

(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

(43) 国际公布日
2021 年 3 月 4 日 (04.03.2021)



(10) 国际公布号
WO 2021/036313 A1

(51) 国际专利分类号:
G06F 15/78 (2006.01)

(21) 国际申请号: PCT/CN2020/086987

(22) 国际申请日: 2020 年 4 月 26 日 (26.04.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910804096.X 2019年8月28日 (28.08.2019) CN

(71) 申请人: 华为技术有限公司 (HUAWEI TECHNOLOGIES CO., LTD.) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

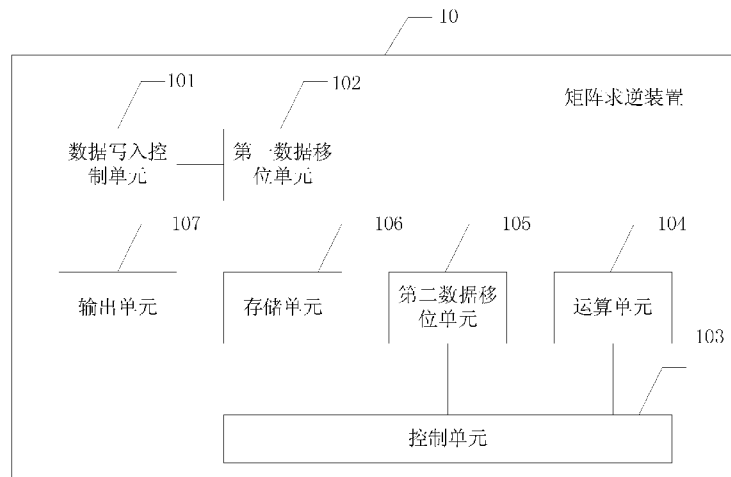
(72) 发明人: 张应松 (ZHANG, Yingsong); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 矫渊培 (JIAO, Yuanpei); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 深圳市深佳知识产权代理事务所 (普通合伙) (SHENPAT INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市罗湖区南湖街道春风路庐山大厦B座18C2、18D、18E、18E2, Guangdong 518001 (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, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS,

(54) Title: CHOLESKY DECOMPOSITION-BASED MATRIX INVERSION APPARATUS

(54) 发明名称: 一种基于乔列斯基分解的矩阵求逆装置



10 MATRIX INVERSION APPARATUS
101 DATA WRITE CONTROL UNIT
102 FIRST DATA SHIFT UNIT
103 CONTROL UNIT
104 CALCULATION UNIT
105 SECOND DATA SHIFT UNIT
106 MEMORY UNIT
107 OUTPUT UNIT

图 1

(57) Abstract: Disclosed in the present application is a Cholesky decomposition-based matrix inversion apparatus, comprising a data write control unit, a first data shift unit, a control unit, a calculation unit, a second data shift unit, a memory unit and an output unit, the calculation unit comprising eight single-precision complex multiplier-accumulators (CMAC), said CMACs having a four-stage pipeline calculation structure, the calculation unit being connected to the control unit, the first data shift unit, the control unit, the second data shift unit and the output unit each being connected to the memory unit, the control unit being connected to the second data shift unit,

JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, 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, KM, ML, MR, NE, SN, TD, TG)。

根据细则4.17的声明:

- 关于申请人有权要求在先申请的优先权 (细则4.17(iii))

本国际公布:

- 包括国际检索报告 (条约第21条(3))。

and the data write control unit being connected to the first data shift unit. Technical solutions of the present application, by means of using an eight-CMAC calculation unit, solve the problems of internal computing resources of current vector processors being relatively few and calculation resource utilization being relatively low, decrease processing delay for Cholesky decomposition-based inversion, and increase network functionality.

(57) 摘要: 本申请公开了一种基于乔列斯基分解的矩阵求逆装置, 包括数据写入控制单元, 第一数据移位单元、控制单元、运算单元、第二数据移位单元、存储单元和输出单元, 该运算单元包括8个单精度复数乘加单元CMAC, 该CMAC具备四级流水运算结构, 运算单元与控制单元相连, 第一数据移位单元、控制单元、第二数据移位单元和输出单元分别与存储单元相连, 控制单元与第二数据移位单元相互连接, 数据写入控制单元与第一数据移位单元相连。本申请技术方案通过使用8个CMAC的运算单元解决当前矢量处理器内部计算资源较少和计算资源的利用率较低的问题, 降低基于乔列斯基的分解求逆处理时延, 提升网络性能。

一种基于乔列斯基分解的矩阵求逆装置

本申请要求于 2019 年 8 月 28 日提交中国专利局、申请号为 201910804096.X、发明名称为“一种基于乔列斯基分解的矩阵求逆装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及数字信号处理领域，具体涉及一种基于乔列斯基分解的矩阵求逆装置。

背景技术

基于乔列斯基分解是一种常用的正定矩阵求逆的方法，该矩阵求逆的原理是：对于一个 n 阶对称正定矩阵 A ，存在一个下三角矩阵 L ，使得 $A=L*L^T$ ，那么该正定矩阵 A 的逆 $A^{-1}=(L*L^T)^{-1}=(L^T)^{-1}*L^{-1}=(L^{-1})^T*L^{-1}$ 。业界的通用做法是采用矢量处理器来实现基于乔列斯基的分解求逆。

基于乔列斯基分解的矩阵求逆运算过程中存在大量的迭代和交织运算，其内部计算量非常大。矢量处理器主要依靠其内部的矢量处理单元来进行基于乔列斯基的分解求逆。然而，当前矢量处理器中只包含 16 个半精度的复数乘加单元 (complex signal processor, CMAC)，只等效于 4 个单精度的 CMAC，即理想情况下每一次只能做 4 次单精度复数运算，使得即使计算资源的利用率能达到百分之百，处理能力也依然较弱。与此同时，矢量处理器在做基于乔列斯基的分解求逆时，先做乔列斯基分解，待分解结果全部求出后再进行求逆操作。由于运算过程中的数据依赖关系，分解的过程是随着迭代的进行所需要的 CMAC 越来越少，而求逆则相反，随着迭代额进行所需要的 CMAC 越来越多。无论是分解或者是求逆，随着迭代的进行，都存在 CMAC 利用率变好或者变差的情景，即 CMAC 平均利用率较低。

综上所述，当前矢量处理器内部计算资源较少和计算资源的利用率较低，使得基于乔列斯基的分解求逆处理时延较长，导致其容易成为链路的瓶颈，影响网络性能。

发明内容

本申请实施例提供一种基于乔列斯基分解的矩阵求逆装置，能够降低基于乔列斯基的分解求逆处理时延，提升网络性能。

本申请提供一种基于乔列斯基分解的矩阵求逆装置，包括数据写入控制单元，第一数据移位单元、控制单元、运算单元、第二数据移位单元、存储单元和输出单元，其中，运算单元包括 8 个单精度复数乘加单元 CMAC，每个 CMAC 具备四级流水运算结构，运算单元与控制单元相连，第一数据移位单元、控制单元、第二数据移位单元和输出单元分别与存储单元相连，控制单元与第二数据移位单元相互连接，数据写入控制单元与第一数据移位单元相连。

该基于乔列斯基分解的矩阵求逆装置中，数据写入控制单元，用于完成矩阵的写入控制，该矩阵为 N 阶正定矩阵， N 为大于 1 且小于或等于 32 的整数。

-2-

该基于乔列斯基分解的矩阵求逆装置中，第一数据移位单元，用于将矩阵的对角线数据移位至每一列的第一位，以得到第一移位数据。

该基于乔列斯基分解的矩阵求逆装置中，控制单元，用于存储单元、第二数据移位单元和运算单元之间的通信与控制。

5 该基于乔列斯基分解的矩阵求逆装置中，运算单元，用于根据控制单元的控制信息，对矩阵进行N次并行迭代运算，以得到N次并行迭代运算中每一次并行迭代运算的运算结果，其中，第x次并行迭代运算的运算结果是根据N阶正定矩阵的第x列分量数据和前(x-1)次迭代运算的运算结果得到的，第x次并行迭代运算的运算结果包含矩阵基于乔列斯基分解的下三角矩阵的第x列的列分量数据和矩阵的逆矩阵的第x行的行分量数据，第x列的列分量数据不包含下三角矩阵的对角线数据，x为大于0小于等于N的整数。

10 该基于乔列斯基分解的矩阵求逆装置中，第二数据移位单元，用于对每一次并行迭代运算的运算结果进行数据移位，以得到每一次迭代运算的运算结果的第二移位数据，第二移位数据用于下一次并行迭代运算的输入。

15 该基于乔列斯基分解的矩阵求逆装置中，存储单元，用于存储第一移位数据和每一次迭代运算的运算结果的第二移位数据。存储单元是本地的缓存，最大可缓存32x32x64bit的矩阵。

该基于乔列斯基分解的矩阵求逆装置中，输出单元，用于根据存储单元中存储的每一次迭代运算的运算结果的第二移位数据，输出逆矩阵。

20 对于N=8的正定矩阵，该矩阵求逆装置可以同时支持4个矩阵的交织运算。对于N=4的正定矩阵，该矩阵求逆装置可以同时支持8个矩阵的交织运算。对于N=16的矩阵，该矩阵求逆装置可以同时支持2个矩阵的交织运算。对于N=32的正定矩阵，无需矩阵的交织运算。当N不为8的整数倍时，可以采用大靠近的方式，交织方式与当N为8的整数倍的正定矩阵的交织方式一致。

25 本申请实施例提供一种基于乔列斯基分解的矩阵求逆装置，能够解决当前矢量处理器内部计算资源较少和计算资源的利用率较低的问题，从而能够降低基于乔列斯基的分解求逆处理时延，提升网络性能。

附图说明

图1为本申请实施例提供的基于乔列斯基分解的矩阵求逆装置的实施例示意图；

图2(a)为本申请实施例提供的数据写入控制单元接收的矩阵格式的示意图；

30 图2(b)为本申请实施例提供的数据写入控制单元输出的矩阵格式的示意图；

图3为本申请实施例提供的第一数据移位单元对矩阵进行移位后的第一移位数据的示意图；

图4(a)为本申请实施例提供的4组交织的8阶矩阵在运算单元中进行第一次并行迭代运算过程中的数据变化示意图；

35 图4(b)为本申请实施例提供的4组交织的8阶矩阵在运算单元中进行第二次并行迭代运算过程中的数据变化示意图；

图4(c)为本申请实施例提供的4组交织的8阶矩阵在运算单元中进行第三次并行迭代

-3-

运算过程中的数据变化示意图;

图 5 为本申请实施例提供的 4 组交织的 8 阶矩阵的每一次并行迭代运算的运算结果的存储格式示意图。

具体实施方式

本申请实施例提供一种基于乔列斯基分解的矩阵求逆装置,能够降低基于乔列斯基的分解求逆处理时延,提升网络性能。

为了使本技术领域的人员更好地理解本申请方案,下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本申请一部分的实施例,而不是全部的实施例。基于本申请中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都应当属于本申请保护的范围。

本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换,以便这里描述的实施例能够以除了在这里图示或描述的内容以外的顺序实施。此外,术语“包括”和“具有”以及他们的任何变形,意图在于覆盖不排他的包含,例如,包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元,而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

在在矩阵解方程运算过程中,需要求出矩阵的逆,比如 $AX=B$,要解出 X 的值,需要先求出 A^{-1} ,之后再根据 $X=A^{-1}*B$ 求出 X 的值。

乔列斯基分解算法是一种很常见的矩阵分解方法,其基本原理是:对于一个 n 阶对称正定矩阵 A ,存在一个下三角矩阵 L ,使得 $A=L*L^T$, L 矩阵对角线上的数都是正实数, L^T 表示下三角矩阵 L 的共轭转置矩阵:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} = L \bullet L^T = \begin{bmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{bmatrix} \bullet \begin{bmatrix} l_{11} & l_{21}^* & \dots & l_{n1}^* \\ 0 & l_{22} & \dots & l_{n2}^* \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & l_{nn}^* \end{bmatrix} \quad (1)$$

乔列斯基分解算法的基本公式为:

$$l_{jj} = \left(a_{jj} - \sum_{p=1}^{j-1} l_{jp} l_{jp}^* \right)^{1/2}, l_{ij} = \left(a_{ij} - \sum_{p=1}^{j-1} l_{ip} l_{jp}^* \right) / l_{jj} \quad (2)$$

其中, $j=1, 2, \dots, n$; l_{jj} 和 l_{ij} 的初值为:

$$l_{11} = \sqrt{a_{11}}, l_{i1} = a_{i1} / l_{11} \quad (3)$$

最后再根据 $A^{-1} = (L * L^T)^{-1} = (L^T)^{-1} * L^{-1} = (L^{-1})^T * L^{-1}$,求得 A 的逆矩阵。

—4—

矢量处理器实现基于乔列斯基分解的矩阵求逆的方法主要包括如下两个步骤：首先进行分解运算，即根据 $A=L*L^T$ 得到下三角矩阵 L 。由于数据的依赖关系，矢量处理器在对矩阵 A 进行乔列斯基分解的过程中，必须一列一列的进行计算，首先计算第一列，待第一列计算完后再计算第二列，每一列的计算依赖前面所有列的计算结果。乔列斯基分解完成得到矩阵 A 的下三角矩阵之后，再进行求逆操作，即根据 $A^{-1}=(L^{-1})^T*L^{-1}$ 计算出 A^{-1} 。由于数据的依赖关系，矢量处理器基于分解结果对矩阵 A 进行求逆的过程中，也需要按照行进行计算，一行计算完成后再计算下一行。

因此，针对当前矢量处理器做分解求逆过程中，待分解结果全部求出后再进行求逆操作，CMAC 利用率较低和矢量处理器 CMAC 计算资源较少的情景，本申请实施例提供一种基于乔列斯基分解的矩阵求逆装置，能够降低基于乔列斯基的分解求逆处理时延，提升网络性能，请参阅图 1。

图 1 为本申请实施例提供的基于乔列斯基分解的矩阵求逆装置 10 的一个实施例示意图。

参阅图 1，本申请实施例提供的基于乔列斯基分解的矩阵求逆装置 10，包括数据写入控制单元 101，第一数据移位单元 102、控制单元 103、运算单元 104、第二数据移位单元 105、存储单元 106、和输出单元 107，其特征在于，所述运算单元 104 包括 8 个单精度复数乘加单元 CMAC，所述 CMAC 具备四级流水运算结构，所述运算单元 104 与所述控制单元 103 相连，所述第一数据移位单元 102、所述控制单元 103、所述第二数据移位单元 105 和所述输出单元 107 分别与所述存储单元 106 相连，所述控制单元 103 与所述第二数据移位单元 105 相互连接，所述数据写入控制单元 101 与所述第一数据移位单元 102 相连。

需要说明的是，本申请实施例提供的基于乔列斯基分解的矩阵求逆装置 10，包含 8 个单精度 CMAC，且每个 CMAC 具备四级流水运算结构，能够直接支持 N 小于或等于 32 的 N 阶正定矩阵的分解求逆的运算，当 N 大于 32 时，可以通过软件的形式将其拆解为 32 以下的维度进行计算，本申请实施例对此不做限定。

由于本申请实施例提供的矩阵求逆装置 10 中包含 8 个单精度 CMAC，且每个 CMAC 具备四级流水运算结构，因此，对于 $N=8$ 的正定矩阵，该矩阵求逆装置 10 可以同时支持 4 个矩阵的交织运算。对于 $N=4$ 的正定矩阵，该矩阵求逆装置 10 可以同时支持 8 个矩阵的交织运算。对于 $N=16$ 的矩阵，该矩阵求逆装置可以同时支持 2 个矩阵的交织运算。对于 $N=32$ 的正定矩阵，无需矩阵的交织运算。上述这几种情况可以实现 CMAC 的百分百利用。需要说明的是，当 N 不为 8 的整数倍时，可以采用大靠近的方式，交织方式与当 N 为 8 的整数倍的正定矩阵的交织方式一致。

具体地，本申请实施例中，矩阵求逆装置 10 中包含的各个功能单元的作用如下：

数据写入控制单元 101，用于完成矩阵的写入控制，需要说明的是，本申请实施例中的矩阵为 N 阶正定矩阵， N 为大于 1 且小于或等于 32 的整数。

第一数据移位单元 102，用于将矩阵的对角线数据移位至每一列的第一位得到第一移位数据。

控制单元 103，用于整个计算任务的调度及控制。

运算单元104, 用于根据控制单元的控制信号, 对矩阵进行N次并行迭代运算, 以得到N次并行迭代运算中每一次并行迭代运算的运算结果, 其中, 第x次并行迭代运算的运算结果是根据N阶正定矩阵的第x列分量数据和前(x-1)次迭代运算的运算结果得到的, 第x次并行迭代运算的运算结果包含矩阵基于乔列斯基分解的下三角矩阵的第x列的列分量数据和矩阵的逆矩阵的第x行的行分量数据, x为大于0小于等于N的整数。

第二数据移位单元105, 用于对每一次并行迭代运算的运算结果进行数据移位, 以得到每一次迭代运算的运算结果的第二移位数据, 用于下一次并行迭代运算的输入。

存储单元106, 用于存储第一移位数据和每一次迭代运算的运算结果的第二移位数据。

需要说明的是, 本申请实施例中, 存储单元106优先响应运算单元104, 在运算单元104没有需求的时候才会响应外部输入; 同时存储单元106还可以保证接收的矩阵个数, 当内部缓存满时反压前级。需要说明的是, 本申请实施例中, 存储单元是本地的缓存, 最大可缓存 $32 \times 32 \times 64 \text{bit}$ 的矩阵。可选地, 存储单元106的带宽为 $64 \times 8 = 512 \text{bit}$, 可以通过 $128 \text{bits} \times 128 \text{depth} \times 4 \text{bank}$ 的方式实现。

输出单元107, 用于根据存储单元106中存储的所述每一次迭代运算的运算结果的第二移位数据, 输出矩阵的逆矩阵。

可选地, 本申请实施例以 $N=8$, 4组矩阵交织的方式对本申请实施例中矩阵求逆装置10中的各个单元模块的功能进行具体的介绍。

本申请实施例中, 当 $N=8$, 采用4组交织的方式进行运算时, 该4个矩阵可以为不同的8阶正定矩阵。数据写入控制单元101首先完成外部输入的矩阵的写入控制, 数据写入控制单元101接收的矩阵均按行传输。如图2(a)所示, 矩阵1-矩阵4为所述数据写入控制单元101按行接收的4个矩阵。具体的, 数据写入控制单元101对外部输入的矩阵的写入控制可以是指对按行传输的矩阵做共轭以实现矩阵的行列转置, 数据写入控制单元101向第一数据移位单元102输出的矩阵格式如图2(b)所示。

本申请实施例中, 根据乔列斯基分解的公式, 需要首先根据矩阵的对角线数据求出矩阵的逆矩阵的对角线数据 d_{ii} , 然后才能继续每一列接下来的分解求逆操作, 因此每一列最开始的运算一定包括对角线数据。为了解决后续运算过程中寻找对角线数据导致的寻址复杂度问题, 本申请实施例通过第一数据移位单元102将每个矩阵的每一列的对角线数据移位至第一位, 从而得到第一移位数据, 存储在存储单元106中, 以作为后续计算的输入。

本申请实施例中, 第一数据移位单元102在接收到数据写入控制单元101传输的矩阵后, 将矩阵的对角线数据均移位至每一列的第一位, 从而简化后续计算的寻址复杂度。移位后的数据为第一移位数据, 该第一移位数据被存储在存储单元106中。例如, 图3示出了矩阵1-矩阵4在经过第一数据移位单元102移位之后所得到的每个矩阵对应的第一移位数据。

本申请实施例中, 控制单元103完成运算单元104、第二数据移位单元105和存储单元106之间的通信, 执行存储单元106中数据的地址计算, 运算单元104中每一次迭代运算的迭代控制、第二数据移位单元105的数据移位控制和多个矩阵交织计算的控制。整个计算任务可以具备较少的时序逻辑, 以及较多的组合逻辑。

—6—

例如，本申请实施例中，当 $N=8$ ，采用4组交织的方式进行运算时，每个矩阵都对应8次并行迭代运算。图4(a)-图4(b)示出了4组交织的8阶矩阵在运算单元104中的8次并行迭代运算过程中数据的变化。

如图4(a)所示，第一次并行迭代运算过程中，控制单元103控制矩阵1-矩阵4中每个矩阵的第一移位数据中第一列分量数据在运算单元104中的输入，通过8个CMAC的运算得到每个矩阵对应的第一次并行迭代运算结果，第一次并行迭代运算结果中包含每个矩阵对应的下三角矩阵的第一列向量数据 $L10-L70$ ，以及逆矩阵的第一行向量数据 $d00$ 。

在得到第一次并行迭代运算结果之后，控制单元103控制第二数据移位单元105对第一次并行迭代运算结果进行数据移位，以得到每个矩阵对应的第二移位数据，作为下一次迭代的输入，如图4(b)中所示的第一次并行迭代结果的第二移位数据。

图4(b)示出了第二次并行迭代运算过程中，控制单元103控制矩阵1-矩阵4中每个矩阵的第一移位数据中第二列分量数据以及第一次并行迭代运算的运算结果对应的第二移位数据在运算单元104中的输入，通过8个CMAC的运算得到每个矩阵对应的第二次并行迭代运算结果，第二次并行迭代运算结果中包含每个矩阵对应的下三角矩阵的第二列向量数据 $L21-L71$ ，以及逆矩阵的第一行向量数据 $d10$ 和 $d11$ 。

对应地，在得到第二次并行迭代运算结果之后，控制单元103控制第二数据移位单元105对第二次并行迭代运算结果进行数据移位，以得到每个矩阵对应的第二移位数据，作为下一次迭代的输入，如图4(c)中所示的第二次并行迭代结果的第二移位数据。

图4(c)示出了第三次并行迭代运算过程中，控制单元103控制矩阵1-矩阵4中每个矩阵的第一移位数据中第三列分量数据以及第一次与第二次并行迭代运算的运算结果分别对应的第二移位数据在运算单元104中的输入，通过8个CMAC的运算得到每个矩阵对应的第三次并行迭代运算结果，第三次并行迭代运算结果中包含每个矩阵对应的下三角矩阵的第三列向量数据 $L32-L72$ ，以及逆矩阵的第一行向量数据 $d20$ 、 $d21$ 和 $d22$ 。

对应地，在得到第三次并行迭代运算结果之后，控制单元103控制第二数据移位单元105对第三次并行迭代运算结果进行数据移位，以得到每个矩阵对应的第二移位数据，作为下一次迭代的输入，如图4(c)中所示的第三次并行迭代结果的第二移位数据。

以此类推，到第八次并行迭代运算时，控制单元103控制矩阵1-矩阵4中每个矩阵的第一移位数据中第八列分量数据以及前7次并行迭代运算的运算结果分别对应的第二移位数据在运算单元104中的输入，通过8个CMAC的运算得到每个矩阵对应的第八次并行迭代运算的运算结果，第八次并行迭代运算的运算结果中包含每个矩阵对应的逆矩阵的第八行向量数据 $d70-d77$ 。

在八次并行迭代运算结束后，存储单元106中所存储的每一次并行迭代运算的运算结果的存储格式如图5所示。最终，输出单元107读取存储单元106中的每一次并行迭代运算的运算结果，并将该运算结果输出。

上述以 $N=8$ ，4组矩阵交织的方式为例对本申请实施例中矩阵求逆装置10中的各个单元模块的功能进行具体的介绍。应理解，对于 N 不等于8的正定矩阵，本申请实施例中矩阵求逆装置10也可以采用相同的原理进行矩阵的分解求逆。例如， $N=4$ ，8个矩阵交织的运算、

—7—

N=16, 2个矩阵交织的运算、N=32, 无需矩阵交织的运算, 以及N不为8的整数倍时, 采用大靠近的方式, 交织方式与当N为8的整数倍的正定矩阵的交织方式一致的运算, 均为在本申请所保护的范围之内。

5 本申请实施例提供的矩阵求逆装置, 能够解决当前矢量处理器内部计算资源较少和计算资源的利用率较低的问题, 从而能够降低基于乔列斯基的分解求逆处理时延, 提升网络性能。

可以理解的是, 在本申请的实施例中涉及的各种数字或字母编号仅为描述方便进行的区分, 并不用来限制本申请的实施例的范围。上述各过程的序号的大小并不意味着执行顺序的先后, 各过程的执行顺序应以其功能和内在逻辑确定。

10 以上对本申请实施例所提供的基于乔列斯基分解的矩阵求逆装置进行了详细介绍, 本文中应用了具体个例对本发明的原理及实施方式进行了阐述, 以上实施例的说明只是用于帮助理解本发明的方法及其核心思想; 同时, 对于本领域的一般技术人员, 依据本发明的思想, 在具体实施方式及应用范围上均会有改变之处, 综上所述, 本说明书内容不应理解为对本发明的限制。

15

权 利 要 求

1、一种基于乔列斯基分解的矩阵求逆装置，包括数据写入控制单元，第一数据移位单元、控制单元、运算单元、第二数据移位单元、存储单元和输出单元，其特征在于，所述运算单元包括8个单精度复数乘加单元CMAC，所述CMAC具备四级流水运算结构，所述运算单元与5 所述控制单元相连，所述第一数据移位单元、所述控制单元、所述第二数据移位单元和所述输出单元分别与所述存储单元相连，所述控制单元与所述第二数据移位单元相互连接，所述数据写入控制单元与所述第一数据移位单元相连。

2、根据权利要求1所述的装置，其特征在于，所述数据写入控制单元，用于完成矩阵的写入控制，所述矩阵为N阶正定矩阵，所述N为大于1且小于或等于32的整数。

10 3、根据权利要求2所述的装置，其特征在于，所述第一数据移位单元，用于将所述矩阵的对角线数据移位至每一列的第一位，以得到第一移位数据。

4、根据权利要求2或3所述的装置，其特征在于，所述控制单元，用于所述存储单元、所述第二数据移位单元和所述运算单元之间的通信与控制。

5、根据权利要求4所述的装置，其特征在于，所述运算单元，用于根据所述控制单元15 的控制信息，对所述矩阵进行N次并行迭代运算，以得到所述N次并行迭代运算中每一次并行迭代运算的运算结果，其中，第x次并行迭代运算的运算结果是根据所述N阶正定矩阵的第x列分量数据和前(x-1)次迭代运算的运算结果得到的，所述第x次并行迭代运算的运算结果包含所述矩阵基于乔列斯基分解的下三角矩阵的第x列的列分量数据和所述矩阵的逆矩阵的第x行的行分量数据，所述第x列的列分量数据不包含所述下三角矩阵的对角线数据，20 所述x为大于0小于等于N的整数。

6、根据权利要求5所述的装置，其特征在于，所述第二数据移位单元，用于对所述每一次并行迭代运算的运算结果进行数据移位，以得到所述每一次迭代运算的运算结果的第二移位数据，所述第二移位数据用于下一次并行迭代运算的输入。

7、根据权利要求6所述的装置，其特征在于，所述存储单元，用于存储所述第一移位25 数据和所述每一次迭代运算的运算结果的第二移位数据。

8、根据权利要求7所述的装置，其特征在于，所述输出单元，用于根据所述存储单元中存储的所述每一次迭代运算的运算结果的第二移位数据，输出所述逆矩阵。

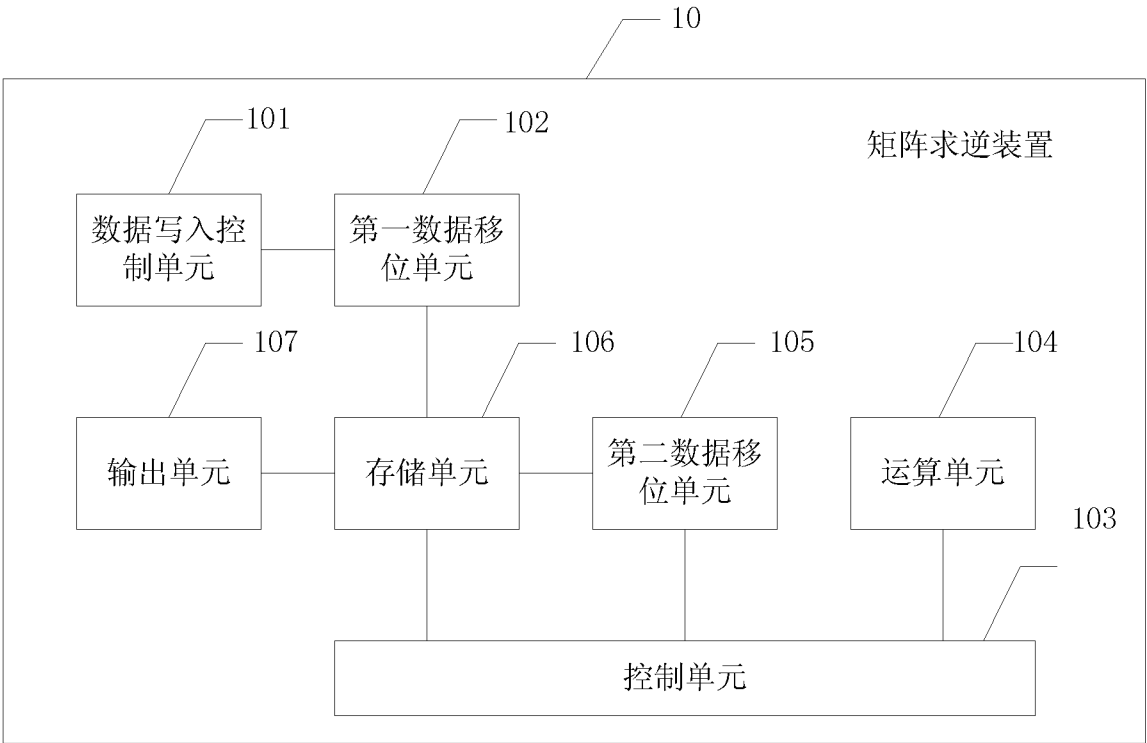


图 1

-2/5-

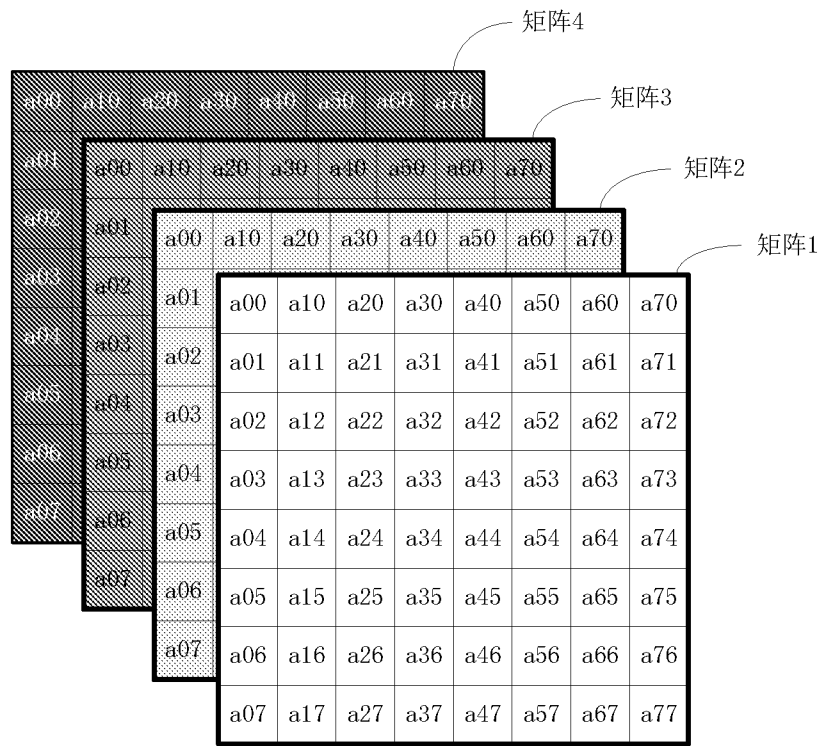


图 2 (a)



图 2 (b)

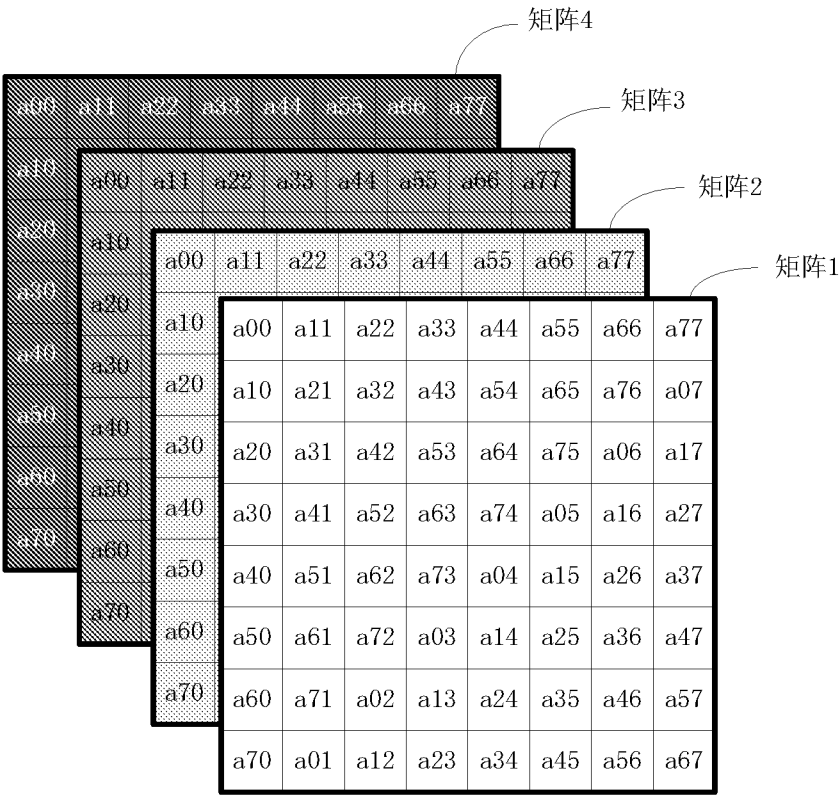


图 3

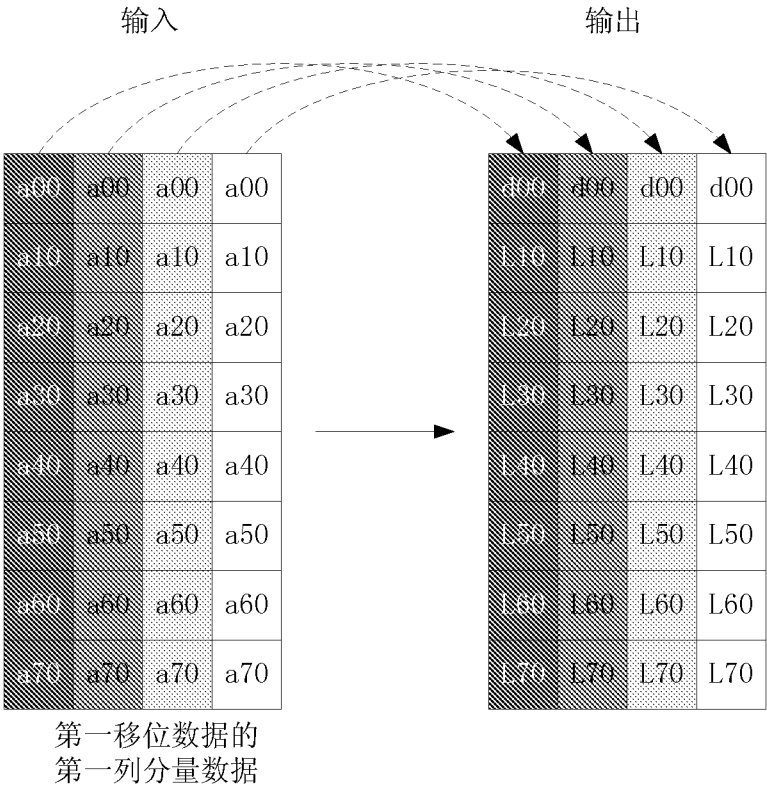


图 4 (a)

-4/5-

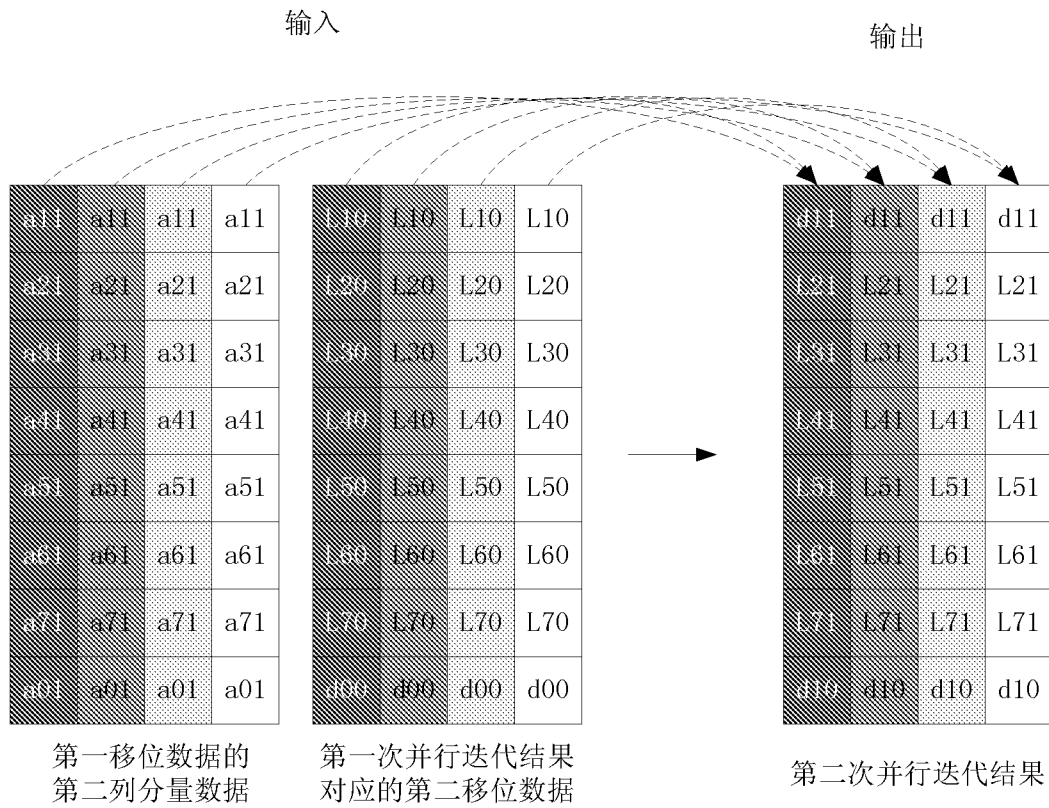


图 4 (b)

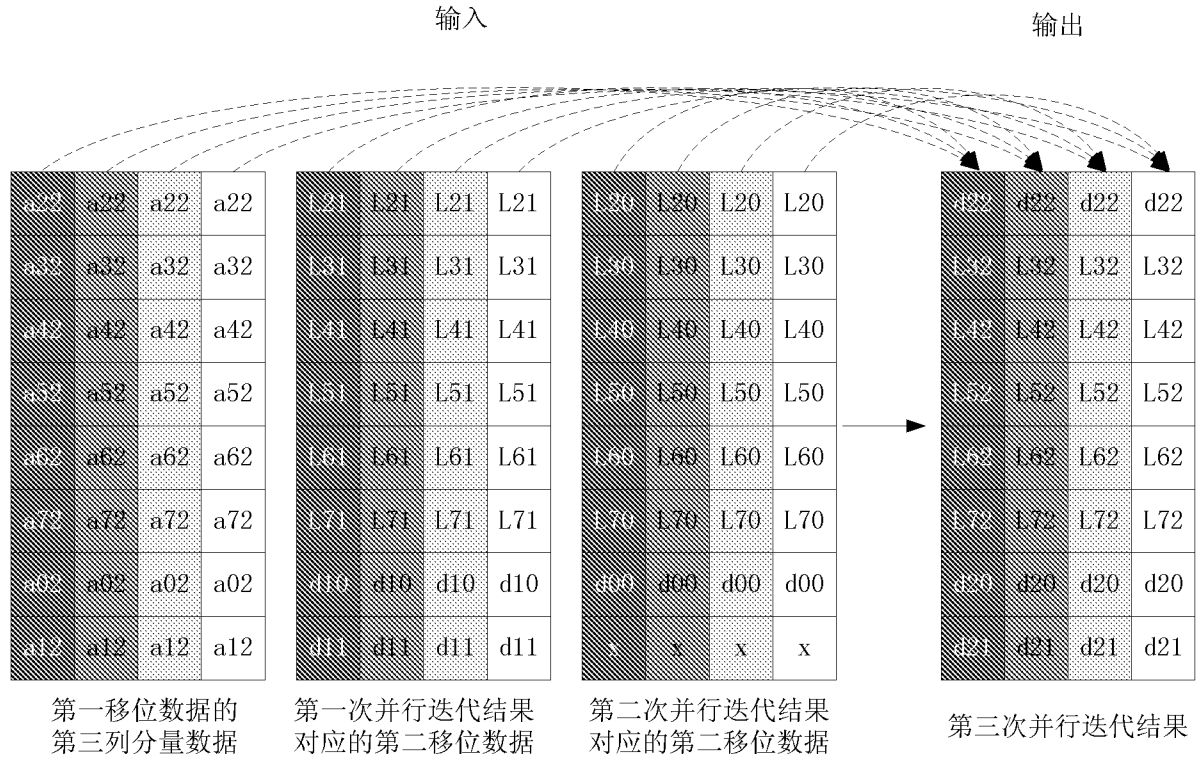


图 4 (c)

L10	L21	L32	L43	L54	L65	L76	d70		
L20	L10	L21	L32	L43	L54	L65	L76	d70	
L30	L20	L10	L21	L32	L43	L54	L65	L76	d70
L40	L30	L20	L10	L21	L32	L43	L54	L65	L76
L50	L40	L30	L20	L10	L21	L32	L43	L54	L65
L60	L50	L40	L30	L20	L10	L21	L32	L43	L54
L70	L60	L50	L40	L30	L20	L10	L21	L32	L43
d00	L70	L60	L50	L40	L30	L20	L10	L21	L32
	d00	L70	L60	L50	L40	L30	L20	L10	L21
		d00	L70	L60	L50	L40	L30	L20	L10
			d00	L70	L60	L50	L40	L30	L20
				d00	L70	L60	L50	L40	L30
					d00	L70	L60	L50	L40
						d00	L70	L60	L50
							d00	L70	L60
								d00	L70
									d00

图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/086987

A. CLASSIFICATION OF SUBJECT MATTER

G06F 15/78(2006.01)i

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)

G06F

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)

DWPI; SIPOABS; CNABS; CNTXT; CNKI: 乔列斯基, 分解, 矩阵, 求逆, cholesky, decomposition, matrix, inverse

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 105701068 A (NANJING UNIVERSITY) 22 June 2016 (2016-06-22) description, paragraphs 2-61	1-4
A	CN 1783060 A (BEIJING T3G TECHNOLOGY CO., LTD.) 07 June 2006 (2006-06-07) entire document	1-8
A	CN 109635241 A (THE 10TH RESEARCH INSTITUTE OF CHINA ELECTRONICS TECHNOLOGY GROUP CORPORATION) 16 April 2019 (2019-04-16) entire document	1-8
A	CN 108733627 A (NANJING UNIVERSITY) 02 November 2018 (2018-11-02) entire document	1-8
A	US 8396914 B1 (Altera Corporation) 12 March 2013 (2013-03-12) entire document	1-8
A	US 8775496 B1 (Xilinx, Inc.) 08 July 2014 (2014-07-08) entire document	1-8



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“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)

“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

“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

“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

“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

“&” document member of the same patent family

Date of the actual completion of the international search

10 June 2020

Date of mailing of the international search report

29 July 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/086987

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	105701068	A	22 June 2016	CN 105701068 B	19 June 2018
CN	1783060	A	07 June 2006	CN 100383781 C	23 April 2008
CN	109635241	A	16 April 2019	None	
CN	108733627	A	02 November 2018	None	
US	8396914	B1	12 March 2013	None	
US	8775496	B1	08 July 2014	None	

国际检索报告

国际申请号

PCT/CN2020/086987

A. 主题的分类

G06F 15/78 (2006.01) i

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)

G06F

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))

DWPI; SIPOABS; CNABS; CNTXT; CNKI: 乔列斯基, 分解, 矩阵, 求逆, cholesky, decomposition, matrix, inverse

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	CN 105701068 A (南京大学) 2016年 6月 22日 (2016 - 06 - 22) 说明书第2-61段	1-4
A	CN 1783060 A (北京天碁科技有限公司) 2006年 6月 7日 (2006 - 06 - 07) 全文	1-8
A	CN 109635241 A (西南电子技术研究所) 2019年 4月 16日 (2019 - 04 - 16) 全文	1-8
A	CN 108733627 A (南京大学) 2018年 11月 2日 (2018 - 11 - 02) 全文	1-8
A	US 8396914 B1 (Altera Corporation) 2013年 3月 12日 (2013 - 03 - 12) 全文	1-8
A	US 8775496 B1 (Xilinx, Inc.) 2014年 7月 8日 (2014 - 07 - 08) 全文	1-8

☐ 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

2020年 6月 10日

国际检索报告邮寄日期

2020年 7月 29日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)

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

王艳坤

电话号码 86-(10)-62412330

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/086987

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	105701068	A	2016年 6月 22日	CN 105701068 B	2018年 6月 19日
CN	1783060	A	2006年 6月 7日	CN 100383781 C	2008年 4月 23日
CN	109635241	A	2019年 4月 16日	无	
CN	108733627	A	2018年 11月 2日	无	
US	8396914	B1	2013年 3月 12日	无	
US	8775496	B1	2014年 7月 8日	无	