(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116484947 B
(45)授权公告日 2023.09.08

(21)申请号 202310744779.7
(22)申请日 2023.06.25
(65)同一申请的已公布的文献号
申请公布号 CN 116484947 A
(43)申请公布日 2023.07.25
(73)专利权人 上海燧原科技有限公司
地址 201306 上海市浦东新区中国(上海)自由贸易试验区临港新片区业盛路188号A-522室
(72)发明人 翟鑫奕 陈禹东 谭磊 田野
(74)专利代理机构 北京品源专利代理有限公司 11332
专利代理师 高艳红
(51)Int.Cl.
G06N 3/10(2006.01)
G06F 15/78(2006.01)
G06F 13/28(2006.01)
G06T 1/40(2006.01)
G06T 1/60(2006.01)

(56)对比文件
CN 116227566 A,2023.06.06
CN 109359732 A,2019.02.19
CN 112947932 A,2021.06.11
CN 113283613 A,2021.08.20
CN 114217807 A,2022.03.22
US 11467811 B1,2022.10.11
CN 1083644 A,1994.03.09
US 5522045 A,1996.05.28
孙佳;柴玉梅.基于同层节点集划分的模糊概念格并行构造算法.《计算机应用与软件》.2016,第33卷(第07期),第261-265,286页.

审查员 蒋亮

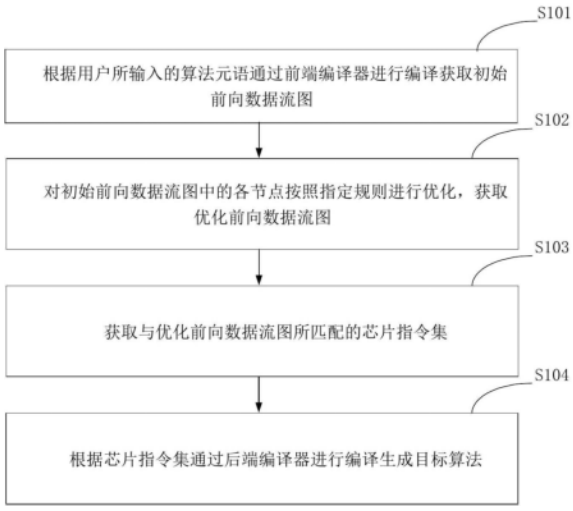
权利要求书3页 说明书10页 附图6页

(54)发明名称

算子的自动生成方法、装置、设备及介质

(57)摘要

本发明公开了一种算子的自动生成方法、装置、设备及介质,包括:根据用户所输入的算法元语通过前端编译器进行编译获取初始前向数据流图;对初始前向数据流图中的各节点按照指定规则进行优化,获取优化前向数据流图;获取与优化前向数据流图所匹配的芯片指令集;根据芯片指令集通过后端编译器进行编译生成目标算法。通过前端编译器根据用户所输入的算法元语进行编译获取的初始前向数据流图后,对所获取的初始前向数据流图进行优化,并获取与优化后的前向数据流图所对应的芯片指令集,并通过后端编译器根据芯片指令集自动生成目标算法,从而在无需用户参数的情况下就可以基于DMA的多级缓存处理器设备自动生成算法。



CN 116484947 B

1.一种算子的自动生成方法,其特征在于,应用于基于DMA的多级缓存处理器设备,包括:

根据用户所输入的算法元语通过前端编译器进行编译获取初始前向数据流图,其中,所述初始前向数据流图中包括计算流节点和数据流节点;

对所述初始前向数据流图中的各节点按照指定规则进行优化,获取优化前向数据流图,其中,所述指定规则包括节点策略标记、节点合并和节点重排;

获取与所述优化前向数据流图所匹配的芯片指令集,其中,所述芯片指令集中包括与所述优化前向数据流图中各节点对应的芯片指令;

根据所述芯片指令集通过后端编译器进行编译生成目标算法;

所述对所述初始前向数据流图中的各节点按照指定规则进行优化,获取优化前向数据流图,包括:对所述初始前向数据流图中的所述计算流节点进行张量化处理,对所述初始前向数据流图中的所述数据流节点进行DMA策略标记,以获取第一优化前向数据流图;

将所述第一优化前向数据流图中具有关联关系的数据流节点采用DMA插入的方式进行多维融合,将所述第一优化前向数据流图中的计算流节点进行保留,以获取第二优化前向数据流图;

确定所述第二优化前向数据流图中各节点的运行时间,并根据所述运行时间将与计算流节点所相邻的数据流节点进行位置重排,以获取第三优化前向数据流图,其中,进行位置重排的所述数据流节点采用预取指令进行标记;

将所述第三优化前向数据流图中指定类型的计算流节点进行合并,将所述第三优化前向数据流图中的数据流节点进行保留,以获取第四优化前向数据流图。

2.根据权利要求1所述的方法,其特征在于,所述对所述初始前向数据流图中的所述数据流节点进行DMA策略标记,包括:

获取预先配置的编译器芯片搜索空间,其中,所述编译器芯片搜索空间中包括各芯片类型和存储结构的对应关系;

确定适配所述前端编译器的当前应用芯片的类型,并通过遍历所述编译器芯片搜索空间确定与所述当前应用芯片所对应的目标存储结构,其中,所述目标存储结构中包括存储层级数量以及存储层级容量;

获取所述初始前向数据流图中各数据流节点位于所述当前应用芯片的位置层级,根据所述位置层级和所述目标存储结构确定针对所述数据流节点的搬运策略;

将所述搬运策略标记在所述初始前向数据流图中的对应数据流节点上,以进行DMA策略标记。

3.根据权利要求2所述的方法,其特征在于,所述根据所述位置层级和所述目标存储结构确定针对所述数据流节点的搬运策略,包括:

确定与所述数据流节点存在逻辑关系的关联计算流节点,并获取所述关联计算流节点的属性信息,其中,所述属性信息包括操作数、张量化信息和类型;

根据所述关联计算流节点的属性信息对所述目标存储结构中的最低存储层级进行切分,以获取计算空间容量;

根据所述位置层级和所述目标存储结构中的所述存储层级数目,确定针对所述数据流节点的搬运方向;

根据所述计算空间容量和所述目标存储结构中的所述存储层级容量，确定针对所述数据流节点在各所述搬运方向上的搬运次数；

将所述搬运方向和所述搬运次数作为针对所述数据流节点的搬运策略。

4.根据权利要求2所述的方法，其特征在于，所述编译器芯片搜索空间中还包括与各芯片类型所对应的空间映射；

其中，所述空间映射中包括节点类型和芯片指令的对应关系。

5.根据权利要求4所述的方法，其特征在于，所述获取与所述优化前向数据流图所匹配的芯片指令集，包括：

提取所述第四优化前向数据流图中的各节点，并根据所提取的节点构建节点集合，其中，所述节点集合中标注有各节点的类型；

根据所述节点集合查询所述空间映射，以获取与所述节点集合中的各节点所对应的芯片指令；

根据所获取的芯片指令构建所述芯片指令集合，其中，所述后端编译器支持所述芯片指令集合。

6.根据权利要求1所述的方法，其特征在于，所述根据所述芯片指令集通过后端编译器进行编译生成目标算法，包括：

根据所述芯片指令通过所述后端编译进行编译获取可执行的二进制文件；

根据所述二进制文件生成所述目标算法。

7.一种算子的自动生成装置，其特征在于，包括：

初始前向数据流图获取模块，用于根据用户所输入的算法元语通过前端编译器进行编译获取初始前向数据流图，其中，所述初始前向数据流图中包括计算流节点和数据流节点；

优化前向数据流图获取模块，用于对所述初始前向数据流图中的各节点按照指定规则进行优化，获取优化前向数据流图，其中，所述指定规则包括节点策略标记、节点合并和节点重排；

芯片指令集获取模块，用于获取与所述优化前向数据流图所匹配的芯片指令集，其中，所述芯片指令集中包括与所述优化前向数据流图中各节点对应的芯片指令；

目标算法生成模块，用于根据所述芯片指令集通过后端编译器进行编译生成目标算法；

所述优化前向数据流图获取模块，用于对所述初始前向数据流图中的所述计算流节点进行张量化处理，对所述初始前向数据流图中的所述数据流节点进行DMA策略标记，以获取第一优化前向数据流图；

将所述第一优化前向数据流图中具有关联关系的数据流节点采用DMA插入的方式进行多维融合，将所述第一优化前向数据流图中的计算流节点进行保留，以获取第二优化前向数据流图；

确定所述第二优化前向数据流图中各节点的运行时间，并根据所述运行时间将与计算流节点所相邻的数据流节点进行位置重排，以获取第三优化前向数据流图，其中，进行位置重排的所述数据流节点采用预取指令进行标记；

将所述第三优化前向数据流图中指定类型的计算流节点进行合并，将所述第三优化前向数据流图中的数据流节点进行保留，以获取第四优化前向数据流图。

8.一种计算机设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,其特征在于,所述处理器执行所述程序时实现如权利要求1-6中任一项所述的方法。

9.一种计算机可执行指令的存储介质,其上存储有计算机程序,其特征在于,该程序被处理器执行时实现如权利要求1-6中任一项所述方法。

算子的自动生成方法、装置、设备及介质

技术领域

[0001] 本发明实施例涉及数据处理技术领域，尤其涉及一种算子的自动生成方法、装置、设备及介质。

背景技术

[0002] 在开发深度学习模型中需要设计对应的算法模型，由于深度学习发展的日新月异，算子层出不穷，如何能够平衡算法和开发效率变得尤为重要。

[0003] 目前绝大多数的深度学习框架还停留在人工优化对应算子的阶段，这种逐个实现算子的方式，定义时间长、维护较为困难，并且通常需要依赖专业深度学习资深算法工程师，并且主要针对的是带缓存芯片架构，例如，CPU或GPU的优化策略，因此并不能实现对基于直接内存访问(Direct Memory Access，DMA)的多级缓存芯片架构下的算法生成。

发明内容

[0004] 本发明实施例提供一种算子的自动生成方法、装置、设备及介质，以实现基于DMA的多级缓存处理器架构下算法的自动生成。

[0005] 第一方面，本发明实施例提供了一种算子的自动生成方法，包括：根据用户所输入的算法元语通过前端编译器进行编译获取初始前向数据流图，其中，所述初始前向数据流图中包括计算流节点和数据流节点；

[0006] 对所述初始前向数据流图中的各节点按照指定规则进行优化，获取优化前向数据流图，其中，所述指定规则包括节点策略标记、节点合并和节点重排；

[0007] 获取与所述优化前向数据流图所匹配的芯片指令集，其中，所述芯片指令集中包括与所述优化前向数据流图中各节点对应的芯片指令；

[0008] 根据所述芯片指令集通过后端编译器进行编译生成目标算法。

[0009] 第二方面，本发明实施例提供了一种算子的自动生成装置，包括：

[0010] 初始前向数据流图获取模块，用于根据用户所输入的算法元语通过前端编译器进行编译获取初始前向数据流图，其中，所述初始前向数据流图中包括计算流节点和数据流节点；

[0011] 优化前向数据流图获取模块，用于对所述初始前向数据流图中的各节点按照指定规则进行优化，获取优化前向数据流图，其中，所述指定规则包括节点策略标记、节点合并和节点重排；

[0012] 芯片指令集获取模块，用于获取与所述优化前向数据流图所匹配的芯片指令集，其中，所述芯片指令集中包括与所述优化前向数据流图中各节点对应的芯片指令；

[0013] 目标算法生成模块，用于根据所述芯片指令集通过后端编译器进行编译生成目标算法。

[0014] 第三方面，本发明实施例提供了一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现如上所述的方

法。

[0015] 第四方面,本发明实施例提供了一种计算机可执行指令的存储介质,其上存储有计算机程序,该程序被处理器执行时实现如上所述的方法。

[0016] 本申请通过前端编译器根据用户所输入的算法元语进行编译获取的初始前向数据流图后,对所获取的初始前向数据流图进行优化,并获取与优化后的前向数据流图所对应的芯片指令集,并通过后端编译器根据芯片指令集自动生成目标算法,从而在无需用户参数的情况下就可以基于DMA的多级缓存处理器设备自动生成算法。

附图说明

[0017] 图1是本发明实施例一提供的一种算子的自动生成方法的流程图;

[0018] 图2是本发明实施例一提供的一种算子的自动生成方法的整体示意图;

[0019] 图3是本发明实施例一提供的基于2的幂次逼近的DMA自动切分示意图;

[0020] 图4是本发明实施例一提供的采用DMA插入的方式进行多维融合的示意图;

[0021] 图5是本发明实施例二提供的一种算子的自动生成方法的流程图;

[0022] 图6是本发明实施例三提供的一种算子的自动生成装置的结构示意图;

[0023] 图7是本发明实施例四提供的一种计算机设备的结构示意图。

具体实施方式

[0024] 下面结合附图和实施例对本发明作进一步的详细说明。可以理解的是,此处所描述的具体实施例仅仅用于解释本发明,而非对本发明的限定。另外还需要说明的是,为了便于描述,附图中仅示出了与本发明相关的部分而非全部结构。

[0025] 实施例一

[0026] 图1为本发明实施例一提供的一种算子的自动生成方法的流程图,本实施例可适用于基于DMA的多级缓存处理器设备进行自动算法生成的情况,该方法可以由算子的自动生成装置来执行,该该装置可以由软件和/或硬件的方式实现,算子的自动生成方法包括:

[0027] 步骤S101,根据用户所输入的算法元语通过前端编译器进行编译获取初始前向数据流图。

[0028] 具体的说,本实施方式中用户可以预先定义一套适合基于DMA的多级缓存处理器架构下编译器的算法元语,并且用户所定义的算法元语可以包括计算流元语和数据流元语。其中,计算流元语可以包括基本运算或者算子循环变量等,例如,加减乘除等;而数据流元语包括标记数据在片上存储的位置、数据搬运、张量切分或者向量化等,例如,将低速存储—中低速存储—高速存储器之间的多级存储访问搬运策略等,并且数据流元语需尤其适配基于DMA的多级缓存处理器架构的要求。

[0029] 其中,基于DMA的多级缓存处理器设备中的前端编译器会接收用户所定义的算法元语,并采用基于循环优化的文本分析策略,即loop stmt策略,将数据流元语和计算流元语编译为初始前向数据流图,而在初始前向数据流图中包括计算流节点和数据流节点,而关于loop stmt策略的具体原理并不是本申请的重点,因此本实施方式中不再进行赘述。

[0030] 需要说明的是,本发明的实施方式相比于其他的算子生成框架而言,拓展了特定架构方案下的算子生成策略、算子描述元语等,补足了在基于DMA的多级缓存芯片架构下,

没有算子生成策略的这一短板。可以定义描述各种基于DMA的多级缓存架构的芯片特性，根据存储效率将芯片中的不同存储模块命名为L1、L2、L3等多个级别，例如，定义高速存储级别为L1或者任何其他名字，从而可以将计算流的数据标记在存储架构上的特定位置。同时针对不同的芯片搬运方式，可以采取我们定义的DMA策略，将不同等级之间搬运简单理解为赋值，而无需考虑其跨存储等级的问题，将DMA的搬运流程交由编译器来分配以及管理。

[0031] 步骤S102，对初始前向数据流图中的各节点按照指定规则进行优化，获取优化前向数据流图。

[0032] 可选的，对初始前向数据流图中的各节点按照指定规则进行优化，获取优化前向数据流图，包括：对初始前向数据流图中的计算流节点进行张量化处理，对初始前向数据流图中的数据流节点进行DMA策略标记，以获取第一优化前向数据流图；将第一优化前向数据流图中具有关联关系的数据流节点采用DMA插入的方式进行多维融合，将第一优化前向数据流图中的计算流节点进行保留，以获取第二优化前向数据流图；确定第二优化前向数据流图中各节点的运行时间，并根据运行时间将与计算流节点所相邻的数据流节点进行位置重排，以获取第三优化前向数据流图，其中，进行位置重排的数据流节点采用预取指令进行标记；将第三优化前向数据流图中指定类型的计算流节点进行合并，将第三优化前向数据流图中的数据流节点进行保留，以获取第四优化前向数据流图。

[0033] 可选的，对初始前向数据流图中的数据流节点进行DMA策略标记，包括：获取预先配置的编译器芯片搜索空间，其中，编译器芯片搜索空间中包括各芯片类型和存储结构的对应关系；确定适配前端编译器的当前应用芯片的类型，并通过遍历编译器芯片搜索空间确定与当前应用芯片所对应的目标存储结构，其中，目标存储结构中包括存储层级数量以及存储层级容量；获取初始前向数据流图中各数据流节点位于当前应用芯片的位置层级，根据位置层级和目标存储结构确定针对数据流节点的搬运策略；将搬运策略标记在初始前向数据流图中的对应数据流节点上，以进行DMA策略标记。

[0034] 可选的，根据位置层级和目标存储结构确定针对数据流节点的搬运策略，包括：确定与数据流节点存在逻辑关系的关联计算流节点，并获取关联计算流节点的属性信息，其中，属性信息包括操作数、张量化信息和类型；根据关联计算流节点的属性信息对目标存储结构中的最低存储层级进行切分，以获取计算空间容量；根据位置层级和目标存储结构中的存储层级数目，确定针对数据流节点的搬运方向；根据计算空间容量和目标存储结构中的存储层级容量，确定针对数据流节点在各搬运方向上的搬运次数；将搬运方向和搬运次数作为针对数据流节点的搬运策略。

[0035] 具体的说，如图2所示为本实施方式中一种算子的自动生成方法的整体示意图，前端编译器在根据用户所输入的算法元语进行编译获取到初始前向流程图之后，由于在初始前向数据流程图中包括多个数据流节点和计算流节点，此时前端编译器会对数据流节点和计算流节点分别进行处理，以获取第一优化前向数据流图。其中，针对计算流节点会执行如图2中的编译优化a步骤，对计算流节点进行张量化处理，以获取并行化张量化格式的计算流节点。例如，计算流节点为y=a+b的加法计算，则会对各参数分别进行张量化处理获得张量化结果[Y]=[A]+[B]，其中，[A]=[a a....a]，[B]=[b b....b]，每个参数的并行度是相同的，而关于每个参数的并行度可以由用户进行确定，本实施方式中并不对其进行限定。

[0036] 另外，针对初始前向数据流图中的数据流节点会进行DMA策略标记，在进行DMA策

略标记时，需要获取预先配置的编译器芯片搜索空间，在搜索空间中包括各芯片类型和存储结构的对应关系，如下表1为搜索空间的示例：

[0037] 表1

[0038]

芯片类型	存储结构
a 类型	存储层数 2 级，L1 层级容量 1024k，L2 层级容量 10M
b 类型	存储层数 3 级，L1 层级容量 512k，L2 层级容量 512M，L3 层级容量 8G
....	

[0039] 其中，由于篇幅限制，表1中仅是以两个芯片类型所对应的存储结构进行举例说明，而并不对每个芯片类型所对应的存储结构的具体形式进行限定。因此在确定出当前适配前端编译器的当前应用芯片的类型为b类型后，通过遍历编译器芯片搜索空间可以确定出与当前所对应的目标存储结构为：存储层数3级，L1层级容量512k，L2层级容量30M，L3层级容量10G，并且每个层级所支持的搬运能力也是不相同的。由于之前已经获取到了初始前向数据流图中各数据流节点位于当前应用芯片的位置层级，则可以确定出与数据流节点存在逻辑关系的关联计算流节点，并获取关联计算流节点的属性信息，并根据关联计算流节点的属性信息对目标存储结构中的最低存储层级进行切分，以获取计算空间容量。如图3所示，为基于2的幂次逼近的DMA自动切分示意图，当确定与数据流节点F所对应的关联计算流节点为一个W(x,y)=A(x,y)+B(x,y)，并且针对关联计算流节点的约束条件为：maximizef(x,y) st.0≤x≤M,0≤y≤M,3*x*y*sizeof(float)≤512kb, y mod vectorsize=0, vectorsize=32B，则如下表2所示为该关联计算流节点的属性信息：

[0040] 表2

[0041]

函数名	形状类型	最终大小
A	（M，N），Float(32)	
B	（M，N），Float(32)	
C	（M，N），Float(32)	

[0042] 其中，针对该关联计算流节点可以确定包括三个操作数，张量并行化信息为32，类型为float，而最终大小则是该关联计算流节点中单个参数所需要的计算空间容量，则可以采用如下公式(1)中基于2的幂次逼近进行计算，获取单个参数所需要的计算空间容量：

[0043] $2^5*2^n \leqslant 512kb/3 \leqslant 2^5*2^{n+1}$ (1)

[0044] 其中，通过对公式1进行n值求解，可以获取到n=7，则将2^7=128作为单个参数所需要的计算空间容量，从而上述表2中针对函数名A、函数名B和函数名C所对应的最终大小分别填写的是128kb，则针对关联计算流节点的的总的计算空间容量为128*3=382kb。如图3所示为基于2的幂次逼近的DMA自动切分示意图，当确定目标存储结构中的最低存储层级L1为512K时，则将L1切分成384kb的当前算子计算空间，以及128kb的子图并行预留空间。当确定L1=512k，L2=512M，L3=8G时，由于已知数据流节点位于L1层级，则可以获取由L2搬运到L1所需的搬运次数为512M/384k=1365次，由L3搬运到L2所需的搬运次数为8G/1365*384k=16次，从而将上述所获取的搬运方向和搬运次数作为该数据流节点F的搬运策略，并将上述所获取的搬运策略标记在初始前向数据流图中的数据流节点F上，以进行DMA策略标记。本实施方式中根据张量化处理的计算流节点以及进行DMA策略标记的数据流节点获取到第一优化

前向数据流图。

[0045] 值得一提的是，在获取到第一优化前向数据流图之后，执行图2中的优化b以对流程图中具有关联关系的数据流节点采用DM插入的方式进行多维融合，如图4所示为采用DMA插入的方式进行多维融合的示意图，从而通过图4中所示的技术，将多个维度的DMA进行合并，比如 DMAL2->L1 + DMAL2->L1，对于上下两个节点，如果遇见了相同的DMA的等级和位置，我们就可以分析其中的位置及大小，从而将这些DMA操作合并，合并之后的操作，所需要的执行次数就从2次变为了1次，从而节省了执行的速度。另外，上述的多维融合仅针对的是数据流节点，而针对第一优化前向数据流图中的计算流节点则会对应的进行保留，从而获取第二优化前向数据流节点，由于对数据流节点进行了融合，因此第二优化前向数据流节点相对于第一优化前向数据流节点来说，节点数目是减少的。

[0046] 具体的说，在获取到第二优化前向数据流节点之后，由于计算流和数据流流是分离存在的，并且可以同时进行操作，因此需要将访问的延时，隐藏在计算之后，具体是通过修改第二优化前向数据流图中数据流节点的位置，以获取第三优化前向数据流图。具体是通过执行图2中的编译优化c来进行位置重排。

[0047] 例如，当确定第二优化前向数据流图中包含四个节点，并且分布顺序为：数据流节点m，计算流节点n，数据流节点p，计算流节点q，其中，节点m、n、p和q的运行时间分别为1s、5s、2s和2s，如果按照原有的顺序执行，则总共占用的运行时间为1+5+2+2=10s。由于计算流和数据流是分离存在的，因此数据流节点p的计算与计算流节点n无关，所以可以将数据流节点p放置到计算流节点n之前，获取到新的节点分布顺序：数据流节点m，数据流节点p，计算流节点n，计算流节点q，即数据流节点p在计算执行的时候，计算流节点n也在同步进行执行，从而会有2s的时间重叠，因此可以获取位置重排后所获取的总共占用的运行时间为1+5+2=8s，本实施方式中会根据节点的位置重排获取到第三优化前向数据流图，并且会将进行位置重排的数据流节点采用预取指令进行标记，例如，针对上述的数据流节点p采用预取指令进行标记。

[0048] 其中，在获取到第三优化前向数据流图之后，会执行图2中的图优化d操作，以对指定类型的计算流节点进行合并，例如，将一个加法和乘法进行合并，形成一个乘加算法，同时将第三优化前向数据流图中的数据流节点进行保留，以获取第四优化前向数据流图。因此，前端编译器在获取到初始前向数据流图后通过采用节点策略标记、节点合并和节点重排等方式可以实现对初始前向数据流图的简化，以便于后续后端编译器自动生成更加简略的算法。

[0049] 步骤S103，获取与优化前向数据流图所匹配的芯片指令集。

[0050] 可选的，编译器芯片搜索空间中还包括与各芯片类型所对应的空间映射；其中，空间映射中包括节点类型和芯片指令的对应关系。

[0051] 可选的，获取与优化前向数据流图所匹配的芯片指令集，包括：提取第四优化前向数据流图中的各节点，并根据所提取的节点构建节点集合，其中，节点集合中标注有各节点的类型；根据节点集合查询空间映射，以获取与节点集合中的各节点所对应的芯片指令；根据所获取的芯片指令构建芯片指令集合，其中，后端编译器支持芯片指令集合。

[0052] 具体的说，在获取到第四优化前向数据流图之后，会根据第四优化前向数据流图中的各节点构建节点集合，并且在节点集合中标有各节点的类型。由于在之前在图1所示的

搜索空间中还包括各类芯片所对应的空间映射，即节点类型和芯片指令的对应关系，例如，节点类型AddOp对应的芯片指令为120.vadd，并且该芯片指令能够被后端编译器识别，因此通过查询上述的空间映射可以获取到与节点集合所对应的芯片指令集合，芯片指令集中包括与第四优化前向数据流图中各节点对应的芯片指令。

[0053] 步骤S104，根据芯片指令集通过后端编译器进行编译生成目标算法。

[0054] 可选的，根据芯片指令集通过后端编译器进行编译生成目标算法，包括：根据芯片指令通过后端编译进行编译获取可执行的二进制文件；根据二进制文件生成目标算法。

[0055] 具体的说，由于上述所获取的芯片指令集中包括能够被后端编译器所能够识别的芯片指令，因此后端编译器会根据芯片指令集进行二进制的编译，以生成可执行的二进制文件，并且根据上述的二进制文件能够自动生成目标算法。因此本实施方式中通过定义特定元语，支持DMA的I/O操作，使用编译器替代传统缓存架构下硬件设计所做的工作，为突破I/O传输的带宽瓶颈，通过预取策略，减小在计算流过程中I/O的延迟，提升传输的带宽；将计算和传输并行，提升最终生成代码的运行效率。因此对于算子开发人员而言，使用本发明提出的元语以及策略后，这些方案将会变得透明，这极大简化了开发算子所需要的门槛，在确保算子开执行速度的前提下，提升了算子开发的效率。

[0056] 本申请通过前端编译器根据用户所输入的算法元语进行编译获取的初始前向数据流图后，对所获取的初始前向数据流图进行优化，并获取与优化后的前向数据流图所对应的芯片指令集，并通过后端编译器根据芯片指令集自动生成目标算法，从而在无需用户参数的情况下就可以基于DMA的多级缓存处理器设备自动生成算法。

[0057] 实施例二

[0058] 图5为本发明实施例二提供的一种算子的自动生成方法的流程图，本实施例以上述实施例为基础，在根据芯片指令集通过后端编译器进行编译生成目标算法之后，还包括对目标算法进行校验，具体包括：

[0059] 步骤S201，根据用户所输入的算法元语通过前端编译器进行编译获取初始前向数据流图。

[0060] 步骤S202，对初始前向数据流图中的各节点按照指定规则进行优化，获取优化前向数据流图。

[0061] 可选的，对初始前向数据流图中的各节点按照指定规则进行优化，获取优化前向数据流图，包括：对初始前向数据流图中的计算流节点进行张量化处理，对初始前向数据流图中的数据流节点进行DMA策略标记，以获取第一优化前向数据流图；将第一优化前向数据流图中具有关联关系的数据流节点采用DMA插入的方式进行多维融合，将第一优化前向数据流图中的计算流节点进行保留，以获取第二优化前向数据流图；确定第二优化前向数据流图中各节点的运行时间，并根据运行时间将与计算流节点所相邻的数据流节点进行位置重排，以获取第三优化前向数据流图，其中，进行位置重排的数据流节点采用预取指令进行标记；将第三优化前向数据流图中指定类型的计算流节点进行合并，将第三优化前向数据流图中的数据流节点进行保留，以获取第四优化前向数据流图。

[0062] 可选的，对初始前向数据流图中的数据流节点进行DMA策略标记，包括：获取预先配置的编译器芯片搜索空间，其中，编译器芯片搜索空间中包括各芯片类型和存储结构的对应关系；确定适配前端编译器的当前应用芯片的类型，并通过遍历编译器芯片搜索空间

确定与当前应用芯片所对应的目标存储结构，其中，目标存储结构中包括存储层级数量以及存储层级容量；获取初始前向数据流图中各数据流节点位于当前应用芯片的位置层级，根据位置层级和目标存储结构确定针对数据流节点的搬运策略；将搬运策略标记在初始前向数据流图中的对应数据流节点上，以进行DMA策略标记。

[0063] 可选的，根据位置层级和目标存储结构确定针对数据流节点的搬运策略，包括：确定与数据流节点存在逻辑关系的关联计算流节点，并获取关联计算流节点的属性信息，其中，属性信息包括操作数、张量化信息和类型；根据关联计算流节点的属性信息对目标存储结构中的最低存储层级进行切分，以获取计算空间容量；根据位置层级和目标存储结构中的存储层级数目，确定针对数据流节点的搬运方向；根据计算空间容量和目标存储结构中的存储层级容量，确定针对数据流节点在各搬运方向上的搬运次数；将搬运方向和搬运次数作为针对数据流节点的搬运策略。

[0064] 步骤S203，获取与优化前向数据流图所匹配的芯片指令集。

[0065] 可选的，编译器芯片搜索空间中还包括与各芯片类型所对应的空间映射；其中，空间映射中包括节点类型和芯片指令的对应关系。

[0066] 可选的，获取与优化前向数据流图所匹配的芯片指令集，包括：提取第四优化前向数据流图中的各节点，并根据所提取的节点构建节点集合，其中，节点集合中标注有各节点的类型；根据节点集合查询空间映射，以获取与节点集合中的各节点所对应的芯片指令；根据所获取的芯片指令构建芯片指令集合，其中，后端编译器支持芯片指令集合。

[0067] 步骤S204，根据芯片指令集通过后端编译器进行编译生成目标算法。

[0068] 可选的，根据芯片指令集通过后端编译器进行编译生成目标算法，包括：根据芯片指令通过后端编译进行编译获取可执行的二进制文件；根据二进制文件生成目标算法。

[0069] 步骤S205，对目标算法进行校验。

[0070] 具体的说，本实施方式中在生成目标算法之后，会对目标算法进行校验，具体是检测目标算法所对应的二进制文件是否存在明显错误的地方，例如存在乱码或者语法错误的地方，当确定存在明显错误的地方时，则确定与二进制文件所对应的目标算法是无效算法，因此校验失败。

[0071] 需要说明的是，当确定校验失败的情况下，会生成校验失败提示，例如，“当前目标算法无效，请注意调整”出现上述的原因可能是前端编译器的优化过程无效，由于所获取的优化前向数据流图错误所造成的，也可能是后端编译器根据空间映射所获取的芯片指令错误所造成的，本实施方式中并不对其进行限定。并且当确定校验失败时，通过将校验失败提示信息进行展示，可以及时用户对前端编译器、后端编译器或者软件进行检修，以进一步提高算子的自动生成效率和准确性。

[0072] 本申请通过前端编译器根据用户所输入的算法元语进行编译获取的初始前向数据流图后，对所获取的初始前向数据流图进行优化，并获取与优化后的前向数据流图所对应的芯片指令集，并通过后端编译器根据芯片指令集自动生成目标算法，从而在无需用户参数的情况下就可以基于DMA的多级缓存处理器设备自动生成算法。并且当确定校验失败时，通过将校验失败提示信息进行展示，可以及时用户对前端编译器、后端编译器或者软件进行检修，以进一步提高算子的自动生成效率和准确性。

[0073] 实施例三

[0074] 图6为本发明实施例三提供的一种算子的自动生成装置的结构示意图，该装置可以执行上述各实施例中涉及到的算子的自动生成方法。该装置可采用软件和/或硬件的方式实现，如图6所示，算子的自动生成装置具体包括：初始前向数据流图获取模块310、优化前向数据流图获取模块320、芯片指令集获取模块330和目标算法生成模块340。

[0075] 初始前向数据流图获取模块310，用于根据用户所输入的算法元语通过前端编译器进行编译获取初始前向数据流图，其中，初始前向数据流图中包括计算流节点和数据流节点；

[0076] 优化前向数据流图获取模块320，用于对初始前向数据流图中的各节点按照指定规则进行优化，获取优化前向数据流图，其中，指定规则包括节点策略标记、节点合并和节点重排；

[0077] 芯片指令集获取模块330，用于获取与优化前向数据流图所匹配的芯片指令集，其中，芯片指令集中包括与优化前向数据流图中各节点对应的芯片指令；

[0078] 目标算法生成模块340，用于根据芯片指令集通过后端编译器进行编译生成目标算法。

[0079] 可选的，优化前向数据流图获取模块，包括：第一优化前向数据流图获取单元，用于对初始前向数据流图中的计算流节点进行张量化处理，对初始前向数据流图中的数据流节点进行DMA策略标记，以获取第一优化前向数据流图；

[0080] 第二优化前向数据流图获取单元，用于将第一优化前向数据流图中具有关联关系的数据流节点采用DMA插入的方式进行多维融合，将第一优化前向数据流图中的计算流节点进行保留，以获取第二优化前向数据流图；

[0081] 第三优化前向数据流图获取单元，用于确定第二优化前向数据流图中各节点的运行时间，并根据运行时间将与计算流节点所相邻的数据流节点进行位置重排，以获取第三优化前向数据流图，其中，进行位置重排的数据流节点采用预取指令进行标记；

[0082] 第四优化前向数据流图获取单元，用于将第三优化前向数据流图中指定类型的计算流节点进行合并，将第三优化前向数据流图中的数据流节点进行保留，以获取第四优化前向数据流图。

[0083] 可选的，第一优化前向数据流图获取单元，用于获取预先配置的编译器芯片搜索空间，其中，编译器芯片搜索空间中包括各芯片类型和存储结构的对应关系；

[0084] 确定适配前端编译器的当前应用芯片的类型，并通过遍历编译器芯片搜索空间确定与当前应用芯片所对应的目标存储结构，其中，目标存储结构中包括存储层级数量以及存储层级容量；

[0085] 获取初始前向数据流图中各数据流节点位于当前应用芯片的位置层级，根据位置层级和目标存储结构确定针对数据流节点的搬运策略；

[0086] 将搬运策略标记在初始前向数据流图中的对应数据流节点上，以进行DMA策略标记。

[0087] 可选的，第一优化前向数据流图获取单元，还用于确定与数据流节点存在逻辑关系的关联计算流节点，并获取关联计算流节点的属性信息，其中，属性信息包括操作数、张量化信息和类型；

[0088] 根据关联计算流节点的属性信息对目标存储结构中的最低存储层级进行切分，以

获取计算空间容量；

[0089] 根据位置层级和目标存储结构中的存储层级数目，确定针对数据流节点的搬运方向；

[0090] 根据计算空间容量和目标存储结构中的存储层级容量，确定针对数据流节点在各搬运方向上的搬运次数；

[0091] 将搬运方向和搬运次数作为针对数据流节点的搬运策略。

[0092] 可选的，编译器芯片搜索空间中还包括与各芯片类型所对应的空间映射；

[0093] 其中，空间映射中包括节点类型和芯片指令的对应关系。

[0094] 可选的，芯片指令集获取模块，用于提取第四优化前向数据流图中的各节点，并根据所提取的节点构建节点集合，其中，节点集合中标注有各节点的类型；

[0095] 根据节点集合查询空间映射，以获取与节点集合中的各节点所对应的芯片指令；

[0096] 根据所获取的芯片指令构建芯片指令集合，其中，后端编译器支持芯片指令集合。

[0097] 可选的，目标算法生成模块，用于根据芯片指令通过后端编译进行编译获取可执行的二进制文件；

[0098] 根据二进制文件生成目标算法。

[0099] 实施例四

[0100] 图7为本发明实施例四提供的一种计算机设备的结构示意图，如图7所示，该计算机设备包括处理器610、存储器620、输入装置630和输出装置640；计算机设备中处理器610的数量可以是一个或多个，图6中以一个处理器610为例；计算机设备中的处理器610、存储器620、输入装置630和输出装置640可以通过总线或其他方式连接，图7中以通过总线连接为例。

[0101] 存储器620作为一种计算机可读存储介质，可用于存储软件程序、计算机可执行程序以及模块，如本发明实施例中的算子的自动生成方法对应的程序指令/模块。处理器610通过运行存储在存储器620中的软件程序、指令以及模块，从而执行计算机设备的各种功能应用以及数据处理，即实现上述的算子的自动生成方法。

[0102] 算子的自动生成方法，应用于基于DMA的多级缓存处理器设备，包括：

[0103] 根据用户所输入的算法元语通过前端编译器进行编译获取初始前向数据流图，其中，初始前向数据流图中包括计算流节点和数据流节点；

[0104] 对初始前向数据流图中的各节点按照指定规则进行优化，获取优化前向数据流图，其中，指定规则包括节点策略标记、节点合并和节点重排；

[0105] 获取与优化前向数据流图所匹配的芯片指令集，其中，芯片指令集中包括与优化前向数据流图中各节点对应的芯片指令；

[0106] 根据芯片指令集通过后端编译器进行编译生成目标算法。

[0107] 存储器620可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序；存储数据区可存储根据终端的使用所创建的数据等。此外，存储器620可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实例中，存储器620可进一步包括相对于处理器610远程设置的存储器，这些远程存储器可以通过网络连接至计算机设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0108] 输入装置630可用于接收输入的数字或字符信息,以及产生与计算机设备的用户设置以及功能控制有关的键信号输入。输出装置640可包括显示屏等显示设备。

[0109] 实施例五

[0110] 本发明实施例五还提供一种包含计算机可执行指令的存储介质,计算机可执行指令在由计算机处理器执行时用于执行一种算子的自动生成方法;

[0111] 算子的自动生成方法,应用于基于DMA的多级缓存处理器设备,包括:

[0112] 根据用户所输入的算法元语通过前端编译器进行编译获取初始前向数据流图,其中,初始前向数据流图中包括计算流节点和数据流节点;

[0113] 对初始前向数据流图中的各节点按照指定规则进行优化,获取优化前向数据流图,其中,指定规则包括节点策略标记、节点合并和节点重排;

[0114] 获取与优化前向数据流图所匹配的芯片指令集,其中,芯片指令集中包括与优化前向数据流图中各节点对应的芯片指令;

[0115] 根据芯片指令集通过后端编译器进行编译生成目标算法。

[0116] 当然,本发明实施例所提供的一种包含计算机可执行指令的存储介质,其计算机可执行指令不限于如上的方法操作,还可以执行本发明任意实施例所提供的算子的自动生成方法中的相关操作。

[0117] 通过以上关于实施方式的描述,所属领域的技术人员可以清楚地了解到,本发明可借助软件及必需的通用硬件来实现,当然也可以通过硬件实现,但很多情况下前者是更佳的实施方式。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可以以软件产品的形式体现出来,该计算机软件产品可以存储在计算机可读存储介质中,如计算机的软盘、只读存储器(Read-Only Memory, ROM)、随机存取存储器(Random Access Memory, RAM)、闪存(FLASH)、硬盘或光盘等,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的方法。

[0118] 值得注意的是,上述算子的自动生成装置的实施例中,所包括的各个单元和模块只是按照功能逻辑进行划分的,但并不局限于上述的划分,只要能够实现相应的功能即可;另外,各功能单元的具体名称也只是为了便于相互区分,并不用于限制本发明的保护范围。

[0119] 注意,上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解,本发明不限于这里所述的特定实施例,对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此,虽然通过以上实施例对本发明进行了较为详细的说明,但是本发明不仅仅限于以上实施例,在不脱离本发明构思的情况下,还可以包括更多其他等效实施例,而本发明的范围由所附的权利要求范围决定。

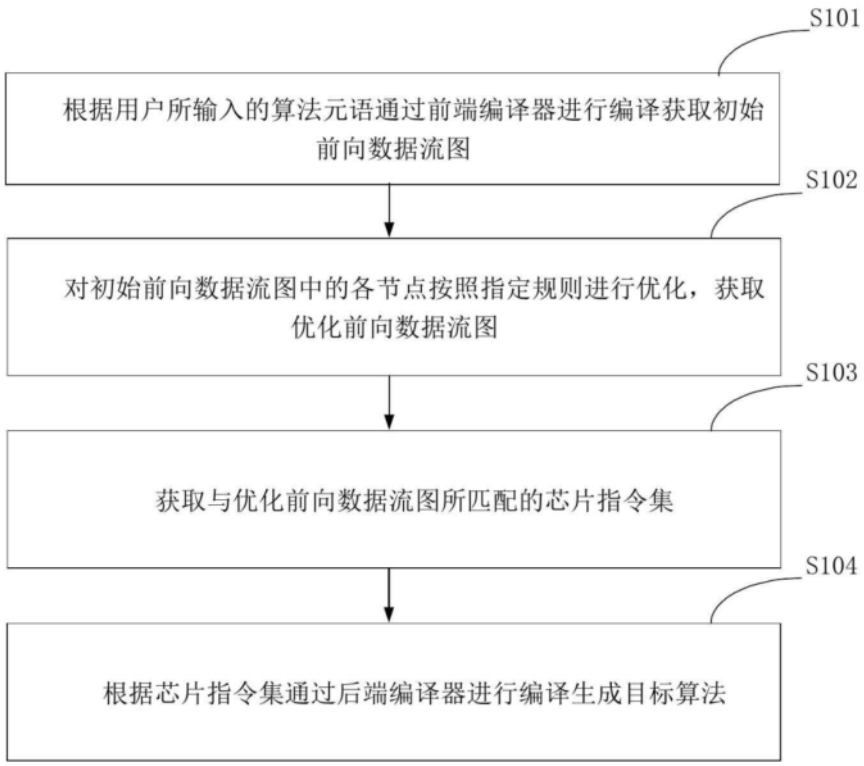

S101
根据用户所输入的算法元语通过前端编译器进行编译获取初始前向数据流图
S102
对初始前向数据流图中的各节点按照指定规则进行优化，获取优化前向数据流图
S103
获取与优化前向数据流图所匹配的芯片指令集
S104
根据芯片指令集通过后端编译器进行编译生成目标算法

图1

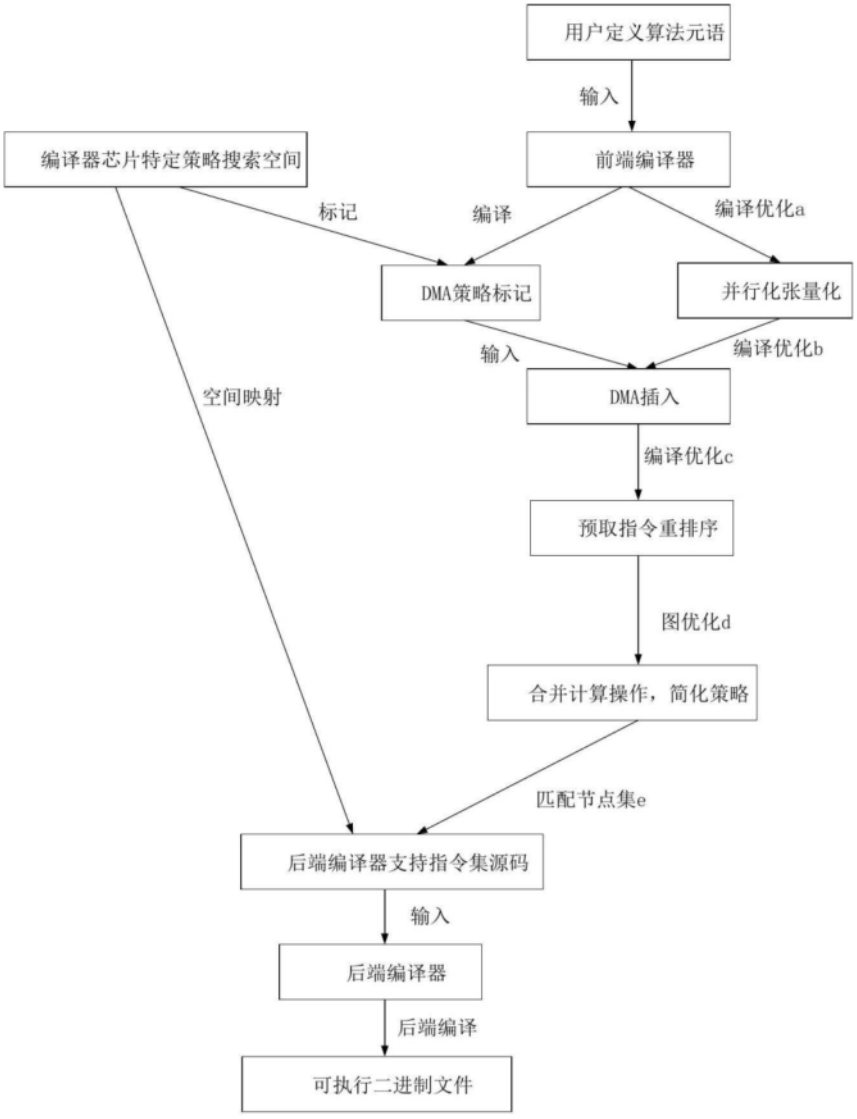

用户定义算法元语
输入
前端编译器
编译器芯片特定策略搜索空间
标记
编译
编译优化a
DMA策略标记
并行化张量化
输入
编译优化b
空间映射
DMA插入
编译优化c
预取指令重排序
图优化d
合并计算操作，简化策略
匹配节点集e
后端编译器支持指令集源码
输入
后端编译器
后端编译
可执行二进制文件

图2

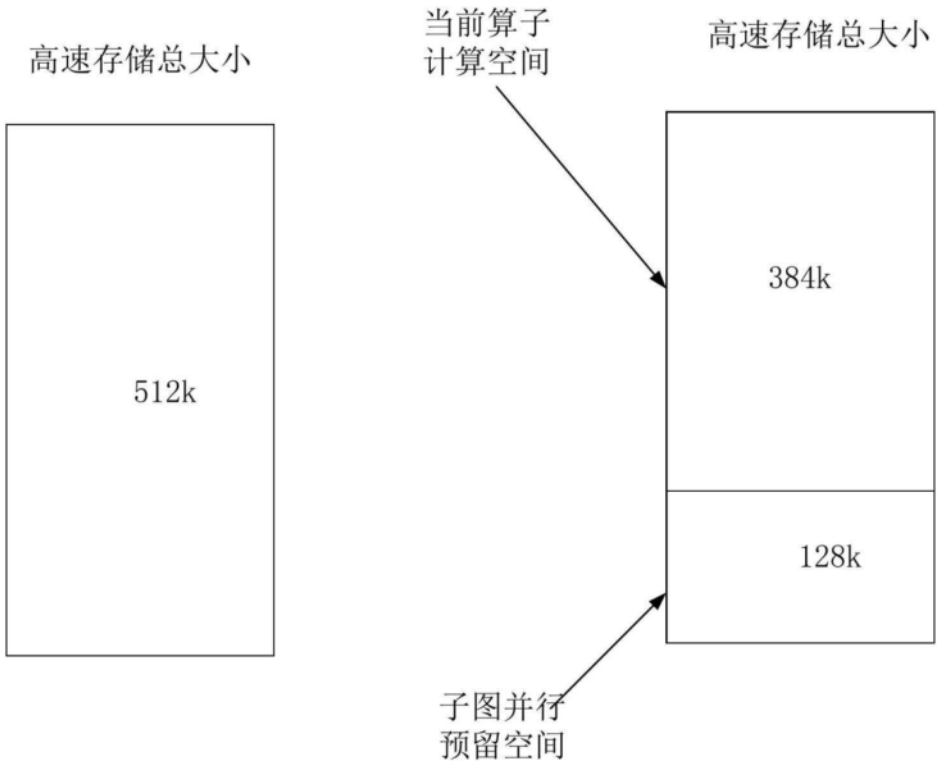


图3

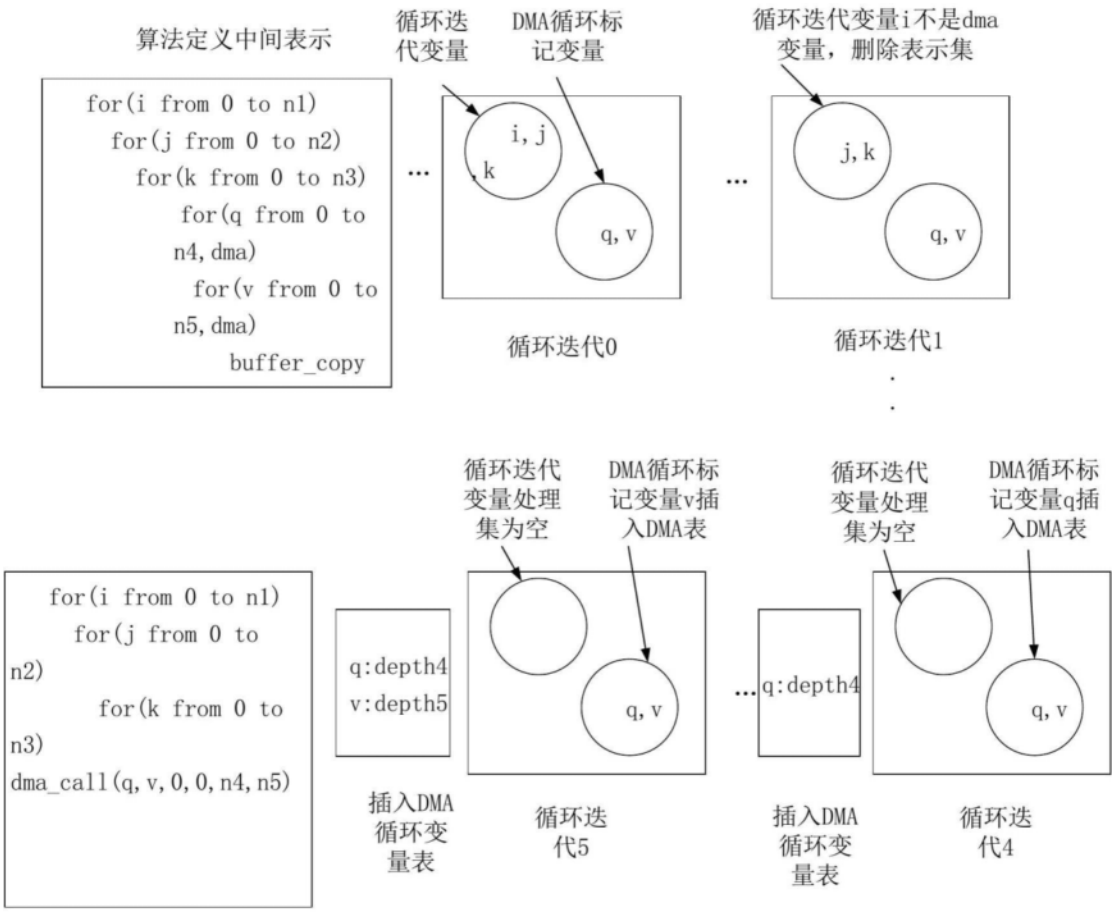

算法定义中间表示
for(i from 0 to n1)
for(j from 0 to n2)
for(k from 0 to n3)
for(q from 0 to n4,dma)
for(v from 0 to n5,dma)
buffer_copy
循环迭代变量
DMA循环标记变量
i,j
,k
q,v
循环迭代0
循环迭代变量i不是dma变量，删除表示集
j,k
q,v
循环迭代1
循环迭代变量处理集为空
DMA循环标记变量v插入DMA表
q:depth4
v:depth5
插入DMA循环变量表
循环迭代5
循环迭代变量处理集为空
DMA循环标记变量q插入DMA表
q:depth4
插入DMA循环变量表
循环迭代4
for(i from 0 to n1)
for(j from 0 to n2)
for(k from 0 to n3)
dma_call(q,v,0,0,n4,n5)

图4

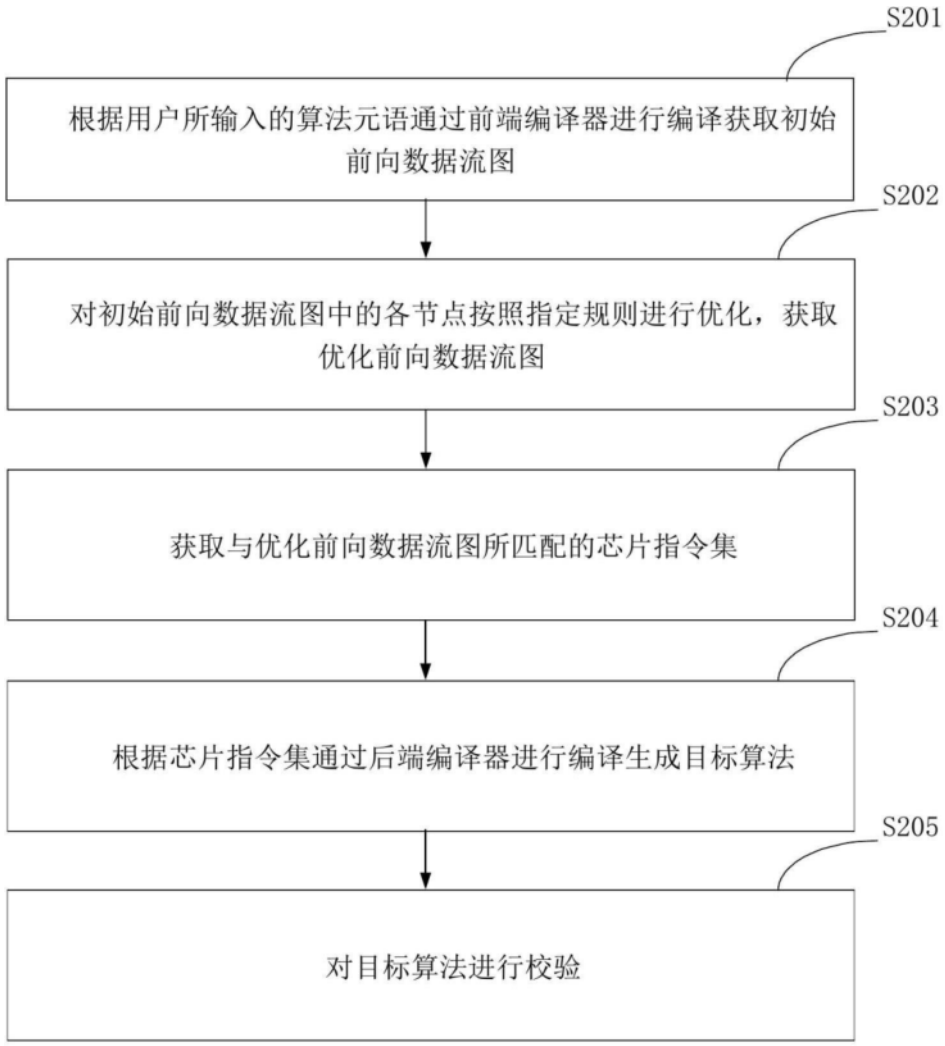

S201
根据用户所输入的算法元语通过前端编译器进行编译获取初始前向数据流图
S202
对初始前向数据流图中的各节点按照指定规则进行优化，获取优化前向数据流图
S203
获取与优化前向数据流图所匹配的芯片指令集
S204
根据芯片指令集通过后端编译器进行编译生成目标算法
S205
对目标算法进行校验

图5

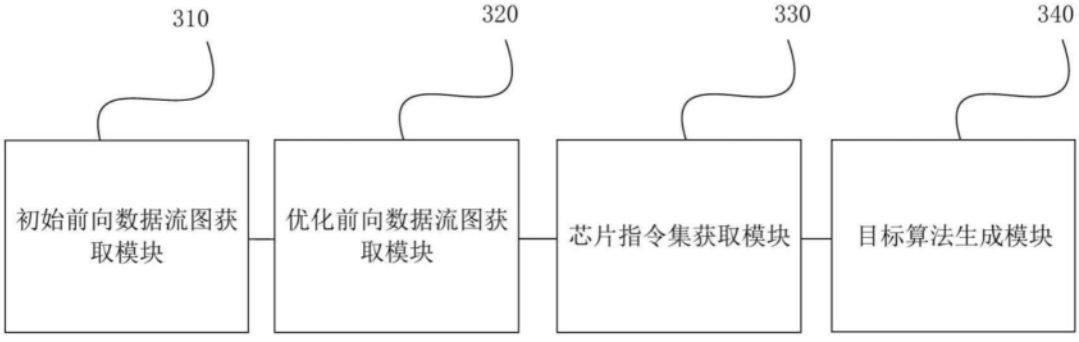

310
初始前向数据流图获取模块
320
优化前向数据流图获取模块
330
芯片指令集获取模块
340
目标算法生成模块

图6

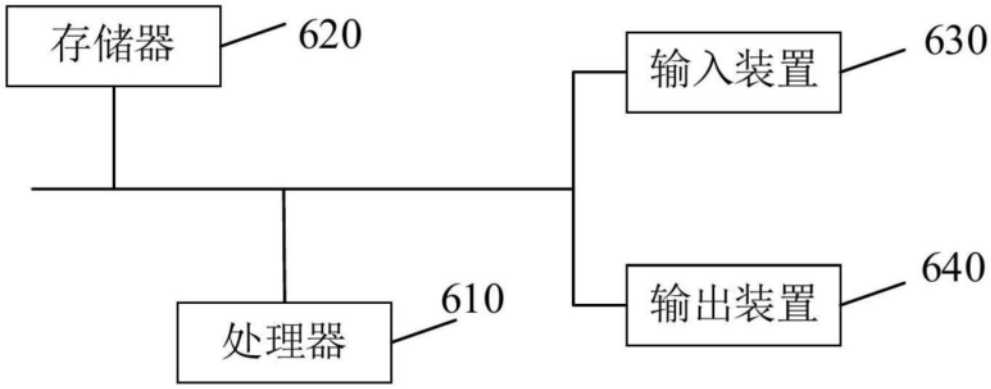


图7