

(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
2017 年 11 月 16 日 (16.11.2017)



(10) 国际公布号
WO 2017/193922 A1

- (51) 国际专利分类号:
G06F 17/16 (2006.01)
- (21) 国际申请号: PCT/CN2017/083682
- (22) 国际申请日: 2017 年 5 月 10 日 (10.05.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201610311399.4 2016年5月11日 (11.05.2016) CN
- (71) 申请人: 广州广电运通金融电子股份有限公司 (GRG BANKING EQUIPMENT CO., LTD.) [CN/CN]; 中国广东省广州市高新技术产业开发区科学城科林路9、11号, Guangdong 510663 (CN)。
- (72) 发明人: 辛遥 (XIN, Yao); 中国广东省广州市萝岗区科学城科林路9号, Guangdong 510663 (CN)。梁添才 (LIANG, Tiancai); 中国广东省广州市萝岗区科学城科林路9号, Guangdong 510663 (CN)。龚文川 (GONG, Wenchuan); 中国广东省广州市萝岗区科学城科林路9号,

Guangdong 510663 (CN)。刘道余 (LIU, Daoyu); 中国广东省广州市萝岗区科学城科林路9号, Guangdong 510663 (CN)。

(74) 代理人: 北京集佳知识产权代理有限公司 (UNITALEN ATTORNEYS AT LAW); 中国北京市朝阳区建国门外大街22号赛特广场7层, Beijing 100004 (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, 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, 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,

(54) Title: PARALLEL HARDWARE ARCHITECTURE AND PARALLEL COMPUTING METHOD FOR FLOATING POINT MATRIX INVERSION

(54) 发明名称: 浮点矩阵求逆的并行硬件架构和并行计算方法

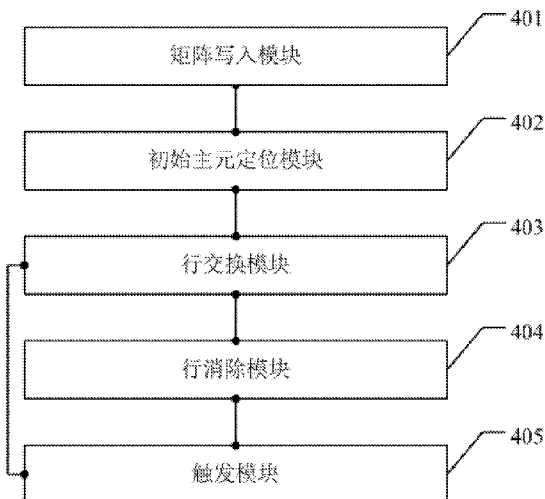


图 4

- 401 MATRIX WRITING MODULE
402 INITIAL PIVOT POSITIONING MODULE
403 ROW SWITCHING MODULE
404 ROW ELIMINATION MODULE
405 TRIGGERING MODULE

(57) Abstract: Disclosed in embodiments of the present invention is a parallel hardware architecture for floating point matrix inversion, for use in resolving the problem of fixed matrix dimensions when an existing matrix operation is implemented by hardware. The parallel hardware architecture for floating point matrix inversion in the embodiments of the present invention comprises: a matrix writing module, used for writing matrix data of an augmented matrix in a first memory and a second memory, the first memory and the second memory dynamically allocating storage spaces to the matrix data; an initial pivot positioning module, used for positioning an initial pivot row of the augmented matrix; a row switching module, used for switching the first row of the augmented matrix with the pivot row; a row elimination module, used for carrying out normalization and row elimination calculation, and positioning a next pivot row; and a triggering module, used for repeatedly and alternately triggering the row switching module and the row elimination module, and outputting the matrix data in the second memory until when the matrix to be inverted stored in the first memory becomes a unit matrix. The embodiments of the present invention also provide a parallel computing method for floating point matrix inversion.

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)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

(57) 摘要：本发明实施例公开了浮点矩阵求逆的并行硬件架构，用于解决现有矩阵运算通过硬件实现时矩阵维数固定的问题。本发明实施例中一种浮点矩阵求逆的并行硬件架构包括：矩阵写入模块，用于将增广矩阵的矩阵数据写入第一存储器和第二存储器中，第一存储器和第二存储器为矩阵数据动态分配存储空间；初始主元定位模块，用于定位增广矩阵的初始主元行；行交换模块，用于将增广矩阵的首行与主元行进行交换；行消除模块，用于进行归一化和行消除计算，并定位下一主元行；触发模块，用于反复交替触发行交换模块和行消除模块，直到第一存储器中存储的待求逆矩阵成为单位矩阵时，输出第二存储器中的矩阵数据。本发明实施例还提供浮点矩阵求逆的并行计算方法。

浮点矩阵求逆的并行硬件架构和并行计算方法

本申请要求于 2016 年 5 月 11 日提交中国专利局、申请号为 201610311399.4、发明名称为“浮点矩阵求逆的并行硬件架构和并行计算方法”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。

5 技术领域

本发明涉及数字电路领域，尤其涉及浮点矩阵求逆的并行硬件架构和并行计算方法。

背景技术

10 矩阵运算是科学计算领域中的一个重要方面，科学计算中的核心问题往往是求解一系列大型线性方程组，这些线性方程组的求解是解决整个问题的基础和核心，其计算量占整个系统计算量的比重很高，而引入矩阵理论来处理和解决科学研究和工程技术上的方程组求解问题已经很普遍，不仅为矩阵理论和方法开辟了广阔的研究前景，也使科学与工程技术的研

15 究发生新的变化，开拓新的研究途径。矩阵运算已成为众多学科领域的数学工具，它不仅在数值分析、最优化方法、数学模型等数学分支上有极其重要的应用，还在信号处理、图像处理、通信技术

20 等科学领域中广泛应用。矩阵求逆是矩阵运算中重要的运算，如何有效的完成矩阵求逆也是数学领域的热点研究。

目前大部分矩阵运算都是利用软件来实现，但是随着矩阵维数的增长，软件处理的速度也会大幅度变慢。矩阵运算的硬件实现能够充分发挥硬件的速度和并行性，目前公开的设计中大部分都是固定维数的矩阵运算，尚无法完成任意维矩阵的硬件实现，尤其是大规模的矩阵求逆的硬件设计。

发明内容

本发明实施例提供了浮点矩阵求逆的并行硬件架构和并行计算方法，能够解决现有矩阵运算通过硬件实现时矩阵维数固定的问题。

25 本发明实施例提供的一种浮点矩阵求逆的并行硬件架构，包括：

矩阵写入模块，用于将增广矩阵的矩阵数据写入第一存储器和第二存储器中，其中，所述第一存储器中写入所述增广矩阵中的待求逆矩阵，所述第二存储器中写入所述增广矩阵中的单位矩阵，所述第一存储器和第二存储器为所述矩阵数据动态分配存储空间；

-2-

初始主元定位模块, 用于定位所述增广矩阵的初始主元行;

行交换模块, 用于将所述增广矩阵的首行与主元行进行交换, 第一次交换时, 所述主元行为所述初始主元行;

行消除模块, 用于对行交换后的所述增广矩阵进行归一化和行消除计算,

5 并定位下一主元行;

触发模块, 用于反复交替触发所述行交换模块和行消除模块, 直到所述第一存储器中存储的所述待求逆矩阵成为单位矩阵时, 输出所述第二存储器中的矩阵数据。

10 可选地, 还包括动态存储分配模块, 用于在所述矩阵写入模块将增广矩阵的矩阵数据写入第一存储器和第二存储器时, 为所述矩阵数据动态分配存储空间;

所述动态存储分配模块包括:

行分配单元, 用于根据所述矩阵数据的大小将存储空间划分为 m 个离散块, 每块所述离散块存储一行矩阵数据, 所述矩阵数据的大小为 $m*n$;

15 列分配单元, 用于将每块所述离散块划分为 $\lceil n/D_p \rceil$ 个行地址空间, 每个所述行地址空间存储 D_p 个浮点数, D_p 等于运算的并行度。

可选地, 所述并行硬件架构还包括:

存储重分配模块, 用于若检测发现运算的并行度发生改变, 则获取改变后的并行度, 并重新为所述矩阵数据动态分配存储空间。

20 可选地, 所述初始主元定位模块包括:

行首数据定位单元, 用于在写入所述增广矩阵的矩阵数据时, 定位所述增广矩阵每一行的首数据;

主元行确定单元, 用于比较每一行的所述首数据, 确定所述首数据最大的所述增广矩阵的所在行为初始主元行。

25 可选地, 所述行消除模块包括:

归一化单元, 用于对行交换后的所述增广矩阵的主元行进行归一化;

缓存单元, 用于将归一化后的所述主元行的元素与主元所在列的元素分别缓存在两个第三存储器中;

行消除计算单元, 用于对所述主元行以外的所有行进行行消除计算;

-3-

主元行重定位单元，用于重新定位行消除计算后的所述增广矩阵的主元行。

本发明实施例提供的一种浮点矩阵求逆的并行计算方法，包括：

5 步骤一、将增广矩阵的矩阵数据写入第一存储器和第二存储器中，其中，所述第一存储器中写入所述增广矩阵中的待求逆矩阵，所述第二存储器中写入所述增广矩阵中的单位矩阵，所述第一存储器和第二存储器为所述矩阵数据动态分配存储空间；

步骤二、定位所述增广矩阵的初始主元行；

10 步骤三、将所述增广矩阵的首行与主元行进行交换，第一次交换时，所述主元行为所述初始主元行；

步骤四、对行交换后的所述增广矩阵进行归一化和行消除计算，并定位下一主元行；

步骤五、反复交替执行所述步骤三和步骤四，直到所述第一存储器中存储的所述待求逆矩阵成为单位矩阵时，输出所述第二存储器中的矩阵数据。

15 可选地，将增广矩阵的矩阵数据写入第一存储器和第二存储器时，为所述矩阵数据动态分配存储空间的方法具体包括：

根据所述矩阵数据的大小将存储空间划分为 m 个离散块，每块所述离散块存储一行矩阵数据，所述矩阵数据的大小为 $m*n$ ；

20 将每块所述离散块划分为 $\lceil n/D_p \rceil$ 个行地址空间，每个所述行地址空间存储 D_p 个浮点数， D_p 等于运算的并行度。

可选地，所述并行计算方法还包括：

若检测发现运算的并行度发生改变，则获取改变后的并行度，并重新为所述矩阵数据动态分配存储空间。

可选地，所述定位所述增广矩阵的初始主元行具体包括：

25 在写入所述增广矩阵的矩阵数据时，定位所述增广矩阵每一行的首数据；
比较每一行的所述首数据，确定所述首数据最大的所述增广矩阵的所在行为初始主元行。

可选地，对行交换后的所述增广矩阵进行归一化和行消除计算，并定位下一主元行具体包括：

—4—

对行交换后的所述增广矩阵的主元行进行归一化;

将归一化后的所述主元行的元素与主元所在列的元素分别缓存在两个第三存储器中;

对所述主元行以外的所有行进行行消除计算;

5 重新定位行消除计算后的所述增广矩阵的主元行。

从以上技术方案可以看出, 本发明实施例具有以下优点:

本发明实施例中, 一种浮点矩阵求逆的并行硬件架构包括: 矩阵写入模块, 用于将增广矩阵的矩阵数据写入第一存储器和第二存储器中, 其中, 所述第一存储器中写入所述增广矩阵中的待求逆矩阵, 所述第二存储器中写入所述增广
10 矩阵中的单位矩阵, 所述第一存储器和第二存储器为所述矩阵数据动态分配存储空间; 初始主元定位模块, 用于定位所述增广矩阵的初始主元行; 行交换模块, 用于将所述增广矩阵的首行与主元行进行交换, 第一次交换时, 所述主元行为所述初始主元行; 行消除模块, 用于对行交换后的所述增广矩阵进行归一化和行消除计算, 并定位下一主元行; 触发模块, 用于反复交替触发所述行交
15 换模块和行消除模块, 直到所述第一存储器中存储的所述待求逆矩阵成为单位矩阵时, 输出所述第二存储器中的矩阵数据。在本发明实施例中, 通过为矩阵数据动态分配存储空间, 所述并行硬件构架适用任意维数的矩阵求逆, 可以满足大规模的矩阵求逆要求, 在不同维数的矩阵求逆之间无需更改硬件设置。同时, 配合并行处理方式, 可以实现任意维度矩阵的求逆处理, 提高求逆运算的
20 速度。

附图说明

图 1 为本发明实施例中一种浮点矩阵求逆的并行计算方法一个实施例流程图;

25 图 2 为本发明实施例中一种浮点矩阵求逆的并行计算方法另一个实施例流程图;

图 3 为本发明实施例中动态存储分配算法的原理示意图;

图 4 为本发明实施例中一种浮点矩阵求逆的并行硬件架构一个实施例结构图;

图 5 为本发明实施例中一种浮点矩阵求逆的并行硬件架构另一个实施例

结构图;

图 6 为本发明的一种浮点矩阵求逆的并行硬件架构一个应用场景下的结构示意图;

图 7 为本发明的一种浮点矩阵求逆的并行硬件架构的行消除模块一个应用场景下的原理示意图。

具体实施方式

本发明实施例提供了浮点矩阵求逆的并行硬件架构和并行计算方法,用于解决现有矩阵运算通过硬件实现时矩阵维数固定的问题。

10 为使得本发明的发明目的、特征、优点能够更加的明显和易懂,下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,下面所描述的实施例仅仅是本发明一部分实施例,而非全部的实施例。基于本发明中的实施例,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它实施例,都属于本发明保护的范围。

15 请参阅图 1,本发明实施例中一种浮点矩阵求逆的并行计算方法一个实施例包括:

101、将增广矩阵的矩阵数据写入第一存储器和第二存储器中;

20 首先,将增广矩阵的矩阵数据写入第一存储器和第二存储器中,其中,该第一存储器中写入该增广矩阵中的待求逆矩阵,该第二存储器中写入该增广矩阵中的单位矩阵,该第一存储器和第二存储器为该矩阵数据动态分配存储空间。

102、定位该增广矩阵的初始主元行;

在将增广矩阵的矩阵数据写入第一存储器和第二存储器中之后,可以定位该增广矩阵的初始主元行。

25 103、将该增广矩阵的首行与主元行进行交换;

在定位该增广矩阵的初始主元行之后,可以将该增广矩阵的首行与主元行进行交换,第一次交换时,该主元行为该初始主元行。

104、对行交换后的该增广矩阵进行归一化和行消除计算,并定位下一主元行;

-6-

在将该增广矩阵的首行与主元行进行交换之后,可以对行交换后的该增广矩阵进行归一化和行消除计算,并定位下一主元行。

105、反复交替执行步骤 103 和步骤 104,直到该第一存储器中存储的该待求逆矩阵成为单位矩阵时,输出该第二存储器中的矩阵数据。

5 反复交替执行步骤 103 和步骤 104,直到该第一存储器中存储的该待求逆矩阵成为单位矩阵时,输出该第二存储器中的矩阵数据。

本实施例中,首先,将增广矩阵的矩阵数据写入第一存储器和第二存储器中,其中,该第一存储器中写入该增广矩阵中的待求逆矩阵,该第二存储器中写入该增广矩阵中的单位矩阵,该第一存储器和第二存储器为该矩阵数据动态分配存储空间;然后,定位该增广矩阵的初始主元行;接着,将该增广矩阵的首行与主元行进行交换,第一次交换时,该主元行为该初始主元行;再之,对行交换后的该增广矩阵进行归一化和行消除计算,并定位下一主元行;最后,反复交替执行该步骤三和步骤四,直到该第一存储器中存储的该待求逆矩阵成为单位矩阵时,输出该第二存储器中的矩阵数据。在本实施例中,通过为矩阵数据动态分配存储空间,该并行计算方法适用任意维数的矩阵求逆,可以满足大规模的矩阵求逆要求,在不同维数的矩阵求逆之间无需更改硬件设置。

10 15

为便于理解,下面对本发明实施例中的一种浮点矩阵求逆的并行计算方法进行详细描述,请参阅图 2,本发明实施例中一种浮点矩阵求逆的并行计算方法另一个实施例包括:

20 201、将增广矩阵的矩阵数据写入第一存储器和第二存储器中;

首先,将增广矩阵的矩阵数据写入第一存储器和第二存储器中,其中,该第一存储器中写入该增广矩阵中的待求逆矩阵,该第二存储器中写入该增广矩阵中的单位矩阵,该第一存储器和第二存储器为该矩阵数据动态分配存储空间。

25 需要说明的是,将增广矩阵的矩阵数据写入第一存储器和第二存储器时,为该矩阵数据动态分配存储空间的方法具体包括:

1、根据该矩阵数据的大小将存储空间划分为 m 个离散块,每块该离散块存储一行矩阵数据,该矩阵数据的大小为 $m*n$;

2、将每块该离散块划分为 $\lceil n/D_p \rceil$ 个行地址空间,每个该行地址空间存储

-7-

D_p 个浮点数, D_p 等于运算的并行度, $\lceil n/D_p \rceil$ 表示向上取整 n/D_p 的结果。

如图 3 所示, 将一个存储器的空间划分为 m 个矩阵行空间 (也即离散块), 然后每个矩阵行空间又被划分为 $\lceil n/D_p \rceil$ 个行地址空间, 其中每个行地址空间存储 D_p 个浮点数 (数据), 从而每个矩阵行空间存储 n 个浮点数。

5 另外, 为了保证动态分配的存储空间与运算的并行度保持匹配, 存储空间分配可以根据并行度实时调整: 若检测发现运算的并行度发生改变, 则获取改变后的并行度, 并重新为该矩阵数据动态分配存储空间。

202、在写入该增广矩阵的矩阵数据时, 定位该增广矩阵每一行的首数据;

10 为了定位初始主元行, 可以在写入该增广矩阵的矩阵数据时, 定位该增广矩阵每一行的首数据。

203、比较每一行的该首数据, 确定该首数据最大的该增广矩阵的所在行为初始主元行;

15 在定位该增广矩阵每一行的首数据之后, 可以比较每一行的该首数据, 确定该首数据最大的该增广矩阵的所在行为初始主元行。例如可以是, 在数据写入时定位每一行的首数据, 每两行之间进行大小比较并存储较大的数据, 当写入结束后其存储的即为最大的首数据, 从而可以确定最大首数据对应的行为初始主元行。

204、将该增广矩阵的首行与主元行进行交换;

20 在确定初始主元行或定位下一主元行 (见步骤 205) 之后, 可以将该增广矩阵的首行与主元行进行交换, 可以理解的是, 第一次交换时, 该主元行为该初始主元行。

205、对行交换后的该增广矩阵进行归一化和行消除计算, 并定位下一主元行;

25 在将该增广矩阵的首行与主元行进行交换之后, 可以对行交换后的该增广矩阵进行归一化和行消除计算, 并定位下一主元行。

需要说明的是, 步骤 205 具体可以包括:

A、对行交换后的该增广矩阵的主元行进行归一化: 举例说明, 增广矩阵为 $[A | I]$, 方阵 $A = (a_{ij}^{(1)})_{n \times n}$ 为待求逆矩阵, a_{ij} 代表第 i 行第 j 列元素, I 为 $n \times n$ 单位矩阵, $a_{pj}^{(k+1)} = a_{pj}^{(k)} (1/a_{pp}^{(k)})$, p 为主元行数, j 为列元素标号, k 为迭代次

—8—

数, 首先计算出主元的倒数 $(1/a_{pp}^{(k)})$, 后将主元行元素与倒数一一相乘, 这样尽可能避免了除法操作;

B、将归一化后的该主元行的元素 $a_{pj}^{(k+1)}$ 与主元所在列的元素 $a_{ip}^{(k)}$ 分别缓存在两个第三存储器中;

5 C、对该主元行以外的所有行进行行消除计算: $a_{ij}^{(k+1)} = a_{ij}^{(k)} - a_{ip}^{(k)} a_{pj}^{(k+1)}$;

Dp、重新定位行消除计算后的该增广矩阵的主元行。

归一化和消除操作作为纯流水线设计, 因为步骤 C 中, $a_{ij}^{(k)}$, $a_{pj}^{(k+1)}$, $a_{ip}^{(k)}$ 分别来自三个存储器, 所以可以同时运算 (即并行计算)。实际上每个位置的数有 Dp 个, 采用水平扫描操作, 即同一行 Dp 个相邻元素参与运算, 直到所有元素完成运算。

206、反复交替执行该步骤三和步骤四, 直到该第一存储器中存储的该待求逆矩阵成为单位矩阵时, 输出该第二存储器中的矩阵数据。

反复交替执行该步骤三和步骤四, 直到该第一存储器中存储的该待求逆矩阵成为单位矩阵时, 输出该第二存储器中的矩阵数据。由高斯-约当部分主元消除法 (Gauss-Jordan Elimination with Partial Pivoting) 的原理可知, 开始时, 第一存储器存储待求逆矩阵, 第二存储器存储单位矩阵, 在进行求逆运算后, 当原待求逆矩阵成为单位矩阵后, 第二存储器中存储的单位矩阵即为待求逆矩阵的逆矩阵, 输出该第二存储器中的矩阵数据即输出所需的逆矩阵。需要说明的是, 所述存储待求逆矩阵和单位矩阵的存储器不是固定的, 当待求逆矩阵成为单位矩阵后, 下一轮存储器交替转换, 由另一个存储器来存放待求逆矩阵。

上面主要描述了一种浮点矩阵求逆的并行计算方法, 下面将对一种浮点矩阵求逆的并行硬件架构进行详细描述, 请参阅图 4, 本发明实施例中一种浮点矩阵求逆的并行硬件架构一个实施例包括:

25 矩阵写入模块 401, 用于将增广矩阵的矩阵数据写入第一存储器和第二存储器中, 其中, 该第一存储器中写入该增广矩阵中的待求逆矩阵, 该第二存储器中写入该增广矩阵中的单位矩阵, 该第一存储器和第二存储器为该矩阵数据动态分配存储空间;

初始主元定位模块 402, 用于定位该增广矩阵的初始主元行;

行交换模块 403, 用于将该增广矩阵的首行与主元行进行交换, 第一次交

换时, 该主元行为该初始主元行;

行消除模块 404, 用于对行交换后的该增广矩阵进行归一化和行消除计算, 并定位下一主元行;

触发模块 405, 用于反复交替触发该行交换模块 403 和行消除模块 404,
5 直到该第一存储器中存储的该待求逆矩阵成为单位矩阵时, 输出该第二存储器中的矩阵数据。

本实施例中, 矩阵写入模块 401 将增广矩阵的矩阵数据写入第一存储器和第二存储器中, 其中, 该第一存储器中写入该增广矩阵中的待求逆矩阵, 该第二存储器中写入该增广矩阵中的单位矩阵, 该第一存储器和第二存储器为该矩阵数据动态分配存储空间; 初始主元定位模块 402 定位该增广矩阵的初始主元行; 行交换模块 403 将该增广矩阵的首行与主元行进行交换, 第一次交换时, 该主元行为该初始主元行; 行消除模块 404 对行交换后的该增广矩阵进行归一化和行消除计算, 并定位下一主元行; 触发模块反复交替触发该行交换模块 403 和行消除模块 404, 直到该第一存储器中存储的该待求逆矩阵成为单位矩阵时, 输出该第二存储器中的矩阵数据。在本实施例中, 通过为矩阵数据动态分配存储空间, 该并行硬件构架适用任意维数的矩阵求逆, 可以满足大规模的矩阵求逆要求, 在不同维数的矩阵求逆之间无需更改硬件设置。

为便于理解, 下面对本发明实施例中的一种浮点矩阵求逆的并行硬件架构进行详细描述, 请参阅图 5, 本发明实施例中一种浮点矩阵求逆的并行硬件架构另一个实施例包括:

矩阵写入模块 501, 用于将增广矩阵的矩阵数据写入第一存储器和第二存储器中, 其中, 该第一存储器中写入该增广矩阵中的待求逆矩阵, 该第二存储器中写入该增广矩阵中的单位矩阵, 该第一存储器和第二存储器为该矩阵数据动态分配存储空间;

25 初始主元定位模块 502, 用于定位该增广矩阵的初始主元行;

行交换模块 503, 用于将该增广矩阵的首行与主元行进行交换, 第一次交换时, 该主元行为该初始主元行;

行消除模块 504, 用于对行交换后的该增广矩阵进行归一化和行消除计算, 并定位下一主元行;

触发模块 505, 用于反复交替触发该行交换模块 503 和行消除模块 504, 直到该第一存储器中存储的该待求逆矩阵成为单位矩阵时, 输出该第二存储器中的矩阵数据。

5 本实施例中, 该并行硬件架构还包括动态存储分配模块 506, 用于在该矩阵写入模块将增广矩阵的矩阵数据写入第一存储器和第二存储器时, 为该矩阵数据动态分配存储空间;

该动态存储分配模块 506 包括:

行分配单元 5061, 用于根据该矩阵数据的大小将存储空间划分为 m 个离散块, 每块该离散块存储一行矩阵数据, 该矩阵数据的大小为 $m*n$;

10 列分配单元 5062, 用于将每块该离散块划分为 $\lceil n/D_p \rceil$ 个行地址空间, 每个该行地址空间存储 D_p 个浮点数, D_p 等于运算的并行度。

本实施例中, 该并行硬件架构还包括:

存储重分配模块 507, 用于若检测发现运算的并行度发生改变, 则获取改变后的并行度, 并重新为该矩阵数据动态分配存储空间。

15 本实施例中, 该初始主元定位模块 502 包括:

行首数据定位单元 5021, 用于在写入该增广矩阵的矩阵数据时, 定位该增广矩阵每一行的首数据;

主元行确定单元 5022, 用于比较每一行的该首数据, 确定该首数据最大的该增广矩阵的所在行为初始主元行。

20 本实施例中, 该行消除模块 504 包括:

归一化单元 5041, 用于对行交换后的该增广矩阵的主元行进行归一化;

缓存单元 5042, 用于将归一化后的该主元行的元素与主元所在列的元素分别缓存在两个第三存储器中;

行消除计算单元 5043, 用于对该主元行以外的所有行进行行消除计算;

25 主元行重定位单元 5044, 用于重新定位行消除计算后的该增广矩阵的主元行。

需要说明的是, 对于本实施例中的行消除模块 504, 在具体一应用场景下, 如图 7 所示, 归一化和消除操可以作为纯流水线设计, 因为在行消除计算过程中, $a_{ij}^{(k)}$, $a_{pj}^{(k+1)}$, $a_{ip}^{(k)}$ 分别来自三个存储器 (如图 7 中的归一化主元存储器、主

元列存储器和矩阵存储器), 所以可以同时通过可扩展浮点运算单元进行运算 (即并行计算), 运算后通过比较器比较重新定位主元行。实际上每个位置的数有 D_p 个, 采用水平扫描操作, 即同一行 D_p 个相邻元素参与运算, 直到所有元素完成运算。

5 为便于理解, 一个应用场景中的浮点矩阵求逆的并行硬件架构如图 6 所示。该并行硬件架构可以封装有 5 个接口 IO 信号:

- 1、矩阵大小: 指示矩阵的规模。
- 2、写地址: 初始时写入矩阵数据的存储地址。
- 3、写数据: 初始时写入存储器的一个地址空间的数据。
- 10 4、读地址: 逆运算完成后读出矩阵数据的存储地址。
- 5、读数据: 逆运算完成后读出存储器的一个地址空间的数据。

此并行硬件架构支持矩阵规模大小可变, 可以进行任意大小的矩阵求逆运算, 唯一的限制为第一存储器和第二存储器的容量和资源大小。该并行硬件架构中的运算模块可以采用多浮点运算单元并行配置, 运算单元数量可以扩展, 支持不同并行度设计, 加速矩阵运算。如并行度 $D_p=4$ 时, 同时有四个数据同时
15 同时进行运算。存储器 (第一存储器、第二存储器或第三存储器等) 的存储空间分配可以根据并行度实时调整, 存储器中内嵌如图 3 所示原理的动态存储分配算法。

所述硬件架构在 FPGA 进行仿真验证, 分别设置并行度 D_p 为 4, 8, 16,
20 32, 最大支持矩阵大小为 256×256 , 最大时钟工作频率分别为 208MHz, 222MHz, 218MHz, 169MHz。完成矩阵求逆时钟数估算为:
 $(n^2 + 13n) \lceil n / D_p \rceil + n^2$, n 为方阵的大小, 计算复杂度由 $O(n^3)$ 减少为 $O(n^3/D_p)$ 。

所属领域的技术人员可以清楚地了解到, 为描述的方便和简洁, 上述描述的系统, 装置和单元的具体工作过程, 可以参考前述方法实施例中的对应过程,
25 在此不再赘述。

在本申请所提供的几个实施例中, 应该理解到, 所揭露的系统, 装置和方法, 可以通过其它的方式实现。例如, 以上所描述的装置实施例仅仅是示意性的, 例如, 所述单元的划分, 仅仅为一种逻辑功能划分, 实际实现时可以有另外的划分方式, 例如多个单元或组件可以结合或者可以集成到另一个系统, 或

-12-

一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

5 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

10 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

15 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、磁碟或者光盘等各种可以
20 存储程序代码的介质。

25 以上所述，以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

权 利 要 求

1、一种浮点矩阵求逆的并行硬件架构，其特征在于，包括：

矩阵写入模块，用于将增广矩阵的矩阵数据写入第一存储器和第二存储器中，其中，所述第一存储器中写入所述增广矩阵中的待求逆矩阵，所述第二存储器中写入所述增广矩阵中的单位矩阵，所述第一存储器和第二存储器为所述
5 矩阵数据动态分配存储空间；

初始主元定位模块，用于定位所述增广矩阵的初始主元行；

行交换模块，用于将所述增广矩阵的首行与主元行进行交换，第一次交换时，所述主元行为所述初始主元行；

10 行消除模块，用于对行交换后的所述增广矩阵进行归一化和行消除计算，并定位下一主元行；

触发模块，用于反复交替触发所述行交换模块和行消除模块，直到所述第一存储器中存储的所述待求逆矩阵成为单位矩阵时，输出所述第二存储器中的矩阵数据。

15 2、根据权利要求 1 所述的并行硬件架构，其特征在于，还包括动态存储分配模块，用于在所述矩阵写入模块将增广矩阵的矩阵数据写入第一存储器和第二存储器时，为所述矩阵数据动态分配存储空间；

所述动态存储分配模块包括：

20 行分配单元，用于根据所述矩阵数据的大小将存储空间划分为 m 个离散块，每块所述离散块存储一行矩阵数据，所述矩阵数据的大小为 $m*n$ ；

列分配单元，用于将每块所述离散块划分为 $\lceil n/D_p \rceil$ 个行地址空间，每个所述行地址空间存储 D_p 个浮点数， D_p 等于运算的并行度。

3、根据权利要求 2 所述的并行硬件架构，其特征在于，所述并行硬件架构还包括：

25 存储重分配模块，用于若检测发现运算的并行度发生改变，则获取改变后的并行度，并重新为所述矩阵数据动态分配存储空间。

4、根据权利要求 1 所述的并行硬件架构，其特征在于，所述初始主元定位模块包括：

行首数据定位单元，用于在写入所述增广矩阵的矩阵数据时，定位所述增

广矩阵每一行的首数据;

主元行确定单元,用于比较每一行的所述首数据,确定所述首数据最大的所述增广矩阵的所在行为初始主元行。

5 5、根据权利要求 1 所述的并行硬件架构,其特征在于,所述行消除模块包括:

归一化单元,用于对行交换后的所述增广矩阵的主元行进行归一化;

缓存单元,用于将归一化后的所述主元行的元素与主元所在列的元素分别缓存在两个第三存储器中;

行消除计算单元,用于对所述主元行以外的所有行进行行消除计算;

10 主元行重定位单元,用于重新定位行消除计算后的所述增广矩阵的主元行。

6、一种浮点矩阵求逆的并行计算方法,其特征在于,包括:

步骤一、将增广矩阵的矩阵数据写入第一存储器和第二存储器中,其中,所述第一存储器中写入所述增广矩阵中的待求逆矩阵,所述第二存储器中写入
15 所述增广矩阵中的单位矩阵,所述第一存储器和第二存储器为所述矩阵数据动态分配存储空间;

步骤二、定位所述增广矩阵的初始主元行;

步骤三、将所述增广矩阵的首行与主元行进行交换,第一次交换时,所述主元行为所述初始主元行;

20 步骤四、对行交换后的所述增广矩阵进行归一化和行消除计算,并定位下一主元行;

步骤五、反复交替执行所述步骤三和步骤四,直到所述第一存储器中存储的所述待求逆矩阵成为单位矩阵时,输出所述第二存储器中的矩阵数据。

25 7、根据权利要求 6 所述的并行计算方法,其特征在于,将增广矩阵的矩阵数据写入第一存储器和第二存储器时,为所述矩阵数据动态分配存储空间的方法具体包括:

根据所述矩阵数据的大小将存储空间划分为 m 个离散块,每块所述离散块存储一行矩阵数据,所述矩阵数据的大小为 $m*n$;

将每块所述离散块划分为 $\lceil n/D_p \rceil$ 个行地址空间,每个所述行地址空间存储

-15-

D_p 个浮点数, D_p 等于运算的并行度。

8、根据权利要求 7 所述的并行计算方法, 其特征在于, 所述并行计算方法还包括:

5 若检测发现运算的并行度发生改变, 则获取改变后的并行度, 并重新为所述矩阵数据动态分配存储空间。

9、根据权利要求 6 所述的并行计算方法, 其特征在于, 所述定位所述增广矩阵的初始主元行具体包括:

在写入所述增广矩阵的矩阵数据时, 定位所述增广矩阵每一行的首数据;

10 比较每一行的所述首数据, 确定所述首数据最大的所述增广矩阵的所在行为初始主元行。

10、根据权利要求 6 所述的并行计算方法, 其特征在于, 对行交换后的所述增广矩阵进行归一化和行消除计算, 并定位下一主元行具体包括:

对行交换后的所述增广矩阵的主元行进行归一化;

15 将归一化后的所述主元行的元素与主元所在列的元素分别缓存在两个第三存储器中;

对所述主元行以外的所有行进行行消除计算;

重新定位行消除计算后的所述增广矩阵的主元行。

— 1/5 —

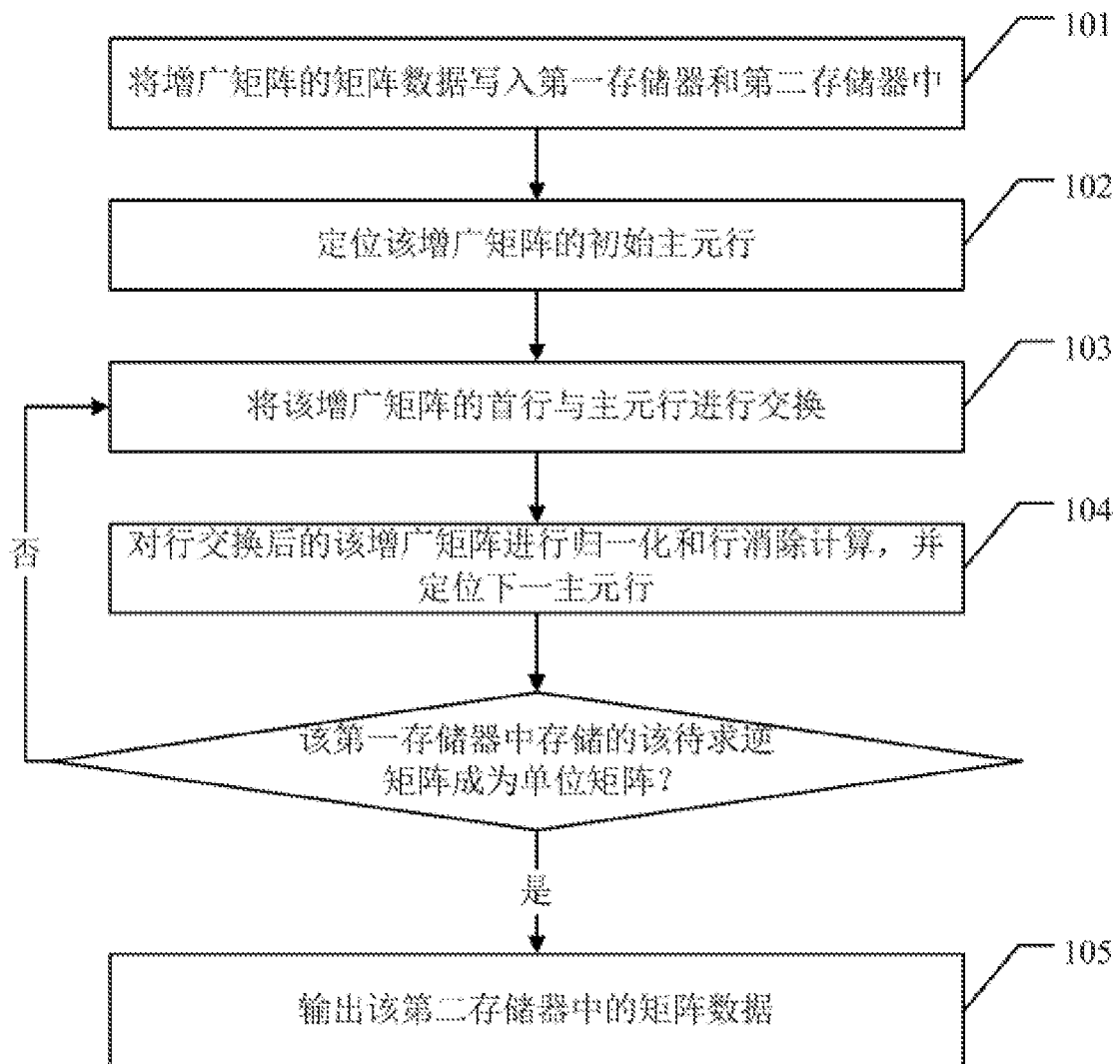


图 1

- 2/5 -

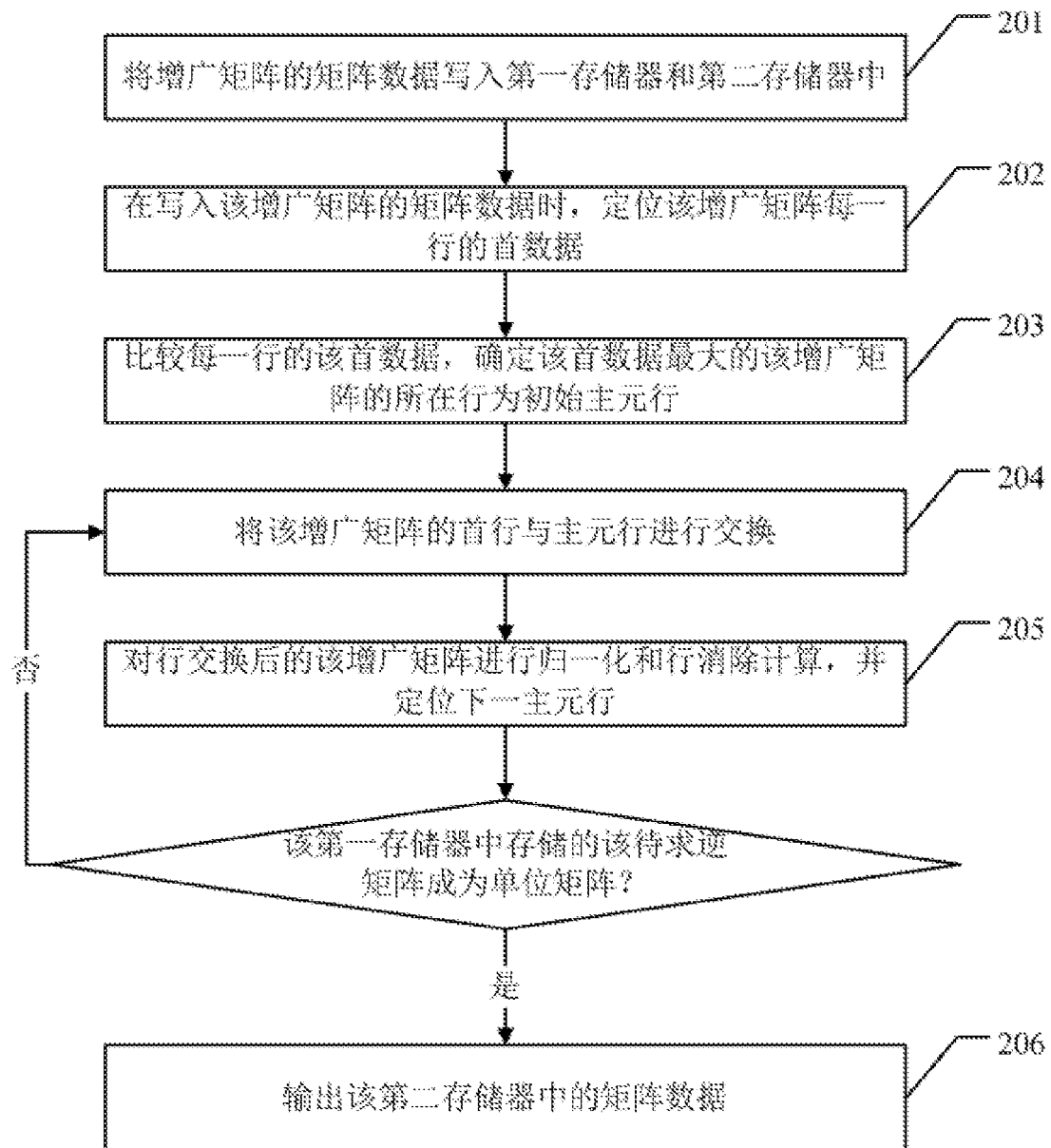


图 2

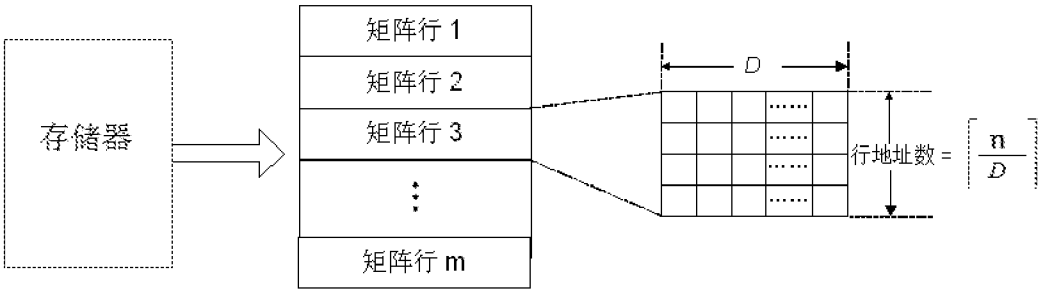


图 3

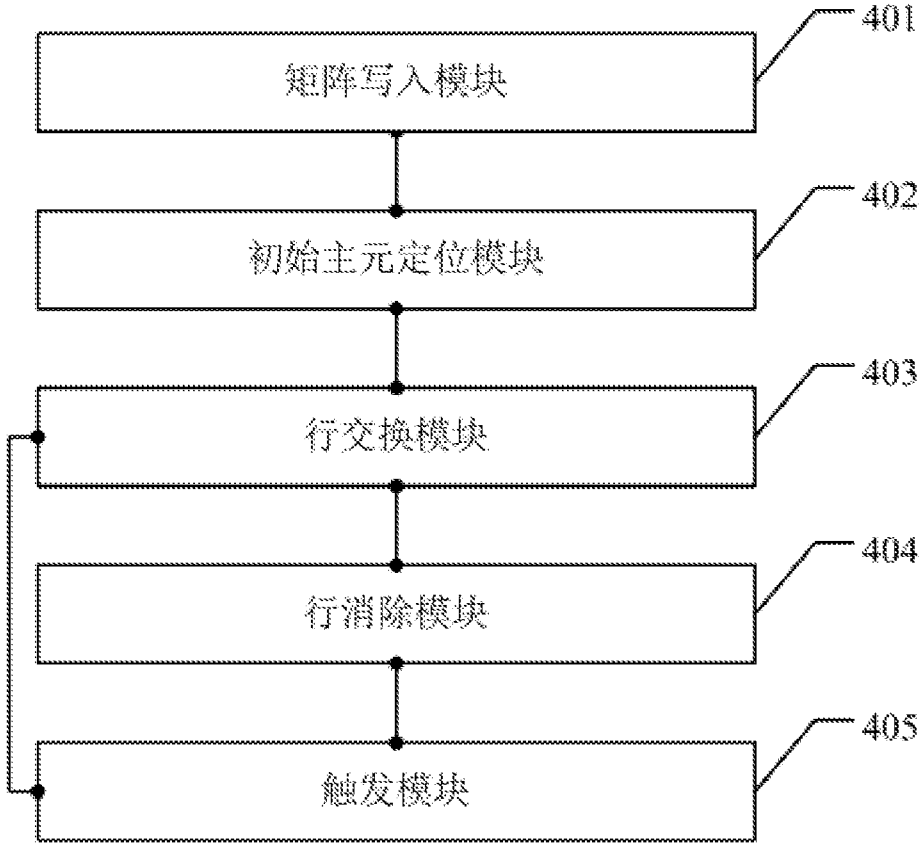


图 4

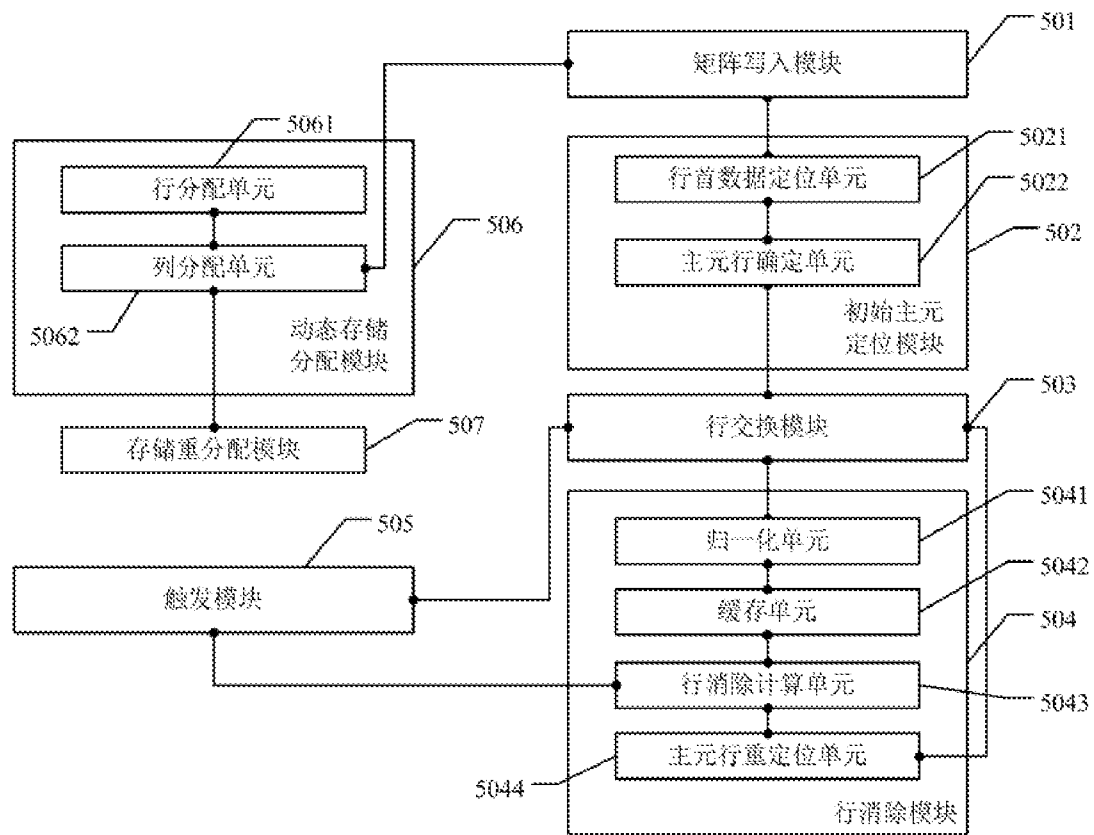


图 5

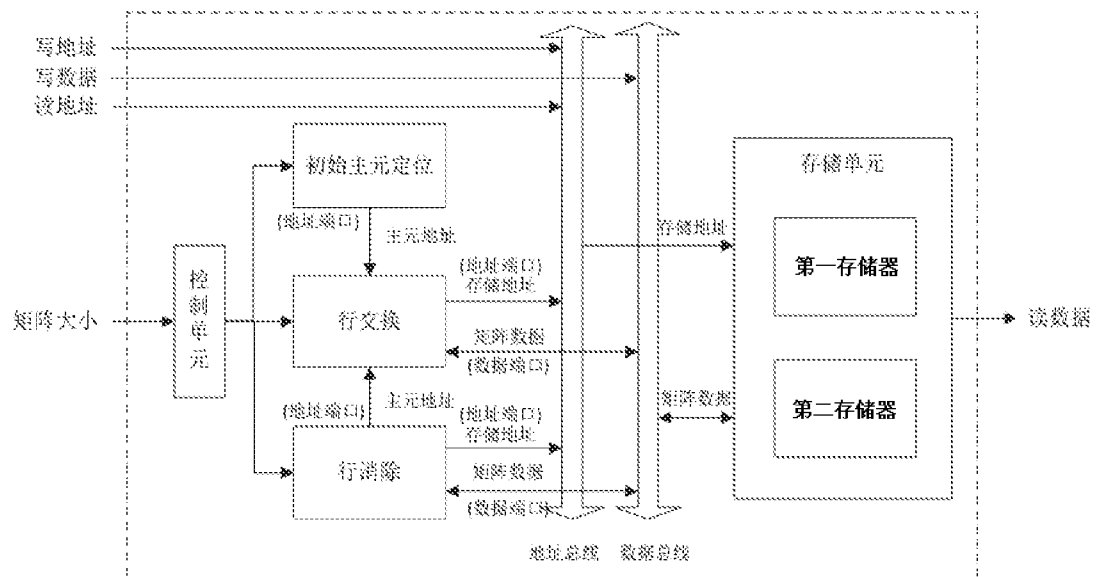


图 6

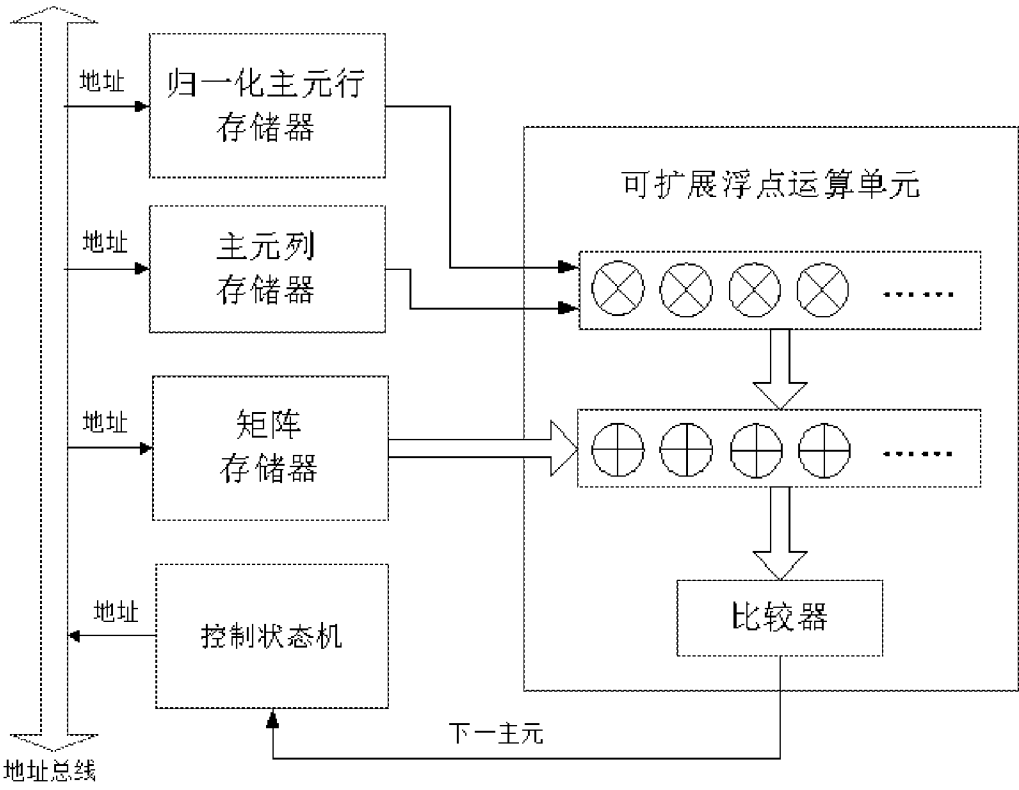


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2017/083682

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/16 (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)

CNABS, VEN, CNKI, IEEE: floating-point matrix, floating point matrix, inversion, parallel, augmented matrix, identity matrix, storage, Gauss-Jordan Elimination with Partial Pivoting

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 106021188 A (GRG BANKING EQUIPMENT CO., LTD.), 12 October 2016 (12.10.2016), claims 1-10, and description, paragraphs [0004]-[0043]	1-10
Y	TIAN, Xishan; "Solving linear equations in parallel by using Gauss elimination with partial pivoting", COMPUTER KNOWLEDGE AND TECHNOLOGY, vol. 7, no. 16, 05 June 2011 (05.06.2011), pages 3960-3963 and 3966	1-10
Y	CN 104572588 A (CHINA ELECTRONICS TECHNOLOGY GROUP CORPORATION NO. 38 RESEARCH INSTITUTE), 29 April 2015 (29.04.2015), description, paragraphs [0026]-[0037]	1-10
A	CN 104239280 A (NANCHANG UNIVERSITY), 24 December 2014 (24.12.2014), the whole document	1-10
A	CN 103678257 A (SHANGHAI JIAO TONG UNIVERSITY), 26 March 2014 (26.03.2014), the whole document	1-10
A	CN 103631761 A (RAINTREE SCIENTIFIC INSTRUMENTS (SHANGHAI) CORPORATION), 12 March 2014 (12.03.2014), 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 July 2017 (07.07.2017)	Date of mailing of the international search report 09 August 2017 (09.08.2017)
Name and mailing address of the ISA/CN: 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 LIU, Yingying Telephone No.: (86-10) 62089925

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2017/083682

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 9001924 B1 (XILINX, INC.), 07 April 2015 (07.04.2015), the whole document	1-10

International application No.
PCT/CN2017/083682

Form PCT/ISA/210 (patent family annex) (July 2009)

A. 主题的分类 G06F 17/16 (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)) CNABS, VEN, CNKI, IEEE: 浮点矩阵, 求逆, 并行, 增广矩阵, 单位矩阵, 存储器, 高斯约当部分主元消除法, floating-point matrix, floating point matrix, inversion, parallel, augmented matrix, identity matrix, storage, Gauss-Jordan Elimination with Partial Pivoting		
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
PX	CN 106021188 A (广州广电运通金融电子股份有限公司) 2016年 10月 12日 (2016 - 10 - 12) 权利要求1-10, 说明书第[0004]-[0043]段	1-10
Y	田希山. "用部分选主元的高斯消去法并行求解线性方程组" 电脑知识与技术, 第7卷, 第16期, 2011年 6月 5日 (2011 - 06 - 05), 第3960-3963, 3966页	1-10
Y	CN 104572588 A (中国电子科技集团公司第三十八研究所) 2015年 4月 29日 (2015 - 04 - 29) 说明书第[0026]-[0037]段	1-10
A	CN 104239280 A (南昌大学) 2014年 12月 24日 (2014 - 12 - 24) 全文	1-10
A	CN 103678257 A (上海交通大学) 2014年 3月 26日 (2014 - 03 - 26) 全文	1-10
A	CN 103631761 A (睿励科学仪器上海有限公司) 2014年 3月 12日 (2014 - 03 - 12) 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		国际检索报告邮寄日期
2017年 7月 7日		2017年 8月 9日
ISA/CN的名称和邮寄地址		受权官员
中华人民共和国国家知识产权局 (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		刘莹莹 电话号码 (86-10)62089925

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
A	US 9001924 B1 (XILINX, INC.) 2015年 4月 7日 (2015 - 04 - 07) 全文	1-10

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/083682

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	106021188	A	2016年 10月 12日	无	
CN	104572588	A	2015年 4月 29日	无	
CN	104239280	A	2014年 12月 24日	无	
CN	103678257	A	2014年 3月 26日	CN 103678257 B	2016年 9月 28日
CN	103631761	A	2014年 3月 12日	无	
US	9001924	B1	2015年 4月 7日	无	