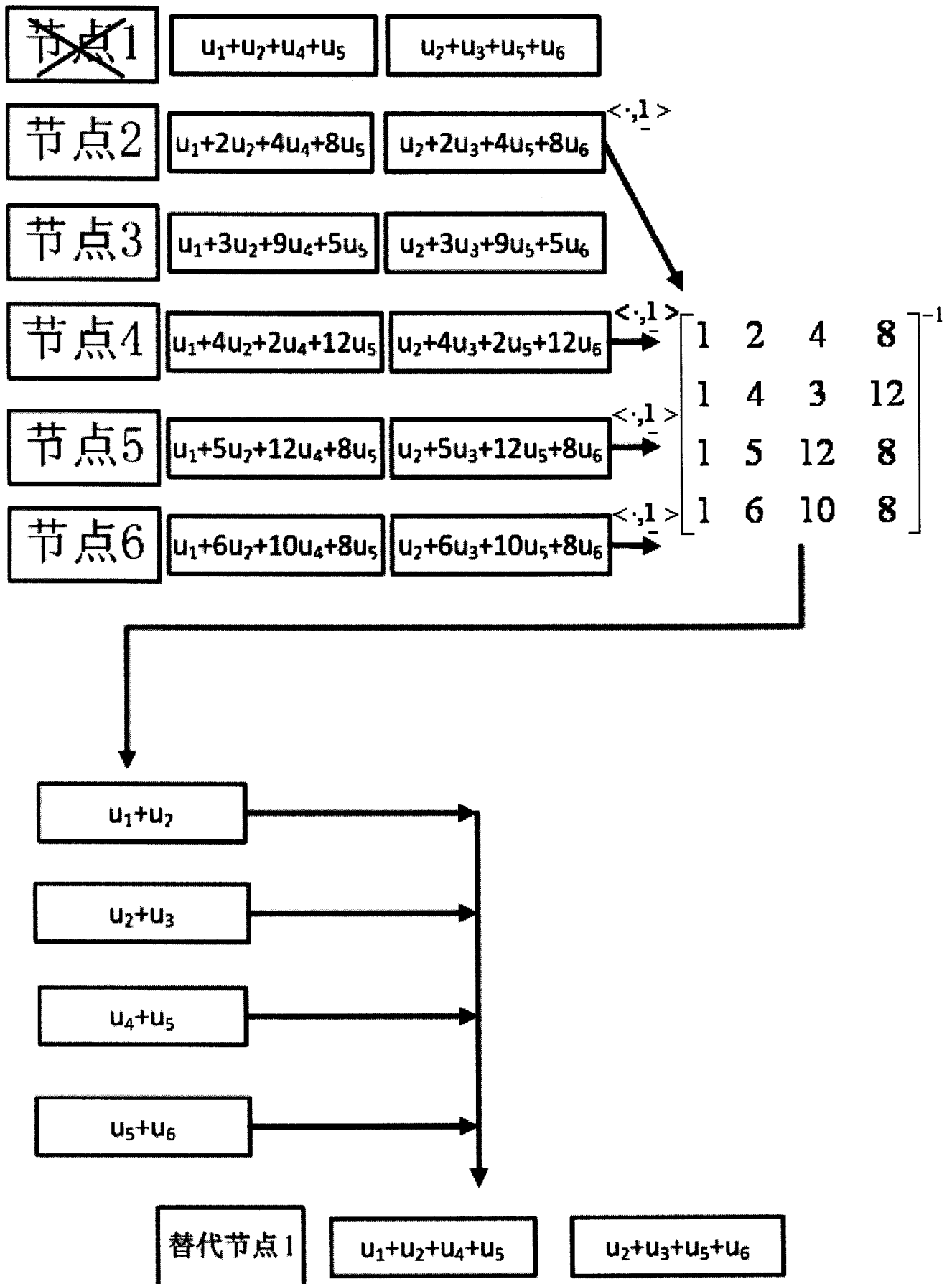


图 6

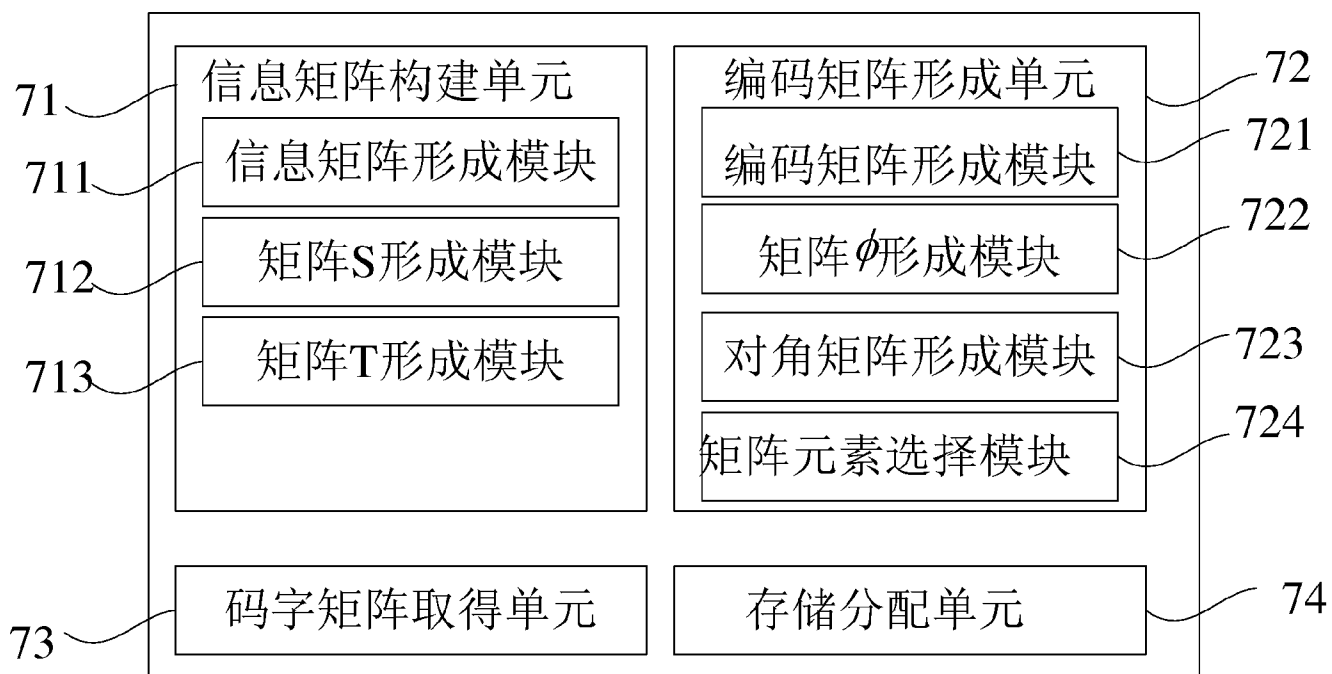


图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/078100

A. CLASSIFICATION OF SUBJECT MATTER

See the extra sheet

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G06F11/-; H04L12/-; H04L29/-; H03M13/-

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, CPRS, CJFD, CNKI, DWPI: MSR, MBR, distributed, storage, regenerat+, code, node?, resum+, renew+, repair+, restor+, matrix, minimum

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WANG, Yu, Research on Data Redundancy and Maintenance Technology in Distributed Storage System, China Doctoral Dissertations Full-text Database, Information Technology Series, 15 June 2012, no.6, ch.3, ISSN 1674-022X	1-10
A	CN 1859115 A (HUAWEI THCHNOLOGIES CO LTD) 08 November 2006 (08.11.2006) see the whole document	1-10
A	CN 101834898 A (CHENGDU INFORMATION TECHNOLOGY CO LTD CA) 15 September 2010 (15.09.2010) see the whole document	1-10
A	EP 2413506 A1 (THOMSON LICENSING) 01 February 2012 (01.02.2012) see the whole document	1-10

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent but published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 07 January 2013 (07.01.2013)	Date of mailing of the international search report 24 January 2013 (24.01.2013)
---	--

Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer SUN, Zehong Telephone No. (86-10) 62412082
---	---

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/CN2012/078100

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 1859115 A	08.11.2006	CN 100579016 C	06.01.2010
CN 101834898 A	15.09.2010	None	
EP 2413506 A1	01.02.2012	US 20120023385 A1	26.01.2012

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2012/078100

CLASSIFICATION OF SUBJECT MATTER

G06F 12/00 (2006.01) i

G06F 11/14 (2006.01) i

国际检索报告

国际申请号

PCT/CN2012/078100

A. 主题的分类

参见附加页

按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC:G06F11/-,H04L12/-,H04L29/-,H03M13/-

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNABS,CPRS,CJFD,CNKI,DWPI:分布式,存储,再生,码,节点,恢复,修复,矩

阵,MSR,MBR,distributed,storage,regenerat+,code,node?,resum+,renew+,repair+,restor+,matrix,minimum

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	王禹,分布式存储系统中的数据冗余与维护技术研究,中国博士学位论文全文数据库信息科技辑,15.6 月 2012,第 6 期,第 3 章,ISSN 1674-022X	1-10
A	CN1859115A (华为技术有限公司) 08.11 月 2006 (08.11.2006) 全文	1-10
A	CN101834898A (中科院成都信息技术有限公司) 15.9 月 2010 (15.09.2010) 全文	1-10
A	EP2413506A1 (THOMSON LICENSING) 01.2 月 2012 (01.02.2012) 全文	1-10



其余文件在 C 栏的续页中列出。



见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

07.1 月 2013 (07.01.2013)

国际检索报告邮寄日期

24.1 月 2013 (24.01.2013)

ISA/CN 的名称和邮寄地址:

中华人民共和国国家知识产权局

中国北京市海淀区蓟门桥西土城路 6 号 100088

传真号: (86-10)62019451

受权官员

孙泽竑

电话号码: (86-10) 62412082

关于同族专利的信息

PCT/CN2012/078100

PCT/ISA/210 表(同族专利附件) (2009 年 7 月)

主题的分类:

G06F12/00 (2006.01) i

G06F11/14 (2006.01) i