(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
2023 年 6 月 29 日 (29.06.2023)

WIPO | PCT

(10) 国际公布号
WO 2023/116431 A1

(51) 国际专利分类号:
G06F 17/16 (2006.01)

(21) 国际申请号: PCT/CN2022/137086

(22) 国际申请日: 2022 年 12 月 7 日 (07.12.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202111567014.8 2021年12月20日 (20.12.2021) CN
202210460849.1 2022年4月28日 (28.04.2022) 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: **傅光宁 (FU, Guangning)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **林腾毅 (LIN, Tengyi)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, 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, CV, 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, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF,

(54) **Title:** MATRIX CALCULATION METHOD, CHIP, AND RELATED DEVICE

(54) **发明名称**: 一种矩阵计算方法、芯片以及相关设备

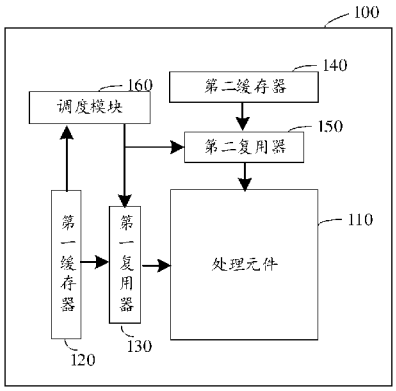

图 1

110 Processing element
120 First cache
130 First multiplexer
140 Second cache
150 Second multiplexer
160 Scheduling module

(57) **Abstract:** The invention provides a matrix calculation method, a chip and a related device; the chip comprises a first cache, a second cache, a scheduling module and a processing element; the first cache is used for caching a first vector, and the second cache is used for caching a second vector; the scheduling module generates a strobe signal on the basis of a bitmap of the first vector, and the strobe signal can enable the processing element to acquire a group of non-zero elements in the first vector from the first cache, and enable the processing element to acquire a group of elements in the second vector from the second cache; then, performing an operation on the first vector and the second vector on the basis of the group of non-zero elements in the first vector and the group of elements in the second vector; wherein, the bitmap of the first vector is used for indicating a non-zero element in the first vector. With the chip provided by the invention, when operations are performed on two vectors, elements with the value of 0 in one vector do not participate in calculations, so that the amount of calculating can be reduced, improving calculation efficiency.

WO 2023/116431 A1

[见续页]

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: 本申请提供一种矩阵计算方法、芯片及相关设备；该芯片包括第一缓存器、第二缓存器、调度模块和处理元件；第一缓存器用于缓存第一向量，第二缓存器用于缓存第二向量；调度模块根据第一向量的位图生成选通信号，选通信号能够使处理元件从第一缓存器中获取第一向量中的一组非零元素，使处理元件从第二缓存器中获取第二向量中的一组元素；然后根据第一向量中的一组非零元素和第二向量中的一组元素进行第一向量和第二向量的运算，其中，第一向量的位图用于指示第一向量中的非零元素。通过本申请提供的芯片进行两个向量的运算时，能够使一个向量中值为0的元素不参与计算，从而能够在降低计算量，提高计算效率。

一种矩阵计算方法、芯片以及相关设备

本申请要求于2021年12月20日提交的申请号为202111567014.8、发明名称为“矩阵计算系统、方法和乘法器”的中国专利申请的优先权，以及于2022年4月28日提交中国专利局、申请号为202210460849.1、申请名称为“一种矩阵计算方法、芯片以及相关设备”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

本申请涉及计算机领域，尤其设计一种矩阵计算方法、芯片以及相关设备。

背景技术

当前人工智能（Artificial Intelligence，AI）、高性能计算（High performance computing，HPC）等领域中，会涉及大量的矩阵乘法运算，例如AI模型训练、AI推理等场景。为了加快计算速度，提高计算效率，通常是通过剪枝技术对矩阵进行稀疏化，然后使用稀疏化后的矩阵进行计算，从而降低计算过程中的计算量。例如在AI训练完成后，将训练后的权重矩阵进行结构化剪枝完成权重矩阵的稀疏化，在AI推理时采用结构化剪枝后的权重矩阵进行推理。但是上述方法只支持对一个确定的矩阵进行剪枝后，使用剪枝后的矩阵进行计算，不支持计算过程中矩阵稀疏程度动态变化的场景。

发明内容

本申请提供一种矩阵计算方法、芯片以及相关设备，能够支持稀疏程度动态变化的矩阵的计算，降低计算中的计算量，提高计算效率的同时，不降低计算的精度。

第一方面，本申请提供一种芯片，该芯片包括：第一缓存器、第二缓存器、第一调度模块和第一处理元件，其中，第一缓存器用于缓存第一向量；第二缓存器用于缓存第二向量；第一调度模块用于根据第一向量的位图生成第一选通信号，该第一选通信号能够使第一处理元件从第一缓存器中获取第一向量中的第一组非零元素，使第一处理元件从第二缓存器中获取第二向量中的第二组元素；第一处理元件用于根据第一向量中的第一组非零元素和第二向量中的第二组元素实现第一向量和所述第二向量的运算，其中，第一向量的位图指示第一向量中的非零元素。

第一向量的位图中每个比特位对应第一向量中的一个元素，每个比特位的值指示第一向量中对应元素是否为0。例如一个比特位的值为0，则表示第一向量中对应的元素的值为0，如果一个比特位的值为1，则表示第一向量中对应的元素的值不为0。第一调度模块能够根据第一向量的位图确定第一向量中的哪些元素是非零元素，从而使第一处理元件能够只获取第一向量中的非零元素，然后从第二向量中获取对应位置的元素，执行第一向量和第二向量之间的运算，从而能够降低计算量，提高计算效率。并且在进行向量的点积的过程中，一个向量中元素值为0的元素与另一个向量中非零元素相乘值为0，不会影响两个向量进行点积的结果，因此只获取第一向量中的非零元素进行第一向量和第二向量的运算，不会降低计算精度。

在一种可能的实现方式中，上述芯片还包括第一复用器和第二复用器；第一复用器用于根据第一选通信号从第一缓存器中获取上述第一向量中的第一组非零元素，并输入第一处理元件；第二复用器用于根据第一选通信号从第二缓存器中获取上述第二向量中的第二组元素，并输入第一处理元件。

在一种可能的实现方式中，上述第一复用器和第二复用器各自包括 K 个多路复用器；第一缓存器和第二缓存器各自包括 W 行 K 列数据单元，每个数据单元用于缓存一个向量或者矩阵中的元素；第一复用器中的每个多路复用器与第一缓存器中的多个数据单元连接，且每个数据单元至少与一个多路复用器连接；第二复用器中的第 i 个多路复用器与第二缓存器中数据单元的连接关系，与第一复用器中的第 i 个多路复用器与第一缓存器中数据单元的连接关系相同。

第一向量与第二向量的运算中，第一向量的第 r 个元素与第二向量的第 r 个元素需要做乘法运算，在将第一向量的各个元素存入第一缓存器以及将第二向量的各个元素存入第二缓存器中时，第一向量中的第 r 个元素存入的第一缓存器的相对位置和第二向量中的第 r 个元素存入第二缓存器的相对位置相同。例如，将第一向量的第 r 个元素存入第一缓存器的第 1 行的第 r 个数据单元，将第二向量的第 r 个元素存入第二缓存器的第 1 行的第 r 个数据单元中。同时，第一复用器中的第 i 个多路复用器连接了第一缓存器中的多个数据单元。这多个数据单元在第一缓存器中的相对位置，与第二复用器中的第 i 个多路复用器连接的第二缓存器中的多个数据单元在第二缓存器中的相对位置相同。例如，第一复用器中第 i 个多路复用器连接了第一缓存器中第一行的第 2 个数据单元和第二行的第 5 个数据单元，则第二复用器中第 i 个多路复用器连接第二缓存器中第一行的第 2 个数据单元和第二行的第 5 个数据单元。这样能够使第一复用器的第 i 个多路复用器与第二复用器的第 i 个多路复用器根据同一个选通信号，读取的是第一缓存器和第二缓存器中相对位置相同的数据单元中的数据，使第一复用器和第二复用器根据相同的选通信号读取的正好是第一向量和第二向量中需要进行乘法运算的元素。

在一种可能的实现方式中，上述第一调度模块具体用于：根据第一向量的位图确定第一复用器的第 j 个多路复用器连接的数据单元中，第 k 个数据单元存储的元素为非零元素，则第一调度模块生成第一复用器的第 j 个多路复用器的选通信号，将第一复用器的第 j 个多路复用器的选通信号发送给第一复用器的第 j 个多路复用器和第二复用器的第 j 个多路复用器，上述第一选通信号包括第一复用器的第 j 个多路复用器的选通信号。

第一复用器包括 K 个多路复用器，则第一调度模块在一个周期内生成这 K 个多路复用器各自对应的选通信号，即第一选通信号包括这 K 个多路复用器各自对应的选通信号。以使第一复用器和第二复用器中各个多路复用器根据接收到的选通信号分别从连接的数据单元中读取一个数据；

上述一个多路复用器所连接的多个数据单元具有不同的优先级，第一调度模块在生成一个多路复用器的选通信号之前，先根据第一向量的位图确定该多路复用器连接的数据单元中，优先级最高的数据单元中存储的第一向量的元素是否是 0。如果优先级最高的数据单元中存储的元素不是 0，则生成该优先级第一的数据单元对应的选通信号；如果优先级最高的数据单元中存储的元素是 0，则再根据第一向量的位图确定优先级第二的数据单元中存储的元素是否为 0，如果优先级第二的数据单元中存储的元素不为 0，则生成优先级第二的数据单元对应的选通信号，如果优先级第二的数据单元中存储的元素是 0，则再根据第一向量的

位图确定优先级第三的数据单元中存储的元素是否为 0，以此类推，直至生成该多路复用器的选通信号。通过设置优先级，能够有序的读取各个多路复用器连接的多个数据单元中的元素。

需要说明的是，第一调度模块在确定一个数据单元中存储的元素不为 0，并生成该数据单元对应的选通信号之后，需要将该数据单元存储的元素在位图中对应的比特位置为 0，以防止该数据单元中的元素再次被读取，导致计算错误。

在一种可能的实现方式中，第一复用器具体用于：根据第一复用器的第 j 个多路复用器的选通信号，通过第一复用器的第 j 个多路复用器获取第一复用器的第 j 个多路复用器连接的数据单元中第 k 个数据单元中的第一元素，并将第一元素输入到第一处理元件，该第一元素是上述第一组非零元素中的一个；第二复用器具体用于：根据上述第一复用器的第 j 个多路复用器的选通信号，通过第二复用器的第 j 个多路复用器，获取第二复用器的第 j 个多路复用器连接的数据单元中第 k 个数据单元中的第二元素，并将第二元素输入到第一处理元件，该第二元素是上述第二组元素中的一个。

在一个周期内，第一调度模块依次生成第一复用器中 K 个多路复用器各自对应的选通信号，使得第一复用器和第二复用器分别从连接的缓存器各取出 K 各元素，进而基于取出的 K 对数据实现第一向量和第二向量的点积。应理解，第一复用器一个周期内可能不能取出第一缓存器中的全部非零元素，但是第一复用器一次能取出 K 个非 0 元素，第一缓存器中存储有 W 行 K 列元素，因此最多经过 W 个周期，即可取出第一缓存器中所有的非零元素，完成第一向量和第二向量之间的运算。

在一种可能的实现方式中，第一处理元件在完成第一向量和第二向量的运算之后，第一调度模块会生成擦除信号，该擦除信号用于指示第一缓存器和第二缓存器将当前缓存的数据擦除，以用于缓存下一次计算需要的数据。

在一个可能的实现方式中，上述第一向量可能是一个向量的一部分，第二向量可能是另一个向量的一部分；或者上述第一向量是一个行向量，第二向量是一个矩阵中的一列；又或者，上述第一向量属于第一矩阵中的任意一行，上述第二向量属于第二矩阵中的任意一列。则该芯片通过多次上述计算过程，能够实现向量与向量的运算、向量与矩阵的运算或者矩阵与矩阵的运算。

在一种可能的实现方式中，上述芯片还包括第三缓存器和第二处理元件，第三缓存器用于缓存第三向量，第三向量属于第二矩阵中除上述第二向量所在的列之外的一列；第一选通信号还用于使第二处理元件从第三缓存器中获取第三向量中的第三组元素；第二处理元件用于根据第一组非零元素和第三组元素进行第一向量和第三向量的运算。

通过增加第三缓存器和第二处理元件，通过第二处理元件能够完成第一向量和第三向量的运算，通过第一处理元件能够完成第一向量和第二向量的运算，而第二向量和第三向量都属于第二矩阵，从而能够使芯片实现向量和矩阵的运算。

在一种可能的实现方式中，上述芯片还包括第三复用器，第三复用器用于根据第一选通信号从第三缓存器中获取第三向量中的第三组元素，并输入上述第二处理元件。

在一种可能的实现方式中，上述第三复用器包括 K 个多路复用器，上述第三缓存器包括 W 行 K 列数据单元，每个数据单元用于缓存一个元素；上述第三复用器中的第 i 个多路复用器与上述第三缓存器中数据单元的连接关系，与第一复用器中的第 i 个多路复用器与第一缓存器中数据单元的连接关系相同。

第三复用器用于从第三缓存器中获取第三向量的元素，并输入第二处理元件，以使第二处理元件实现第一向量与第三向量的运算。第一向量与第三向量的运算中，第一向量的第 r 个元素与第三向量的第 r 个元素需要做乘法运算，在将第一向量的各个元素存入第一缓存器以及将第三向量的各个元素存入第三缓存器中时，第一向量中的第 r 个元素存入的第一缓存器的相对位置和第三向量中的第 r 个元素存入第三缓存器的相对位置相同。例如，将第一向量的第 r 个元素存入第一缓存器的第 1 行的第 r 个数据单元，将第三向量的第 r 个元素存入第三缓存器的第 1 行的第 r 个数据单元中。同时，第一复用器中的第 i 个多路复用器连接了第一缓存器中的多个数据单元。这多个数据单元在第一缓存器中的相对位置，与第三复用器中的第 i 个多路复用器连接的第三缓存器中的多个数据单元在第三缓存器中的相对位置相同。例如，第一复用器中第 i 个多路复用器连接了第一缓存器中第一行的第 2 个数据单元和第二行的第 5 个数据单元，则第三复用器中第 i 个多路复用器连接第三缓存器中第一行的第 2 个数据单元和第二行的第 5 个数据单元。这样能够使第一复用器的第 i 个多路复用器与第三复用器的第 i 个多路复用器根据同一个选通信号，读取的是第一缓存器和第三缓存器中相对位置相同的数据单元中的数据，使第一复用器和第三复用器根据相同的选通信号读取的正好是第一向量和第三向量中需要进行乘法运算的元素。

在一种可能的实现方式中，上述第三复用器具体用于：根据第一复用器的第 j 个多路复用器的选通信号，通过第三复用器的第 j 个多路复用器，获取所述第三复用器的第 j 个多路复用器连接的数据单元中第 k 个数据单元中的第三元素，并将第三元素输入到上述第二处理元件，该第三元素是第三组元素中的一个。

在一种可能的实现方式中，上述芯片还包括第四缓存器、第二调度模块和第三处理元件；其中，第四缓存器用于缓存第四向量，该第四向量属于上述第一矩阵中除第一向量所在的行之外的一行；

第二调度模块用于根据第四向量的位图生成第二选通信号，该第二选通信号用于使第三处理元件从第四缓存器中获取第四向量中的第四组非零元素；使第三处理元件从上述第二缓存器中获取第二向量中的第五组元素；其中，第四向量的位图指示第四向量中的非零元素；

第三处理元件用于根据第四组非零元素和第五组元素实现第四向量和第二向量之间的运算。

通过在增加第三缓存器和第二处理元件之后，再次增加第四缓存器、第二调度模块和第三处理元件；通过第三处理元件能够完成第一矩阵中的第四向量与第二矩阵中的第二向量的运算，通过第二处理元件能够完成第一矩阵中的第一向量和第二矩阵中第三向量的运算，通过第二元件能够完成第一矩阵中的第一向量和第二矩阵中的第二向量的运算，从而能够使芯片实现矩阵和矩阵的运算。

在一种可能的实现方式中，上述芯片还包括第四复用器和第五复用器，第四复用器，用于根据第二选通信号从第四缓存器中获取第四向量中的第四组非零元素，并输入第三处理元件；第五复用器，用于根据第二选通信号从第二缓存器中获取第二向量中的第五组元素，并输入第三处理元件。

在一种可能的实现方式中，第四复用器包括 K 个多路复用器，第五复用器包括 K 个多路复用器；第四缓存器包括 W 行 K 列数据单元，每个数据单元用于缓存一个元素；

上述第四复用器中的每个多路复用器与第四缓存器中的多个数据单元连接，且每个数据单元至少与一个多路复用器连接；上述第四复用器中的第 i 个多路复用器与所述第四缓存器

中数据单元的连接关系，与上述第一复用器中的第 i 个多路复用器与所述第一缓存器中数据单元的连接关系相同；第五复用器中的第 i 个多路复用器与第二缓存器中数据单元的连接关系，与上述第一复用器中的第 i 个多路复用器与第一缓存器中数据单元的连接关系相同。

需要说明的是，第五复用器也可以连接一个第五缓存器，但是由于第五复用器用于获取第二向量的元素，而第二缓存器中缓存有第二向量，因此第五复用器可以和第二复用器共享第二复用器，从而能够降低芯片的复杂度、减小芯片体积以及降低成本。

第四复用器用于从第四缓存器中获取第三向量的元素，并输入第三处理元件，以使第三处理元件实现第四向量与第二向量的运算。第四向量与第二向量的运算中，第四向量的第 r 个元素与第二向量的第 r 个元素需要做乘法运算，在将第四向量的各个元素存入第四缓存器以及将第二向量的各个元素存入第二缓存器中时，第四向量中的第 r 个元素存入的第四缓存器的相对位置和第二向量中的第 r 个元素存入第二缓存器的相对位置相同。例如，将第二向量的第 r 个元素存入第二缓存器的第 1 行的第 r 个数据单元，将第四向量的第 r 个元素存入第四缓存器的第 1 行的第 r 个数据单元中。同时，第四复用器中的第 i 个多路复用器连接了第四缓存器中的多个数据单元。这多个数据单元在第四缓存器中的相对位置，与第五复用器中的第 i 个多路复用器连接的第二缓存器中的多个数据单元在第二缓存器中的相对位置相同。这样能够使第四复用器的第 i 个多路复用器与第五复用器的第 i 个多路复用器根据同一个选通信号，读取的是第四缓存器和第二缓存器中相对位置相同的数据单元中的数据，使第四复用器和第五复用器根据相同的选通信号读取的正好是第四向量和第二向量中需要进行乘法运算的元素。

在一种可能的实现方式中，上述第二调度模块具体用于：根据第四向量的位图确定第四复用器的第 j 个多路复用器连接的数据单元中，第 m 个数据单元存储的元素为非零元素，则第一调度模块生成第 j 个多路复用器的选通信号，将第 j 个多路复用器的选通信号发送给第四复用器的第 j 个多路复用器和第五复用器的第 j 个多路复用器，第二选通信号包括第四复用器的第 j 个多路复用器的选通信号。

在一种可能的实现方式中，上述第四复用器具体用于：根据第四复用器的第 j 个多路复用器的选通信号，通过第四复用器的第 j 个多路复用器，获取第四复用器的第 j 个多路复用器连接的数据单元中第 m 个数据单元中的第三元素，并将第三元素输入到所述第三处理元件，第三元素是所述第四组非零元素中的一个；第五复用器具体用于：根据第四复用器的第 j 个多路复用器的选通信号，通过第五复用器的第 j 个多路复用器，获取第五复用器的第 j 个多路复用器连接的数据单元中第 m 个数据单元中的第四元素，并将第四元素输入到第三处理元件，上述第四元素是第五组元素中的一个。

在一个周期内，第二调度模块依次生成第四复用器中 K 个多路复用器各自对应的选通信号，使得第四复用器和第五复用器分别从连接的缓存器各取出 K 各元素，进而基于取出的 K 对数据实现第四向量和第二向量的点积。应理解，第二复用器一个周期内可能不能取出第四缓存器中的全部非零元素，但是第四复用器一次能取出 K 个非 0 元素，第四缓存器中存储有 W 行 K 列元素，因此最多经过 W 个周期，即可取出第四缓存器中所有的非零元素，完成第四向量和第二向量之间的运算。

第二方面，本申请提供一种矩阵计算方法，应用于芯片，该芯片包括第一缓存器、第二缓存器、第一调度模块和第一处理元件，该方法包括：芯片通过第一缓存器缓存第一向量，通过第二缓存器缓存第二向量；芯片在缓存第一向量和第二向量之后，芯片的第一调度模块

根据第一向量的位图生成第一选通信号，该第一选通信号能够使第一处理元件从第一缓存器中获取第一向量中的第一组非零元素，使第一处理元件从第二缓存器中获取第二向量中的第二组元素，其中，第一向量的位图指示第一向量中的非零元素；然后该芯片通过第一处理元件根据第一组非零元素和第二组元素实现第一向量和第二向量之间的运算。

在一种可能的实现方式中，上述芯片还包括第一复用器和第二复用器；该芯片的第一复用器根据第一选通信号从第一缓存器中获取第一向量中的第一组非零元素，并输入第一处理元件；芯片的第二复用器根据第一选通信号从第二缓存器中获取第二向量中的第二组元素，并输入第一处理元件。

在一种可能的实现方式中，第一复用器和第二复用器各自包括 K 个多路复用器；第一缓存器和第二缓存器均包括 W 行 K 列数据单元，每个数据单元用于缓存一个元素；第一复用器中的每个多路复用器与第一缓存器中的多个数据单元连接，且每个数据单元至少与一个多路复用器连接；第二复用器中的第 i 个多路复用器与第二缓存器中数据单元的连接关系，和第一复用器中的第 i 个多路复用器与第一缓存器中数据单元的连接关系相同。

在一种可能的实现方式中，上述芯片的第一调度模块根据第一向量的位图生成第一选通信号，包括：芯片根据第一向量的位图确定第一复用器的第 j 个多路复用器连接的数据单元中，第 k 个数据单元存储的元素为非零元素，第一调度模块生成第 j 个多路复用器的选通信号，将第 j 个多路复用器的选通信号发送给第一复用器的第 j 个多路复用器和第二复用器的第 j 个多路复用器，其中，第一选通信号包括第一复用器的第 j 个多路复用器的选通信号。

在一种可能的实现方式中，上述芯片的第一复用器根据第一选通信号从第一缓存器中获取第一向量中的第一组非零元素，并输入所述第一处理元件，芯片的第二复用器根据第一选通信号从第二缓存器中获取第二向量中的第二组元素，并输入第一处理元件，包括：芯片的第一复用器根据第一复用器的第 j 个多路复用器的选通信号，通过第一复用器的第 j 个多路复用器，获取第一复用器的第 j 个多路复用器连接的数据单元中第 k 个数据单元中的第一元素，并将第一元素输入到第一处理元件；该芯片的第二复用器根据第一复用器的第 j 个多路复用器的选通信号，通过第二复用器的第 j 个多路复用器，获取第二复用器的第 j 个多路复用器连接的数据单元中第 k 个数据单元中的第二元素，并将第二元素输入到第一处理元件，其中，第一元素是第一组非零元素中的一个，第二元素是第二组元素中的一个。

在一种可能的实现方式中，上述第一向量属于第一矩阵中的任意一行，上述第二向量属于第二矩阵中的任意一列。即上述第一向量可以是一个矩阵中的任意一行中的部分或者全部元素，上述第二向量可以是另一个矩阵中的任意一列中的部分或者全部元素。

在一种可能的实现方式中，上述芯片还包括第三缓存器和第二处理元件；上述方法还包括：上述芯片通过第三缓存器缓存第三向量，该第三向量属于上述第二矩阵中除第二向量所在的列之外的一列，上述第一选通信号还能够使第二处理元件从第三缓存器中获取第三向量中的第三组元素；上述芯片的第二处理元件根据第一组非零元素和第三组元素实现第一向量和第三向量的运算。

在一种可能的实现方式中，上述芯片还包括第三复用器，上述方法还包括：上述芯片的第三复用器根据第一选通信号从第三缓存器中获取第三向量中的第三组元素，并输入第二处理元件。

在一种可能的实现方式中，上述第三复用器包括 K 个多路复用器，第三缓存器包括 W 行 K 列数据单元，每个数据单元用于缓存一个元素；第三复用器中的第 i 个多路复用器与第

三缓存器中数据单元的连接关系，和上述第一复用器中的第 i 个多路复用器与第一缓存器中数据单元的连接关系相同。

在一种可能的实现方式中，上述芯片的所述第三复用器根据所述第一选通信号从第三缓存器中获取所述第三向量中的所述第三组元素，并输入第二处理元件，包括：

上述芯片的第三复用器根据第一复用器的第 j 个多路复用器的选通信号，通过第三复用器的第 j 个多路复用器，获取第三复用器的第 j 个多路复用器连接的数据单元中第 k 个数据单元中的第三元素，并将第三元素输入到第二处理元件，其中，第三元素是上述第三组元素中的一个。

在一种可能的实现方式中，上述芯片还包括第四缓存器、第二调度模块和第三处理元件；上述方法还包括：芯片通过第四缓存器缓存第四向量，该第四向量属于第一矩阵中除第一向量所在的行之外的一行；芯片的第二调度模块根据第四向量的位图生成第二选通信号，该第二选通信号用于使第三处理元件从第四缓存器中获取第四向量中的第四组非零元素；使第三处理元件从第二缓存器中获取第二向量中的第五组元素，其中，上述第四向量的位图指示第四向量中的非零元素；上述芯片的第三处理元件根据第四组非零元素和第五组元素实现第四向量和第二向量的运算。

在一种可能的实现方式中，上述芯片还包括第四复用器和第五复用器，该方法还包括：芯片的第四复用器根据上述第二选通信号从第四缓存器中获取第四向量中的第四组非零元素，并输入第三处理元件；芯片的第五复用器根据第二选通信号从第二缓存器中获取第二向量中的第五组元素，并输入第三处理元件。

在一种可能的实现方式中，上述第四复用器和第五复用器各自包括 K 个多路复用器；上述第四缓存器包括 W 行 K 列数据单元，每个数据单元用于缓存一个元素；

上述第四复用器中的每个多路复用器与第四缓存器中的多个数据单元连接，且每个数据单元至少与一个多路复用器连接；上述第四复用器中的第 i 个多路复用器与第四缓存器中数据单元的连接关系，和上述第一复用器中的第 i 个多路复用器与第一缓存器中数据单元的连接关系相同；第五复用器中的第 i 个多路复用器与第五缓存器中数据单元的连接关系，和第一复用器中的第 i 个多路复用器与第一缓存器中数据单元的连接关系相同。

在一种可能的实现方式中，上述芯片的第二调度模块根据第四向量的位图生成第二选通信号，包括：根据第四向量的位图确定第四复用器的第 j 个多路复用器连接的数据单元中，第 m 个数据单元存储的元素为非零元素，则第二调度模块生成所述第 j 个多路复用器的选通信号，将第 j 个多路复用器的选通信号发送给第四复用器的第 j 个多路复用器和第五复用器的第 j 个多路复用器，上述第二选通信号包括所述第四复用器的第 j 个多路复用器的选通信号。

在一种可能的实现方式中，上述芯片的第四复用器根据第二选通信号从第四缓存器中获取第四向量中的所述第四组非零元素，并输入所述第三处理元件；上述芯片的第五复用器根据第二选通信号从第二缓存器中获取第二向量中的第五组元素，并输入第三处理元件，包括：上述芯片的第四复用器根据第四复用器的第 j 个多路复用器的选通信号，通过第四复用器的第 j 个多路复用器，获取第四复用器的第 j 个多路复用器连接的数据单元中第 m 个数据单元中的第四元素，并将第四元素输入到第三处理元件；芯片的第五复用器根据第四复用器的第 j 个多路复用器的选通信号，通过第五复用器的第 j 个多路复用器，获取第五复用器的第 j 个多路复用器连接的数据单元中第 m 个数据单元中的第五元素，并将第五元素输入到第三处

理元件，其中，上述第四元素是第四组非零元素中的一个，上述第五元素是第五组元素中的一个。

第三方面，本申请提供一种矩阵计算装置，包括第一调度单元和第一处理单元，其中，第一调度单元用于根据第一向量的位图生成第一选通信号，该第一选通信号能够使第一处理单元从第一缓存器中获取第一向量中的第一组非零元素，使第一处理单元从第二缓存器中获取第二向量中的第二组元素，第一处理单元用于根据第一向量中的第一组非零元素和第二向量中的第二组元素实现第一向量和第二向量的运算，其中，第一向量的位图指示第一向量中的非零元素。

在一种可能的实现方式中，上述第一处理单元根据第一组非零元素和第二组元素实现第一向量和第二向量之间的运算之后，还包括：第一调度单元生成擦除信号，该擦除信号指示第一缓存器和第二缓存器擦除当前各自缓存的数据。

在一种可能的实现方式中，上述第一向量属于第一矩阵中的任意一行的部分或全部元素，第二向量属于第二矩阵中的任意一列的部分或全部元素。

在一种可能的实现方式中，上述矩阵计算装置还包括第二处理单元，上述第一选通信号还能够使第二处理单元获取第三向量中的第三组元素；第二处理单元用于根据第一组非零元素和第三组元素实现第一向量和第三向量的运算；其中，上述第三向量属于上述第二矩阵中除第二向量所在的列之外的一列。

在一种可能的实现方式中，上述矩阵计算装置还包括第二调度单元和第三处理单元，第二调度单元用于根据第四向量的位图生成第二选通信号，第二选通信号用于使第三处理元件获取第四向量中的第四组非零元素；使第三处理元件获取第二向量中的第五组元素，其中，第四向量的位图指示第四向量中的非零元素，所述第四向量属于所述第一矩阵中除所述第一向量所在的行之外的一行；第三处理单元，用于根据第四组非零元素和第五组元素实现第四向量和第二向量的运算。

第四方面，本申请提供一种计算设备，该计算设备包括芯片和存储器，存储器用于存储代码，所述芯片执行所述代码实现如第二方面以及第二方面任意可能实现方式中所述的方法。

第五方面，本申请提供一种计算机可读存储介质，计算设备可读存储介质中存储有指令，当其在计算设备上运行时，使得计算设备执行第二方面以及第二方面任意可能实现方式中所述的方法。

第六方面，提供了一种包含指令的计算机程序产品，包括计算机程序或指令，当该计算机程序或指令在计算设备上运行时，使得该计算设备执行第二方面以及第二方面任意可能实现方式中所述的方法。

本申请在上述各方面提供的实现方式的基础上，还可以进行进一步组合以提供更多实现方式。

附图说明

为了更清楚地说明本申请实施例技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 为本申请实施例提供的一种芯片的结构示意图；

图 2 为本申请实施例提供的一种处理元件的示意图；

图 3 为本申请实施例提供的一种缓存器和复用器的示意图；

图 4 为本申请实施例提供的一种多路复用器与缓存器连接的示意图；

图 5 为本申请实施例提供的一种数据单元优先级的示意图；

图 6 为本申请实施例提供的一种缓存器中的数据与对应的位图的示意图；

图 7 为本申请实施例提供的一种实现向量与向量运算的示意图；

图 8 为本申请实施例提供的另一种芯片的结构示意图；

图 9 为本申请实施例提供的一种实现向量与矩阵运算的示意图；

图 10 是本申请实施例提供的另一种芯片的结构示意图；

图 11 为本申请实施例提供的一种实现矩阵与矩阵运算的示意图；

图 12 是本申请实施例提供的一种矩阵计算方法的流程示意图；

图 13 是本申请实施例提供的一种矩阵计算装置的示意图；

图 14 是本申请实施例提供的一种计算设备的结构示意图。

具体实施方式

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描述，显然，所描述的实施例仅仅是本申请的一部分实施例，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，都属于本申请保护的范围。

当前在涉及到矩阵乘法运算的场景中，为了提高计算效率，通常采用剪枝技术对矩阵进行稀疏化处理以降低计算量，提高计算效率。例如在 AI 训练完成后，将训练后的权重矩阵进行结构化剪枝完成权重矩阵的稀疏化，在 AI 推理时采用结构化剪枝后的权重矩阵进行推理。但是上述方法仅适合在确定参与计算的矩阵后，对该矩阵进行剪枝操作，例如 AI 推理场景，但是不适合其他矩阵稀疏度动态变化的场景。例如在 AI 训练时，权重（weight）矩阵、梯度（gradients）矩阵或者激活（activation）矩阵等是动态变化的，矩阵的稀疏度也是动态变化的，如果在要在 AI 训练的过程中实现矩阵的稀疏化，需要较大的时间成本，会抵消通过剪枝实现矩阵稀疏化之后带来的加速效果。因此，如何实现矩阵稀疏程度动态变化下矩阵计算的加速是一个亟待解决的技术问题。

本申请提供一种芯片，能够实现稀疏度动态变化的矩阵的运算。如图 1 所示，图 1 是本申请实施例提供的一种芯片的结构示意图。该芯片 100 包括处理元件（Processing Elements，PE）110、第一缓存器 120、第一复用器（Multiplexer，MUX）130、第二缓存器 140、第二复用器 150 以及调度模块 160。其中，第一复用器 130 的输入端与第一缓存器 120 连接，第一复用器 130 的输出端与处理元件 110 连接；第二复用器 150 输入端与第二缓存器 140 连接，第二复用器 150 输出端与处理元件 110 连接；第一复用器 130 和第二复用器 150 的控制端分别与调度模块 160 连接。

当需要进行第一向量和第二向量的运算时，第一缓存器 120 用于缓存第一向量中的全部元素或者部分元素；第二缓存器 140 用于缓存第二向量中的全部元素或者部分元素。调度模块 160 用于根据第一缓存器 120 中的数据，产生选通信号，并将该选通信号发送给第一复用器 130 和第二复用器 150，该选通信号用于使第一复用器 130 从连接的第一缓存器 120 中获取一组非 0 元素，使第二复用器从连接的第二缓存器 140 中获取一组元素；第一复用器 130

和第二复用器 150 将各自从对应缓存器中获取的一组数据输入至处理元件 110 中，以使处理元件 110 进行第一向量与第二向量之间的运算。

下面结合附图对芯片 100 所包括的各个部分分别进行详细介绍。如图 2 所示，图 2 是本申请实施例提供的一种处理元件的示意图。处理元件 110 为一个长度为 K 的点积处理模块，包括 K 个乘积累加（Multiply Accumulate，MAC）单元和 K-1 个加法器（或加法树）；一个处理元件 110 一个周期能够从第一缓存器 120 中获取 K 个元素，并能够从第二缓存器 140 中获取 K 个元素，即一个处理元件 110 一个周期能够获取 K 对值，执行 K 次乘积和 K-1 次加法操作，实现$a_1b_1+a_2b_2+\cdots+a_Kb_K$的操作。其中，$a_1$、$a_2$、$a_3$、…、$a_K$为第一向量中的元素，$b_1$、$b_2$、$b_3$、…、$b_K$为第二向量中的元素。

如图 3 所示，图 3 是本申请实施例提供的一种缓存器和复用器的示意图。每个缓存器包括 W 行 K 列的数据单元，每个数据单元用于缓存一个元素的值，图 3 中以 W 等于 4，K 等于 8 为例。每个缓存器对应有 K 个读端口，每个读端口被实现为一个多路复用器，一个缓存器对应的复用器包括 K 个多路复用器，即上述图 2 中的第一复用器 130 和第二复用器 150 分别包括 K 个多路复用器。每个多路复用器分别与缓存器中的多个数据单元连接，且每个数据单元与至少一个多路复用器连接。每个多路复用器能够用于读取该多路复用器所连接的多个数据单元中缓存的数据。其中，每个多路复用器每次只能根据调度模块 160 的选通信号从连接的多个数据单元中读取其中一个数据单元中的数据。

如图 4 所示，图 4 为本申请实施例提供的一种多路复用器与缓存器连接的示意图。一个复用器包括多个多路复用器，每个多路复用器分别与缓存器中的多个数据单元连接，一个多路复用器连接的多个数据单元具有不同的优先级。本申请实施例中，将每个缓存器中的数据单元表示为（r，c），如图 4 中所示，第 1 行第 3 列的数据单元表示为（1，3），第 4 行第 5 列的数据单元表示为（4，5），以此类推，则 r 为[1，W]的整数，c 为[1，K]的整数。以图 4 中多路复用器与缓存器的连接关系为例，即每个多路复用器与缓存器的 8 个数据单元连接，第 i 个多路复用连接的 8 个数据单元的优先级从高到低依次为：（1，i），（2，i），（3，i），（4，i），（2，i+1），（2，mod（i-1 ,8））,（3，mod（i-2 ,8））,（4，mod（i-3 ,8））。其中，mod 为取模运算。

示例性的，图 5 中示出了 i 等于 1（即第一个多路复用器）和 i 等于 4（即第四个多路复用器）时，两个多路复用器各自连接的 8 个数据单元的优先级顺序。其中，图 5 中数据单元中的数字 1~8 表示优先级，数字越小优先级越高；第一个多路复用器连接的 8 个数据单元的优先级从高到低依次为：（1，1），（2，1），（3，1），（4，1），（2，2），（2，8），（3，7），（4，6）；第四个多路复用器连接的 8 个数据单元的优先级从高到低依次为：（1，4），（2，4），（3，4），（4，4），（2，5），（2，3），（3，2），（4，1）。

需要说明的是，上述图 4 中所示的每个多路复用器连接 8 个数据单元仅是一种示例，不能理解为具体限定，每个多路复用器还可以连接更多或更少的数据单元，例如每个多路复用器连接 7 个数据单元，或者连接 10 个、15 个或 16 个数据单元等，但每个数据单元至少与一个多路复用器连接即可。上述图 4 中每个多路复用器与多个数据单元的连接关系仅是一种示例，不能理解为具体限定，本申请对每个多路复用器与多个数据单元的连接关系不做具体限定。但是在同一个芯片中，每个缓存器所包括的数据单元均应为 W 行 K 列。将一个缓存器以及与该缓存器连接的复用器称为缓存模块，则同一个芯片中任意两个缓存模块中复用器与缓存器中数据单元的连接关系相同，也就是说，如果一个芯片包括多个缓存模块，每个缓

存模块中的多路复用器与缓存器中数据单元的连接关系相同。例如第一复用器 130 中第一个多路复用器连接的多个数据单元在第一缓存器 120 中的相对位置，与第二复用器 150 中第一个多路复用器连接的多个数据单元在第二缓存器 140 中的相对位置相同。示例性的，第一复用器 130 中第 i 个多路复用器连接了第一缓存器 120 中（1，i），（2，i），（3，i），（4，i），（2，i+1），（2，mod（i-1 ,8））， （3，mod（i-2 ,8））， （4，mod（i-3 ,8））这 8 个数据单元，则第二复用器中第 i 个多路复用器同样连接第二缓存器中（1，i），（2，i），（3，i），（4，i），（2，i+1），（2，mod（i-1 ,8））， （3，mod（i-2 ,8））， （4，mod（i-3 ,8））这 8 个数据单元。这样能够使第一复用器的第 i 个多路复用器与第二复用器的第 i 个多路复用器根据同一个选通信号，读取的是第一缓存器 120 和第二缓存器 140 中相对位置相同的数据单元中的数据，使第一复用器和第二复用器根据相同的选通信号读取的正好是第一向量和第二向量中进行乘法运算的元素。

本申请实施例中，每个多路复用器每次从所连接的多个数据单元中读取一个非 0 的元素，调度模块 160 在确定一个多路复用器读取哪个数据单元中的数据时，根据优先级的高低，从优先级最高的数据单元开始，确定优先级为 1 的数据单元中的元素是否为 0，如果优先级为 1 的数据单元中的元素不为 0，则调度模块 160 生成优先级为 1 的数据单元的选通信号，使多路复用器读取优先级为 1 的数据单元中的元素；如果优先级为 1 的数据单元中的元素为 0，则调度模块 160 再确定优先级为 2 的数据单元中的元素是否为 0，如果优先级为 2 的数据单元中的元素不为 0，则调度模块 160 生成优先级为 2 的数据单元的选通信号，使多路复用器读取优先级为 2 的数据单元中的元素；如果优先级为 2 的数据单元中的元素为 0，则调度模块 160 再确定优先级为 3 的数据单元中的元素是否为 0。依次类推，调度模块 160 按照优先级的顺序，直至找到一个存储的元素不为 0 的数据单元，然后生成该数据单元对应的选通信号，使多路复用器读取该数据单元中的元素并发送至处理元件 110，以使处理元件 110 进行点积运算。

上述一个缓存器与 K 个多路复用器连接，调度模块 160 在一个周期内，对一个缓存器对应的 K 个多路复用器，需要产生 K 个调度信号 DS_j，每个多路复用器对应一个选通信号。其中，j =1，2，…，8，j 为正整数。也就是说，每个周期内，每个多路复用器需要根据调度模块 160 的选通信号，从缓存器中读取一个元素发送到处理元件 110 中，一个周期内，复用器能够通过 K 个多路复用器从连接的缓存器中获取 K 个元素。需要说明的是，如果一个多路复用器连接的多个数据单元中的元素都为 0 时，该多路复用器将元素 0 发送到处理元件 110。

本申请实施例中，上述调度模块 160 根据缓存器中存储的数据对应的位图（bitmap）确定每个数据单元中的元素是否为 0。具体的，如图 6 所示，图 6 是缓存器中的数据与对应的位图的示意图。每个缓存器在存入数据之后，会根据数据单元中缓存的数据生成一个位图，该位图中的每一位对应缓存器中的一个数据单元。位图中的每一位的值为 0 或 1，用于指示对应的数据单元中的元素是否为 0。例如，当位图中的一位为 0 时，指示该位对应的数据单元中的元素的值为 0；当位图中的一位为 1 时，指示该位对应的数据单元中的值不为 0；或者当位图中的一位为 0 时，指示该位对应的数据单元中的值不为 0；当位图中的一位为 1 时，指示该位对应的数据单元中的值为 0。本申请实施例中，以位图中的一位为 0 指示该位对应的数据单元中的值为 0，位图中的一位为 1 时指示该位对应的数据单元中的值不为 0 为例，对本申请实施例进行介绍。

调度模块160在需要生成一个多路复用器的选通信号时，调度模块160首先确定位图中，该多路复用器连接的数据单元中优先级为1的数据单元对应的位的值是否为0，如果位图中优先级为1的数据单元对应的位的值为1，则表示优先级为1的数据单元中的元素不为0，调度模块160生成优先级为1的数据单元对应的选通信号发送给该多路复用器。如果位图中优先级为1的数据单元对应的位的值为0，则表示优先级为1的数据单元中的元素为0，调度模块160再确定位图中，该多路复用器连接的数据单元中优先级为2的数据单元对应的位的值是否为0，如果位图中优先级为2的数据单元对应的位的值为1，则表示优先级为2的数据单元中的元素不为0，调度模块生成优先级为2的数据单元对应的选通信号发送给该多路复用器。如果位图中优先级为2的数据单元对应的位的值为0，则表示优先级为2的数据单元中的元素为0，调度模块160再确定位图中，该多路复用器连接的数据单元中优先级为3的数据单元对应的位的值是否为0。以此类推，在此不再赘述。

需要说明的是，调度模块160控制一个多路复用器从一个数据单元中获取一个非0的元素后，调度模块160需要将该数据单元存储的元素在位图中对应的位置0，以防止该数据单元中的元素被重复读取，避免优先级低于该数据单元的数据单元中的非0元素没有被读取并参与运算而导致的计算错误。

上述图2至图6介绍了图1所示的芯片100所包括的各部分的工作原理，下面结合附图介绍利用上述芯片100实现向量与向量的运算、向量与矩阵的运算以及矩阵与矩阵的运算的方法。

如图7所示，图7是本申请实施例提供的一种实现向量与向量运算的示意图。以图1中的芯片100实现向量C与向量D的点积运算为例，其中，芯片100中的第一缓存器120和第二缓存器140均包括W行K列的数据单元，第一复用器130和第二复用器150均包括K个多路复用器。向量C为1*WK的向量，向量D为WK*1的向量，即向量C包括1行W*K列元素，向量D包括W*K行1列元素。

在通过芯片100进行点积运算时，芯片100将向量C包括的元素加载到第一缓存器120的数据单元中，将向量D包括的元素加载到第二缓存器140的数据单元中。在将向量C的数据加载到第一缓存器120和向量D的数据加载到第二缓存器140中时，向量C的第1列到第K列的数据被依次存入第一缓存器120的第一行的K个数据单元中，向量D的第1行到第K行的数据被依次存入第二缓存器140的第一行的K个数据单元中；向量C的第K+1列到第2K列的数据被依次存入第一缓存器120的第二行的K个数据单元中，向量D的第K+1行到第2K行的数据被依次存入第二缓存器140的第二行的K个数据单元中；依此类推，直至将向量C的第（W-1）K列至第WK列的数据依次存入第一缓存器120的第W行的K个数据单元中，将向量D的第（W-1）K行至第WK行的数据依次存入第二缓存器140的第W行的K个数据单元中。

向量C的W*K个数据存入第一缓存器120后，第一缓存器120根据每个数据单元中存储的元素的值生成对应的位图，如果一个数据单元中的元素的值不为0，则将该数据单元在位图中对应的位设置为1，如果一个数据单元中的元素的值为0，则将该数据单元在位图中对应的位设置为0。如图6所示，图6中所示的位图为第一缓存器120缓存的数据对应的位图。图6中以W等于4，K等于8为例，第一复用器130中各个多路复用器与第一缓存器120中数据单元的连接关系如图4中所示。应理解，也可以由第二缓存器140根据每个数据单元

中的值生成对应的位图，本申请实施例不做具体限定。

第一缓存器 120 在生成位图之后，将位图发送给调度模块 160。调度模块 160 在接收到位图之后，需要首先生成第一复用器 130 中第一个多路复用器的选通信号 DS_1。具体的，调度模块 160 根据第一复用器 130 中第一个多路复用器连接的多个数据单元的优先级，先确定第一复用器 130 中第一个多路复用器连接的数据单元中优先级为 1 的数据单元在位图中对应的位的值是否为 0。如果优先级为 1 的数据单元在位图中对应的位的值不为 0，则调度模块 160 生成选通信号 000，将该选通信号 000 发送给第一复用器 130 的第一个多路复用器以及第二复用器 150 的第一个多路复用器，该选通信号 000 用于使第一复用器 130 的第一个多路复用器读取优先级为 1 的数据单元中的元素并发送给处理元件 110，使第二复用器 150 的第一个多路复用器读取优先级为 1 的数据单元中的元素并发送给处理元件 110。

如果调度模块 160 确定第一复用器 130 中第一个多路复用器连接的数据单元中优先级为 1 的数据单元在位图中对应的位的值为 0，则调度模块 160 再确定第一复用器 130 中第一个多路复用器连接的数据单元中优先级为 2 的数据单元在位图中对应的位的值是否为 0。如果优先级为 2 的数据单元在位图中对应的位的值不为 0，则调度模块 160 生成选通信号 001，将该选通信号 001 发送给第一复用器 130 的第一个多路复用器以及第二复用器 150 的第一个多路复用器，该选通信号 001 用于使第一复用器 130 的第一个多路复用器读取优先级为 2 的数据单元中的元素并发送给处理元件 110，使第二复用器 150 的第一个多路复用器读取优先级为 2 的数据单元中的元素并发送给处理元件 110。

如果调度模块 160 确定第一复用器 130 中第一个多路复用器连接的数据单元中优先级为 2 的数据单元在位图中对应的位的值为 0，则调度模块 160 再确定第一复用器 130 中第一个多路复用器连接的数据单元中优先级为 3 的数据单元在位图中对应的位的值是否为 0。依次类推，直至第一复用器 130 中的第一个多路复用器从第一缓存器 120 中读取一个数据 c_1 发送给处理元件 110，以及第二复用器 150 中的第一个多路复用器从第二缓存器 140 中读取一个数据 d_1 发送给处理元件 110，从而使处理元件 110 执行 c_1*d_1 的操作。需要说明的是，如果第一复用器 130 中第一个多路复用器连接的多个数据单元所缓存的元素中存在不为 0 的数据，则 c_1 的值不为 0，d_1 可能为 0，也可能不为 0；如果第一复用器 130 中第一个多路复用器连接的多个数据单元所缓存的数据均为 0，则 c_1 的值为 0，d_1 可能为 0，也可能不为 0。

对于第一复用器 130 的第 2 个至第 K 个多路复用器，调度模块 160 依次通过上述相同的方法生成对应的选通信号 DS_2~DS_K，使第一复用器 130 和第二复用器 150 中的每个多路复用器输出一个数据至处理元件 110。在第一个周期内，第一复用器 130 和第二复用器 150 中的每个多路复用器各输出一个数据至处理元件 110，使处理元件 110 完成 K 次乘积操作和 K-1 次加法操作。上述 K 次乘积操作和 K-1 次加法操作为：$e_1 = c_1d_1 + c_2d_2 + \cdots + c_td_t + \cdots + c_Kd_K$，其中，$c_t$ 表示第一复用器 130 中第 t 个多路复用器输出的数据，d_t 表示第二复用器 150 中第 t 个多路复用器输出的数据，t 为大于 0 小于等于 K 的正整数。

需要说明的是，调度模块 160 在生成一个选通信号发送给第一复用器 130，第一复用器 130 从一个数据单元中读取一个数据后，调度模块 160 将该数据单元在位图中对应的位置 0。

在第二个周期内以及后续的每个周期内，调度模块 160 继续执行第一周期中所执行的操作，使第一复用器 130 和第二复用器 150 中的每个多路复用器输出一个数据至处理元件 110，处理元件 110 完成 K 次乘积操作和 K-1 次加法操作。直至位图中所有位的值均为 0。最后将

每个周期内处理元件 110 完成 K 次乘积操作和 K-1 次加法操作后得到的值相加，即为向量 C 和向量 D 的点积。

由于一个缓存器中包括 W*K 个数据单元，而一个周期复用器能够读取 K 个数据单元中的数据参与计算，因此最多经过 W 个周期，即可完成向量 C 和向量 D 的点积运算。如果向量 C 具有一定的稀疏度，即向量 C 中存在部分值为 0 的元素，通过上述提供的芯片，即使每次输入到缓存器中的向量 C 的稀疏度变换，也能够使向量 C 中元素值为 0 的元素不参与计算，从而能够在降低计算量，提高计算效率的同时，不降低计算的精度。

需要说明的是，在调度模块 160 确定位图中的值都为 0 之后，调度模块 160 会产生一个擦除信号，并发送给第一缓存器 120 和第二缓存器 140，以使第一缓存器 120 和第二缓存器 140 将当前缓存的数据擦除，以便用于缓存下一批数据。

应理解，上述向量 C 可以是向量 X 的一部分，向量 D 可以是向量 Y 的一部分，例如向量 X 是一个 1*Z 的向量，向量 Y 是一个 Z*1 的向量，其中，Z 大于 W*K。由于芯片 100 中的一个缓存器每次只能存入 W*K 个数据，因此将向量 X 和向量 Y 进行切分，每次最多将 W*K 个元素存入芯片 100 的缓存器中进行计算。上述向量 C 可以是 1 行 W*K 列的行向量，向量 D 可以是一个 W*K 行 T 列的矩阵中的任意一列，向量 C 与该矩阵的运算结果是一个 1 行 T 列的向量，在进行向量 C 与该矩阵的运算的过程中，每次将该矩阵中的一列缓存至第二缓存器 140 中，得到运算结果中的一个元素。

如图 8 所示，图 8 是本申请实施例提供的另一种芯片的示意图。该芯片 200 包括 N 个处理元件 PE_1~PE_N、N+1 个缓存器 B_0~B_N、N+1 个复用器 M_0~M_N 以及一个调度模块 210。其中，N 为大于或等于 2 的整数。缓存器 B_0 与调度模块 210 连接，调度模块 210 与每个复用器连接，N+1 个缓存器与 N+1 个复用器一一对应，每个复用器与一个缓存器连接。复用器 M_0 与 N 个处理元件均存在连接，复用器 M_0 每次读取的数据都会同步发送给 N 个处理元件。其中，每个复用器与对应缓存器的连接关系可以参照上述图 3 和图 4 所对应的描述，在此不再赘述。

本申请实施例中，该芯片 200 能够实现向量与矩阵的乘法运算。以图 8 所示的芯片 200 实现向量 C 与矩阵 B 的乘法运算为例，其中，芯片 200 中的缓存器均包括 W 行 K 列的数据单元，复用器均包括 K 个多路复用器。向量 C 为 1*WK 的向量，矩阵 B 为 WK*N 的矩阵，即向量 C 包括 1 行 W*K 列元素，矩阵 B 包括 W*K 行 N 列元素。应理解，矩阵 B 相当于 N 个 WK*1 的向量，即矩阵 B 相当于 N 个上述向量 D；向量 C 与矩阵 B 的乘法运算相当于将向量 C 分别与矩阵 B 的每一列进行向量与向量的点积运算，因此，可以将矩阵 B 看做是包括向量 D_1~D_N 的 N 个向量，每个向量分别对应矩阵 B 中的一列。

如图 9 所示，图 9 是本申请实施例提供的一种实现向量与矩阵运算的示意图。其中，图 9 中向量 C 和矩阵 B 中的黑色方块表示非 0 元素。在通过芯片 200 进行向量 C 和矩阵 B 的乘法运算时，芯片 200 将向量 C 包括的元素加载到缓存器 B_0 的数据单元中，将矩阵 B 的第 1 列对应的向量 D_1 的元素加载到缓存器 B_1 的数据单元中，将矩阵 B 的第 2 列对应的向量 D_2 的元素加载到缓存器 B_2 的数据单元中，将矩阵 B 的第 3 列对应的向量 D_3 包括的元素加载到缓存器 B_3 的数据单元中，依次类推，将矩阵 B 的第 N 列对应的向量 D_N 的元素加载到缓存器 B_N 的数据单元中。其中，将向量 C 加载到缓存器 B_0 的方法可以参照上述图 7 所对应的实施例中将向量 C 加载到第一缓存器 120 中的方法，将向量 D_1~D_N 分别加载到对应缓存器的方法可以参照上述图 7 所对应的实施例中将向量 D 加载到第二缓存器 140 中的方法，在此不再赘述。

向量 C 的 WK 个数据存入缓存器 B_0 后，缓存器 B_0 生成对应的位图，缓存器 B_0 生成对应位图的方法可以参照上述第一缓存器 120 生成位图的方法，在此不再赘述。缓存器 B_0 在生成位图之后，将该位图发送给调度模块 210。调度模块 210 在接收到位图之后，首先生成复用器 M_0 中第一个多路复用器的选通信号 DS_1。调度模块 210 生成选通信号 DS_1 的方法可以参照上述调度模块 160 生成选通信号 DS_1 的方法，在此不再赘述。

本申请实施例中，调度模块 210 在生成选通信号 DS_1 之后，将选通信号 DS_1 发送给复用器 M_0~M_N 中的第一个多路复用器，复用器 M_0 中的第一个多路复用器根据选通信号 DS_1 读取一个数据，将该数据发送给 N 个处理元件 PE_0~PE_N。复用器 M_1~M_N 中的第一个多路复用器根据选通信号 DS_1 各读取一个数据，并发送给各自连接的处理元件。

对于复用器 M_0~M_N 的第 2 个至第 K 个多路复用器，调度模块 160 依次通过上述相同的方法生成对应的选通信号 DS_2~DS_K，使复用器 M_0~M_N 中的每个多路复用器输出一个数据至处理元件 110。调度模块 210 在一个周期内依次生成选通信号 DS_1~DS_K 之后，每个处理元件会获取 K 对数据，完成 K 次乘积操作和 K-1 次加法操作。

需要说明的是，调度模块 210 在生成一个选通信号发送给复用器 M_0，复用器 M_0 从一个数据单元中读取一个数据后，调度模块 210 将该数据单元在位图中对应的位置 0。

在第二个周期内以及后续的每个周期内，调度模块 210 继续执行第一周期中所执行的操作，使复用器 M_0~M_N 中的每个多路复用器输出一个数据至处理元件 110，处理元件 110 完成 K 次乘积操作和 K-1 次加法操作，直至位图中所有位的值均为 0。对于任意处理元件 PE_h，将每个周期内完成 K 次乘积操作和 K-1 次加法操作后得到的值相加，即为向量 C 和向量 D_h 的点积。其中，h 为大于等于 1 小于等于 N 的正整数。

向量 C 与矩阵 B 的乘法运算结果为 1*N 的向量 H，其中，处理元件 PE_h 完成向量 C 和向量 D_h 的点积运算后输出的值即为向量 H 中第 h 个元素的值。

应理解，上述向量 C 可以是向量 X 的一部分，也可以是一个矩阵 E 其中一行的部分或者全部元素；矩阵 B 可以是矩阵 F 的一部分，例如向量 X 是一个 1*Z 的向量，矩阵 F 是一个 Z*N 的向量，其中，Z 大于 W*K。由于芯片 200 中的一个缓存器每次只能存入 W*K 个数据，因此将向量 X 和矩阵 F 进行切分，每次将向量 X 的 W*K 个元素存入芯片 200 的缓存器 B_0 中，将矩阵 F 中的第 h 列的 W*K 个元素存入芯片 200 的缓存器 B_h 中，即，将矩阵 F 中的第 1 列到第 N 列的 W*K 个元素分布存入芯片 200 的缓存器 B_0~B_N 中。

如图 10 所示，图 10 是本申请实施例提供的另一种芯片的示意图。该芯片 300 包括 M*N 个处理元件 PE、M+N 个缓存器、M*（N+1）个复用器以及 M 个调度模块 S_1~S_M。其中，上述芯片 300 所包括的各个部分的连接关系如图 10 中所示，M*N 个处理元件呈矩阵式分布，M*N 个处理元件分布于 M 行，每一行包括 N 个处理元件。图 10 中芯片 300 的结构中每一行相当于图 8 所示的芯片 200 的结构，芯片 300 相当于包括 M 个图 8 所示的芯片 200。

本申请实施例中，该芯片 300 能够实现矩阵与矩阵的乘法运算。以图 10 所示的芯片 300 实现矩阵 A 与矩阵 B 的乘法运算为例。其中，芯片 300 中的每个缓存器均包括 W 行 K 列的数据单元，每个复用器均包括 K 个多路复用器。矩阵 A 为 M*WK 的矩阵，矩阵 B 为 WK*N 的矩阵，即矩阵 A 包括 M 行 W*K 列元素，矩阵 B 包括 W*K 行 N 列元素。应理解，矩阵 A 相当于 M 个 1*WK 的向量，即矩阵 A 相当于 M 个上述向量 C；因此矩阵 A 与矩阵 B 的乘法运算相当于将矩阵 A 的每一行分别与矩阵 B 进行向量与矩阵的乘法运算。因此，可以将矩阵

A 看做是包括向量 C_1~C_M 的 M 个向量，每个向量分别对应矩阵 A 中的一行。

上述图 9 所示的芯片 200 能够完成向量 C 与矩阵 B 的乘法运算，而矩阵 A 相当于 M 个向量 C，芯片 300 相当于 M 个图 8 或图 9 所示的芯片 200 的结构。因此，如图 11 所示，图 11 为本申请实施例提供的一种实现矩阵与矩阵运算的示意图。芯片 300 能够按照芯片 200 计算向量 C 与矩阵 B 的乘法运算的方法，同时计算 M 个向量与矩阵 B 的乘法运算，即，芯片 300 能够计算矩阵 A 所包括的 M 个向量（向量 C_1~C_M）与矩阵 B 的乘法运算，进而得到矩阵 A 与矩阵 B 的乘法运算结果。其中，矩阵 A 为 M*WK 的矩阵，矩阵 B 为 WK*N 的矩阵，矩阵 A 与矩阵 B 的乘法运算结果为 M*N 的矩阵 Q，芯片 300 第 g 行的 N 个处理元件完成向量 C_g 和矩阵 B 的乘法运算后输出的 1*N 的向量即为矩阵 Q 中的第 g 行的元素，其中，g 为大于等于 1 小于等于 M 的正整数。上述芯片 300 中每一行所包括的缓存器、复用器、调度模块以及处理元件实现向量与矩阵乘法的过程可以参照上述图 9 所示的芯片 200 实现向量 C 和矩阵 B 的乘法运算的过程，在此不再赘述。

需要说明的是，芯片 300 中的缓存器 B_{10}、B_{20}、…B_{g0}、…B_{M0} 中的每个缓存器在分别获取到矩阵 A 中的元素后，会分别生成各自缓存的数据对应的位图，并发送给各自连接的调度模块。例如，B_{10} 生成位图 1 发送给调度模块 S_1、B_{20} 生成位图 2 发送给调度模块 S_2，B_{g0} 生成位图 1 发送给调度模块 S_g 等等。

应理解，上述矩阵 A 可以是矩阵 G 的一部分；矩阵 B 可以是矩阵 F 的一部分，例如矩阵 A 是一个 M*Z 的向量，矩阵 F 是一个 Z*N 的向量，其中，Z 大于 W*K。由于芯片 300 中的一个缓存器每次只能存入 W*K 个数据，因此将矩阵 G 和矩阵 F 进行切分，每次将矩阵 G 的第 g 行的 W*K 个元素存入芯片 200 的缓存器 B_{g0} 中，将矩阵 F 中的第 h 列的 W*K 个元素存入芯片 200 的缓存器 B_{1h} 中。

通过上述芯片 300，能够在执行矩阵与矩阵的乘法运算的过程中，使矩阵中元素值为 0 的元素不参与计算，从而能够在降低计算量，提高计算效率的同时，不降低计算的精度。

下面结合上述图 1 至图 11，介绍本申请实施例提供的一种矩阵计算方法，该方法应用于能够实现矩阵计算的芯片中，例如上述图 1 所示的芯片 100 中，其中，芯片 100 的相关结构可以参照上述图 1 至图 5 的相关描述，在此不再赘述。如图 12 所示，图 12 是本申请实施例提供的一种矩阵计算方法的流程示意图，该矩阵计算方法包括以下 S121 至 S123。

S121.芯片缓存第一向量和第二向量。

其中，芯片可以是上述图 1 中的芯片 100。芯片 100 将第一向量缓存于第一缓存器，将第二向量缓存与第二缓存器，芯片 100 将第一向量缓存到第一缓存器中的方法可以参照上述芯片 100 将向量 C 缓存到第一缓存器 120 中的方法，将第二向量缓存到第二缓存器中的方法可以参照上述芯片 100 将向量 D 缓存到第二缓存器 140 中的方法，在此不再赘述。

需要说明的是，第一向量可以是上述图 7 所示的实施例中的向量 C，第二向量可以是上述图 7 所示的实施例中的向量 D。第一向量可以是向量 X 的一部分，第二向量可以是向量 Y 的一部分，例如向量 X 是一个 1*Z 的向量，向量 Y 是一个 Z*1 的向量，其中，Z 大于 W*K。由于芯片 100 中的一个缓存器每次只能存入 W*K 个数据，因此将向量 X 和向量 Y 进行切分，每次最多将 W*K 个元素存入芯片 100 的缓存器中进行计算。第一向量也可以是第一矩阵中的一行或者第一矩阵中某一行的一部分元素，第二向量也可以是第二矩阵中的一列或者第二矩阵中某一列的一部分元素，例如，上述向量 X 是第一矩阵中的一行，向量 Y 是第二矩阵中

的一列。

S122. 芯片的第一调度模块根据第一向量的位图生成第一选通信号。

其中，第一向量的位图指示第一向量中的非零元素，向量的位图可以参照上述图 6 所对应的相关描述，在此不再赘述。上述第一选通信号能够使第一处理元件从第一缓存器中获取第一向量中的第一组非零元素，使第一处理元件从第二缓存器中获取第二向量中的第二组元素。其中，从第二向量中获取的第二组元素可能包括非零元素，也可能是全零的元素。

本申请实施例中，芯片 100 还包括第一复用器和第二复用器，第一处理元件通过第一复用器从第一缓存器中获取第一向量的第一组非零元素，通过第二复用器从第二缓存器中获取第二向量的第二组非零元素。其中，第一调度模块可以是图 7 中的调度模块 160，第一复用器可以是图 7 中的第一复用器 130，第二复用器可以是图 7 中的第二复用器 150，则第一选通信号是上述图 7 所对应的实施例中的选通信号 DS_1~DS_K，第一调度模块根据第一向量的位图生成第一选通信号的过程可以参照上述图 7 对应的实施例中，调度模块 160 根据向量 C 的位图生成选通信号 DS_1~DS_K 的过程，在此不再赘述。

S123. 芯片的第一处理元件根据第一组非零元素和第二组元素实现第一向量和第二向量之间的运算。

第一处理元件可以是图 7 中的处理元件 110，第一组非零元素为第一复用器的 K 个多路复用器根据接收到的选通信号从第一缓存器中获取的 K 个数据，第二组元素为第二服用器的 K 个多路复用器根据接收到的选通信号从第二缓存器获取的 K 个数据。第一复用器和第二复用器根据接收到的选通信号获取缓存器中的方法可以参照上述第一复用器 130 和第二复用器 150，第一处理元件根据第一组非零元素和第二组元素能够实现第一向量和第二向量之间的运算的方法可以参照上述处理元件 110 实现向量 C 和向量 D 之间的运算的方法，在此不再赘述。

本申请实施例中，上述芯片还可以包括第三缓存器和第二处理元件，第三缓存器用于缓存第三向量，上述第一选通信号还能够使第二处理元件从第三缓存器中获取第三向量中的第三组元素，使第二处理元件从第一缓存器获取所述第一向量中的第一组非零元素。第二处理元件根据第一向量中的第一组非零元素和第三向量中的第三组元素，实现第一向量和第三向量的运算。其中，第三向量属于上述第二矩阵中除第二向量所在的列之外的一列。

在一种可能的实现方式中，上述第一缓存器可以是图 8 中缓存器 B_0，第二缓存器可以是图 8 中的缓存器 B_1，第一复用器可以是图 8 中的复用器 M_0，第二复用器可以是图 8 中的复用器 M_1，第一处理元件可以是图 8 中的处理元件 PE_1。第三缓存器可以是图 8 中的缓存器 B_2~B_N 中的任意一个或者多个，则第二处理元件是与第三缓存器连接的处理元件，第三复用器是与第三缓存器连接的复用器。即，在芯片包括第三环缓存器和第二处理元件时，芯片可以是上述图 8 所示的芯片 200，芯片 200 实现第一向量和第二向量，以及第一向量和第三向量的运算的方法可以参照上述图 8 和图 9 所对应的实施例中的相关描述，在此不再赘述。

本申请实施例中，上述芯片在包括第一缓存器、第二缓存器、第一调度模块、第一处理元件、第三缓存器以及第二处理元件的基础上，还可以包括第四缓存器、第二调度模块以及第三处理元件。第四缓存器用于缓存第四向量，其中，第四向量属于第一矩阵中除上述第一向量所在的行之外的一行。第二调度模块根据第四向量的位图生成第二选通信号，第二选通信号能够使第三处理元件从第四缓存器中获取第四向量中的第四组非零元素；使第三处理元

件从第二缓存器中获取第二向量中的第五组元素，第二处理元件能够根据第四向量中的第四组非零元素和第二向量中的第五组元素，实现第四向量和第二向量的运算。其中，第四向量的位图指示所述第四向量中的非零元素。

在一种可能的实现方式中，上述第一缓存器可以是图 10 中缓存器 B_{10}，第二缓存器可以是图 10 中的缓存器 B_{11}，第一复用器可以是图 10 中的复用器 M_{10}，第二复用器可以是图 10 中的复用器 M_{11}，第一处理元件可以是图 10 中的处理元件 PE_{11}。第三缓存器可以是图 10 中的缓存器 B_{12}~B_{1N} 中的任意一个或者多个，则第二处理元件是与第三缓存器连接的处理元件，第三复用器是与第三缓存器连接的复用器；第二调度模块可以是图 10 中调度模块 S_2~S_M 中的任意一个或多个，则第四缓存器是与第二调度模块的连接的缓存器，第三处理元件是图 10 与第二调度模块在同一行的一个或多个处理元件。即，上述芯片在包括第一缓存器、第二缓存器、第一调度模块、第一处理元件、第三缓存器以及第二处理元件的基础上，还包括第四缓存器、第二调度模块以及第三处理元件时，芯片可以是上述图 10 所示的芯片 300，芯片 300 实现第一向量和第二向量、第一向量和第三向量以及第四向量与第二向量的运算的方法可以参照上述图 10 和图 11 所对应的实施例中的相关描述，在此不再赘述。

对于上述方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明并不受所描述的动作顺序的限制，其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作并不一定是本发明所必须的。

本领域的技术人员根据以上描述的内容，能够想到的其他合理的步骤组合，也属于本发明的保护范围内。其次，本领域技术人员也应该熟悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作并不一定是本发明所必须的。

上文中结合图 1 至图 12 详细描述了本申请实施例提供的芯片以及根据芯片进行矩阵计算的方法，下面结合图 13 和图 14，介绍本申请实施例提供的矩阵计算装置与设备。如图 13 所示，图 13 是本申请实施例提供的一种矩阵计算装置的示意图。该矩阵计算装置可以是上述芯片 100、芯片 200 或芯片 300。其中，该矩阵计算装置 131 包括第一调度单元 132 和第一处理单元 133，其中，第一调度单元 132 用于根据第一向量的位图生成第一选通信号，该第一选通信号能够使第一处理单元 133 从第一缓存器中获取第一向量中的第一组非零元素，使第一处理单元 133 从第二缓存器中获取第二向量中的第二组元素，第一处理单元 133 用于根据第一向量中的第一组非零元素和第二向量中的第二组元素实现第一向量和第二向量的运算，其中，第一向量的位图指示第一向量中的非零元素。具体的，该矩阵计算装置 131 可以是上述图 7 所示的芯片 100，上述第一调度单元 132 能够用于实现上述调度模块 160 所实现的功能，第一处理单元 133 能够用于实现上述处理元件 110 所实现的功能，在此不再赘述。

在一种可能的实现方式中，上述第一处理单元 133 根据第一组非零元素和第二组元素实现第一向量和第二向量之间的运算之后，还包括：第一调度单元 132 生成擦除信号，该擦除信号指示第一缓存器和第二缓存器擦除当前缓存的数据。

在一种可能的实现方式中，上述第一向量属于第一矩阵中的任意一行的部分或全部元素，第二向量属于第二矩阵中的任意一列的部分或全部元素。

在一种可能的实现方式中，上述矩阵计算装置还包括第二处理单元 134，上述第一选通信号还能够使第二处理单元 134 获取第三向量中的第三组元素，使第二处理单元 134 获取第

一向量中的第一组非零元素；第二处理单元 134 根据第一组非零元素和第三组元素实现第一向量和第三向量的运算；其中，上述第三向量属于上述第二矩阵中除第二向量所在的列之外的一列。具体的，在矩阵计算装置包括第二处理单元 134 时，该矩阵计算装置 131 可以是上述图 8 或图 9 所示的芯片 200，上述第一调度单元 132 能够用于实现上述图 9 中调度模块 210 所实现的功能，第一处理单元 133 能够用于实现上述图 9 中处理元件 PE_1 所实现的功能，第二处理单元 134 用于实现上述图 9 中除处理元件 PE_1 外其他处理元件所实现的功能，在此不再赘述。

在一种可能的实现方式中，上述矩阵计算装置还包括第二调度单元 135 和第三处理单元 136，第二调度单元 135 根据第四向量的位图生成第二选通信号，第二选通信号用于使第三处理元件 136 获取第四向量中的第四组非零元素；使第三处理元件 136 获取第二向量中的第五组元素，其中，第四向量的位图指示第四向量中的非零元素；第三处理单元 136 用于根据第四组非零元素和第五组元素实现第四向量和第二向量的运算。

具体的，在矩阵计算装置还包括第二处理单元 134、第二调度单元 135 和第三处理单元 136 时，该矩阵计算装置 131 可以是上述图 10 或图 11 所示的芯片 300，上述第一调度单元 132 能够用于实现上述图 11 中调度模块 S_1 所实现的功能，第一处理单元 133 能够用于实现上述图 11 中处理元件 PE_{11} 所实现的功能，第二处理单元 134 用于实现上述图 11 中第一行除处理元件 PE_{11} 外其他处理元件所实现的功能，第一调度单元 132 能够用于实现上述图 11 中除调度模块 S_1 之外其他调度模块所实现的功能，在此不再赘述。

图 14 是本申请实施例提供的一种计算设备的示意图，该计算设备 141 包括上述芯片 100、芯片 200 或芯片 300。计算设备 141 还包括处理器 142、芯片 143、存储器 144 和通信接口 145，其中，处理器 142、存储器 144 和通信接口 145 通过总线 146 进行通信。

芯片 143 可以是上述芯片 100、芯片 200 或芯片 300 中的任意一种，能够协助计算设备 141 实现上述芯片 100、芯片 200 或芯片 300 实现的各种功能。

芯片 143 能够在处理器 142 调度下实现上述图 7 至图 12 所对应的实施例中的操作。处理器 142 可以有多种具体实现形式，例如处理器 142 可以为中央处理器（central processing unit，CPU）、图像处理器（graphics processing unit，GPU）、嵌入式神经网络处理器（neural-network processing units，NPU）或张量处理器（tensor processing unit，TPU），处理器 142 还可以是单核处理器或多核处理器。处理器 142 可以由 CPU 和硬件芯片的组合。上述硬件芯片可以是 ASIC，可编程逻辑器件（programmable logic device，PLD）或其组合。上述 PLD 可以是复杂可编程逻辑器件（complex programmable logic device，CPLD），现场可编程逻辑门阵列（field-programmable gate array，FPGA），通用阵列逻辑（generic array logic，GAL）或其任意组合。处理器 142 也可以单独采用内置处理逻辑的逻辑器件来实现，例如 FPGA 或数字信号处理器（digital signal processor，DSP）等。

存储器 144 可以是非易失性存储器，例如，只读存储器（read-only memory，ROM）、可编程只读存储器（programmable ROM，PROM）、可擦除可编程只读存储器（erasable PROM，EPROM）、电可擦除可编程只读存储器（electrically EPROM，EEPROM）或闪存。存储器 144 也可以是易失性存储器，易失性存储器可以是随机存取存储器（random access memory，RAM），其用作外部高速缓存。通过示例性但不是限制性说明，许多形式的 RAM 可用，例如静态随机存取存储器（static RAM，SRAM）、动态随机存取存储器（dynamic RAM，DRAM）、

同步动态随机存取存储器（synchronous DRAM，SDRAM）、双倍数据速率同步动态随机存取存储器（double data rate SDRAM，DDR SDRAM）、增强型同步动态随机存取存储器（enhanced SDRAM，ESDRAM）、同步连接动态随机存取存储器（synchlink DRAM，SLDRAM）和直接内存总线随机存取存储器（direct rambus RAM，DR RAM）。

存储器 144 可用于存储程序代码和数据，例如缓存上述向量或者矩阵，以便于芯片 143 调用存储器 144 中存储的程序代码执行图 1 至图 12 所对应的实施例中的操作步骤。

通信接口 145 为有线接口（例如以太网接口），为内部接口（例如高速串行计算机扩展总线（Peripheral Component Interconnect express, PCIE）总线接口）、有线接口（例如以太网接口）或无线接口（例如蜂窝网络接口或使用无线局域网接口），用于与其他计算设备或模块进行通信。

总线 146 是快捷外围部件互联标准（Peripheral Component Interconnect Express，PCIE）总线，或扩展工业标准结构（extended industry standard architecture，EISA）总线、统一总线（unified bus，Ubus 或 UB）、计算机快速链接（compute express link，CXL）、缓存一致互联协议（cache coherent interconnect for accelerators，CCIX）等。总线 146 包括带外总线和高速总线等，为了清楚说明起见，在图中将各种总线都标为总线 146。

需要说明的，图 14 仅仅是本申请实施例的一种可能的实现方式，实际应用中，计算设备 141 还包括更多或更少的部件，这里不作限制。

本申请实施例提供一种计算机读存储介质，包括：该计算机读存储介质中存储有计算机指令；当该计算机指令在计算设备上运行时，使得该计算设备执行上述图 1 至图 12 所对应的实施例中的操作。

本申请实施例提供了一种包含指令的计算机程序产品，包括计算机程序或指令，当该计算机程序或指令在计算机上运行时，使得该计算设备执行上述图 1 至图 12 所对应的实施例中的操作。

上述实施例，全部或部分地通过软件、硬件、固件或其他任意组合来实现。当使用软件实现时，上述实施例全部或部分地以计算机程序产品的形式实现。计算机程序产品包括至少一个计算机指令。在计算机上加载或执行计算机程序指令时，全部或部分地产生按照本发明实施例的流程或功能。计算机为通用计算机、专用计算机、计算机网络、或者其他编程装置。计算机指令存储在计算机读存储介质中，或者从一个计算机读存储介质向另一个计算机读存储介质传输，例如，计算机指令从一个网站站点、计算机、服务器或数据中心通过有线（例如同轴电缆、光纤、数字用户线（digital subscriber line, DSL））或无线（例如红外、无线、微波等）方式向另一个网站站点、计算机、服务器或数据中心进行传输。计算机读存储介质是计算机能够存取的任何用介质或者是包含至少一个用介质集合的服务器、数据中心等数据存储节点。用介质是磁性介质（例如，软盘、硬盘、磁带）、光介质（例如，高密度数字视频光盘（digital video disc, DVD）、或者半导体介质。

以上，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，轻易想到各种等效的修复或替换，这些修复或替换都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以权利要求的保护范围为准。

权 利 要 求 书

1.一种芯片，其特征在于，包括：

第一缓存器，用于缓存第一向量；

第二缓存器，用于缓存第二向量；

第一调度模块，用于根据第一向量的位图生成第一选通信号，所述第一选通信号使第一处理元件从所述第一缓存器中获取所述第一向量中的第一组非零元素，使所述第一处理元件从所述第二缓存器中获取所述第二向量中的第二组元素，所述第一向量的位图指示所述第一向量中的非零元素；

所述第一处理元件，用于根据所述第一组非零元素和所述第二组元素实现所述第一向量和所述第二向量之间的运算。

2.根据权利要求 1 所述的芯片，其特征在于，所述芯片还包括第一复用器和第二复用器；

所述第一复用器，用于根据所述第一选通信号从所述第一缓存器中获取所述第一向量中的所述第一组非零元素，并输入所述第一处理元件；

所述第二复用器，用于根据所述第一选通信号从所述第二缓存器中获取所述第二向量中的所述第二组元素，并输入所述第一处理元件。

3.根据权利要求 2 所述的芯片，其特征在于，所述第一复用器包括 K 个多路复用器，所述第二复用器包括 K 个多路复用器；所述第一缓存器和所述第二缓存器均包括 W 行 K 列数据单元，每个数据单元用于缓存一个元素；

所述第一复用器中的每个多路复用器与第一缓存器中的多个数据单元连接，且每个数据单元至少与一个多路复用器连接；所述第二复用器中的第 i 个多路复用器与所述第二缓存器中数据单元的连接关系，和所述第一复用器中的第 i 个多路复用器与所述第一缓存器中数据单元的连接关系相同。

4.根据权利要求 3 所述的芯片，其特征在于，所述第一调度模块具体用于：

根据所述第一向量的位图确定所述第一复用器的第 j 个多路复用器连接的数据单元中，第 k 个数据单元存储的元素为非零元素，所述第一调度模块生成所述第 j 个多路复用器的选通信号，将所述第 j 个多路复用器的选通信号发送给所述第一复用器的第 j 个多路复用器和所述第二复用器的第 j 个多路复用器，所述第一选通信号包括所述第一复用器的第 j 个多路复用器的选通信号。

5.根据权利要求 4 所述的芯片，其特征在于，

所述第一复用器具体用于：根据所述第一复用器的第 j 个多路复用器的选通信号，通过所述第一复用器的第 j 个多路复用器，获取所述第一复用器的第 j 个多路复用器连接的数据单元中第 k 个数据单元中的第一元素，并将所述第一元素输入到所述第一处理元件，所述第一元素是所述第一组非零元素中的一个；

所述第二复用器具体用于：根据所述第一复用器的第 j 个多路复用器的选通信号，通过所述第二复用器的第 j 个多路复用器，获取所述第二复用器的第 j 个多路复用器连接的数据单元中第 k 个数据单元中的第二元素，并将所述第二元素输入到所述第一处理元件，所述第二元素是所述第二组元素中的一个。

6.根据权利要求 1-5 任一项所述的芯片，其特征在于，所述第一调度模块还用于在实现所述第一向量和所述第二向量之间的运算后，生成擦除信号，所述擦除信号用于指示所述第一缓存器和所述第二缓存器擦除当前缓存的数据。

7.根据权利要求 4 或 5 所述的芯片，其特征在于，所述第一向量属于第一矩阵中的任意一行，所述第二向量属于第二矩阵中的任意一列。

8.根据权利要求 7 所述的芯片，其特征在于，所述芯片还包括第三缓存器和第二处理元件；其中，

所述第三缓存器，用于缓存第三向量，所述第三向量属于所述第二矩阵中除所述第二向量所在的列之外的一列；所述第一选通信号还用于使所述第二处理元件从所述第三缓存器中获取所述第三向量中的第三组元素；

所述第二处理元件，用于根据所述第一组非零元素和所述第三组元素实现所述第一向量和所述第三向量之间的运算。

9.根据权利要求 8 所述的芯片，其特征在于，所述芯片还包括第三复用器，用于根据所述第一选通信号从第三缓存器中获取所述第三向量中的所述第三组元素，并输入所述第二处理元件。

10.根据权利要求 9 所述的芯片，其特征在于，所述第三复用器包括 K 个多路复用器，所述第三缓存器包括 W 行 K 列数据单元，每个数据单元用于缓存一个元素；所述第三复用器中的第 i 个多路复用器与所述第三缓存器中数据单元的连接关系，和所述第一复用器中的第 i 个多路复用器与所述第一缓存器中数据单元的连接关系相同。

11.根据权利要求 10 所述的芯片，其特征在于，所述第三复用器具体用于：根据所述第一复用器的第 j 个多路复用器的选通信号，通过所述第三复用器的第 j 个多路复用器，获取所述第三复用器的第 j 个多路复用器连接的数据单元中第 k 个数据单元中的第三元素，并将所述第三元素输入到所述第二处理元件，所述第三元素是所述第三组元素中的一个。

12.根据权利要求 7-11 任一项所述的芯片，其特征在于，所述芯片还包括第四缓存器、第二调度模块和第三处理元件；其中，

所述第四缓存器，用于缓存第四向量，所述第四向量属于所述第一矩阵中除所述第一向量所在的行之外的一行；

所述第二调度模块，用于根据所述第四向量的位图生成第二选通信号，所述第二选通信号用于使所述第三处理元件从所述第四缓存器中获取所述第四向量中的第四组非零元素；使

所述第三处理元件从所述第二缓存器中获取所述第二向量中的第五组元素，所述第四向量的位图指示所述第四向量中的非零元素；

所述第三处理元件，用于根据所述第四组非零元素和所述第五组元素实现所述第四向量和所述第二向量之间的运算。

13.根据权利要求 12 所述的芯片，其特征在于，所述芯片还包括第四复用器和第五复用器，

所述第四复用器，用于根据所述第二选通信号从所述第四缓存器中获取所述第四向量中的所述第四组非零元素，并输入所述第三处理元件；

所述第五复用器，用于根据所述第二选通信号从所述第二缓存器中获取所述第二向量中的所述第五组元素，并输入所述第三处理元件。

14.根据权利要求 13 所述的芯片，其特征在于，所述第四复用器包括 K 个多路复用器，所述第五复用器包括 K 个多路复用器；所述第四缓存器包括 W 行 K 列数据单元，每个数据单元用于缓存一个元素；

所述第四复用器中的每个多路复用器与所述第四缓存器中的多个数据单元连接，且每个数据单元至少与一个多路复用器连接；所述第四复用器中的第 i 个多路复用器与所述第四缓存器中数据单元的连接关系，和所述第一复用器中的第 i 个多路复用器与所述第一缓存器中数据单元的连接关系相同；所述第五复用器中的第 i 个多路复用器与所述第二缓存器中数据单元的连接关系，和所述第一复用器中的第 i 个多路复用器与所述第一缓存器中数据单元的连接关系相同。

15.根据权利要求 14 所述的芯片，其特征在于，所述第二调度模块具体用于：

根据所述第四向量的位图确定所述第四复用器的第 j 个多路复用器连接的数据单元中，第 m 个数据单元存储的元素为非零元素，所述第二调度模块生成所述第四复用器的第 j 个多路复用器的选通信号，将所述第四复用器的第 j 个多路复用器的选通信号发送给所述第四复用器的第 j 个多路复用器和所述第五复用器的第 j 个多路复用器，所述第二选通信号包括所述第四复用器的第 j 个多路复用器的选通信号。

16.根据权利要求 15 所述的芯片，其特征在于，

所述第四复用器具体用于：根据所述第四复用器的第 j 个多路复用器的选通信号，通过所述第四复用器的第 j 个多路复用器，获取所述第四复用器的第 j 个多路复用器连接的数据单元中第 m 个数据单元中的第四元素，并将所述第四元素输入到所述第三处理元件，所述第四元素是所述第四组非零元素中的一个；

所述第五复用器具体用于：根据所述第四复用器的第 j 个多路复用器的选通信号，通过所述第五复用器的第 j 个多路复用器，获取所述第五复用器的第 j 个多路复用器连接的数据单元中第 m 个数据单元中的第五元素，并将所述第五元素输入到所述第三处理元件，所述第五元素是所述第五组元素中的一个。

17.一种矩阵计算方法，其特征在于，应用于芯片，包括：

所述芯片缓存第一向量和第二向量，其中，所述第一向量缓存于第一缓存器，所述第二向量缓存于第二缓存器；

所述芯片的第一调度模块根据第一向量的位图生成第一选通信号，所述第一选通信号使第一处理元件从所述第一缓存器中获取所述第一向量中的第一组非零元素，使所述第一处理元件从所述第二缓存器中获取所述第二向量中的第二组元素，所述第一向量的位图指示所述第一向量中的非零元素；

所述芯片的所述第一处理元件根据所述第一组非零元素和所述第二组元素实现所述第一向量和所述第二向量之间的运算。

18.根据权利要求 17 所述的方法，其特征在于，所述芯片还包括第一复用器和第二复用器；所述方法还包括：

所述芯片的所述第一复用器根据所述第一选通信号从所述第一缓存器中获取所述第一向量中的所述第一组非零元素，并输入所述第一处理元件；

所述芯片的所述第二复用器根据所述第一选通信号从所述第二缓存器中获取所述第二向量中的所述第二组元素，并输入所述第一处理元件。

19.根据权利要求 18 所述的方法，其特征在于，所述第一复用器包括 K 个多路复用器，所述第二复用器包括 K 个多路复用器；所述第一缓存器和所述第二缓存器均包括 W 行 K 列数据单元，每个数据单元用于缓存一个元素；

所述第一复用器中的每个多路复用器与第一缓存器中的多个数据单元连接，且每个数据单元至少与一个多路复用器连接；所述第二复用器中的第 i 个多路复用器与所述第二缓存器中数据单元的连接关系，和所述第一复用器中的第 i 个多路复用器与所述第一缓存器中数据单元的连接关系相同。

20.根据权利要求 19 所述的方法，其特征在于，所述芯片的第一调度模块根据第一向量的位图生成第一选通信号，包括：

所述芯片的第一调度模块根据所述第一向量的位图确定所述第一复用器的第 j 个多路复用器连接的数据单元中，第 k 个数据单元存储的元素为非零元素，生成所述第 j 个多路复用器的选通信号，将所述第 j 个多路复用器的选通信号发送给所述第一复用器的第 j 个多路复用器和所述第二复用器的第 j 个多路复用器，所述第一选通信号包括所述第一复用器的第 j 个多路复用器的选通信号。

21. 根据权利要求 20 所述的方法，其特征在于，所述芯片的所述第一复用器根据所述第一选通信号从所述第一缓存器中获取所述第一向量中的所述第一组非零元素，并输入所述第一处理元件，所述芯片的所述第二复用器根据所述第一选通信号从所述第二缓存器中获取所述第二向量中的所述第二组元素，并输入所述第一处理元件，包括：

所述芯片的所述第一复用器根据所述第一复用器的第 j 个多路复用器的选通信号，通过所述第一复用器的第 j 个多路复用器，获取所述第一复用器的第 j 个多路复用器连接的数据单元中第 k 个数据单元中的第一元素，并将所述第一元素输入到所述第一处理元件，所述第一元素是所述第一组非零元素中的一个；

所述芯片的所述第二复用器根据所述第一复用器的第 j 个多路复用器的选通信号，通过所述第二复用器的第 j 个多路复用器，获取所述第二复用器的第 j 个多路复用器连接的数据单元中第 k 个数据单元中的第二元素，并将所述第二元素输入到所述第一处理元件，所述第二元素是所述第二组元素中的一个。

22.根据权利要求 17-21 任一项所述的方法，其特征在于，所述芯片的所述第一处理元件根据所述第一组非零元素和所述第二组元素实现所述第一向量和所述第二向量之间的运算之后，还包括：

所述芯片的第一调度模块生成擦除信号，所述擦除信号指示所述第一缓存器和所述第二缓存器擦除当前缓存的数据。

23.根据权利要求 20 或 21 所述的方法，其特征在于，所述第一向量属于第一矩阵中的任意一行，所述第二向量属于第二矩阵中的任意一列。

24.根据权利要求 23 所述的方法，其特征在于，所述芯片还包括第三缓存器和第二处理元件；所述方法还包括：

所述芯片缓存第三向量，所述第三向量缓存于所述第三缓存器，所述第三向量属于所述第二矩阵中除所述第二向量所在的列之外的一列，所述第一选通信号还用于使所述第二处理元件从所述第三缓存器中获取所述第三向量中的第三组元素；

所述芯片的所述第二处理元件根据所述第一组非零元素和所述第三组元素实现所述第一向量和所述第三向量之间的运算。

25.根据权利要求 24 所述的方法，其特征在于，所述芯片还包括第三复用器，所述方法还包括：

所述芯片的所述第三复用器根据所述第一选通信号从第三缓存器中获取所述第三向量中的所述第三组元素，并输入所述第二处理元件。

26.根据权利要求 25 所述的方法，其特征在于，所述第三复用器包括 K 个多路复用器，所述第三缓存器包括 W 行 K 列数据单元，每个数据单元用于缓存一个元素；所述第三复用器中的第 i 个多路复用器与所述第三缓存器中数据单元的连接关系，和所述第一复用器中的第 i 个多路复用器与所述第一缓存器中数据单元的连接关系相同。

27.根据权利要求 26 所述的方法，其特征在于，所述芯片的所述第三复用器根据所述第一选通信号从第三缓存器中获取所述第三向量中的所述第三组元素，并输入所述第二处理元件，包括：

所述芯片的所述第三复用器根据所述第一复用器的第 j 个多路复用器的选通信号，通过所述第三复用器的第 j 个多路复用器，获取所述第三复用器的第 j 个多路复用器连接的数据单元中第 k 个数据单元中的第三元素，并将所述第三元素输入到所述第二处理元件，所述第三元素是所述第三组元素中的一个。

28. 根据权利要求 23-27 任一项所述的方法，其特征在于，所述芯片还包括第四缓存器、第二调度模块和第三处理元件；所述方法还包括：

所述芯片缓存第四向量，所述第四向量缓存于所述第四缓存器，所述第四向量属于所述第一矩阵中除所述第一向量所在的行之外的一行；

所述芯片的所述第二调度模块根据所述第四向量的位图生成第二选通信号，所述第二选通信号用于使所述第三处理元件从所述第四缓存器中获取所述第四向量中的第四组非零元素；使所述第三处理元件从所述第二缓存器中获取所述第二向量中的第五组元素，所述第四向量的位图指示所述第四向量中的非零元素；

所述芯片的所述第三处理元件根据所述第四组非零元素和所述第五组元素实现所述第四向量和所述第二向量的运算。

29.根据权利要求 28 所述的方法，其特征在于，所述芯片还包括第四复用器和第五复用器，所述方法还包括：

所述芯片的所述第四复用器根据所述第二选通信号从所述第四缓存器中获取所述第四向量中的所述第四组非零元素，并输入所述第三处理元件；

所述芯片的所述第五复用器根据所述第二选通信号从所述第二缓存器中获取所述第二向量中的所述第五组元素，并输入所述第三处理元件。

30.根据权利要求 29 所述的方法，其特征在于，所述第四复用器包括 K 个多路复用器，所述第五复用器包括 K 个多路复用器；所述第四缓存器包括 W 行 K 列数据单元，每个数据单元用于缓存一个元素；

所述第四复用器中的每个多路复用器与所述第四缓存器中的多个数据单元连接，且每个数据单元至少与一个多路复用器连接；所述第四复用器中的第 i 个多路复用器与所述第四缓存器中数据单元的连接关系，和所述第一复用器中的第 i 个多路复用器与所述第一缓存器中数据单元的连接关系相同；所述第五复用器中的第 i 个多路复用器与所述第五缓存器中数据单元的连接关系，和所述第一复用器中的第 i 个多路复用器与所述第一缓存器中数据单元的连接关系相同。

31.根据权利要求 30 所述的方法，其特征在于，所述芯片的所述第二调度模块根据所述第四向量的位图生成第二选通信号，包括：

所述芯片的所述第二调度模块根据所述第四向量的位图确定所述第四复用器的第 j 个多路复用器连接的数据单元中，第 m 个数据单元存储的元素为非零元素，生成所述第四复用器的第 j 个多路复用器的选通信号，将所述第四复用器的第 j 个多路复用器的选通信号发送给所述第四复用器的第 j 个多路复用器和所述第五复用器的第 j 个多路复用器，所述第二选通信号包括所述第四复用器的第 j 个多路复用器的选通信号。

32.根据权利要求 31 所述的方法，其特征在于，所述芯片的所述第四复用器根据所述第二选通信号从所述第四缓存器中获取所述第四向量中的所述第四组非零元素，并输入所述第三处理元件；所述芯片的所述第五复用器根据所述第二选通信号从所述第二缓存器中获取所述第二向量中的所述第五组元素，并输入所述第三处理元件，包括：

所述芯片的所述第四复用器根据所述第四复用器的第 j 个多路复用器的选通信号，通过所述第四复用器的第 j 个多路复用器，获取所述第四复用器的第 j 个多路复用器连接的数据单元中第 m 个数据单元中的第四元素，并将所述第四元素输入到所述第三处理元件，所述第四元素是所述第四组非零元素中的一个；

所述芯片的所述第五复用器根据所述第四复用器的第 j 个多路复用器的选通信号，通过所述第五复用器的第 j 个多路复用器，获取所述第五复用器的第 j 个多路复用器连接的数据单元中第 m 个数据单元中的第五元素，并将所述第五元素输入到所述第三处理元件，所述第五元素是所述第五组元素中的一个。

33.一种矩阵计算装置，其特征在于，所述矩阵计算装置包括第一调度单元和第一处理单元，其中，

所述第一调度单元，用于根据第一向量的位图生成第一选通信号，所述第一选通信号使第一处理单元获取所述第一向量中的第一组非零元素，使所述第一处理单元获取第二向量中的第二组元素，所述第一向量的位图指示所述第一向量中的非零元素；

所述第一处理单元，用于根据所述第一组非零元素和所述第二组元素实现所述第一向量和所述第二向量之间的运算。

34.根据权利要求 33 所述的装置，其特征在于，所述第一调度单元，还用于在所述第一处理单元完成第一向量和第二向量之间的运算之后，生成擦除信号，该擦除信号指示第一缓存器和第二缓存器擦除当前缓存的数据。

35.根据权利要求 33 或 34 所述的装置，其特征在于，所述第一向量属于第一矩阵中的任意一行，所述第二向量属于第二矩阵中的任意一列。

36.根据权利要求 35 所述的装置，其特征在于，所述矩阵计算装置还包括第二处理单元，所述第一选通信号还用于使所述第二处理单元获取所述第三向量中的第三组元素，所述第三向量属于所述第二矩阵中除所述第二向量所在的列之外的一列，

所述第二处理单元，用于根据所述第一组非零元素和所述第三组元素实现所述第一向量和所述第三向量之间的运算。

37.根据权利要求 35 或 36 所述的装置，其特征在于，所述矩阵计算装置还包括第二调度单元和第三处理单元，

所述第二调度单元，用于根据所述第四向量的位图生成第二选通信号，所述第二选通信号用于使第三处理单元获取所述第四向量中的第四组非零元素；使所述第三处理单元获取所述第二向量中的第五组元素，所述第四向量的位图指示所述第四向量中的非零元素，所述第四向量属于所述第一矩阵中除所述第一向量所在的行之外的一行；

所述第三处理单元，用于根据所述第四组非零元素和所述第五组元素实现所述第四向量和所述第二向量之间的运算。

38.一种计算设备，其特征在于，所述计算设备包括芯片和存储器，所述存储器用于存储代码，所述芯片用于执行所述代码实现如权利要求 17 至 32 任一项所述的方法。

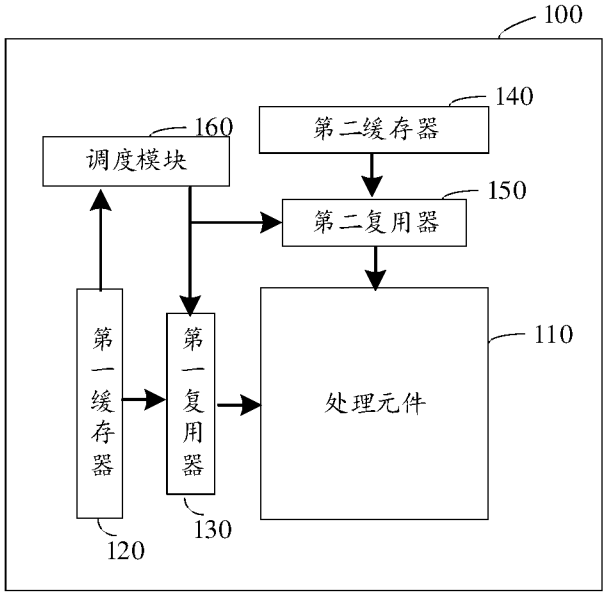

图 1

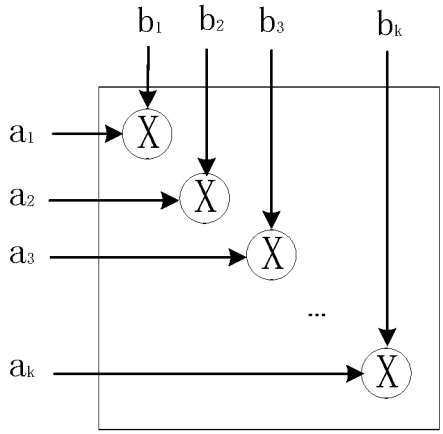

图 2

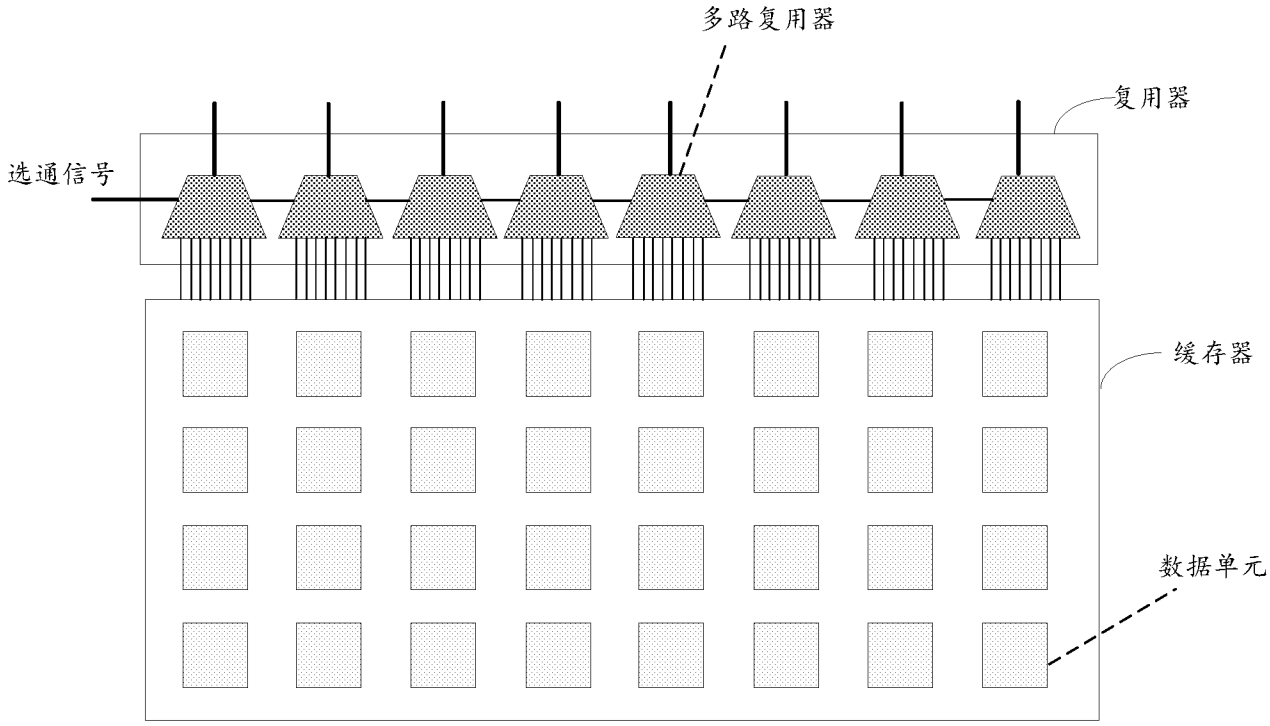

图 3

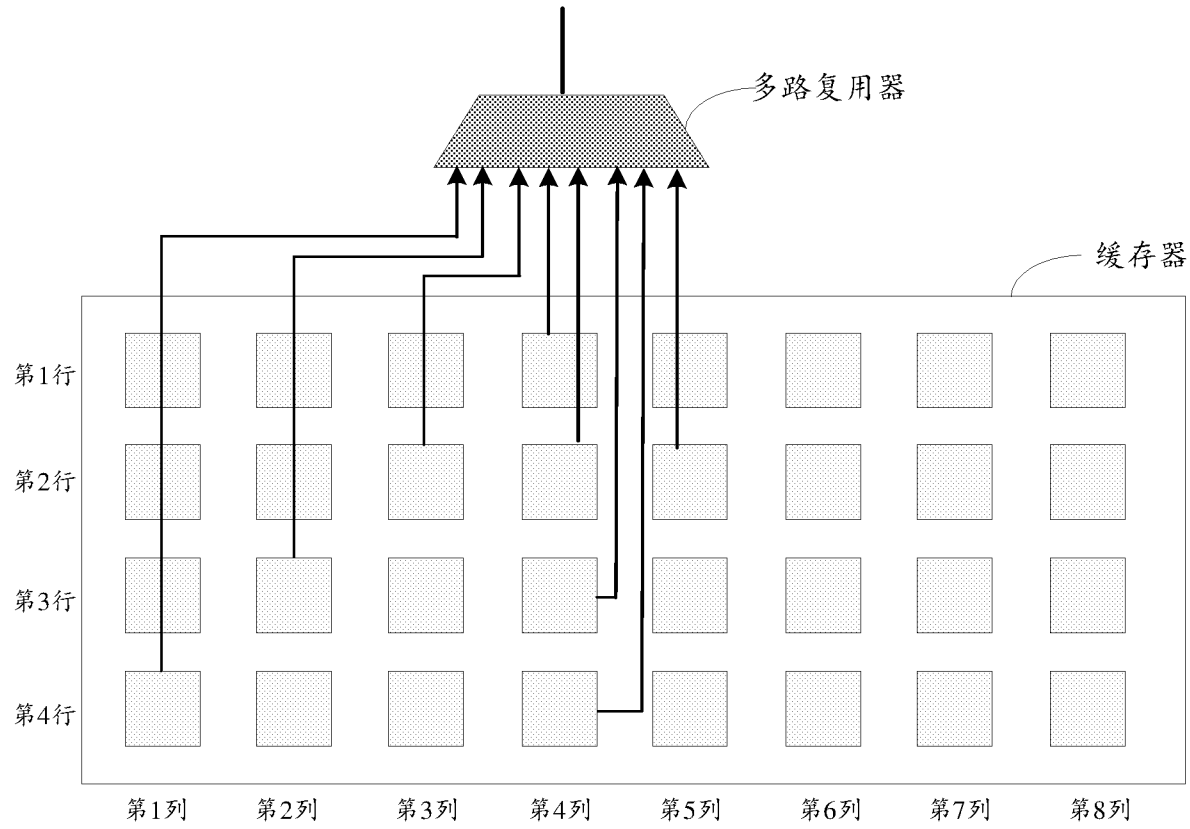
多路复用器
缓存器
第1行
第2行
第3行
第4行
第1列
第2列
第3列
第4列
第5列
第6列
第7列
第8列

图 4

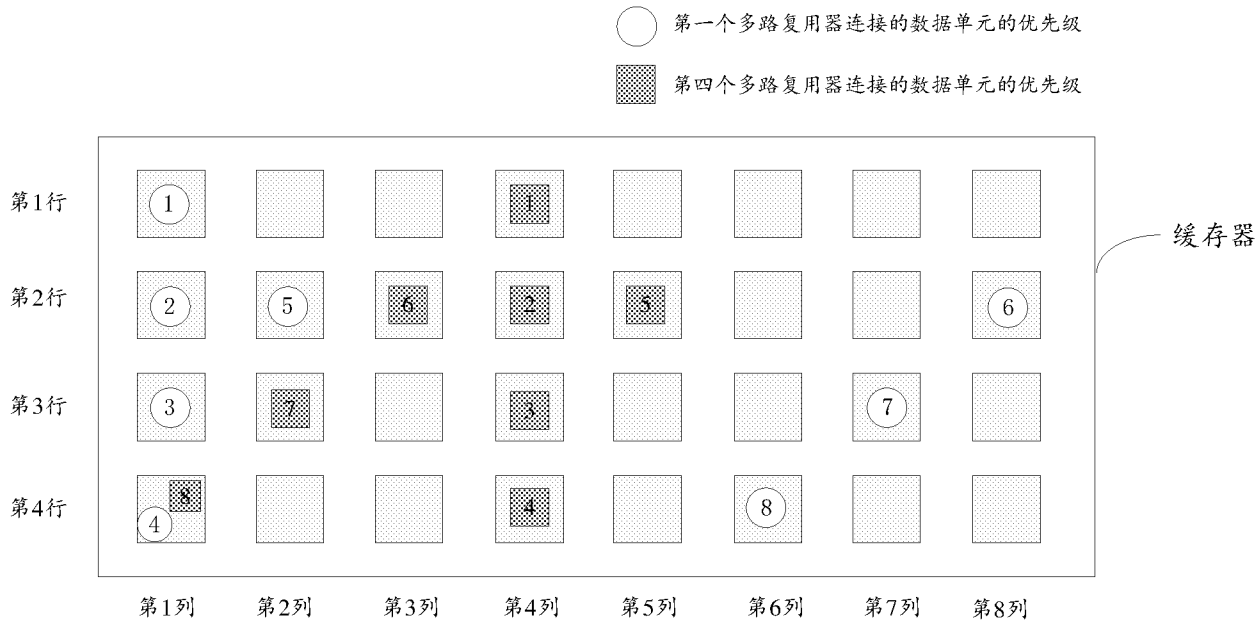
第一个多路复用器连接的数据单元的优先级
第四个多路复用器连接的数据单元的优先级
缓存器
第1行
第2行
第3行
第4行
第1列
第2列
第3列
第4列
第5列
第6列
第7列
第8列

图 5

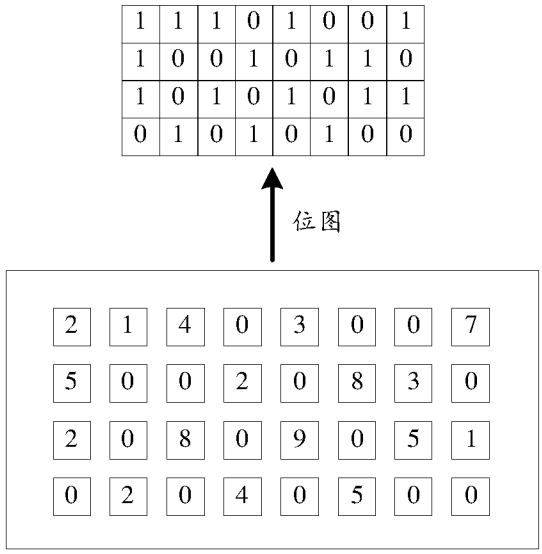
1 1 1 0 1 0 0 1
1 0 0 1 0 1 1 0
1 0 1 0 1 0 1 1
0 1 0 1 0 1 0 0
位图
2 1 4 0 3 0 0 7
5 0 0 2 0 8 3 0
2 0 8 0 9 0 5 1
0 2 0 4 0 5 0 0

图 6

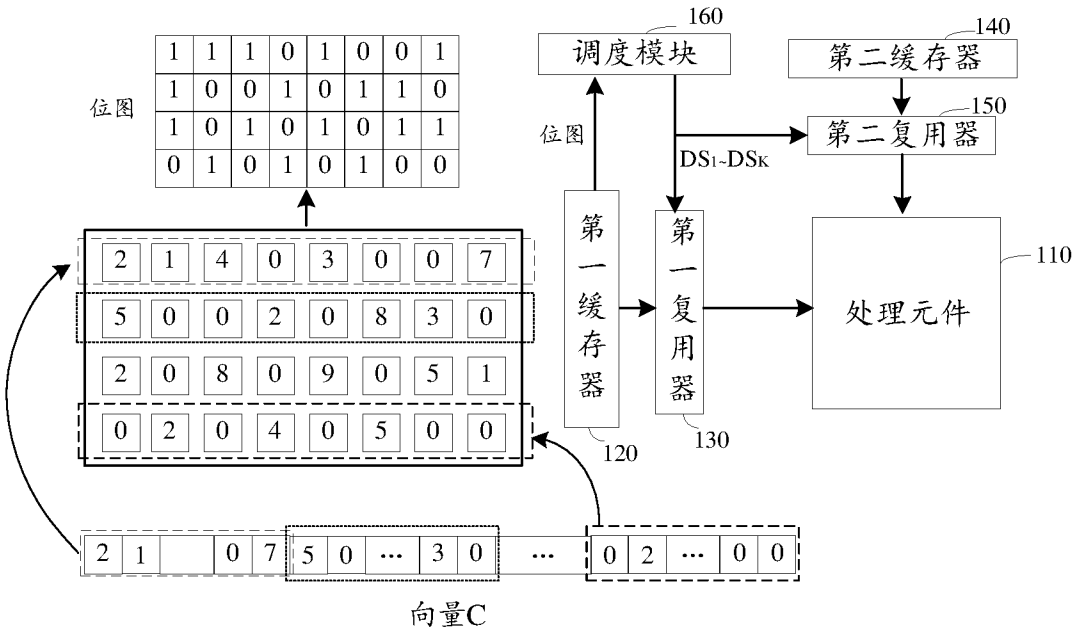
160
调度模块
140
第二缓存器
150
第二复用器
位图
DS1~DSK
第一缓存器
第一复用器
处理元件
110
120
130
位图
2 1 4 0 3 0 0 7
5 0 0 2 0 8 3 0
2 0 8 0 9 0 5 1
0 2 0 4 0 5 0 0
向量C

图 7

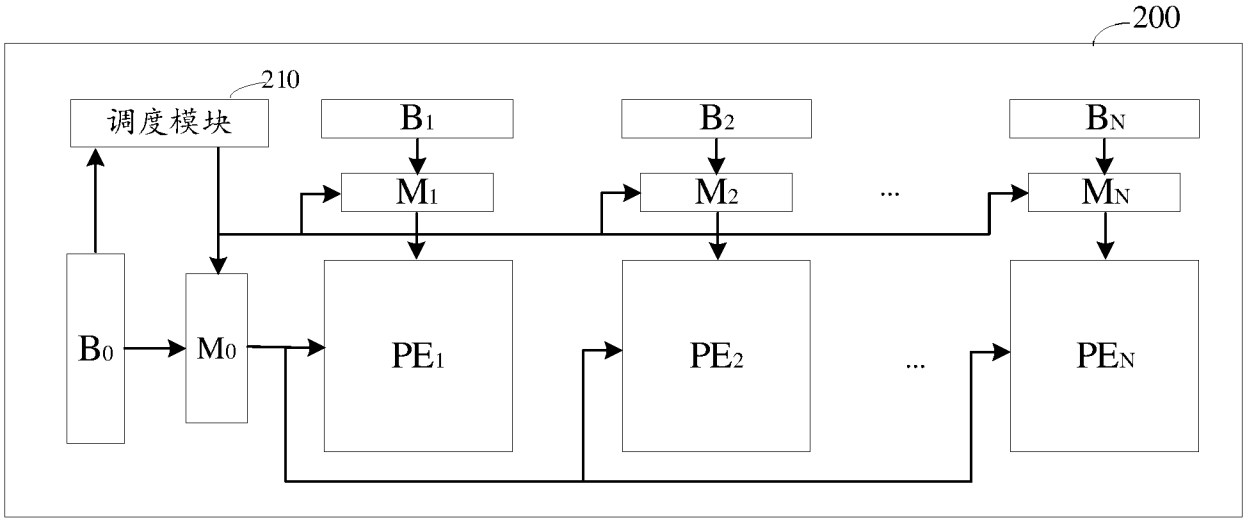
200
210
调度模块
B1
B2
BN
M1
M2
MN
B0
M0
PE1
PE2
PEN

图 8

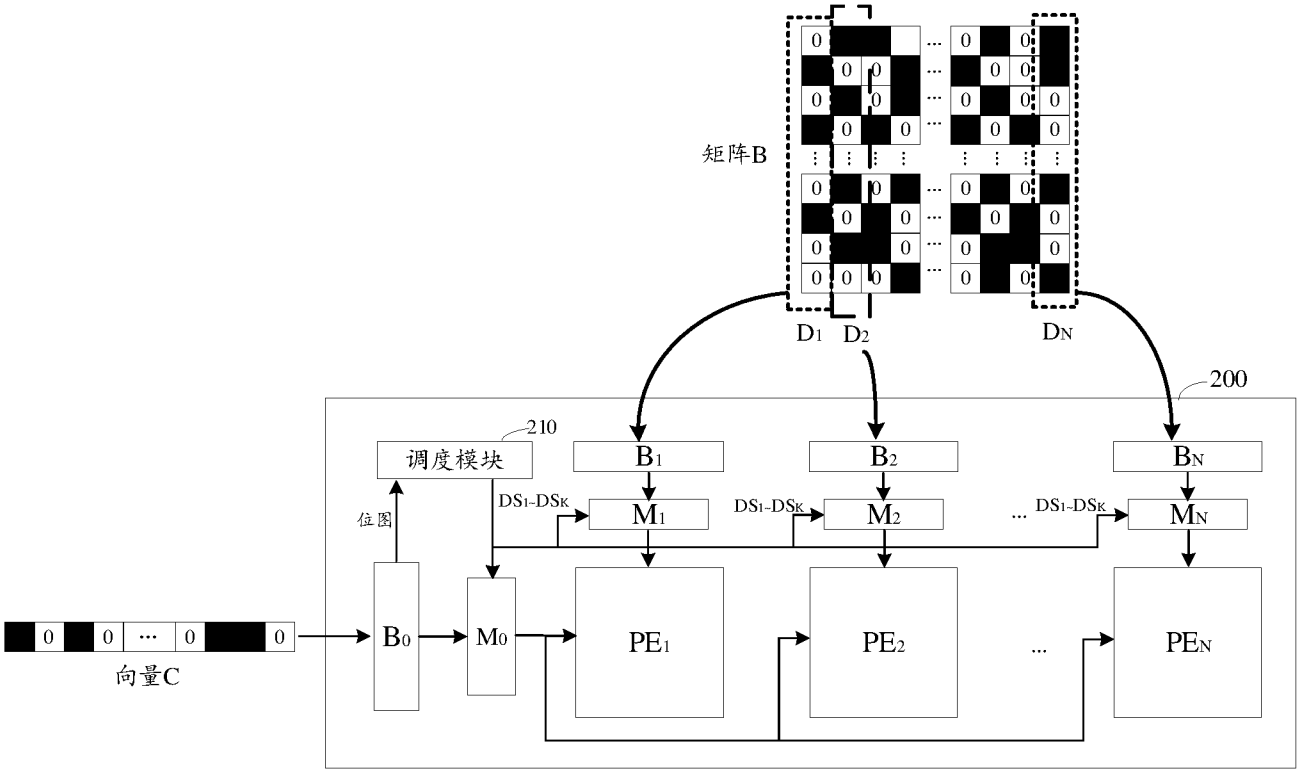
矩阵B
D1
D2
DN
200
210
调度模块
位图
DS1-DSK
B1
B2
BN
M1
M2
MN
B0
M0
PE1
PE2
PEN
向量C

图 9

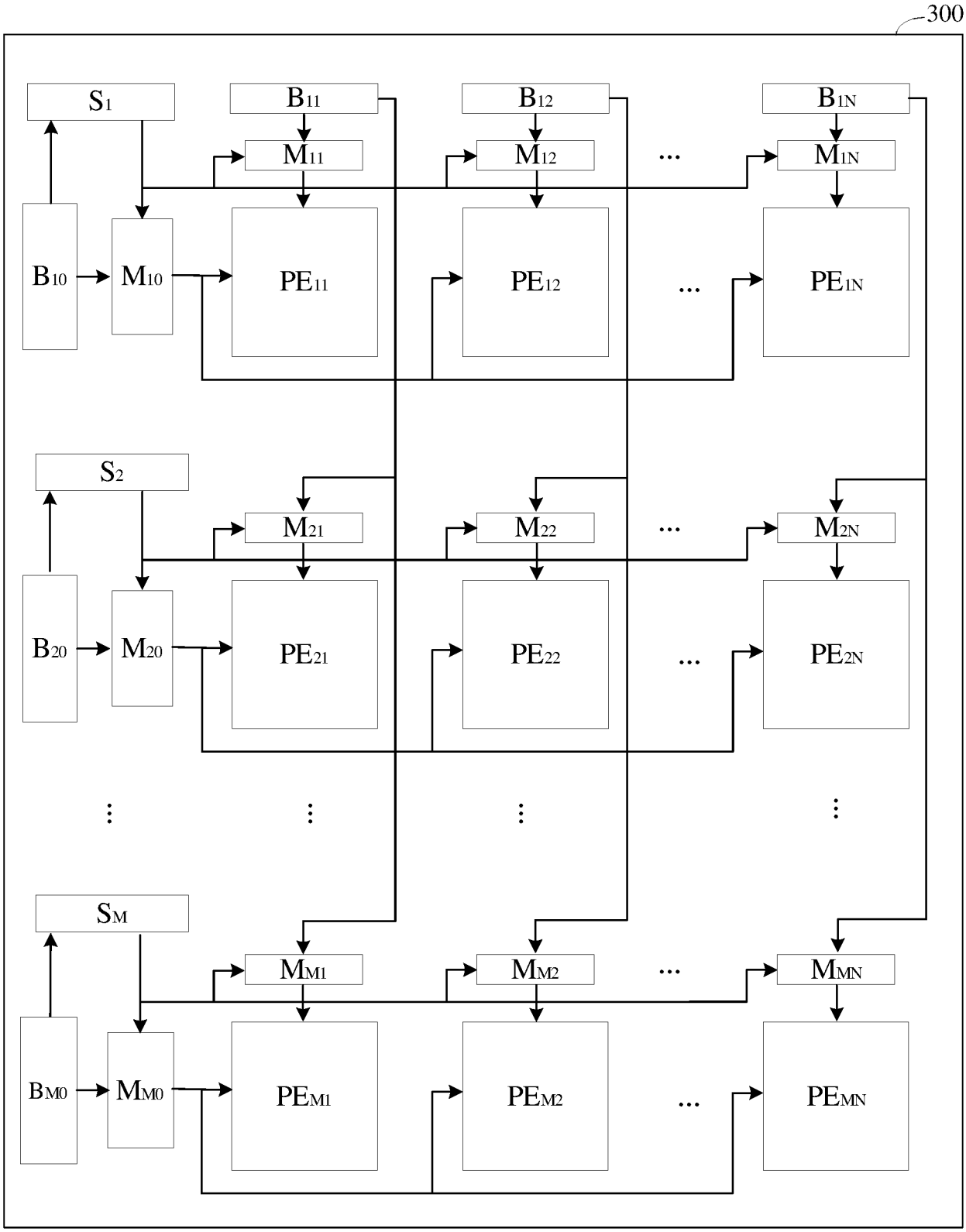
300
S1
B11
B12
B1N
M11
M12
M1N
B10
M10
PE11
PE12
PE1N
S2
M21
M22
M2N
B20
M20
PE21
PE22
PE2N
SM
MM1
MM2
MMN
BM0
MM0
PEM1
PEM2
PEMN

图 10

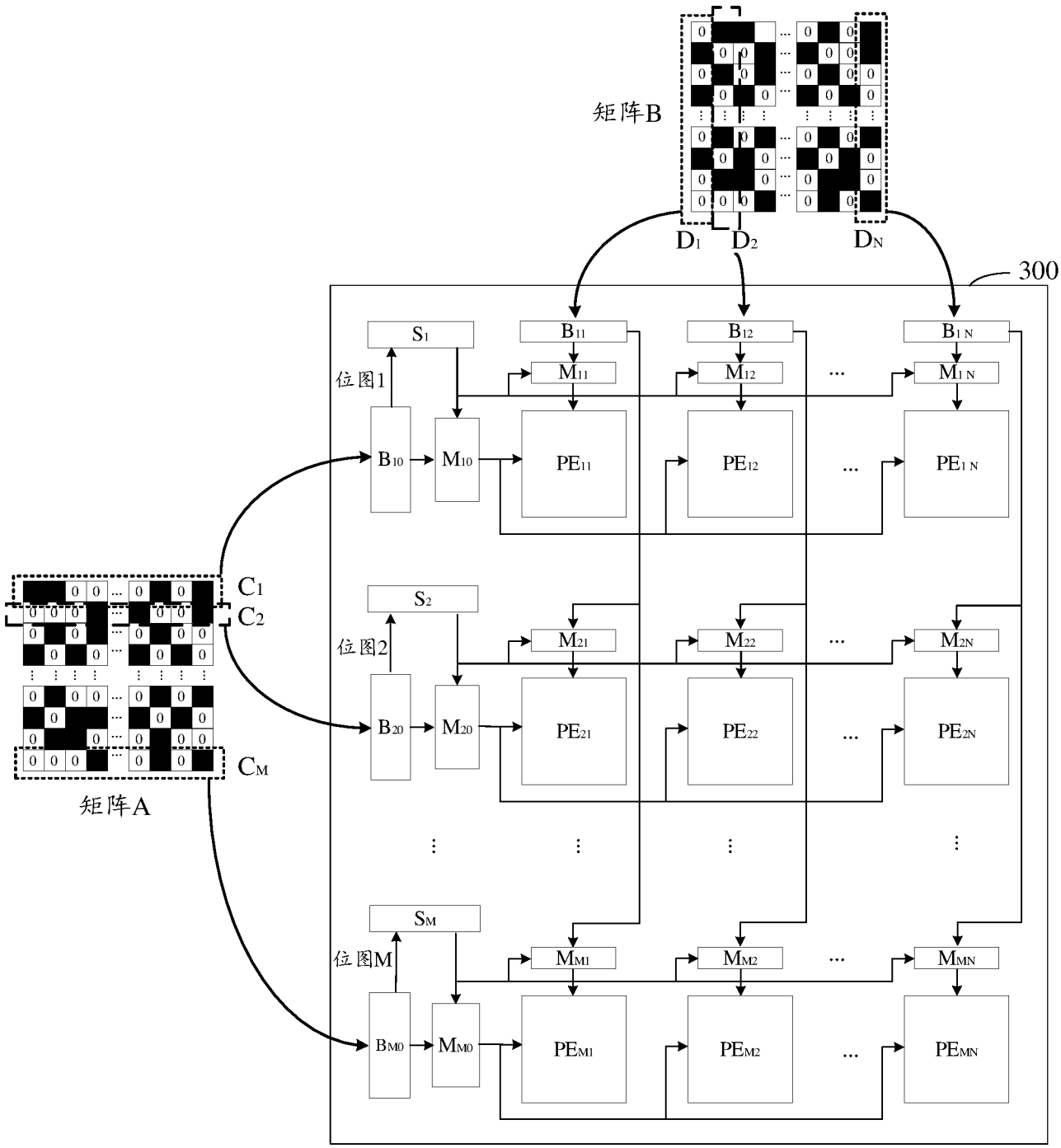
矩阵B
D1
D2
DN
300
S1
B11
B12
B1 N
位图1
M11
M12
M1 N
B10
M10
PE11
PE12
PE1 N
C1
C2
CM
矩阵A
S2
位图2
M21
M22
M2N
B20
M20
PE21
PE22
PE2N
SM
位图M
MM1
MM2
MMN
BM0
MM0
PEM1
PEM2
PEMN

图 11

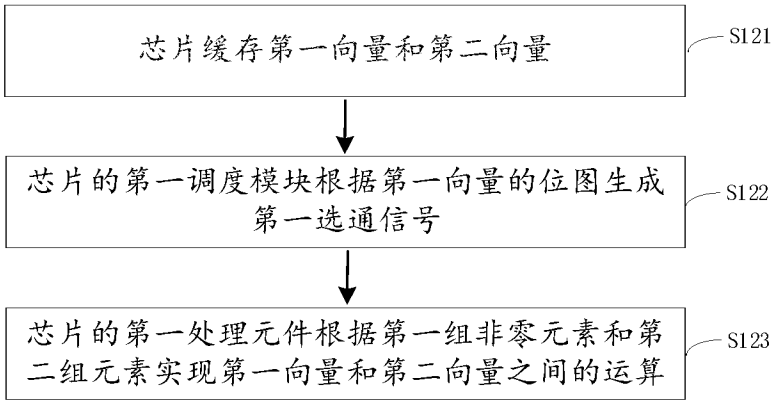
芯片缓存第一向量和第二向量
S121
芯片的第一调度模块根据第一向量的位图生成第一选通信号
S122
芯片的第一处理元件根据第一组非零元素和第二组元素实现第一向量和第二向量之间的运算
S123

图 12

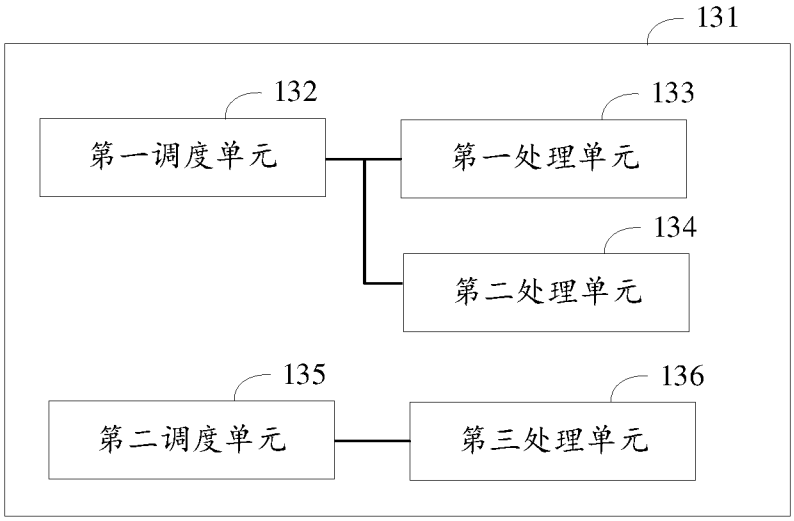
131
132
第一调度单元
133
第一处理单元
134
第二处理单元
135
第二调度单元
136
第三处理单元

图 13

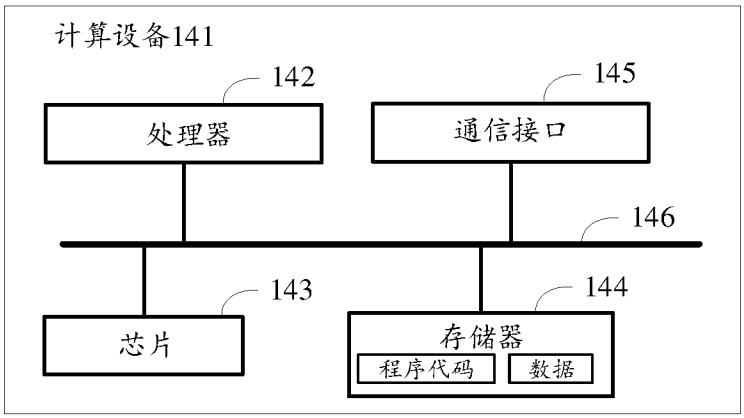
计算设备141
142
处理器
145
通信接口
146
143
芯片
144
存储器
程序代码
数据

图 14

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/137086

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)

IPC: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)

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 矩阵, 乘, 点积, 非0, 非零, 复用器, 缓存, 位图, 选通, 选择, matrix, multiplication, bitmap, non-zero, multiplexer, buffer, select

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2013073599 A1 (LINEAR ALGEBRA TECHNOLOGIES, LIMITED) 21 March 2013 (2013-03-21) description, paragraphs [0015]-[0085]	1-38
A	CN 113486298 A (NANJING UNIVERSITY) 08 October 2021 (2021-10-08) entire document	1-38
A	CN 113506589 A (HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY) 15 October 2021 (2021-10-15) entire document	1-38
A	US 2021097130 A1 (ARM LIMITED) 01 April 2021 (2021-04-01) entire document	1-38
A	US 2021240684 A1 (ALIBABA GROUP HOLDING LIMITED) 05 August 2021 (2021-08-05) entire document	1-38

☐ 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
"D" document cited by the applicant in the international application
"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	Date of mailing of the international search report
01 February 2023	**07 February 2023**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/137086

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US 2013073599 A1	21 March 2013	GB 2476800 A	13 July 2011
		WO 2011083152 A1	14 July 2011
		KR 20120113777 A	15 October 2012
		US 9104633 B2	11 August 2015
		EP 2521968 A1	14 November 2012
		IN 2054KOLNP2012 A	22 March 2013
CN 113486298 A	08 October 2021	None	
CN 113506589 A	15 October 2021	None	
US 2021097130 A1	01 April 2021	None	
US 2021240684 A1	05 August 2021	WO 2021158374 A1	12 August 2021
		CN 115066692 A	16 September 2022

国际检索报告

国际申请号
PCT/CN2022/137086

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)

IPC: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））

VEN,CNABS,CNTXT,WOTXT,EPTXT,USTXT,CNKI,IEEE:矩阵,乘,点积,非0,非零,复用器,缓存,位图,选通,选择, matrix, multiplication,bitmap,non-zero,multiplexer,buffer,select

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	US 2013073599 A1 (LINEAR ALGEBRA TECHNOLOGIES, LIMITED) 2013年3月21日 (2013-03-21) 说明书第[0015]-[0085]段	1-38
A	CN 113486298 A (南京大学) 2021年10月8日 (2021-10-08) 全文	1-38
A	CN 113506589 A (华中科技大学) 2021年10月15日 (2021-10-15) 全文	1-38
A	US 2021097130 A1 (ARM LIMITED) 2021年4月1日 (2021-04-01) 全文	1-38
A	US 2021240684 A1 (ALIBABA GROUP HOLDING LIMITED) 2021年8月5日 (2021-08-05) 全文	1-38

☐ 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

“D” 申请人在国际申请中引证的文件

“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
2023年2月1日	2023年2月7日

ISA/CN的名称和邮寄地址	受权官员
中国国家知识产权局 中国北京市海淀区蓟门桥西土城路6号 100088	孙国辉
传真号 (86-10)62019451	电话号码 (+86) 010-53961538

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2022/137086

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)
US	2013073599	A1	2013年3月21日	GB	2476800	A	2011年7月13日
				WO	2011083152	A1	2011年7月14日
				KR	20120113777	A	2012年10月15日
				US	9104633	B2	2015年8月11日
				EP	2521968	A1	2012年11月14日
				IN	2054KOLNP2012	A	2013年3月22日
CN	113486298	A	2021年10月8日		无		
CN	113506589	A	2021年10月15日		无		
US	2021097130	A1	2021年4月1日		无		
US	2021240684	A1	2021年8月5日	WO	2021158374	A1	2021年8月12日
				CN	115066692	A	2022年9月16日