(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105224467 A
(43)申请公布日 2016.01.06

(21)申请号 201410240235.8

(22)申请日 2014.05.30

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼
申请人 北京航空航天大学

(72)发明人 王丽娜 史晓华 常玉立

(74)专利代理机构 北京中博世达专利商标代理有限公司 11274
代理人 申健

(51)Int.Cl.
G06F 12/02(2006.01)
G06F 17/30(2006.01)

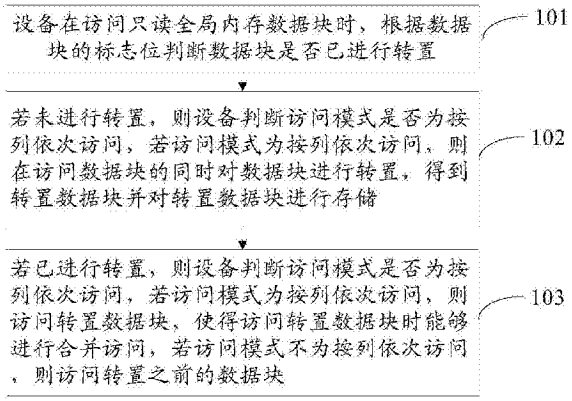
权利要求书4页 说明书15页 附图3页

(54)发明名称

一种全局内存访问的方法和设备

(57)摘要

本发明实施例提供一种全局内存访问的方法和设备，涉及通信领域，解决了全局内存访问中可能出现的非合并访问情况，从而提高全局内存的访问带宽。具体方案为：在访问只读全局内存数据块时，根据数据块的标志位判断数据块是否已进行转置；若未进行转置，则判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则在访问数据块的同时对数据块进行转置，得到转置数据块并对转置数据块进行存储；若已进行转置，则判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则访问转置数据块，使得访问转置数据块时能够进行合并访问，若访问模式不为按列依次访问，则访问转置之前的数据块。本发明用于只读全局内存的访问。



CN 105224467 A

1. 一种全局内存访问的方法，其特征在于，包括：

在访问只读全局内存数据块时，根据所述数据块的标志位判断所述数据块是否已进行转置；

若未进行转置，则判断访问模式是否为按列依次访问，若所述访问模式为所述按列依次访问，则在访问所述数据块的同时对所述数据块进行转置，得到转置数据块并对所述转置数据块进行存储；

若已进行转置，则判断所述访问模式是否为所述按列依次访问，若所述访问模式为所述按列依次访问，则访问所述转置数据块，使得访问所述转置数据块时能够进行合并访问，若所述访问模式不为所述按列依次访问，则访问转置之前的数据块。

2. 根据权利要求 1 所述的方法，其特征在于，所述判断访问模式是否为按列依次访问包括：

判断所述访问模式是否为按列访问；

若判断所述访问模式为按列访问，则再判断所述访问模式是否为依次访问。

3. 根据权利要求 1 或 2 所述的方法，其特征在于，所述数据块的标志位为第一标识；

所述在访问所述数据块的同时对所述数据块进行转置，得到转置数据块并对所述转置数据块进行存储包括：

将所述数据块的标志位从所述第一标识更新为第二标识，并将所述当前 half-warp 线程束访问的元素通过局部内存进行转置，并存至新的数据空间中。

4. 根据权利要求 3 所述的方法，其特征在于，所述判断访问模式是否为按列访问包括：

获取当前 half-warp 线程束访问所述数据块时所访问的每个元素的索引值，根据所述索引值并按照第一公式获取每个元素对应的列号；

若每个元素对应的列号相等，且相邻索引值之间相差为 N，N 表示所述数据块的列数，则确定所述访问模式为所述按列访问；

若每个元素对应的列号中有两个列号相差为 1，同时列号相等的相邻索引值相差为 N，其中的列号大者对应的每个元素按照第二公式得出的行号中最小值为 0，列号小者对应的每个元素按照所述第二公式得出的行号中最大值为 M-1，M 表示所述数据块的行数，则确定所述访问模式为所述按列访问；

其中，所述第一公式包括：$columnIndex = index - N * \left\lfloor \frac{index}{N} \right\rfloor$，

所述第二公式包括：$m = \left\lfloor \frac{index}{N} \right\rfloor$，m 表示所述行号，columnIndex 表示所述列号，index 表示所述索引值，N 表示所述数据块的列数。

5. 根据权利要求 4 所述的方法，其特征在于，所述判断所述访问模式是否为依次访问包括：

将此次 half-warp 线程束所访问每个元素对应的索引值中的最小索引值按照所述第一公式得到的第一列号，与上一次 half-warp 线程束所访问每个元素对应的索引值中的最大索引值按照所述第一公式得到的第二列号进行比较；

若所述第一列号与所述第二列号相等，且所述最小索引值和所述最大索引值满足第三

公式，则确定所述访问模式为所述依次访问；

若所述第一列号与所述第二列号相差为1，且所述最小索引值按照所述第二公式得到的行号为0，所述最大索引值按照所述第二公式得到的行号为M-1，则确定所述访问模式为所述依次访问；

所述第三公式包括：

$\left\lfloor \frac{\max Index}{N} \right\rfloor - \left\lfloor \frac{\min Index}{N} \right\rfloor = 1$，max Index表示所述最大索引值，min Index表示所述最小索引值。

6. 根据权利要求3所述的方法，其特征在于，所述将所述当前half-warp线程束访问的元素通过局部内存进行转置，并存至新的数据空间中包括：

获取所述数据块的空间大小，在全局内存中分配同等大小的数据空间，同时分配局部内存用于存储待转置的元素；

将所述当前half-warp线程束访问的元素进行转置，并将转置后的元素存储在局部内存中；

将转置后的元素形成的转置数据块写回所述全局内存分配的同等大小的数据空间；

其中，所述局部内存的大小为：

Block_dim*(Block_dim+1)*sizeof(type of Data)

Block_dim表示所述当前half-warp线程束的线程个数，sizeof(type of Data)表示所述数据块中的一个元素的存储空间大小。

7. 根据权利要求6所述的方法，其特征在于，所述方法还包括：

在访问所述数据块或所述转置数据块时，根据当前half-warp线程束访问的每个元素的索引值中的最大值，判断此次访问是否结束；

若所述当前half-warp线程束访问的每个元素的索引值中的最大值满足：maxIndex＝M*N-1，则确定此次访问结束；

其中，maxIndex表示所述当前half-warp线程束访问的每个元素的索引值中的最大值。

8. 根据权利要求1所述的方法，其特征在于，所述根据所述数据块的标志位判断所述数据块是否已进行转置包括：

若所述标志位为所述第一标识，则确定所述数据块未进行转置；

若所述标志位为所述第二标识，则确定所述数据块已进行转置。

9. 一种设备，其特征在于，包括：

第一判断单元，用于在访问只读全局内存数据块时，根据所述数据块的标志位判断所述数据块是否已进行转置；

第二判断单元，用于若未进行转置，则判断访问模式是否为按列依次访问；

转置单元，用于若所述访问模式为所述按列依次访问，则在访问所述数据块的同时对所述数据块进行转置，得到转置数据块并对所述转置数据块进行存储；

所述第二判断单元，还用于若已进行转置，则判断所述访问模式是否为所述按列依次访问；

访问单元，用于若所述访问模式为所述按列依次访问，则访问所述转置数据块，使得访

问所述转置数据块时能够进行合并访问，若所述访问模式不为所述按列依次访问，则访问转置之前的数据块。

10. 根据权利要求 9 所述的设备，其特征在于，所述判断单元具体用于：

判断所述访问模式是否为按列访问；

若判断所述访问模式为按列访问，则再判断所述访问模式是否为依次访问。

11. 根据权利要求 9 或 10 所述的设备，其特征在于，所述数据块的标志位为第一标识；

所述转置单元具体用于：

将所述数据块的标志位从所述第一标识更新为第二标识，并将所述当前 half-warp 线程束访问的元素通过局部内存进行转置，并存至新的数据空间中。

12. 根据权利要求 11 所述的设备，其特征在于，所述判断单元具体用于：

获取当前 half-warp 线程束访问所述数据块时所访问的每个元素的索引值，根据所述索引值并按照第一公式获取每个元素对应的列号；

若每个元素对应的列号相等，且相邻索引值之间相差为 N，N 表示所述数据块的列数，则确定所述访问模式为所述按列访问；

若每个元素对应的列号中有两个列号相差为 1，同时列号相等的相邻索引值相差为 N，其中的列号大者对应的每个元素按照第二公式得出的行号中最小值为 0，列号小者对应的每个元素按照所述第二公式得出的行号中最大值为 M-1，M 表示所述数据块的行数，则确定所述访问模式为所述按列访问；

其中，所述第一公式包括：$columnIndex = index - N * \left\lfloor \frac{index}{N} \right\rfloor$，

所述第二公式包括：$m = \left\lfloor \frac{index}{N} \right\rfloor$，columnIndex 表示所述列号，index 表示所述索引值，m 表示所述行号，N 表示所述数据块的列数。

13. 根据权利要求 12 所述的设备，其特征在于，所述判断单元具体用于：

将此次 half-warp 线程束所访问每个元素对应的索引值中的最小索引值按照所述第一公式得到的第一列号，与上一次 half-warp 线程束所访问每个元素对应的索引值中的最大索引值按照所述第一公式得到的第二列号进行比较；

若所述第一列号与所述第二列号相等，且将所述最小索引值和所述最大索引值满足第三公式，则确定所述访问模式为所述依次访问；

若所述第一列号与所述第二列号相差为 1，且所述最小索引值按照所述第二公式得到的行号为 0，所述最大索引值按照所述第二公式得到的行号为 M-1，则确定所述访问模式为所述依次访问；

所述第三公式包括：

$\left\lfloor \frac{\max Index}{N} \right\rfloor - \left\lfloor \frac{\min Index}{N} \right\rfloor = 1$，max Index 表示所述最大索引值，min Index 表示所述最小索引值。

14. 根据权利要求 11 所述的设备，其特征在于，所述转置单元具体用于：

获取所述数据块的空间大小，在全局内存中分配同等大小的数据空间，同时分配局部

内存用于存储待转置的元素；

将所述当前 half-warp 线程束访问的元素进行转置，并将转置后的元素存储在局部内存中；

将转置后的元素形成的转置数据块写回所述全局内存分配的同等大小的数据空间；

其中，所述局部内存的大小为：

Block_dim*(Block_dim+1)*sizeof(type of Data)，

Block_dim 表示所述当前 half-warp 线程束的线程个数，sizeof(type of Data) 表示所述数据块中的一个元素的存储空间大小。

15. 根据权利要求 14 所述的设备，其特征在于，所述判断单元还用于：

在访问所述数据块或所述转置数据块时，根据当前 half-warp 线程束访问的每个元素的索引值中的最大值，判断此次访问是否结束；

若所述当前 half-warp 线程束访问的每个元素的索引值中的最大值满足：maxIndex = M*N-1，则确定此次访问结束；

其中，maxIndex 表示所述当前 half-warp 线程束访问的每个元素的索引值中的最大值。

16. 根据权利要求 9 所述的设备，其特征在于，所述判断单元具体用于：

若所述标志位为所述第一标识，则确定所述数据块未进行转置；

若所述标志位为所述第二标识，则确定所述数据块已进行转置。

一种全局内存访问的方法和设备

技术领域

[0001] 本发明涉及计算机领域，尤其涉及一种全局内存访问的方法和设备。

背景技术

[0002] 图形处理器(Graphic Processing Unit，GPU)在对全局内存进行访问时，通常有两种情况：一种是按行的顺序访问数据块，另一种是按列的顺序访问数据块。在按行访问数据块时，一般情况下，各个线程访问的数据地址是连续的，通常会进行合并访问，但是在按列访问数组时，由于访问的数据地址不连续，会出现非合并访问的情况。其中，合并访问是指当访问的数据地址连续时，GPU 通常将多个线程的内存访问尽量合并到较少的内存请求命令中，存储器进行一次传输就可以处理多个线程的访存请求。

[0003] 其中，GPU 全局内存的访问是否满足合并访问条件，是对图形处理器通用计算技术(General Purpose Computing on Graphics Processing Units，GPGPU)程序性能影响最明显的因素之一。在计算能力 1.0/1.1 的 GPU 硬件上，是否满足合并访问条件在很多情况下会使 GPGPU 程序的速度产生高达一个数量级的差异，对存储器带宽性能有很大影响。

[0004] 现有技术中，对于计算能力为 1.x 的设备，half-warp(由 warp 中的前 16 个或者后 16 个线程组成)的 16 个线程对全局内存进行装载或者存储访问时，当按列依次访问某一块连续的只读全局内存地址空间时，由于线程束依次访问的数据地址不连续，会出现非合并访问的情况，就会造成 half-warp 中的 16 个线程会访问 16 次全局内存，使得全局内存的访问带宽会降到最低。

发明内容

[0005] 本发明的实施例提供一种全局内存访问的方法和设备，能够解决现有技术中按列访问时非合并访问导致的存储器访问带宽低下的问题。

[0006] 为达到上述目的，本发明的实施例采用如下技术方案：

[0007] 第一方面，提供了一种全局内存访问的方法，包括：

[0008] 在访问只读全局内存数据块时，根据所述数据块的标志位判断所述数据块是否已进行转置；

[0009] 若未进行转置，则判断访问模式是否为按列依次访问，若所述访问模式为所述按列依次访问，则在访问所述数据块的同时对所述数据块进行转置，得到转置数据块并对所述转置数据块进行存储；

[0010] 若已进行转置，则判断所述访问模式是否为所述按列依次访问，若所述访问模式为所述按列依次访问，则访问所述转置数据块，使得访问所述转置数据块时能够进行合并访问，若所述访问模式不为所述按列依次访问，则访问转置之前的数据块。

[0011] 结合第一方面，在第一方面的第一种可能的实现方式中，所述判断访问模式是否为按列依次访问包括：

[0012] 判断所述访问模式是否为按列访问；

[0013] 若判断所述访问模式为按列访问，则再判断所述访问模式是否为依次访问。

[0014] 结合第一方面或第一方面的第一种可能的实现方式，在第二种可能的实现方式中，所述数据块的标志位为第一标识；

[0015] 所述在访问所述数据块的同时对所述数据块进行转置，得到转置数据块并对所述转置数据块进行存储包括：

[0016] 将所述数据块的标志位从所述第一标识更新为第二标识，并将所述当前half-warp线程束访问的元素通过局部内存进行转置，并存至新的数据空间中。

[0017] 结合第一方面的第二种可能的实现方式，在第三种可能的实现方式中，判断访问模式是否为按列访问包括：

[0018] 获取当前half-warp线程束访问所述数据块时所访问的每个元素的索引值，根据所述索引值并按照第一公式获取每个元素对应的列号；

[0019] 若每个元素对应的列号相等，且相邻索引值之间相差为N，N表示所述数据块的列数，则确定所述访问模式为所述按列访问；

[0020] 若每个元素对应的列号中有两个列号相差为1，同时列号相等的相邻索引值相差为N，其中的列号大者对应的每个元素按照第二公式得出的行号中最小值为0，列号小者对应的每个元素按照所述第二公式得出的行号中最大值为M-1，M表示所述数据块的行数，则确定所述访问模式为所述按列访问；

[0021] 其中，所述第一公式包括：$columnIndex = index - N * \left\lfloor \frac{index}{N} \right\rfloor$，columnIndex表示所述列号，index表示所述索引值，N表示所述数据块的列数；

[0022] 所述第二公式包括：$m = \left\lfloor \frac{index}{N} \right\rfloor$，m表示所述行号，index表示所述索引值，N表示所述数据块的列数。

[0023] 结合第一方面的第三种可能的实现方式，在第四种可能的实现方式中，所述判断所述访问模式是否为依次访问包括：

[0024] 将此次half-warp线程束所访问每个元素对应的索引值中的最小索引值按照所述第一公式得到的第一列号，与上一次half-warp线程束所访问每个元素对应的索引值中的最大索引值按照所述第一公式得到的第二列号进行比较；

[0025] 若所述第一列号与所述第二列号相等，且所述最小索引值和所述最大索引值满足第三公式，则确定所述访问模式为所述依次访问；

[0026] 若所述第一列号与所述第二列号相差为1，且所述最大索引值按照所述第二公式得到的行号为0，所述最小索引值按照所述第二公式得到的行号为M-1，则确定所述访问模式为所述依次访问；

[0027] 所述第三公式包括：

[0028] $\left\lfloor \frac{\max Index}{N} \right\rfloor - \left\lfloor \frac{\min Index}{N} \right\rfloor = 1$，max Index表示所述最大索引值，min Index表示所述最小索引值。

[0029] 结合第一方面的第二种可能的实现方式，在第五种可能的实现方式中，所述将所

述当前 half-warp 线程束访问的元素通过局部内存进行转置，并存至新的数据空间中包括：

[0030] 获取所述数据块的空间大小，在全局内存中分配同等大小的数据空间，同时分配局部内存用于存储待转置的元素；

[0031] 将所述当前 half-warp 线程束访问的元素进行转置，并将转置后的元素存储在局部内存中；

[0032] 将转置后的元素形成的转置数据块写回所述全局内存分配的同等大小的数据空间；

[0033] 其中，所述局部内存的大小为：

[0034] Block_dim*(Block_dim+1)*sizeof(type of Data)

[0035] Block_dim 表示所述当前 half-warp 线程束的线程个数，sizeof(type of Data) 表示所述数据块中的一个元素的存储空间大小。

[0036] 结合第一方面的第五种可能的实现方式，在第六种可能的实现方式中，所述方法还包括：

[0037] 在访问所述数据块或所述转置数据块时，根据当前 half-warp 线程束访问的每个元素的索引值中的最大值，判断此次访问是否结束；

[0038] 若所述当前 half-warp 线程束访问的每个元素的索引值中的最大值满足：maxIndex ＝ M*N-1，则确定此次访问结束；

[0039] 其中，maxIndex 表示所述当前 half-warp 线程束访问的每个元素的索引值中的最大值。

[0040] 结合第一方面，在第七种可能的实现方式中，根据数据块的标志位判断所述数据块是否已进行转置包括：

[0041] 若所述标志位为所述第一标识，则确定所述数据块未进行转置；

[0042] 若所述标志位为所述第二标识，则确定所述数据块已进行转置。

[0043] 第二方面，提供了一种设备，包括：

[0044] 第一判断单元，用于在访问只读全局内存数据块时，根据所述数据块的标志位判断所述数据块是否已进行转置；

[0045] 第二判断单元，还用于若未进行转置，则判断访问模式是否为按列依次访问；

[0046] 转置单元，用于若所述访问模式为所述按列依次访问，则在访问所述数据块的同时对所述数据块进行转置，得到转置数据块并对所述转置数据块进行存储；

[0047] 所述第二判断单元，还用于若已进行转置，则判断所述访问模式是否为所述按列依次访问；

[0048] 访问单元，用于若所述访问模式为所述按列依次访问，则访问所述转置数据块，使得访问所述转置数据块时能够进行合并访问，若所述访问模式不为所述按列依次访问，则访问转置之前的数据块。

[0049] 结合第二方面，在第二方面的第一种可能的实现方式中，所述判断单元具体用于：

[0050] 判断所述访问模式是否为按列访问；

[0051] 若判断所述访问模式为按列访问，则再判断所述访问模式是否为依次访问。

[0052] 结合第二方面或第二方面的第一种可能的实现方式，在第二种可能的实现方式中，所述数据块的标志位为第一标识；

[0053] 所述转置单元具体用于：

[0054] 将所述数据块的标志位从所述第一标识更新为第二标识，并将所述当前half-warp线程束访问的元素通过局部内存进行转置，并存至新的数据空间中。

[0055] 结合第二方面的第二种可能的实现方式，在第三种可能的实现方式中，所述判断单元具体用于：

[0056] 获取当前half-warp线程束访问所述数据块时所访问的每个元素的索引值，根据所述索引值并按照第一公式获取每个元素对应的列号；

[0057] 若每个元素对应的列号相等，且相邻索引值之间相差为N，N表示所述数据块的列数，则确定所述访问模式为所述按列访问；

[0058] 若每个元素对应的列号中有两个列号相差为1，同时列号相等的相邻索引值相差为N，其中的列号大者对应的每个元素按照第二公式得出的行号中最小值为0，列号小者对应的每个元素按照所述第二公式得出的行号中最大值为M-1，M表示所述数据块的行数，则确定所述访问模式为所述按列访问；

[0059] 其中，所述第一公式包括：$columnIndex = index - N * \left\lfloor \frac{index}{N} \right\rfloor$，columnIndex表示所述列号，index表示所述索引值，N表示所述数据块的列数；

[0060] 所述第二公式包括：$m = \left\lfloor \frac{index}{N} \right\rfloor$，m表示所述行号，index表示所述索引值，N表示所述数据块的列数。

[0061] 结合第二方面的第三种可能的实现方式，在第四种可能的实现方式中，所述判断单元具体用于：

[0062] 将此次half-warp线程束所访问每个元素对应的索引值中的最小索引值按照所述第一公式得到的第一列号，与上一次half-warp线程束所访问每个元素对应的索引值中的最大索引值按照所述第一公式得到的第二列号进行比较；

[0063] 若所述第一列号与所述第二列号相等，且将所述最小索引值和所述最大索引值满足第三公式，则确定所述访问模式为所述依次访问；

[0064] 若所述第一列号与所述第二列号相差为1，且所述最小索引值按照所述第二公式得到的行号为0，所述最大索引值按照所述第二公式得到的行号为M-1，则确定所述访问模式为所述依次访问；

[0065] 所述第三公式包括：

[0066] $\left\lfloor \frac{\max Index}{N} \right\rfloor - \left\lfloor \frac{\min Index}{N} \right\rfloor = 1$，max Index表示所述最大索引值，min Index表示所述最小索引值。

[0067] 结合第二方面的第二种可能的实现方式，在第五种可能的实现方式中，所述转置单元具体用于：

[0068] 获取所述数据块的空间大小，在全局内存中分配同等大小的数据空间，同时分配

局部内存用于存储待转置的元素；

[0069] 将所述当前 half-warp 线程束访问的元素进行转置，并将转置后的元素存储在局部内存中；

[0070] 将转置后的元素形成的转置数据块写回所述全局内存分配的同等大小的数据空间；

[0071] 其中，所述局部内存的大小为：

[0072] Block_dim*(Block_dim+1)*sizeof(type of Data)

[0073] Block_dim 表示所述当前 half-warp 线程束的线程个数，sizeof(type of Data) 表示所述数据块中的一个元素的存储空间大小。

[0074] 结合第二方面的第五种可能的实现方式，在第六种可能的实现方式中，所述判断单元还用于：

[0075] 在所述判断单元判断所述访问模式是否为按列依次访问之前，根据当前 half-warp 线程束访问的每个元素的索引值中的最大值，判断此次访问是否结束；

[0076] 若所述当前 half-warp 线程束访问的每个元素的索引值中的最大值满足：maxIndex ＝ M*N-1，则确定此次访问结束；

[0077] 其中，maxIndex 表示所述当前 half-warp 线程束访问的每个元素的索引值中的最大值。

[0078] 结合第二方面，在第二方面的第七种可能的实现方式中，所述判断单元具体用于：

[0079] 若所述标志位为所述第一标识，则确定所述数据块未进行转置；

[0080] 若所述标志位为所述第二标识，则确定所述数据块已进行转置。

[0081] 本发明实施例提供的全局内存访问的方法和设备，在访问只读全局内存数据块时，根据数据块的标志位判断数据块是否已进行转置；若未进行转置，则判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则在访问数据块的同时对数据块进行转置，得到转置数据块并对转置数据块进行存储；若已进行转置，则判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则访问转置数据块，使得访问转置数据块时能够进行合并访问，若访问模式不为按列依次访问，则访问转置之前的数据块，解决了访问全局内存时，可能会出现非合并访问而导致的全局内存访问带宽降低的问题。

附图说明

[0082] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，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。

[0083] 图 1 为本发明实施例提供的一种全局内存访问的方法流程示意图；

[0084] 图 2 为本发明实施例提供的一种全局内存访问的方法流程示意图；

[0085] 图 3 为本发明实施例提供的一种设备结构框图；

[0086] 图 4 为本发明实施例提供的一种设备结构框图。

具体实施方式

[0087] 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，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。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他实施例，都属于本发明保护的范围。

[0088] 本发明实施例的应用场景可以是由GPGPU、开放运算语言(Open Computing Language，OpenCL)/同一计算设备架构(Compute Unified Device Architecture，CUDA)编译平台、GPU应用程序组成。其中，GPU应用程序通过OpenCL/CUDA编译平台在GPGPU上运行。本发明实施例是针对GPU应用程序对GPGPU全局内存的访问模式的改进，即实现时，需要对OpenCL/CUDA编译平台进行相应的改进，使之能够完成相应的功能。

[0089] 实施例一

[0090] 本发明实施例提供一种全局内存访问的方法，参见图1，其步骤包括：

[0091] 101、设备在访问只读全局内存数据块时，根据数据块的标志位判断数据块是否已进行转置。

[0092] 该设备可以为计算机等。全局内存可以用以存储没有初始化的和初始化为0的全局变量bss、数据data和只读数据rodata。这里的只读全局内存是指全局内存中的只读数据。

[0093] 具体的，在根据数据块的标志位flag判断数据块是否已进行转置时，如果标志位为第一标识false，则确定该数据块未进行转置；如果标志位为第二标识true，则确定该数据块已进行转置。

[0094] 102、若未进行转置，则设备判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则在访问数据块的同时对数据块进行转置，得到转置数据块并对转置数据块进行存储。

[0095] 由于GPU全局内存的访问模式可以为：按行访问的模式、按列访问的模式、以及乱序访问的模式。针对本发明要解决的按列访问数据块出现的非合并访问的情况，在访问全局内存时，首先对访问模式进行判断，确定是否为按列依次访问。这里还要判断是否为依次访问，也就是访问的数据地址是否按列连续，是由于数据地址不连续出现非合并访问时，如果对数据进行转置，其转置后的数据块的数据地址也不连续，再次访问转置数据块时，继续会出现非合并访问的情况。

[0096] 如果为按列依次访问，就访问原数据块，其中每访问一个原数据块中的数据，对该数据进行一次转置，这样访问原数据块完毕后，就同时形成了原数据块的转置数据块，以便于下一次将要按列依次访问原数据块时，直接访问其对应的转置数据块，使得访问的数据地址连续，可以进行合并访问。

[0097] 其中，判断访问模式是否为按列依次访问，是通过先判断访问模式是否为按列访问，若判断访问模式为按列访问，则再判断访问模式是否为依次访问。

[0098] 其中，对数据块进行转置，指的是将数据块的第一行变成第一列，第二行变成第二列，……，最后一行变成最后一列。

[0099] 103、若已进行转置，则设备判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则访问转置数据块，使得访问转置数据块时能够进行合并访问，若访问模式不

为按列依次访问，则访问转置之前的数据块。

[0100] 本发明实施例提供的全局内存访问的方法，在访问只读全局内存数据块时，根据数据块的标志位判断数据块是否已进行转置；若未进行转置，则判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则在访问数据块的同时对数据块进行转置，得到转置数据块并对转置数据块进行存储；若已进行转置，则判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则访问转置数据块，使得访问转置数据块时能够进行合并访问，若访问模式不为按列依次访问，则访问转置之前的数据块，解决了现有技术中，访问全局内存过程中，可能会出现按列访问时非合并访问的情况，而导致的全局内存访问带宽降低的问题。

[0101] 实施例二

[0102] 本发明实施例提供一种全局内存访问的方法，以访问只读全局内存数据块二维矩阵 Data，Data 数据块的大小为 M*N(M 行 N 列)，以行优先的顺序存储为例进行说明，如图 2 所示，包括：

[0103] 201、在访问只读全局内存数据块时，根据数据块的标志位判断数据块是否已进行转置，若未进行转置，则进入步骤 202；若已进行转置，则进入步骤 207。

[0104] 示例性的，可以将全局内存中的数据块通过标志位 flag 进行标识，若数据块的 flag 为第一标识 flase，则确定该数据块没有作出调整，即未做任何处理，未进行转置；若数据块的 flag 为第二标识 true 则确定该数据块已经经过转置。

[0105] 202、判断访问模式是否为按列访问，若为按列访问，则进入步骤 203；若不为按列访问，则进入步骤 206。

[0106] 先对访问模式是否为按列访问进行判断，这里可以通过在 GPU 编译平台中在访问语句前插入桩代码，用以指示判断访问模式是否为按列访问，也可以为其它的指示方式，这里不做限定。

[0107] 其中，在判断是否按列访问时，可以先获取当前 half-warp 线程束访问数据块时所访问的每个元素的索引值，根据索引值并按照第一公式获取当前 half-warp 线程束访问的子数据块中的每个元素的列号，这里的子数据块是指当前 half-warp 线程束访问的该 Data 数据块中的部分元素这里的第一公式包括：

[0108] $columnIndex = index - N * \left\lfloor \frac{index}{N} \right\rfloor$，columnIndex 表示列号，index 表示索引值，N 表示数据块的列数，$\left\lfloor \frac{index}{N} \right\rfloor$ 表示当前计算的元素所在行之前的行数，$N * \left\lfloor \frac{index}{N} \right\rfloor$ 表示当前计算的元素所在行之前的所有行的元素总数。

[0109] 若每个元素对应的列号相等，且相邻索引值之间相差为 N，则可以确定访问模式为按列访问；若每个元素对应的列号中有两个列号相差为 1，同时列号相等的相邻索引值相差为 N，其中的列号大者对应的每个元素按照第二公式得出的行号中最小值为 0，列号小者对应的每个元素按照第二公式得出的行号中最大值为 M-1，M 表示数据块的行数，则确定访问模式为按列访问，这里的第二公式包括：$m = \left\lfloor \frac{index}{N} \right\rfloor$，m 表示行号，index 表示索引值，N

表示数据块的列数。也就是说，判断是否为按列访问，有两种情况，一种是判断是否为同一列的元素，另一种是判断此次访问的是否为相邻两列的元素。

[0110] 其中的索引值 index 为所访问的全局内存数据块元素的标识，本发明的元素的标识为 0、1，……M*N-1。这里的行号是从 0 依次至 M-1 标识的。

[0111] 203、判断访问模式是否为依次访问，若为依次访问，则进入步骤 204；若不为依次访问，则进入步骤 206。

[0112] 若判断了访问模式为按列访问后，再判断是否访问的是数据的地址是否连续，也即此次 half-warp 线程束所访问的子数据块和上一次 half-warp 线程束所访问的子数据块是否为相邻子数据块，这里的相邻子数据块是该 M*N 矩阵中的两部分子数据块。具体可以将此次 half-warp 线程束所访问每个元素对应的索引值中的最小索引值按照第一公式得到的第一列号，与上一次 half-warp 线程束所访问每个元素对应的索引值中的最大索引值按照第一公式得到的第二列号进行比较，若第一列号与第二列号相等，且将最小索引值和最大索引值满足第三公式：

[0113]

$$\left\lfloor \frac{\max Index}{N} \right\rfloor - \left\lfloor \frac{\min Index}{N} \right\rfloor = 1,$$

[0114] maxIndex 表示最大索引值，minIndex 表示最小索引值，则确定访问模式为依次访问，也就是说，此次 half-warp 线程束所访问的最小索引值的元素，和上一次 half-warp 线程束所访问的最大索引值的元素属于同一列，且此次 half-warp 线程束所访问的最小索引值的元素，和上一次 half-warp 线程束所访问的最大索引值的元素位于相邻的两行，那么就确定访问的是连续的子数据块。

[0115] 若第一列号与第二列号相差为 1，且最大索引值按照第二公式得到的行号为 0，0 代表第一行，最小索引值按照第二公式得到的行号为 M-1，M-1 代表最后一行，则确定访问模式为依次访问，为连续的子数据块。

[0116] 204、在访问数据块的同时对数据块进行转置，得到转置数据块并对转置数据块进行存储。

[0117] 具体而言，在确定了此次访问为按列依次访问后，则在此次访问的同时，对该数据块 Data 进行转置，并将该数据块的标志位 flag 更新为第二标识 true，以表示该数据块 Data 存在转置数据块 Data'。

[0118] 其中，对数据块 Data 进行转置，是通过将当前 half-warp 线程束访问的元素通过局部内存进行转置，并存至新的数据空间中实现的。示例性的，先获取该数据块 Data 的空间大小，在全局内存中分配同等大小的数据空间，用来存储转置后的新数据块 Data'，同时分配局部内存 block 用于存储待转置的元素，而后在当前 half-warp 线程束访问 Data 元素的同时将访问的元素进行转置，这里是通过每访问一个元素，对该元素进行转置实现的，并将转置后的元素存储在局部内存中，待此次访问并转置完成后，将转置后的元素形成转置数据块写回全局内存分配的同等大小的数据空间。也即此次访问还是访问的原数据块，为非合合并访问，形成转置数据块，是为了方便再次将要按列访问原数据块时，直接访问其转置数据块即可，也即下一次访问就会转化为合并访问。指的是 CPU 中内存模型的其中一种当事件过程被触发时，局部内存便会分配内存空间给待转置数据块。

[0119] 其中，局部内存的大小可以为：

[0120] Block_dim*(Block_dim+1)*sizeof(type of Data)

[0121] Block_dim 表示当前 half-warp 线程束的线程个数，sizeof(type of Data) 表示数据块中的一个元素的存储空间大小。这里的 Block_dim+1 之所以要加 1 是为了防止局部内存出现存储冲突 (bank conflict) 的情况出现。具体而言，bank 是指局部内存被划分为大小相等，能被同时访问的存储器模块，不同的存储器模块可以互不干扰同时工作，但当 half-warp 请求访问的多个地址位于同一 bank 时，由于存储器模块在一个时刻无法响应多个请求，因此这些请求就必须被串行的完成，会出现 bank conflict 情况。Block-dim+1 之后可以保证 half-warp 请求访问的多个地址位于不同 bank。

[0122] 上述转置过程中，根据 GPU 的内置编程模型，先将 Data 中的数据存放到 block 中可以通过下列语言实现：

[0123] xIndex = blockIdx.x*Block_dim+threadIdx.x；

[0124] yIndex = blockIdx.y*Block_dim+threadIdx.y；

[0125] Index = yIndex*N+xIndex；

[0126] block[threadIdx.y][threadIdx.x] = Data[index]；

[0127] 再将转置后的矩阵写回在全局内存中分配好的 Data’中可以通过以下语言实现：

[0128] xIndex = blockIdx.y*Block_dim+threadIdx.x；

[0129] yIndex = blockIdx.x*Block_dim+threadIdx.y；

[0130] Data’[yIndex*M+xIndex] = block[threadIdx.x][threadIdx.y]。

[0131] 205、判断访问是否结束，若未结束，则进入步骤 202；若结束，则进入步骤 211。

[0132] 在确定了数据块没有发生转置时，在访问数据块时，如果发生访问模式是按列访问的情况，要在访问数据元素的同时进行转置，在转置时同时判断转置是否结束。

[0133] 这里可以根据当前 half-warp 线程束访问的子数据块中的每个元素的索引值中的最大值，判断转置是否结束，若满足 maxIndex = M*N-1，则确定此次访问结束，maxIndex 表示当前 half-warp 线程束访问的每个元素的索引值中的最大值。

[0134] 206、访问未进行转置处理之前的数据块。

[0135] 这里的访问转置处理之前的数据块，可以是由于前述步骤 203 判定了此次访问不是按列访问，或者是由于前述步骤 204 判定了此次访问不是依次访问，都要访问 Data 数据块中的元素，并将 Data 的标志位 flag 设为第一标识 false，标识该数据块未进行转置。

[0136] 207、判断访问模式是否为按列依次访问，若为按列依次访问，则进入步骤 208；若不为按列依次访问，则进入步骤 209。

[0137] 当确定了数据块 Data 的标志位为 true 后时，说明该数据块 Data 存在转置数据块 Data’，这时，再判断当前访问是否为按列依次访问，这里的按列依次访问的实现方式与步骤 203 和步骤 204 类似，不再赘述。

[0138] 208、访问转置数据块，而后进入步骤 210。

[0139] 如果当前访问是按列依次访问，则访问 Data’中的数据。具体可以是：根据当前 half-warp 线程束所获得的 Data 数据块中的元素的索引值 index 获得对应的 Data’数据块中的相应元素的索引值 index’，并访问 Data[index’]：index’ = (int)(index/N)+(index％ N)*M。

[0140] 209、访问转置之前的数据块，而后进入步骤 210。

[0141] 如果当前 half-warp 线程束不是按列依次访问，就访问转置之前的 Data 中的元素，这里包括不是按列访问，或者是按列但不是依次访问的情况。

[0142] 210、判断此次访问是否结束，若未结束，则进入步骤 207；若结束，则进入步骤 211。

[0143] 这里当前 half-warp 线程束访问当前的元素完成后，都要判断访问是否结束，判断的依据是根据记录的当前 half-warp 线程束访问 Data 元素的最大索引值 maxdex，是否满足 maxIndex = M*N-1，如果满足，则访问结束，如果不满足，则继续访问，进入步骤 207。

[0144] 211、结束。

[0145] 这样一来，对于 Data 数据块以行优先存储的情况，当全局内存只读数据块被改变存储布局时，无论此后以何种模式（按列、按行、乱序）访问该 Data 数据块，只要判断其为按列访问，就直接访问其对应的转置后的数据块，避免了非合并访问的情况，提升了存储器的访问带宽。

[0146] 需要说明的是，本发明是针对按列依次访问 GPGPU 内存模型的全局存储器而提出的实施方案，对于全局存储器只读单元可能还有其他的访问方式，如斜对角访问等等，都可以应用本发明的实施思维来解决其它的访问方式对应的问题。

[0147] 本发明实施例提供的全局内存访问的方法，在访问只读全局内存数据块时，根据数据块的标志位判断数据块是否已进行转置；若未进行转置，则判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则在访问数据块的同时对数据块进行转置，得到转置数据块并对转置数据块进行存储；若已进行转置，则判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则访问转置数据块，使得访问转置数据块时能够进行合并访问，若访问模式不为按列依次访问，则访问转置之前的数据块，解决了现有技术中，访问全局内存过程中，可能会出现按列访问时非合并访问的情况，而导致的全局内存访问带宽降低的问题。

[0148] 实施例三

[0149] 本发明实施例提供一种设备 01，如图 3 所示，包括：

[0150] 第一判断单元 011，用于在访问只读全局内存数据块时，根据数据块的标志位判断数据块是否已进行转置。

[0151] 第二判断单元 012，用于若未进行转置，则判断访问模式是否为按列依次访问。

[0152] 转置单元 013，用于若访问模式为按列依次访问，则在访问数据块的同时对数据块进行转置，得到转置数据块并对转置数据块进行存储；

[0153] 第二判断单元 012，还用于若已进行转置，则判断访问模式是否为按列依次访问。

[0154] 访问单元 014，用于若访问模式为按列依次访问，则访问转置数据块，使得访问转置数据块时能够进行合并访问，若访问模式不为按列依次访问，则访问转置之前的数据块。

[0155] 可选的，所提供的设备，还包括：插入单元 015，用于在判断访问模式是否为按列访问之前，在 GPU 编译平台的访问语句前插入桩代码，桩代码用于指示判断访问模式是否为按列依次访问。

[0156] 可选的，第一判断单元 011 可以具体用于：

[0157] 若标志位为第一标识，则确定数据块未进行转置；

[0158] 若标志位为第二标识，则确定数据块已进行转置。

[0159] 可选的，第二判断单元 012 可以具体用于：

[0160] 判断访问模式是否为按列访问；

[0161] 若判断访问模式为按列访问，则再判断访问模式是否为依次访问。

[0162] 可选的，第二判断单元 012 可以具体用于：

[0163] 获取当前 half-warp 线程束访问数据块时所访问的每个元素的索引值，根据索引值并按照第一公式获取每个元素对应的列号；

[0164] 若每个元素对应的列号相等，且相邻索引值之间相差为 N，N 表示数据块的列数，则确定访问模式为按列访问；

[0165] 若每个元素对应的列号中有两个列号相差为 1，同时列号相等的相邻索引值相差为 N，其中的列号大者对应的每个元素按照第二公式得出的行值中最小值为 0，列号小者对应的每个元素按照第二公式得出的行值中最大值为 M-1，M 表示数据块的行数，则确定访问模式为按列访问；

[0166] 其中，第一公式包括：$columnIndex = index - N * \left\lfloor \frac{index}{N} \right\rfloor$，columnIndex 表示列号，index 表示索引值，N 表示数据块的列数；

[0167] 第二公式包括：$m = \left\lfloor \frac{index}{N} \right\rfloor$，m 表示行值，index 表示索引值，N 表示数据块的列数。

[0168] 可选的，第二判断单元 012 可以具体用于：

[0169] 将此次 half-warp 线程束所访问每个元素对应的索引值中的最小索引值按照第一公式得到的第一列号，与上一次 half-warp 线程束所访问每个元素对应的索引值中的最大索引值按照第一公式得到的第二列号进行比较；

[0170] 若第一列号与第二列号相等，且将最小索引值和最大索引值满足第三公式，则确定访问模式为依次访问；

[0171] 若第一列号与第二列号相差为 1，且最大索引值按照第二公式得到的行值为 0，最小索引值按照第二公式得到的行值为 M-1，则确定访问模式为依次访问；

[0172] 第三公式包括：

[0173] $\left\lfloor \frac{\max Index}{N} \right\rfloor - \left\lfloor \frac{\min Index}{N} \right\rfloor = 1$，max Index 表示最大索引值，min Index 表示最小索引值。

[0174] 可选的，转置单元 013 可以具体用于：

[0175] 将数据块的标志位更新为第二标识，并将当前 half-warp 线程束访问的元素通过局部内存进行转置，并存至新的数据空间中。

[0176] 可选的，转置单元 013 可以具体用于：

[0177] 获取数据块的空间大小，在全局内存中分配同等大小的数据空间，同时分配局部内存用于存储待转置的元素；

[0178] 将当前 half-warp 线程束访问的元素进行转置，并将转置后的元素存储在局部内

存中；

[0179] 将转置后的元素形成的转置数据块写回全局内存分配的同等大小的数据空间；

[0180] 其中，局部内存的大小为：

[0181] Block_dim*(Block_dim+1)*sizeof(type of Data)

[0182] Block_dim表示当前half-warp线程束的线程个数，sizeof(type of Data)表示数据块中的一个元素的存储空间大小。

[0183] 可选的，第二判断单元012还可以用于：

[0184] 在判断单元判断访问模式是否为按列依次访问之前，根据当前half-warp线程束访问的每个元素的索引值中的最大值，判断此次访问是否结束；

[0185] 若当前half-warp线程束访问的每个元素的索引值中的最大值满足：maxIndex＝M*N-1，则确定此次访问结束；

[0186] 其中，maxIndex表示当前half-warp线程束访问的每个元素的索引值中的最大值。

[0187] 可选的，访问单元014还可以用于：

[0188] 若访问模式不为按列依次访问，则访问未进行转置处理之前的数据块。

[0189] 本发明实施例提供一种设备，包括第一判断单元、第二判断单元、转置单元以及访问单元，第一判断单元用于在访问只读全局内存数据块时，根据数据块的标志位判断数据块是否已进行转置，第二判断单元用于若未进行转置，则判断访问模式是否为按列依次访问，转置单元用于若访问模式为按列依次访问，则在访问数据块的同时对数据块进行转置，得到转置数据块并对转置数据块进行存储，第二判断单元还用于若已进行转置，则判断访问模式是否为按列依次访问，访问单元用于若访问模式为按列依次访问，则访问转置数据块，若访问模式不为按列依次访问，则访问转置之前的数据块，解决了现有技术中，访问全局内存过程中，可能会出现按列访问时非合并访问的情况，而导致的全局内存访问带宽降低的问题。

[0190] 实施例四

[0191] 本发明实施例提供一种设备02，如图4所示，包括：总线021、连接到总线021的处理器022、存储器023、接收器024和发射器025，其中，该存储器023用于存储指令和数据，其中，处理器022执行该指令用于在访问存储器023的只读全局内存数据块时，根据数据块的标志位判断数据块是否已进行转置；处理器022执行该指令还用于若未进行转置，则判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则在访问存储器023的数据块的同时对数据块进行转置，得到转置数据块并对转置数据块进行存储，处理器022执行该指令还用于若已进行转置，则判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则访问转置数据块，使得访问转置数据块时能够进行合并访问，若访问模式不为按列依次访问，则访问转置之前的数据块。

[0192] 在本发明实施例中，可选的，处理器022在判断访问模式是否为按列访问之前，还用于：

[0193] 在GPU编译平台的访问语句前插入桩代码，桩代码用于指示判断访问模式是否为按列依次访问。

[0194] 在本发明实施例中，可选的，处理器022执行指令根据数据块的标志位判断数据

块是否已进行转置包括：

[0195] 若标志位为第一标识，则确定数据块未进行转置；

[0196] 若标志位为第二标识，则确定数据块已进行转置。

[0197] 在本发明实施例中，可选的，处理器 022 执行指令判断访问模式是否为按列依次访问包括：

[0198] 判断访问模式是否为按列访问；

[0199] 若判断访问模式为按列访问，则再判断访问模式是否为依次访问。

[0200] 在本发明实施例中，可选的，处理器 022 执行指令判断访问模式是否为按列访问包括：

[0201] 获取当前 half-warp 线程束访问数据块时所访问的每个元素的索引值，根据索引值并按照第一公式获取每个元素对应的列号；

[0202] 若每个元素对应的列号相等，且相邻索引值之间相差为 N，N 表示数据块的列数，则确定访问模式为按列访问；

[0203] 若每个元素对应的列号中有两个列号相差为 1，同时列号相等的相邻索引值相差为 N，其中的列号大者对应的每个元素按照第二公式得出的行号中最小值为 0，列号小者对应的每个元素按照第二公式得出的行号中最大值为 M-1，M 表示数据块的行数，则确定访问模式为按列访问；

[0204] 其中，第一公式包括：$columnIndex = index - N * \left\lfloor \frac{index}{N} \right\rfloor$，

[0205] 第二公式包括：$m = \left\lfloor \frac{index}{N} \right\rfloor$，m 表示行号，columnIndex 表示列号，index 表示索引值，N 表示数据块的列数。

[0206] 在本发明实施例中，可选的，处理器 022 执行指令判断访问模式是否为依次访问包括：

[0207] 将此次 half-warp 线程束所访问每个元素对应的索引值中的最小索引值按照第一公式得到的第一列号，与上一次 half-warp 线程束所访问每个元素对应的索引值中的最大索引值按照第一公式得到的第二列号进行比较；

[0208] 若第一列号与第二列号相等，且最小索引值和最大索引值满足第三公式，则确定访问模式为依次访问；

[0209] 若第一列号与第二列号相差为 1，且最小索引值按照第二公式得到的行号为 0，最大索引值按照第二公式得到的行号为 M-1，则确定访问模式为依次访问；

[0210] 第三公式包括：

[0211] $\left\lfloor \frac{\max Index}{N} \right\rfloor - \left\lfloor \frac{\min Index}{N} \right\rfloor = 1$，max Index 表示最大索引值，min Index 表示最小索引值。

[0212] 在本发明实施例中，可选的，处理器 022 执行指令在访问数据块的同时对数据块进行转置，得到转置数据块并对转置数据块进行存储包括：

[0213] 将数据块的标志位更新为第二标识，并将当前 half-warp 线程束访问的元素通过

局部内存进行转置，并存至新的数据空间中。

[0214] 在本发明实施例中，可选的，处理器022执行指令将当前half-warp线程束访问的元素通过局部内存进行转置，并存至新的数据空间中包括：

[0215] 获取数据块的空间大小，在全局内存中分配同等大小的数据空间，同时分配局部内存用于存储待转置的元素；

[0216] 将当前half-warp线程束访问的元素进行转置，并将转置后的元素存储在局部内存中；

[0217] 将转置后的元素形成的转置数据块写回全局内存分配的同等大小的数据空间；

[0218] 其中，局部内存的大小为：

[0219] Block_dim*(Block_dim+1)*sizeof(type of Data)

[0220] Block_dim表示当前half-warp线程束的线程个数，sizeof(type of Data)表示数据块中的一个元素的存储空间大小。

[0221] 在本发明实施例中，可选的，处理器022执行指令还用于：

[0222] 在判断访问模式是否为按列依次访问之前，根据当前half-warp线程束访问的每个元素的索引值中的最大值，判断此次访问是否结束；

[0223] 若当前half-warp线程束访问的每个元素的索引值中的最大值满足：maxIndex＝M*N-1，则确定此次访问结束；

[0224] 其中，maxIndex表示当前half-warp线程束访问的每个元素的索引值中的最大值。

[0225] 在本发明实施例中，可选的，处理器022执行指令还用于：

[0226] 若访问模式不为按列依次访问，则访问未进行转置处理之前的数据块。

[0227] 本发明实施例提供一种设备，包括总线、连接到总线的处理器、存储器、接收器和发射器，其中，该存储器用于存储指令和数据，其中，处理器执行该指令用于在访问存储器的只读全局内存数据块时，根据数据块的标志位判断数据块是否已进行转置；处理器执行该指令还用于若未进行转置，则判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则在访问存储器的数据块的同时对数据块进行转置，得到转置数据块并对转置数据块进行存储，处理器执行该指令还用于若已进行转置，则判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则访问转置数据块，使得访问转置数据块时能够进行合并访问，若访问模式不为按列依次访问，则访问转置之前的数据块，解决了现有技术中，访问全局内存过程中，按列访问时可能会出现非合并访问的情况，而导致的全局内存访问带宽降低的问题。

[0228] 在本申请所提供的几个实施例中，应该理解到，所揭露的设备和方法，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。

[0229] 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包括，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，也可以采用硬件加软件功能单元的形式实现。

[0230] 上述以软件功能单元的形式实现的集成的单元，可以存储在一个计算机可读取存储介质中。上述软件功能单元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 (Read-Only Memory，简称 ROM)、随机存取存储器 (Random Access Memory，简称 RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0231] 最后应说明的是：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。

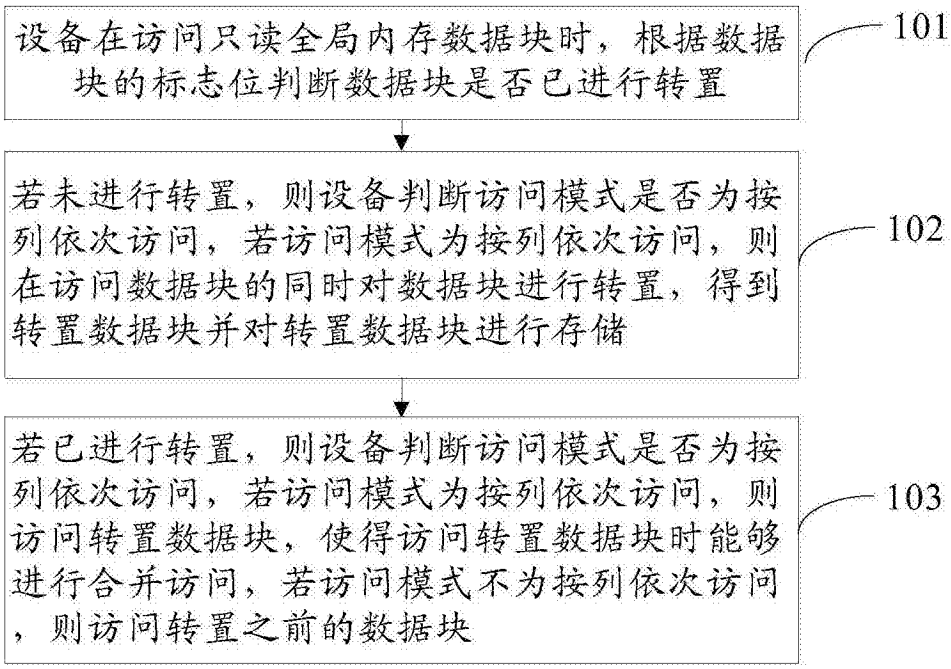

设备在访问只读全局内存数据块时，根据数据块的标志位判断数据块是否已进行转置
101
若未进行转置，则设备判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则在访问数据块的同时对数据块进行转置，得到转置数据块并对转置数据块进行存储
102
若已进行转置，则设备判断访问模式是否为按列依次访问，若访问模式为按列依次访问，则访问转置数据块，使得访问转置数据块时能够进行合并访问，若访问模式不为按列依次访问，则访问转置之前的数据块
103

图 1

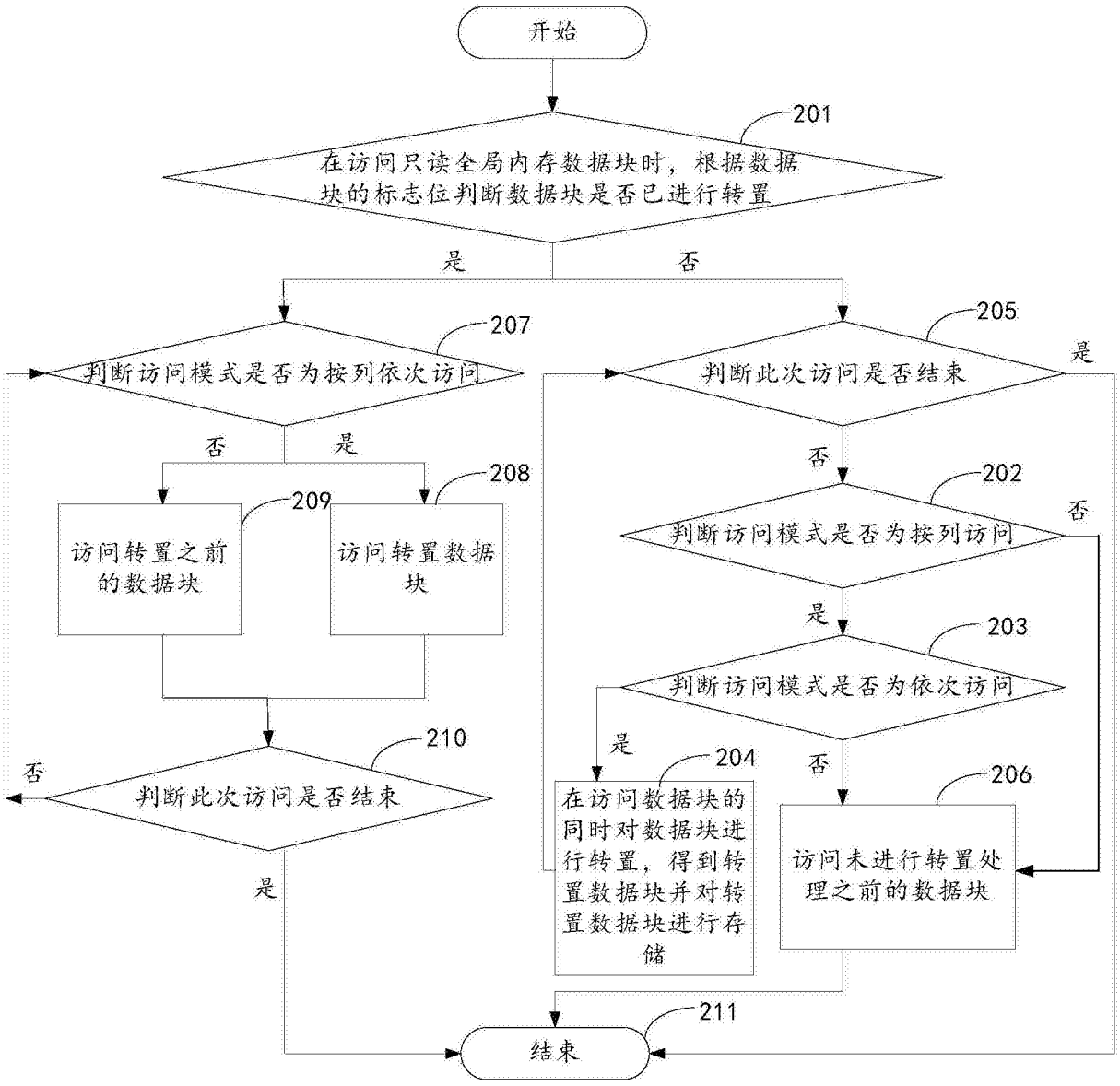

开始
201
在访问只读全局内存数据块时，根据数据块的标志位判断数据块是否已进行转置
是
否
207
判断访问模式是否为按列依次访问
否
是
208
209
访问转置之前的数据块
访问转置数据块
210
否
判断此次访问是否结束
是
205
判断此次访问是否结束
是
否
202
判断访问模式是否为按列访问
否
是
203
判断访问模式是否为依次访问
是
否
204
在访问数据块的同时对数据块进行转置，得到转置数据块并对转置数据块进行存储
206
访问未进行转置处理之前的数据块
211
结束

图 2

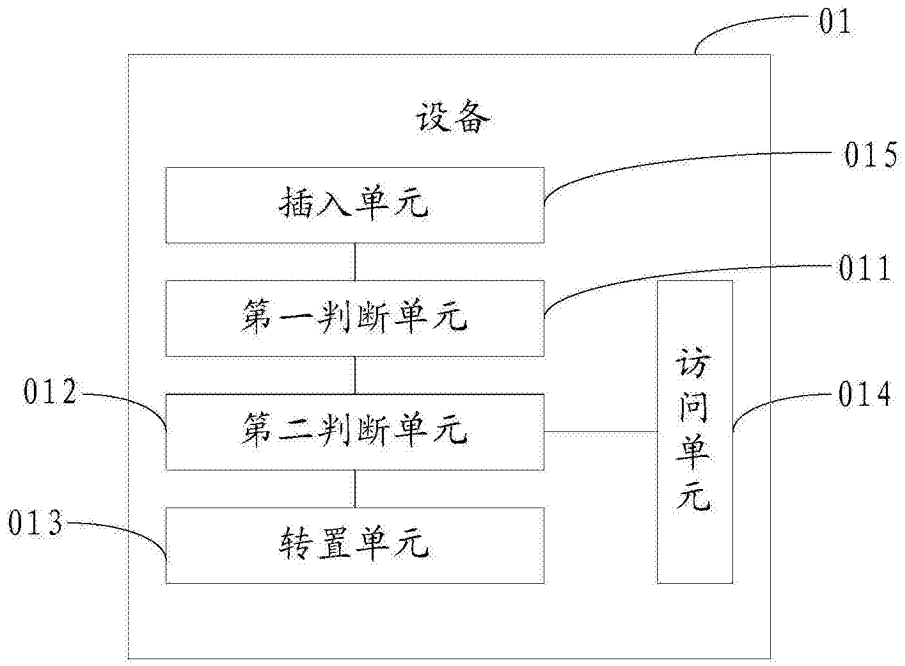

01
设备
015
插入单元
011
第一判断单元
012
第二判断单元
013
转置单元
访问单元
014

图 3

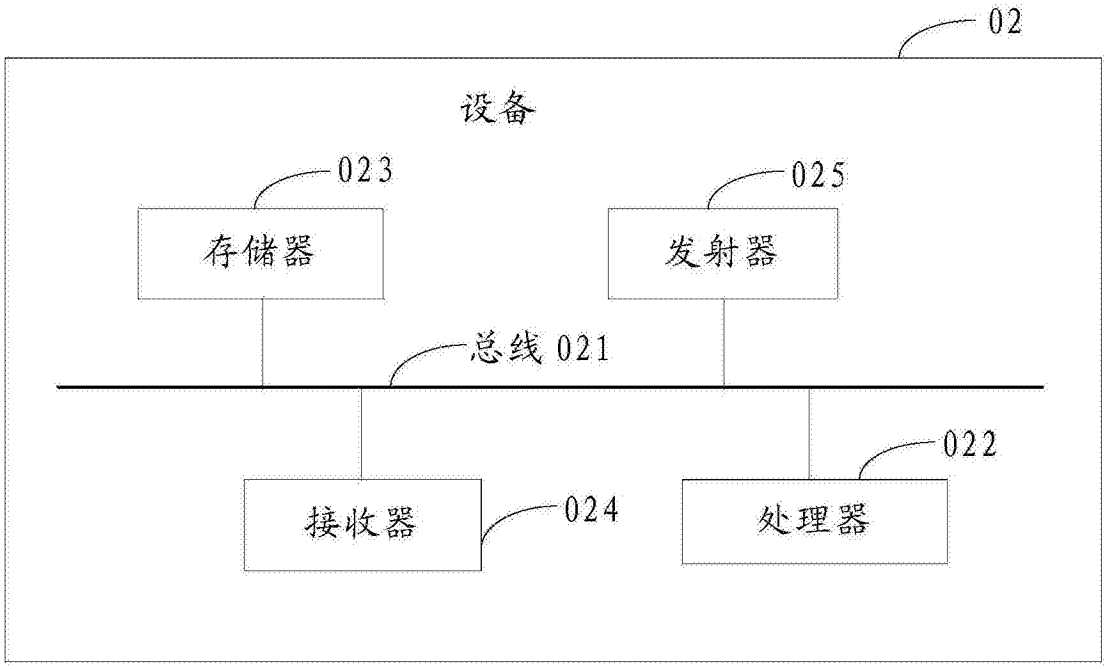

02
设备
023
存储器
025
发射器
总线 021
022
接收器
024
处理器

图 4