(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2021 年 3 月 25 日 (25.03.2021)

WIPO | PCT

(10) 国际公布号
WO 2021/052391 A1

(51) 国际专利分类号:
G06F 8/41 (2018.01)

(21) 国际申请号: PCT/CN2020/115759

(22) 国际申请日: 2020 年 9 月 17 日 (17.09.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910896548.1 2019年9月18日 (18.09.2019) CN
201911271859.5 2019年12月11日 (11.12.2019) CN

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

(72) 发明人: 耿臻 (GENG, Zhen); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 狄鹏 (DI, Peng); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 淡孝强 (DAN, Xiaoqiang); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 深圳市深佳知识产权代理事务所 (普通合伙) (SHENPAT INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市罗湖区南湖街道春风路庐山大厦B座18C2、18D、18E、18E2, Guangdong 518001 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,

(54) **Title:** METHOD FOR CONSTRUCTING INTERMEDIATE REPRESENTATION, COMPILER AND SERVER

(54) 发明名称: 一种构建中间表达的方法、编译器和服务器

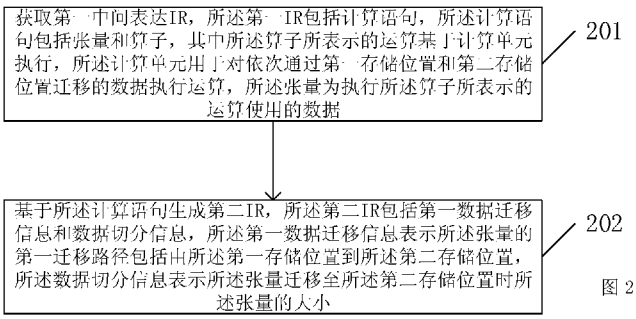


图 2

201 Acquire a first intermediate representation (IR), the first IR comprising a calculation statement, and the calculation statement comprising a tensor and an operator, wherein an operation represented by the operator is executed on the basis of a calculation unit, the calculation unit is used to perform operations on data that is migrated sequentially through a first storage location and a second storage location, and the tensor is data used to execute the operation represented by the operator

202 Generate a second IR on the basis of the calculation statement, wherein the second IR comprises first data migration information and data segmentation information, the first data migration information indicates that a first migration path of the tensor comprises from the first storage location to the second storage location, and the data segmentation information indicates the size of the tensor when the tensor migrates to the second storage location

(57) **Abstract:** A method for constructing an intermediate representation (IR), the method comprising: acquiring a first IR, the first IR comprising a calculation statement, and the calculation statement comprising a tensor and an operator, wherein an operation represented by the operator is executed on the basis of a calculation unit, the calculation unit is used to perform operations on data that is migrated sequentially through a first storage location and a second storage location, and the tensor is data used to execute the operation represented by the operator (201); and generating a second IR on the basis of the calculation statement, wherein the second IR comprises first data

WO 2021/052391 A1

[见续页]

CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

migration information and data segmentation information, the first data migration information indicates that a first migration path of the tensor comprises from the first storage location to the second storage location, and the data segmentation information indicates the size of the tensor when the tensor migrates to the second storage location (202). The described method may construct an IR that is able to represent the migration of a tensor between different storage locations.

(57) 摘要：一种构建中间表达的方法，方法包括：获取第一中间表达IR，所述第一IR包括计算语句，所述计算语句包括张量和算子，其中所述算子所表示的运算基于计算单元执行，所述计算单元用于对依次通过第一存储位置和第二存储位置迁移的数据执行运算，所述张量为执行所述算子所表示的运算使用的数据（201）；基于所述计算语句生成第二IR，所述第二IR包括第一数据迁移信息和数据切分信息，所述第一数据迁移信息表示所述张量的第一迁移路径包括由所述第一存储位置到所述第二存储位置，所述数据切分信息表示所述张量迁移至所述第二存储位置时所述张量的大小（202）。所述方法可以构建出能表达张量在不同存储位置之间的迁移的中间表达。

一种构建中间表达的方法、编译器和服务器

本申请要求于 2019 年 09 月 18 日提交中国专利局、申请号为 201910896548.1、发明名称为“一种构建中间表达的方法、编译器和服务器”，以及于 2019 年 12 月 11 日提交中国专利局、申请号为 201911271859.5、发明名称为“一种构建中间表达的方法、编译器和服务器”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

本申请涉及电子技术领域，尤其涉及一种构建中间表达的方法、编译器和服务器。

背景技术

随着人工智能（artificial intelligence，AI）技术的发展，基于领域特定架构（domain specific architecture，DSA）设计的 AI 芯片不断涌现。不同于现代微处理器，基于 DSA 的 AI 芯片，使用了多个存储位置来缩短数据搬移的距离，以减少数据搬移带来的开销。随着 DSA 的演进，计算单元的增加以及算子复杂度的提升，存储位置的数量成倍增加，此时需要通过多个存储位置将张量搬移至对应的计算单元，来实现算子的执行。

然而现有技术中，并没有一种可以针对于需要通过多个存储位置，将数据迁移至计算单元的中间表达（intermediate representation，IR）的构建方法。

发明内容

本申请实施例提供了一种构建中间表达的方法，包括：

获取第一中间表达 IR，所述第一 IR 包括计算语句，所述计算语句包括张量和算子，其中所述算子所表示的运算基于计算单元执行，所述计算单元用于对依次通过第一存储位置和第二存储位置迁移的数据执行运算，所述张量为执行所述算子所表示的运算使用的数据；

基于所述计算语句生成第二 IR，所述第二 IR 包括第一数据迁移信息和数据切分信息，所述第一数据迁移信息表示所述张量的第一迁移路径包括由所述第一存储位置到所述第二存储位置，其中，第一存储位置和所述第二存储位置之间可以存在多条数据传输通路，有的数据传输通路上可以没有其他存储位置，例如下文描述的第一存储位置和所述第二存储位置之间的数据传输通路是直连的；有的数据通路上可以有其他存储位置，例如下文提及的第三存储位置。所述数据切分信息表示所述张量迁移至所述第二存储位置时所述张量的大小。

可选的，在第一方面的一种可选设计中，所述计算单元还用于对通过第二迁移路径的数据执行运算，所述第二迁移路径包括由所述第一存储位置，经过至少一个第三存储位置，到所述第二存储位置，所述第一迁移路径与所述第二迁移路径为不同的数据迁移路径。

可选的，在第一方面的一种可选设计中，所述至少一个第三存储位置指示片上缓存上的存储位置。

可选的，在第一方面的一种可选设计中，所述方法还包括：

基于所述计算语句生成所述计算语句对应的第一数据流信息，所述第一数据流信息包括所述张量的第二数据迁移信息，所述第二数据迁移信息表示所述张量由所述第一存储位置依次通过至少一个第三迁移位置迁移至所述第二存储位置；

基于所述第一存储位置与所述第二存储位置之间存在直连的数据传输通路，生成第二数据流信息，其中所述第二数据流信息包括所述第一数据迁移信息。

本实施例中，第一存储位置与所述第二存储位置之间存在直连的数据传输通路可以理解为：第一存储位置与所述第二存储位置之间存在不经过其他任意一个存储位置的数据传输通路。

例如，所述张量用于依次通过第一存储位置、至少一个第三存储位置以及第二存储位置迁移至所述计算单元。又例如，所述张量用于依次通过至少一个第四存储位置、第一存储位置以及第二存储位置迁移至所述计算单元。又例如，所述张量用于依次通过至少一个第四存储位置、第一存储位置、至少一个第三存储位置以及第二存储位置迁移至所述计算单元。又例如，所述张量用于依次通过至少一个第四存储位置、第一存储位置、至少一个第三存储位置、第二存储位置以及至少一个第五存储位置迁移至所述计算单元。

可选的，在第一方面的一种可选设计中，所述第一存储位置指示片上缓存上的存储位置，所述第二存储位置指示片上缓存上的存储位置。

可选的，在第一方面的一种可选设计中，所述第一存储位置指示内存，所述第二存储位置指示片上缓存上的存储位置。

可选的，在第一方面的一种可选设计中，所述张量包括多个维度，每个维度对应于一个轴变量，所述轴变量用于表示对应维度的张量大小，所述数据切分信息包括多个轴变量和与每个轴变量对应的切分轴变量，所述切分轴变量用于表示张量迁移时对应维度的张量大小。

可选的，在第一方面的一种可选设计中，所述第二 IR 还包括至少一个目标变量，以及每个目标变量的取值范围，所述轴变量与所述至少一个目标变量线性相关，所述轴变量通过所述至少一个目标变量，以及所述至少一个目标变量的取值范围来表示对应维度的张量大小。

可选的，在第一方面的一种可选设计中，所述第二 IR 为树结构的 IR，所述树结构包括根节点和子节点，所述根节点对应于所述计算语句，所述子节点对应于所述第二存储位置，其中，所述第二 IR 包括与所述第二存储位置对应的子节点的信息，所述与所述第二存储位置对应的子节点的信息包括所述第一数据迁移信息和所述数据切分信息。

可选的，在第一方面的一种可选设计中，所述第二 IR 还包括：数据搬移信息，所述数据搬移信息表示所述张量被从所述第一存储位置搬移至所述第二存储位置。

可选的，在第一方面的一种可选设计中，所述方法还包括：

根据所述计算单元的类型确定第一存储位置和第二存储位置。

第二方面，本申请提供了一种编译器，包括：

获取单元，用于获取第一中间表达 IR，所述第一 IR 包括计算语句，所述计算语句包

括张量和算子，其中所述算子所表示的运算基于计算单元执行，所述计算单元用于对依次通过第一存储位置和第二存储位置迁移的数据执行运算，所述张量为执行所述算子所表示的运算使用的数据；

处理单元，用于基于所述计算语句生成第二 IR，所述第二 IR 包括第一数据迁移信息和数据切分信息，所述第一数据迁移信息表示所述张量的第一迁移路径包括由所述第一存储位置到所述第二存储位置，所述数据切分信息表示所述张量迁移至所述第二存储位置时所述张量的大小。

可选的，在第二方面的一种可选设计中，所述计算单元还用于对通过第二迁移路径的数据执行运算，所述第二迁移路径包括由所述第一存储位置，经过至少一个第三存储位置，到所述第二存储位置，所述第一迁移路径与所述第二迁移路径为不同的数据迁移路径。

可选的，在第二方面的一种可选设计中，所述至少一个第三存储位置指示片上缓存上的存储位置。

可选的，在第二方面的一种可选设计中，所述处理模块还用于：基于所述计算语句生成所述计算语句对应的第一数据流信息，所述第一数据流信息包括所述张量的第二数据迁移信息，所述第二数据迁移信息表示所述张量由所述第一存储位置依次通过至少一个第三迁移位置迁移至所述第二存储位置；

基于所述第一存储位置与所述第二存储位置之间存在直连的数据传输通路，生成第二数据流信息，其中所述第二数据流信息包括所述第一数据迁移信息。

可选的，在第二方面的一种可选设计中，所述第一存储位置指示片上缓存上的存储位置，所述第二存储位置指示片上缓存上的存储位置。

可选的，在第二方面的一种可选设计中，所述第一存储位置指示内存，所述第二存储位置指示片上缓存上的存储位置。

可选的，在第二方面的一种可选设计中，所述张量包括多个维度，每个维度对应于一个轴变量，所述轴变量用于表示对应维度的张量大小，所述数据切分信息包括多个轴变量和与每个轴变量对应的切分轴变量，所述切分轴变量用于表示张量迁移时对应维度的张量大小。

可选的，在第二方面的一种可选设计中，所述第二 IR 还包括至少一个目标变量，以及每个目标变量的取值范围，所述轴变量与所述至少一个目标变量线性相关，所述轴变量通过所述至少一个目标变量，以及所述至少一个目标变量的取值范围来表示对应维度的张量大小。

可选的，在第二方面的一种可选设计中，所述第二 IR 为树结构的 IR，所述树结构包括根节点和子节点，所述根节点对应于所述计算语句，所述子节点对应于所述第二存储位置，其中，所述第二 IR 包括与所述第二存储位置对应的子节点的信息，所述与所述第二存储位置对应的子节点的信息包括所述第一数据迁移信息和所述数据切分信息。

可选的，在第二方面的一种可选设计中，所述第二 IR 还包括：数据搬移信息，所述数据搬移信息表示所述张量被从所述第一存储位置搬移至所述第二存储位置。

可选的，在第二方面的一种可选设计中，所述处理单元还用于：

根据所述计算单元的类型确定第一存储位置和第二存储位置。

第三方面，本申请提供了一种计算机系统，包括：处理器和存储器；

所述处理器和所述存储器电连接；

所述处理器用于调用所述存储器中的代码开执行如上述第一方面任一所述的方法。

第四方面，本申请提供了一种计算机存储介质，所述计算机可读存储介质存储指令，当所述指令在计算机上运行时，使得所述计算机执行如上述第一方面任一所述的方法。

第五方面，本申请实施例提供了一种构建中间表达的方法，包括：

获取第一中间表达 IR，所述第一 IR 包括计算语句，所述计算语句包括张量和算子，所述算子所表示的运算基于计算单元执行，所述张量用于依次通过第一存储位置和第二存储位置迁移至所述计算单元，所述张量为执行所述算子所表示的运算使用的数据；

基于所述计算语句生成第二 IR，所述第二 IR 包括第一数据迁移信息和数据切分信息，所述第一数据迁移信息表示所述张量的迁移路径包括由所述第一存储位置到所述第二存储位置，所述数据切分信息表示每次迁移至所述第二存储位置的所述张量的大小。

可选的，所述第一存储位置可以为所述张量迁移路径上的起始存储位置。

可选的，第一存储位置和第二存储位置可以为相邻的存储位置。

可选的，第一存储位置和第二存储位置之间也可以间隔有其他存储位置。

可选的，第二存储位置与计算单元之间也可以间隔有其他存储位置。

例如，所述张量用于依次通过第一存储位置、至少一个第三存储位置以及第二存储位置迁移至所述计算单元。又例如，所述张量用于依次通过至少一个第四存储位置、第一存储位置以及第二存储位置迁移至所述计算单元。又例如，所述张量用于依次通过至少一个第四存储位置、第一存储位置、至少一个第三存储位置以及第二存储位置迁移至所述计算单元。又例如，所述张量用于依次通过至少一个第四存储位置、第一存储位置、至少一个第三存储位置、第二存储位置以及至少一个第五存储位置迁移至所述计算单元。

可选的，在第五方面的一种可选设计中，所述基于所述计算语句生成第二 IR，包括：基于所述计算语句、所述第一存储位置对应的存储大小以及所述第二存储位置对应的存储大小生成第二 IR，所述第二 IR 包括第一数据迁移信息和数据切分信息，所述第一数据迁移信息表示所述张量的迁移路径包括由所述第一存储位置到所述第二存储位置，所述数据切分信息表示每次迁移至所述第二存储位置的所述张量的大小。

可选的，在第五方面的一种可选设计中，所述第一存储位置指示片上缓存上的存储位置，所述第二存储位置指示片上缓存上的存储位置。

可选的，在第五方面的一种可选设计中，所述第一存储位置为内存，所述第二存储位置指示片上缓存上的存储位置。

可选的，在第五方面的一种可选设计中，所述张量对应于所述计算语句中的目标变量，所述第二 IR 还包括：所述目标变量的取值范围，以及所述目标变量至所述第一存储位置的张量的映射。

可选的，在第五方面的一种可选设计中，所述数据切分信息包括所述张量每次迁移至所述第二存储位置时对应的所述目标变量的取值范围。

可选的，在第五方面的一种可选设计中，所述张量包括目标轴变量，所述切分张量包括所述目标轴变量的切分轴变量，所述数据切分信息包括所述张量每次迁移至所述第二存储位置，所述切分轴变量与所述目标轴变量之间的大小关系。

可选的，在第五方面的一种可选设计中，所述第二 IR 包括与所述第二缓存对应的节点信息，所述节点信息包括所述第一数据迁移信息和所述数据切分信息。

可选的，在第五方面的一种可选设计中，所述第二 IR 还包括：读写信息，所述读写信息表示所述第二存储位置从所述第一存储位置读取所述张量。

可选的，在第五方面的一种可选设计中，所述方法还包括：

生成所述计算语句对应的第一数据流信息，所述第一数据流信息包括与所述张量相关联的第一数据迁移信息，所述第一数据迁移信息表示所述张量由所述第一存储位置迁移至所述第二存储位置。

可选的，在第五方面的一种可选设计中，所述生成所述计算语句对应的第一数据流信息，包括：

生成所述计算语句对应的第一数据流信息，所述第一数据流信息包括所述张量的第二数据迁移信息，所述第二数据迁移信息表示所述张量由第一存储位置依次通过至少一个第三迁移位置迁移至第二存储位置；

基于所述第一存储位置与所述第二存储位置之间存在数据传输通路，生成所述第一数据流信息，所述第一数据流信息包括与所述张量相关联的所述数据迁移信息。

可选的，在第五方面的一种可选设计中，所述第三迁移位置为片上缓存。

可选的，在第五方面的一种可选设计中，所述方法还包括：

根据所述计算单元的类型确定第一存储位置和第二存储位置。

第六方面，本申请提供了一种编译器，包括：

获取单元，用于获取第一中间表达 IR，所述第一 IR 包括计算语句，所述计算语句包括张量和算子，所述算子所表示的运算基于计算单元执行，所述张量用于依次通过第一存储位置和第二存储位置迁移至所述计算单元，所述张量为执行所述算子所表示的运算使用的数据；

处理单元，用于基于所述计算语句生成第二 IR，所述第二 IR 包括第一数据迁移信息和数据切分信息，所述第一数据迁移信息表示所述张量的迁移路径包括由所述第一存储位置到所述第二存储位置，所述数据切分信息表示每次迁移至所述第二存储位置的所述张量的大小。

可选的，在第六方面的一种可选设计中，所述第一存储位置指示片上缓存上的存储位置，所述第二存储位置指示片上缓存上的存储位置。

可选的，在第六方面的一种可选设计中，所述第一存储位置为内存，所述第二存储位置指示片上缓存上的存储位置。

可选的，在第六方面的一种可选设计中，所述张量对应于所述计算语句中的目标变量，所述第二 IR 还包括：所述目标变量的取值范围，以及所述目标变量至所述第一存储位置的张量的映射。

可选的，在第六方面的一种可选设计中，所述数据切分信息包括所述张量每次迁移至所述第二存储位置时对应的所述目标变量的取值范围。

可选的，在第六方面的一种可选设计中，所述张量包括目标轴变量，所述切分张量包括所述目标轴变量的切分轴变量，所述数据切分信息包括所述张量每次迁移至所述第二存储位置，所述切分轴变量与所述目标轴变量之间的大小关系。

可选的，在第六方面的一种可选设计中，所述第二 IR 包括与所述第二缓存对应的节点信息，所述节点信息包括所述第一数据迁移信息和所述数据切分信息。

可选的，在第六方面的一种可选设计中，所述第二 IR 还包括：读写信息，所述读写信息表示所述第二存储位置从所述第一存储位置读取所述张量。

可选的，在第六方面的一种可选设计中，所述处理单元还用于：

生成所述计算语句对应的第一数据流信息，所述第一数据流信息包括与所述张量相关联的第一数据迁移信息，所述第一数据迁移信息表示所述张量由所述第一存储位置迁移至所述第二存储位置。

可选的，在第六方面的一种可选设计中，所述处理单元具体用于：

生成所述计算语句对应的第一数据流信息，所述第一数据流信息包括所述张量的第二数据迁移信息，所述第二数据迁移信息表示所述张量由第一存储位置依次通过至少一个第三迁移位置迁移至第二存储位置；

基于所述第一存储位置与所述第二存储位置之间存在数据传输通路，生成所述第一数据流信息，所述第一数据流信息包括与所述张量相关联的所述数据迁移信息。

可选的，在第六方面的一种可选设计中，所述第三迁移位置为片上缓存。

可选的，在第六方面的一种可选设计中，所述处理单元还用于：

根据所述计算单元的类型确定第一存储位置和第二存储位置。

第七方面，本申请提供了一种计算机系统，包括：处理器和存储器；

所述处理器和所述存储器电连接；

所述处理器用于调用所述存储器中的代码开执行如上述第五方面任一所述的方法。

第八方面，本申请提供了一种计算机存储介质，所述计算机可读存储介质存储指令，当所述指令在计算机上运行时，使得所述计算机执行如上述第五方面任一所述的方法。

本申请实施例提供了一种构建中间表达的方法，包括：获取第一中间表达 IR，所述第一 IR 包括计算语句，所述计算语句包括张量和算子，其中所述算子所表示的运算基于计算单元执行，所述计算单元用于对依次通过第一存储位置和第二存储位置迁移的数据执行运算，所述张量为执行所述算子所表示的运算使用的数据；基于所述计算语句生成第二 IR，所述第二 IR 包括第一数据迁移信息和数据切分信息，所述第一数据迁移信息表示所述张量的第一迁移路径包括由所述第一存储位置到所述第二存储位置，所述数据切分信息表示所述张量迁移至所述第二存储位置时所述张量的大小。通过上述方式，编译器可以构建出能表达出张量在不同存储位置（片上缓存或内存）之间的迁移（包括迁移的方向和迁移的大小）的中间表达，可以应用在针对于 DSA 的 AI 芯片上的 IR 的构建。

附图说明

图 1 为本申请实施例的应用架构示意图；

图 2 为本申请实施例提供的一种构建中间表达的方法的流程图示意；

图 3 为本申请实施例提供的一种 AI 芯片中的 AI 核的结构示意；

图 4 为本申请实施例提供的一种数据流信息的抽象示意；

图 5 为本申请实施例提供的一种构建中间表达的方法的实施例示意图；

图 6 为本申请实施例提供的一种数据流信息的抽象示意；

图 7 为本申请实施例提供的一种数据流信息的抽象示意；

图 8 为本申请实施例提供的一种数据流信息的抽象示意；

图 9 为本申请实施例提供的一种编译器的结构示意图。

具体实施方式

本申请实施例提供了一种构建中间表达的方法、编译器和服务器，可以构建出能表达出张量在不同存储位置（片上缓存或内存）之间的迁移（包括迁移的方向和迁移的大小）的中间表达。

下面结合附图，对本申请的实施例进行描述。本领域普通技术人员可知，随着技术的发展和新场景的出现，本申请实施例提供的技术方案对于类似的技术问题，同样适用。

本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的术语在适当情况下可以互换，这仅仅是描述本申请的实施例中对相同属性的对象在描述时所采用的区分方式。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，以便包含一系列单元的过程、方法、系统、产品或设备不必限于那些单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它单元。

参照图 1，图 1 为本申请实施例的应用架构示意图，如图 1 中示出的那样，本申请可以应用在服务器中，其中服务器可以包括软件部分的 AI 训练和推理框架，AI 训练和推理框架可以包括编译器，本申请实施例中，编译器可以从存储器中获取源代码，并将源代码编译为中间表达，以及可以让 AI 芯片识别并执行的机器语言。

参照图 2，图 2 为本申请实施例提供的一种构建中间表达的方法的流程图示意，如图 2 中示出的那样，本申请提供的构建中间表达的方法包括：

201、编译器获取第一中间表达 IR，所述第一 IR 包括计算语句，所述计算语句包括张量和算子，其中所述算子所表示的运算基于计算单元执行，所述计算单元用于对依次通过第一存储位置和第二存储位置迁移的数据执行运算，所述张量为执行所述算子所表示的运算使用的数据。

可选地，本申请实施例中，编译器可以获取第一中间表示（intermediate representation，IR）。

应理解，可以通过解析以及编译计算机源代码等方式生成第一IR。或者，可通过现有计算机程序的反编译来生成第一IR。或者，从外部获得第一IR。其中，用于生成第一IR的

源代码可使用高级编辑语言的应用编辑接口来编写。该高级编程语言可以是领域专用语言（domain specific language,DSL)。

第一IR可以存储在存储器(服务器的外部存储或内部存储)中，本申请实施例中，编译器可以从存储器(服务器的外部存储或内部存储)读取第一IR。具体的，第一IR可以由DSL语言来描述，本申请实施例中，所述DSL语言可以为：Halide、Graphlt、Spatial、或其他定制的领域特定语言。其中，Halide适用于向量和张量运算，GraphIt适用于图计算领域，Spatial适用于可编程硬件领域，定制的领域特定语言适用于对应的定制领域。

本申请实施例中，编译器可以依次遍历获取的第一IR中的每一条计算语句，并分析每条计算语句的将会使用到的计算单元。

在一种实施例中，计算单元可以至少包括：标量计算单元，向量vector计算单元和张量cube计算单元。

本申请实施例中，针对于不同的算子类型，将会使用到的计算单元不同，例如vector计算单元可以支持加法运算(add),减法运算(sub),乘法运算(mul)，取倒数运算(rec)，取指数运算(exp)，取对数运算(log)、以及量化运算等。cube计算单元可以支持卷积运算。

在一些AI芯片的架构中（例如基于特定领域架构（domain specific architecture，DSA）的AI芯片)，使用了许多专用片上缓存来缩短数据搬移的距离，以减少数据搬移带来的开销，随着DSA架构的演进，计算单元的增加以及算子复杂度的提升，片上缓存类型成倍增加，例如：多用途张量加速器(versatile tensor accelerator，VTA)有3种类型的片上缓存，在其他的处理器中，片上缓存的类型可以为5个甚至更多，本申请并不限定。

参照图3，图3为本申请实施例提供的一种AI芯片中的AI核的结构示意，如图3中示出的那样，AI核包括多个片上缓存（L1 buffer、LOA buffer、LOB buffer、LOC buffer、Unified buffer)、多个计算单元（cube计算单元、vector计算单元)、数据中转处理单元以及总线接口部件。其中，总线接口部件可以获取到AI芯片内存中的张量，通过上述片上缓存，将张量迁移至对应的计算单元，实现对应的运算。

需要说明的是，在确定了计算单元之后，可以基于AI核的具体架构，确定数据（张量）的迁移路线，其中，迁移路线可以不唯一，即数据可以通过多种迁移路线，经过不同的片上缓存迁移至计算单元。

需要说明的是，本申请中的“迁移”可以表示数据的读取，例如张量由buffer1迁移至buffer2，可以表示buffer2读取buffer1中的张量。

需要说明的是，尽管图3中未示出，AI核还可以包括数据控制单元，该数据控制单元可以控制张量在片上缓存的迁移方向。

需要说明的是，图3中的片上缓存与图1中的服务器中的存储器可以理解为不同的存储介质（片上缓存为AI芯片中的存储介质)。

本申请实施例中，第一IR的计算语句可以包括至少一个张量以及至少一个张量相关的的算子，其中张量为数据，例如可以理解为多维度的向量。

本申请实施例中，算子可以表示某一运算规则，例如某一算子为乘法运算，相应的，该算子需要基于vector计算单元来执行，而张量从内存迁移至vector计算单元需要依次通

过：内存、总线接口部件、Unified buffer，才能迁移至vector计算单元。即，所述vector计算用于依次通过内存和Unified buffer的迁移来获取所述张量，进而实现所述算子的运算。

以第一IR为halide IR为例，示例性的，所述第一IR可以参照如下的IR示意：

```
#第一IR示意
for(i1, 0, 64){
   for(i2, 0, 64){
      result(i1,i2)=input_1(i1,i2)*input_2(i1,i2)}}
```

其中，input_1(i1,i2)表示张量A，input_2(i1,i2)表示张量B，result(i1,i2)为结果。for(i1, 0, 64)表示一层for循环，遍历规则为目标变量i1从0开始，每次累加1，累加64次。result(i1,i2)是一个两维的张量，其为张量A和张量B的乘积，其每个维度的大小为64。

需要说明的是，目标变量可以是计算语句的循环变量，其中循环变量的一部分可以为其包括的张量的目标变量。

本申请实施例中，编译器可以获取到上述第一IR中的计算语句，该计算语句包括张量A、张量B、张量（result）和算子（乘法运算），进而，编译器可以确定该算子（乘法运算）对应的计算单元为vector计算单元，所述vector计算用于依次通过内存和Unified buffer获取所述张量，进而实现所述算子（乘法运算）的运算。

可选的，本申请实施例中，编译器可以生成所述计算语句对应的第一数据流信息，所述第一数据流信息包括与所述张量相关联的第一数据迁移信息，所述第一数据迁移信息表示所述张量由第一存储位置迁移至第二存储位置。

可选的，所述第一存储位置指示片上缓存上的存储位置，所述第二存储位置指示片上缓存上的存储位置。

可选的，所述第一存储位置指示内存上的存储位置，所述第二存储位置指示片上缓存上的存储位置。

需要说明的是，本申请实施例中的数据流信息可以表示张量在AI核中的迁移，例如可以是一种数据栈结构，该结构可以表示张量在AI核中的迁移路径，本申请并不限定数据流信息的具体实现方式。

以上述为例，参照图4，图4为本申请实施例提供的一种数据流信息的抽象示意，如图4中示出的那样，编译器可以生成所述计算语句对应的第一数据流信息，第一数据流信息可以表示张量A被从内存迁移至Unified buffer中，张量B被从内存迁移至Unified buffer中，张量A和张量B被迁移至Vector计算单元，Vector计算单元可以对张量A和张量B进行乘积运算得到运算结果result，之后，运算结果result被迁移至并内存。

其中，第一数据流信息中与张量A相关的数据迁移信息为：“内存（A）—UB（A）”，需要说明的是，图4中的UB为上述Unified buffer。与张量B相关的数据迁移信息为：“内存（B）—UB（B）”。与张量result相关的数据迁移信息为：“UB（result）—内存（result）”。

202、编译器基于所述计算语句生成第二IR，所述第二IR包括第一数据迁移信息和数据切分信息，所述第一数据迁移信息表示所述张量的第一迁移路径包括由所述第一存储位置

到所述第二存储位置，所述数据切分信息表示所述张量迁移至所述第二存储位置时所述张量的大小。

本申请实施例中，编译器可以生成第二IR，其中第二IR和第一IR可以是基于不同语言描述的中间表示。可选的，第二IR可以基于多面体编译技术来实现。

本申请实施例中，第二IR可以是基于多面体技术的调度树实现的，其可以包括根节点和子节点，其中，根节点为域节点，包含了算子的计算语句，以及其中的变量的变化范围。特定范围节点(band node)，在第二IR中的标记为child或schedule。

本申请实施例中，第二IR可以包括数据迁移信息，所述数据迁移信息表示所述张量被由第一存储位置迁移至第二存储位置。

参照如下的第二IR示意：

```
#第二IR示意
Domain :{S_0[i1,i2]: 0<=i1<=63 and 0<=i2<=63}  //第一部分
Child:
  Mark: “realize_UB”
    Child:
      Extension:
“{[ i1,i2]->L1read[[[i1,i2]->A[arg0,arg1]]->A_local_L1[arg0’,arg1’]]}”//第二部分
```

其中，第一部分表示第一IR的一条语句，及其相关的目标变量（i1和i2）的变化范围，Mark为标记节点，其中，”realize_UB”表示片上缓存的类型为Unified buffer，即：Mark:“realize_UB”表示与所述第二缓存（UB）对应的节点信息。相应的，与第二缓存（UB）对应的节点信息下包含与第二缓存（UB）对应的数据迁移信息和数据切分信息。

其中，第二部分表示张量的数据迁移信息以及数据切分信息，即从内存到片上缓存UB的搬移映射关系，其中i1,i2表示语句相关的目标变量，arg0,arg1表示张量的轴变量，arg0’,arg1’表示第一张量的切分轴变量。

需要说明的是，由于片上缓存之间的存储量大小可能会不同，因此需要对张量进行切分，使得后续每一次张量的迁移过程都是基于切分后的张量进行的。

具体的，[i1,i2]->L1read[[[i1,i2]->A[arg0,arg1]]->A_local_L1[arg0’,arg1’]]表示第一张量由内存迁移到片上缓存UB映射关系，->表示一层映射关系。其中，[[i1,i2]->A[arg0,arg1]]表示目标变量i1,i2至所述第一存储位置（内存）的张量的映射，[i1,i2]->A[arg0,arg1]]->A_local_L1[arg0’,arg1’]表示所述第一存储位置的张量（内存）到第二存储位置的张量（UB）的映射，其中A_local_L1[arg0’,arg1’]表示所述张量在每次迁移至所述第二存储位置时需要进行数据切分。

可选的，所述张量包括多个维度，每个维度对应于一个轴变量（例如上述的arg0和arg1），所述轴变量用于表示对应维度的张量大小，所述数据切分信息包括多个轴变量和与每个轴变量对应的切分轴变量，所述切分轴变量用于表示张量迁移时对应维度的张量大小

可选的，所述第二IR还包括至少一个目标变量，以及每个目标变量的取值范围，所述轴变量与所述至少一个目标变量线性相关，所述轴变量通过所述至少一个目标变量，以及所述至少一个目标变量的取值范围来表示对应维度的张量大小。

本申请实施例中，轴变量对应的维度的张量大小可以由目标变量的线性组合以及取值范围来表达，示例性的：

```
for（io，0,265）  {
    for（i1，0,512）
    }
```

arg0 >= 64i0 and 0 <= arg0 <= 255 and arg0 <= 63 + 64i0 and arg1 >= 512i1 and 0 <= arg1 <= 511 and arg1 <= 511 + 512i1，上述表达了一种轴变量arg0以及轴变量arg1对应的维度的张量大小的表示方式的示意，其中，arg0对应的维度的张量大小由i0来表达，arg1对应的维度的张量大小由i1来表达。

需要说明的是，上述轴变量的表达方式仅为一种示意，这里并不限定。

可选的，第二IR还可以包括：所述切分轴变量与所述轴变量之间的大小关系。例如，第二IR可以包括每一次迁移过程中切分轴变量arg0’与轴变量arg0之间的大小关系，以及切分轴变量arg1’与轴变量arg1之间的大小关系。

需要说明的是，本申请实施例中的一个轴变量对应于张量的一个维度，多个轴变量可以表示张量的大小。切分轴变量为轴变量进行切分（或划分）后得到的一部分轴变量。其中，切分可以指将轴变量划分为多个子轴变量，例如将64*64的轴变量划分为两个32*64的切分轴变量。

例如，若规定2arg1’=arg1，则相当于张量在迁移过程中，轴变量arg1每次是按照一半的轴变量arg1的大小进行迁移的（轴变量的大小取决于对应的目标变量的取值范围），此时，需要两次才能完成，轴变量arg1的迁移。

通过上述方式，编译器可以构建出能表达出张量在不同存储位置之间的迁移（包括迁移的方向和迁移的大小）的中间表达。

需要说明的是，尽管未示出，第二IR还可以包括与计算语句中的算子相关的运算语句，由于本申请仅关注于张量在片上缓存之间的迁移过程，关于算子的编译本申请不再赘述。

可选的，在一种实施例中，编译器可以获取第一数据流信息中最靠前的数据流信息（例如图4中为内存（A）至UE（A），根据当前获取到的数据流信息，确定该数据流信息在第二IR中的插入位置，如上述实施例中描述的，此时，数据流信息对应的插入位置为“realize UB”，并基于张量的大小确定张量的切分大小，可以是对张量对应的轴变量的切分。

编译器获取在当前切分空间下，张量在当前切分下在不同片上缓存直接的搬移信息，示例性的，搬移信息可以如下所示；

```
[[io,i1]->A[arg0,arg1]]->A_local_l1[arg0’,arg1’];
2arg0’=arg0,2arg1’=arg1
```

其中，i0,i1表示语句的相关目标变量；A[arg0,arg1]表示张量A，及其轴变量；A_local_L1[arg0’,arg1’]表示片上缓存L1上的张量A_local_L1以及该张量的轴变量；

2arg0’=arg0,2arg1’=arg1表示张量A_local_L1与张量A之间的轴变量的关系。

编译器可以基于上述得到的片上缓存间的搬运信息，执行普锐伯自映射运算(presburger operation identity)。普锐伯自映射运算主要实现一个自身到自动的映射关系：I={i->i;i∈S}

示例性的，可以得到如下映射到自身的映射：

[[[io,i1]->A[arg0,arg1]]->A_local_l1[arg0’,arg1’]]->

[[[io,i1]->A[arg0,arg1]]->A_local_l1[arg0’,arg1’]];

编译器可以基于得到的自映射信息，执行普锐伯域乘运算(Presburger operation DFDP)。其中，DFDP是一个domain域的乘法映射关系变换：DFDP={i->k: j:[i->j]->k∈S}

示例性的，可以得到如下乘法映射结果：

[io,i1]->[[[io,i1]->A[arg0,arg1]]->A_local_l1[arg0’,arg1’]]。

得到乘法映射结果后，可以设置范围域的读写信息；

示例性的，可以得到如下数据迁移信息和数据切分信息：

[io,i1]->L1read[[[io,i1]->A[arg0,arg1]]->A_local_l1[arg0’,arg1’]]。

其中，L1read表示数据的读取。

编译器可以将该上述数据迁移信息和数据切分信息插入到“realize UB”的子节点下，得到第二IR。

本申请实施例提供了一种构建中间表达的方法，包括：获取第一中间表达 IR，所述第一 IR 包括计算语句，所述计算语句包括张量和算子，其中所述算子所表示的运算基于计算单元执行，所述计算单元用于对依次通过第一存储位置和第二存储位置迁移的数据执行运算，所述张量为执行所述算子所表示的运算使用的数据；基于所述计算语句生成第二 IR，所述第二 IR 包括第一数据迁移信息和数据切分信息，所述第一数据迁移信息表示所述张量的第一迁移路径包括由所述第一存储位置到所述第二存储位置，所述数据切分信息表示所述张量迁移至所述第二存储位置时所述张量的大小。通过上述方式，编译器可以构建出能表达出张量在不同存储位置（片上缓存或内存）之间的迁移（包括迁移的方向和迁移的大小）的中间表达，可以应用在针对于 DSA 的 AI 芯片上的 IR 的构建。

参照图 5，图 5 为本申请实施例提供的一种构建中间表达的方法的实施例示意图，如图 5 中示出的那样，构建中间表达的方法包括：

501、编译器获取第一中间表达 IR。

步骤 501 的你具体描述可以参照步骤 401，此处不再赘述。

502、编译器基于所述计算语句生成所述计算语句对应的第一数据流信息，所述第一数据流信息包括所述张量的第二数据迁移信息，所述第二数据迁移信息表示所述张量由所述第一存储位置依次通过至少一个第三迁移位置迁移至所述第二存储位置。

503、编译器基于所述第一存储位置与所述第二存储位置之间存在直连的数据传输通路，生成第二数据流信息，其中所述第二数据流信息包括所述第一数据迁移信息。

可选的，编译器可以生成所述计算语句对应的第一数据流信息，所述第一数据流信息

包括所述张量的第二数据迁移信息，所述第二数据迁移信息表示所述张量由第一存储位置依次通过至少一个第三迁移位置迁移至第二存储位置，并基于所述第一存储位置与所述第二存储位置之间存在直连的数据传输通路，其中，（直连的数据传输通路可以理解为：第一存储位置与所述第二存储位置之间没有其他的存储位置，张量从第一存储位置可以不经过其他存储位置，直接传输到第二存储位置），生成所述第一数据流信息，所述第一数据流信息包括所述第一数据迁移信息。可选的，所述第三迁移位置指示片上缓存上的存储位置。

本申请实施例中，在计算语句包括多个算子的场景中，编译器可以生成所述计算语句对应的第二数据流信息，如图 6 中示出的那样，初始化的第二数据流信息包括了很多冗余的迁移过程，如图 7 中示出的那样，虚线指出了可优化的可选迁移路径，例如，从 UB（F(A)）、DDR（F(A)）至 L1（F(A)）的迁移路线中（UB 对应于第一存储位置，DDR 对应于至少一个第二迁移位置，L1 对应于第二存储位置），此时，参照图 3，由于 UB buffer 至 L1 buffer 之间存在数据传输通路，因此，UB buffer 可以直接将张量 F(A)传递至 L1 buffer。相应的，如图 8 中示出的那样，此时第一数据流信息包括包括与所述张量 F(A)相关联的第一数据迁移信息，所述第一数据迁移信息表示所述张量由第一存储位置（UB buffer）迁移至第二存储位置（L1 buffer）。

可选的，可以对每条迁移赋予一定的权值，权值越高，表示迁移对性能影响的代价越大。数据流权值表可以参照如下的表 1 的示意：

表 1

Buffer类型	权值
DDR	3
L1/UB	2
L0A/L0B/L0C	1

其中，DDR 的权值表示从内存中获取张量对性能的代价的量化。L1/UB 的权值表示从 L1 buffer 和从 UB buffer 中获取张量对性能的代价的量化。L01/L0B/L0C 的权值表示从 L01 buffer 、L0B buffer 和 L0C buffer 中获取张量对性能的代价的量化。

需要说明的是，编译器还需要标记数据流信息中每条边的属性是数据搬移，还是数据计算。只有数据搬移的边，在该处理流程中可以优化掉，而对于数据计算的边，在优化时是不能优化掉的。

例如，图 6 中示出的，从 DDR（A）至 UB（F(A)）的路线，由于其实涉及了对张量 A 进行数据计算（得到 F（A)），因此，DDR（A）至 UB（F(A)）的迁移路线不能删掉。

编译器可以遍历数据流图中所有节点，并计算所有入度为零的开始节点集（例如图 7 中的 DDR（A)、DDR（B）和 DDR（D））和出度为零的结束节点集（例如图 7 中的 DDR（RES)），得到该数据流图的所有可行的路径表，该路径表可以表示从开始节点集到结束节点集的可能路径。

计算每条路径对应的权值，该权值为路径上所有迁移过程权值之和，并确定权值之和最小的路径对应的数据流信息为第一数据流信息。例如，如图 8 中的示出的那样，图 8 示

出了优化后的第一数据流图。

本申请实施例中，获取第一中间表达 IR，所述第一 IR 包括计算语句，所述计算语句包括张量；基于所述计算语句生成所述计算语句对应的第一数据流信息，所述第一数据流信息包括所述张量的第二数据迁移信息，所述第二数据迁移信息表示所述张量由所述第一存储位置依次通过至少一个第三迁移位置迁移至所述第二存储位置；基于所述第一存储位置与所述第二存储位置之间存在直连的数据 传输通路，生成第二数据流信息，其中所述第二数据流信息包括所述第一数据迁移信息。通过上述方式，将所述第二数据迁移信息中多余的迁移路径删掉，在保证可以迁移到第二存储位置的基础上，减小了张量的搬移代价，减少了系统的开销。

参照图 9，图 9 为本申请实施例提供的一种编译器的结构示意图，如图 9 中示出的那样，编译器包括：

获取单元 901，用于获取第一中间表达 IR，所述第一 IR 包括计算语句，所述计算语句包括张量和算子，其中所述算子所表示的运算基于计算单元执行，所述计算单元用于对依次通过第一存储位置和第二存储位置迁移的数据执行运算，所述张量为执行所述算子所表示的运算使用的数据；

处理单元 902，用于基于所述计算语句生成第二 IR，所述第二 IR 包括第一数据迁移信息和数据切分信息，所述第一数据迁移信息表示所述张量的第一迁移路径包括由所述第一存储位置到所述第二存储位置，所述数据切分信息表示所述张量迁移至所述第二存储位置时所述张量的大小。

可选的，所述计算单元还用于对通过第二迁移路径的数据执行运算，所述第二迁移路径包括由所述第一存储位置，经过至少一个第三存储位置，到所述第二存储位置，所述第一迁移路径与所述第二迁移路径为不同的数据迁移路径。

可选的，所述至少一个第三存储位置指示片上缓存上的存储位置。

可选的，所述处理模块还用于：基于所述计算语句生成所述计算语句对应的第一数据流信息，所述第一数据流信息包括所述张量的第二数据迁移信息，所述第二数据迁移信息表示所述张量由所述第一存储位置依次通过至少一个第三迁移位置迁移至所述第二存储位置；

基于所述第一存储位置与所述第二存储位置之间存在直连的数据传输通路，生成第二数据流信息，其中所述第二数据流信息包括所述第一数据迁移信息。

可选的，所述第一存储位置指示片上缓存上的存储位置，所述第二存储位置指示片上缓存上的存储位置。

可选的，所述第一存储位置指示内存，所述第二存储位置指示片上缓存上的存储位置。

可选的，所述张量包括多个维度，每个维度对应于一个轴变量，所述轴变量用于表示对应维度的张量大小，所述数据切分信息包括多个轴变量和与每个轴变量对应的切分轴变量，所述切分轴变量用于表示张量迁移时对应维度的张量大小。

可选的，所述第二 IR 还包括至少一个目标变量，以及每个目标变量的取值范围，所述

轴变量与所述至少一个目标变量线性相关，所述轴变量通过所述至少一个目标变量，以及所述至少一个目标变量的取值范围来表示对应维度的张量大小。

可选的，所述第二 IR 为树结构的 IR，所述树结构包括根节点和子节点，所述根节点对应于所述计算语句，所述子节点对应于所述第二存储位置，其中，所述第二 IR 包括与所述第二存储位置对应的子节点的信息，所述与所述第二存储位置对应的子节点的信息包括所述第一数据迁移信息和所述数据切分信息。

可选的，所述第二 IR 还包括：数据搬移信息，所述数据搬移信息表示所述张量被从所述第一存储位置搬移至所述第二存储位置。

可选的，所述处理单元还用于：

根据所述计算单元的类型确定第一存储位置和第二存储位置。

另外需说明的是，以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。另外，本申请提供的装置实施例附图中，模块之间的连接关系表示它们之间具有通信连接，具体可以实现为一条或多条通信总线或信号线。

通过以上的实施方式的描述，所属领域的技术人员可以清楚地了解到本申请可借助软件加必需的通用硬件的方式来实现，当然也可以通过专用硬件包括专用集成电路、专用 CPU、专用存储器、专用元器件等来实现。一般情况下，凡由计算机程序完成的功能都可以很容易地用相应的硬件来实现，而且，用来实现同一功能的具体硬件结构也可以是多种多样的，例如模拟电路、数字电路或专用电路等。但是，对本申请而言更多情况下软件程序实现是更佳的实施方式。基于这样的理解，本申请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，该计算机软件产品存储在可读取的存储介质中，如计算机的软盘、U 盘、移动硬盘、ROM、RAM、磁碟或者光盘等，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，训练设备，或者网络设备等）执行本申请各个实施例所述的方法。

在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。

所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时，全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、训练设备或数据中心通过有线（例如同轴电缆、光纤、数字用户线（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、训练设备或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存储的任何可用介质或者是包含一个或多个可用介质集成的训练设备、数据中心等数据存储设备。所述可用介质可以是磁性介质，（例如，软盘、硬盘、磁带）、光介

质（例如，DVD）、或者半导体介质（例如固态硬盘（Solid State Disk，SSD））等。

权 利 要 求

1、一种构建中间表达的方法，其特征在于，包括：

获取第一中间表达 IR，所述第一 IR 包括计算语句，所述计算语句包括张量和算子，其中所述算子所表示的运算基于计算单元执行，所述计算单元用于对依次通过第一存储位置和第二存储位置迁移的数据执行运算，所述张量为执行所述算子所表示的运算使用的数据；

基于所述计算语句生成第二 IR，所述第二 IR 包括第一数据迁移信息和数据切分信息，所述第一数据迁移信息表示所述张量的第一迁移路径包括由所述第一存储位置到所述第二存储位置，所述数据切分信息表示所述张量迁移至所述第二存储位置时所述张量的大小。

2、根据权利要求 1 所述的方法，其特征在于，所述计算单元还用于对通过第二迁移路径的数据执行运算，所述第二迁移路径包括由所述第一存储位置，经过至少一个第三存储位置，到所述第二存储位置，所述第一迁移路径与所述第二迁移路径为不同的数据迁移路径。

3、根据权利要求 2 所述的方法，其特征在于，所述至少一个第三存储位置指示片上缓存上的存储位置。

4、根据权利要求 2 或 3 所述的方法，其特征在于，所述方法还包括：

基于所述计算语句生成所述计算语句对应的第一数据流信息，所述第一数据流信息包括所述张量的第二数据迁移信息，所述第二数据迁移信息表示所述张量由所述第一存储位置依次通过至少一个第三迁移位置迁移至所述第二存储位置；

基于所述第一存储位置与所述第二存储位置之间存在直连的数据传输通路，生成第二数据流信息，其中所述第二数据流信息包括所述第一数据迁移信息。

5、根据权利要求 1 至 4 任一所述的方法，其特征在于，所述第一存储位置指示片上缓存上的存储位置，所述第二存储位置指示片上缓存上的存储位置。

6、根据权利要求 1 至 5 任一所述的方法，其特征在于，所述第一存储位置指示内存，所述第二存储位置指示片上缓存上的存储位置。

7、根据权利要求 1 至 6 任一所述的方法，其特征在于，所述张量包括多个维度，每个维度对应于一个轴变量，所述轴变量用于表示对应维度的张量大小，所述数据切分信息包括多个轴变量和与每个轴变量对应的切分轴变量，所述切分轴变量用于表示张量迁移时对应维度的张量大小。

8、根据权利要求 7 所述的方法，其特征在于，所述第二 IR 还包括至少一个目标变量，以及每个目标变量的取值范围，所述轴变量与所述至少一个目标变量线性相关，所述轴变量通过所述至少一个目标变量，以及所述至少一个目标变量的取值范围来表示对应维度的张量大小。

9、根据权利要求 1 至 8 任一所述的方法，其特征在于，所述第二 IR 为树结构的 IR，所述树结构包括根节点和子节点，所述根节点对应于所述计算语句，所述子节点对应于所述第二存储位置，其中，所述第二 IR 包括与所述第二存储位置对应的子节点的信息，所述与所述第二存储位置对应的子节点的信息包括所述第一数据迁移信息和所述数据切分信

息。

10、根据权利要求 1 至 9 任一所述的方法，其特征在于，所述第二 IR 还包括：数据搬移信息，所述数据搬移信息表示所述张量被从所述第一存储位置搬移至所述第二存储位置。

11、根据权利要求 1 至 10 任一所述的方法，其特征在于，所述方法还包括：

根据所述计算单元的类型确定第一存储位置和第二存储位置。

12、一种编译器，其特征在于，包括：

获取单元，用于获取第一中间表达 IR，所述第一 IR 包括计算语句，所述计算语句包括张量和算子，其中所述算子所表示的运算基于计算单元执行，所述计算单元用于对依次通过第一存储位置和第二存储位置迁移的数据执行运算，所述张量为执行所述算子所表示的运算使用的数据；

处理单元，用于基于所述计算语句生成第二 IR，所述第二 IR 包括第一数据迁移信息和数据切分信息，所述第一数据迁移信息表示所述张量的第一迁移路径包括由所述第一存储位置到所述第二存储位置，所述数据切分信息表示所述张量迁移至所述第二存储位置时所述张量的大小。

13、根据权利要求 12 所述的编译器，其特征在于，所述计算单元还用于对通过第二迁移路径的数据执行运算，所述第二迁移路径包括由所述第一存储位置，经过至少一个第三存储位置，到所述第二存储位置，所述第一迁移路径与所述第二迁移路径为不同的数据迁移路径。

14、根据权利要求 13 所述的编译器，其特征在于，所述至少一个第三存储位置指示片上缓存上的存储位置。

15、根据权利要求 12 或 13 所述的编译器，其特征在于，所述处理模块还用于：基于所述计算语句生成所述计算语句对应的第一数据流信息，所述第一数据流信息包括所述张量的第二数据迁移信息，所述第二数据迁移信息表示所述张量由所述第一存储位置依次通过至少一个第三迁移位置迁移至所述第二存储位置；

基于所述第一存储位置与所述第二存储位置之间存在直连的数据传输通路，生成第二数据流信息，其中所述第二数据流信息包括所述第一数据迁移信息。

16、根据权利要求 12 至 15 任一所述的编译器，其特征在于，所述第一存储位置指示片上缓存上的存储位置，所述第二存储位置指示片上缓存上的存储位置。

17、根据权利要求 12 至 16 任一所述的编译器，其特征在于，所述第一存储位置指示内存，所述第二存储位置指示片上缓存上的存储位置。

18、根据权利要求 12 至 17 任一所述的编译器，其特征在于，所述张量包括多个维度，每个维度对应于一个轴变量，所述轴变量用于表示对应维度的张量大小，所述数据切分信息包括多个轴变量和与每个轴变量对应的切分轴变量，所述切分轴变量用于表示张量迁移时对应维度的张量大小。

19、根据权利要求 18 所述的编译器，其特征在于，所述第二 IR 还包括至少一个目标变量，以及每个目标变量的取值范围，所述轴变量与所述至少一个目标变量线性相关，所述轴变量通过所述至少一个目标变量，以及所述至少一个目标变量的取值范围来表示对应

维度的张量大小。

20、根据权利要求 12 至 19 任一所述的编译器，其特征在于，所述第二 IR 为树结构的 IR，所述树结构包括根节点和子节点，所述根节点对应于所述计算语句，所述子节点对应于所述第二存储位置，其中，所述第二 IR 包括与所述第二存储位置对应的子节点的信息，所述与所述第二存储位置对应的子节点的信息包括所述第一数据迁移信息和所述数据切分信息。

21、根据权利要求 12 至 20 任一所述的编译器，其特征在于，所述第二 IR 还包括：数据搬移信息，所述数据搬移信息表示所述张量被从所述第一存储位置搬移至所述第二存储位置。

22、根据权利要求 12 至 21 任一所述的编译器，其特征在于，所述处理单元还用于：

根据所述计算单元的类型确定第一存储位置和第二存储位置。

23、一种计算机系统，其特征在于，包括：处理器和存储器；

所述处理器和所述存储器电连接；

所述处理器用于调用所述存储器中的代码开执行如权利要求 1 至 11 任一所述的方法。

24、一种计算机存储介质，其特征在于，所述计算机可读存储介质存储指令，当所述指令在计算机上运行时，使得所述计算机执行权利要求 1 至 11 中任一项所述的方法。

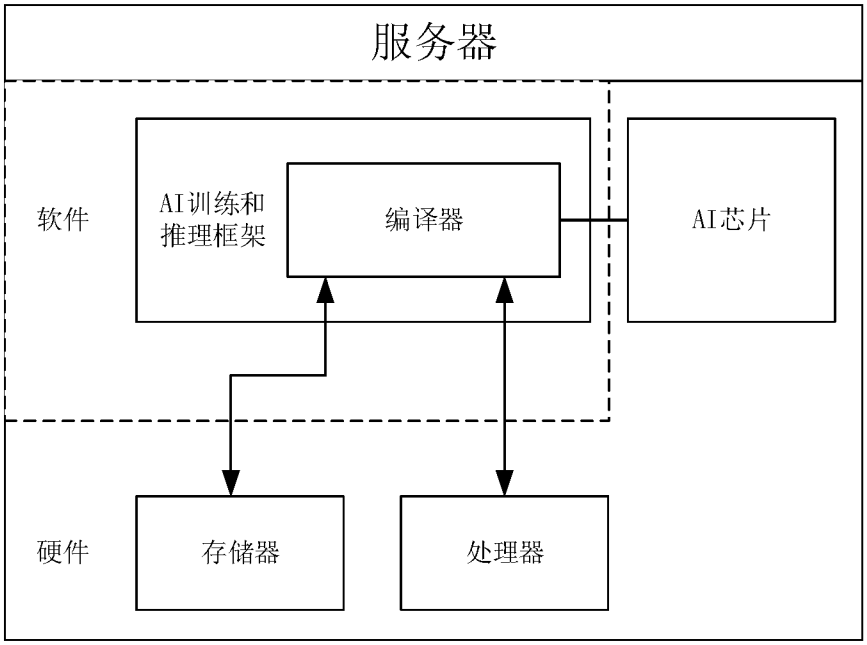


图 1

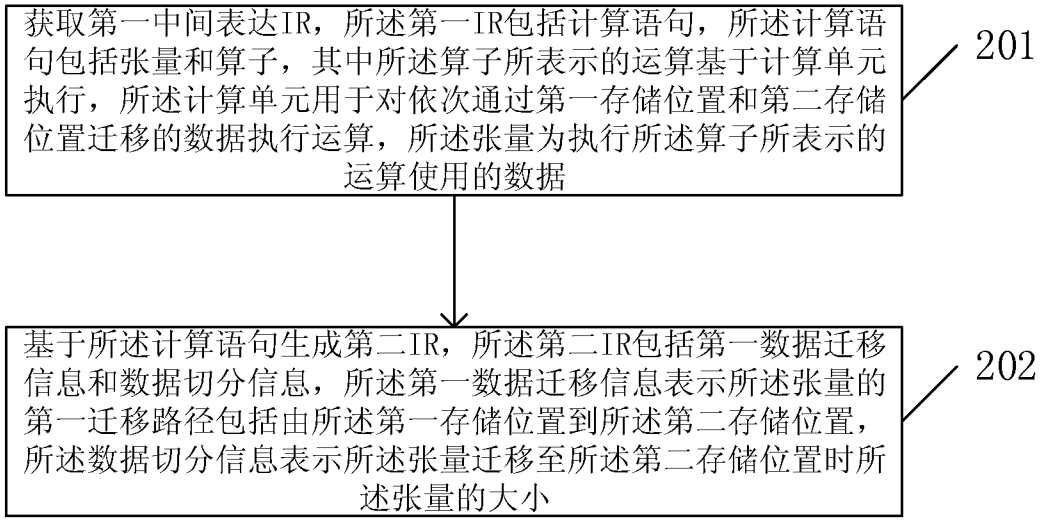

获取第一中间表达IR，所述第一IR包括计算语句，所述计算语句包括张量和算子，其中所述算子所表示的运算基于计算单元执行，所述计算单元用于对依次通过第一存储位置和第二存储位置迁移的数据执行运算，所述张量为执行所述算子所表示的运算使用的数据
201
基于所述计算语句生成第二IR，所述第二IR包括第一数据迁移信息和数据切分信息，所述第一数据迁移信息表示所述张量的第一迁移路径包括由所述第一存储位置到所述第二存储位置，所述数据切分信息表示所述张量迁移至所述第二存储位置时所述张量的大小
202

图 2

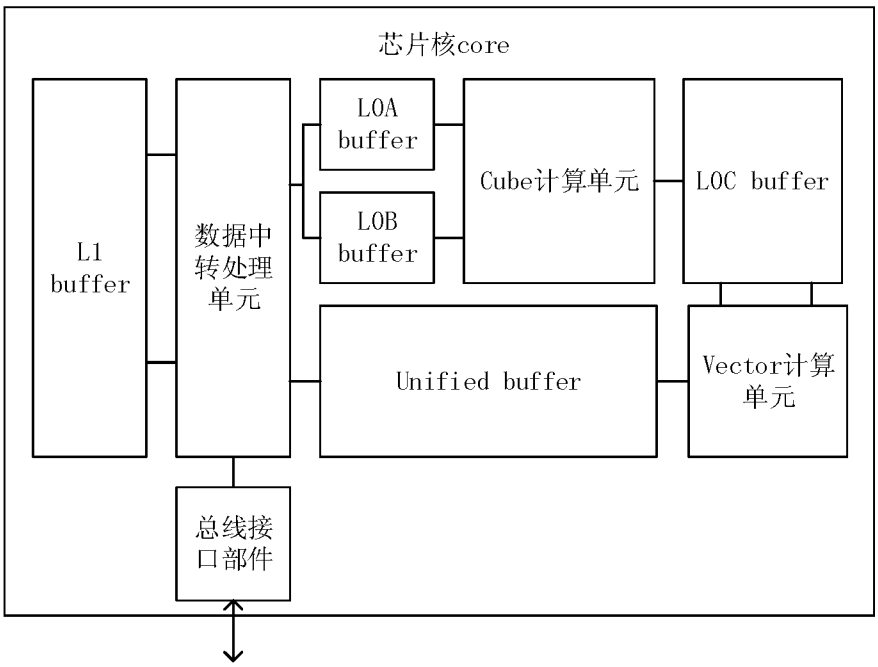

芯片核core
L1 buffer
数据中转处理单元
L0A buffer
L0B buffer
Cube计算单元
L0C buffer
Unified buffer
Vector计算单元
总线接口部件

图 3

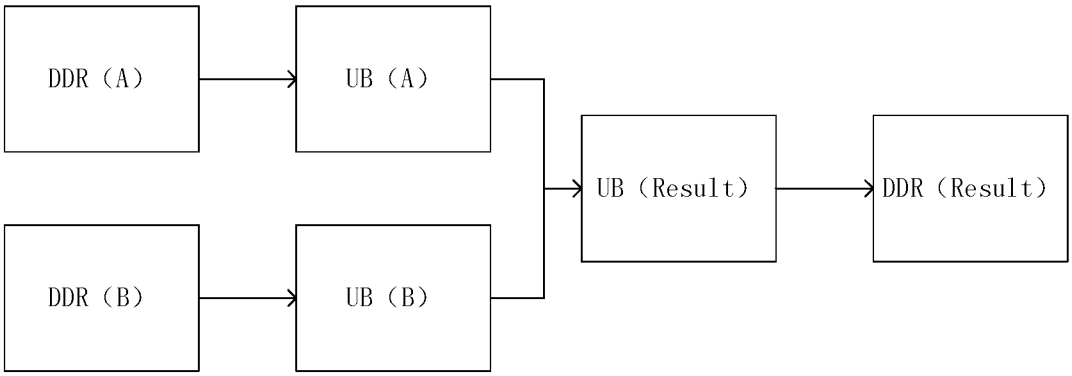

DDR（A）
UB（A）
DDR（B）
UB（B）
UB（Result）
DDR（Result）

图 4

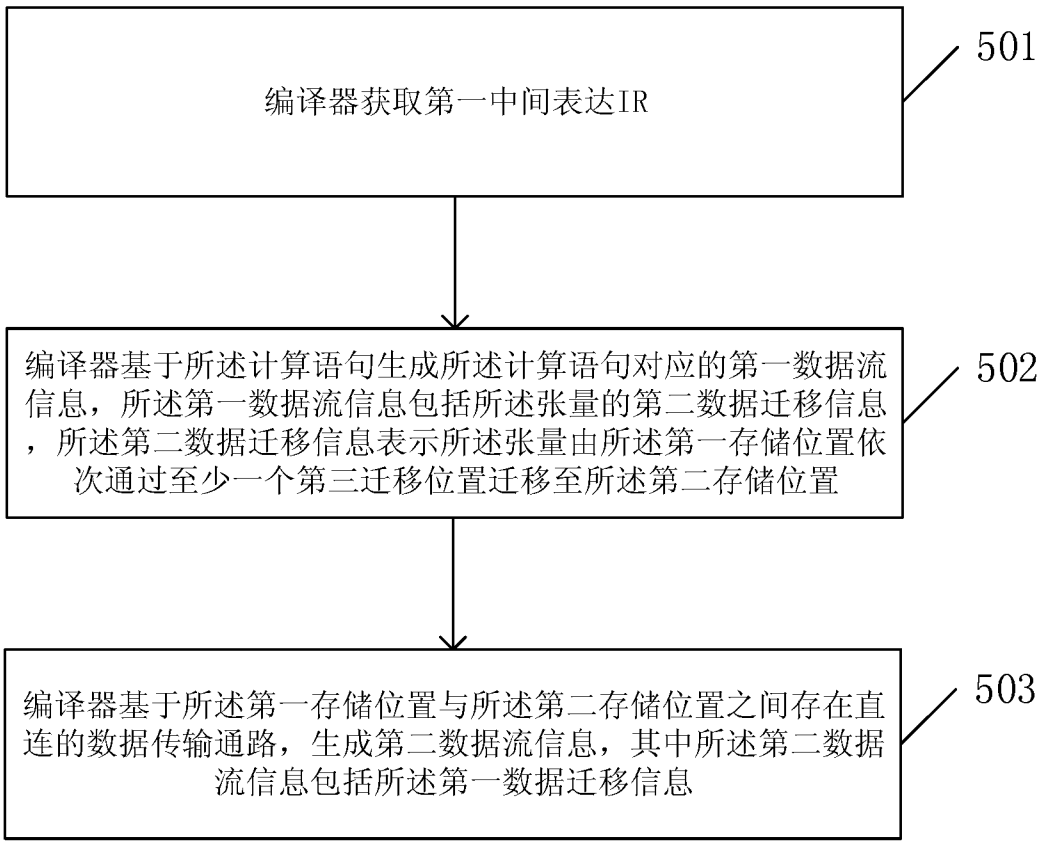

编译器获取第一中间表达IR
501
编译器基于所述计算语句生成所述计算语句对应的第一数据流信息，所述第一数据流信息包括所述张量的第二数据迁移信息，所述第二数据迁移信息表示所述张量由所述第一存储位置依次通过至少一个第三迁移位置迁移至所述第二存储位置
502
编译器基于所述第一存储位置与所述第二存储位置之间存在直连的数据传输通路，生成第二数据流信息，其中所述第二数据流信息包括所述第一数据迁移信息
503

图 5

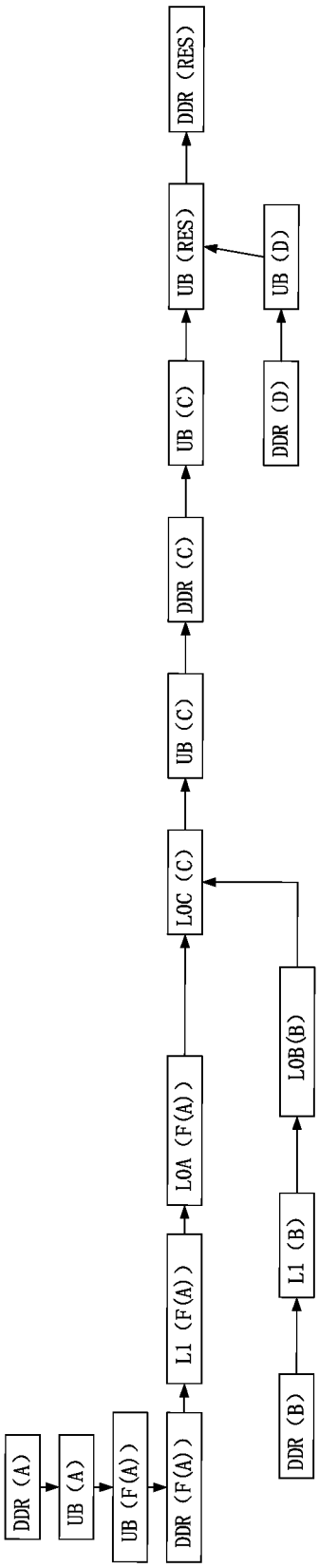

DDR (A)
UB (A)
UB (F(A))
DDR (F(A))
L1 (F(A))
LOA (F(A))
LOC (C)
UB (C)
DDR (C)
UB (C)
UB (RES)
DDR (RES)
DDR (B)
L1 (B)
LOB(B)
DDR (D)
UB (D)

图6

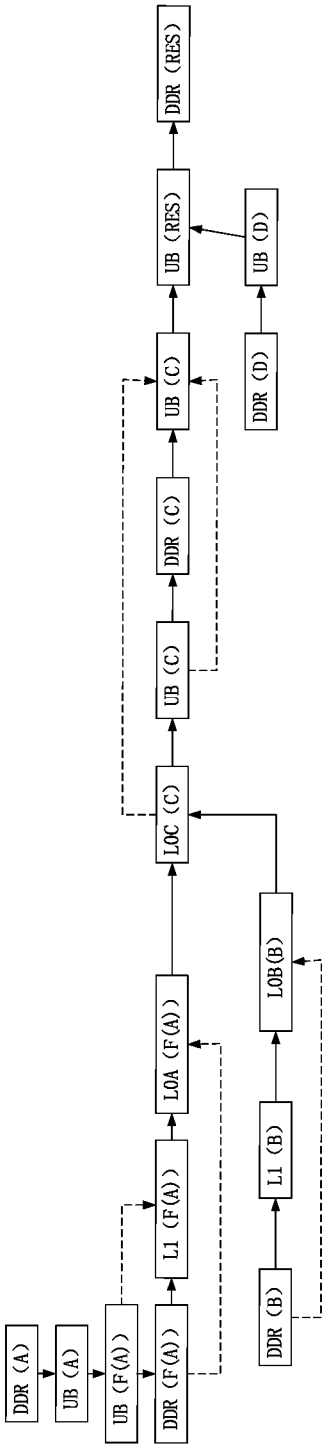

DDR (A)
UB (A)
UB (F(A))
DDR (F(A))
L1 (F(A))
LOA (F(A))
LOC (C)
UB (C)
DDR (C)
UB (C)
UB (RES)
DDR (RES)
DDR (B)
L1 (B)
LOB(B)
DDR (D)
UB (D)

图7

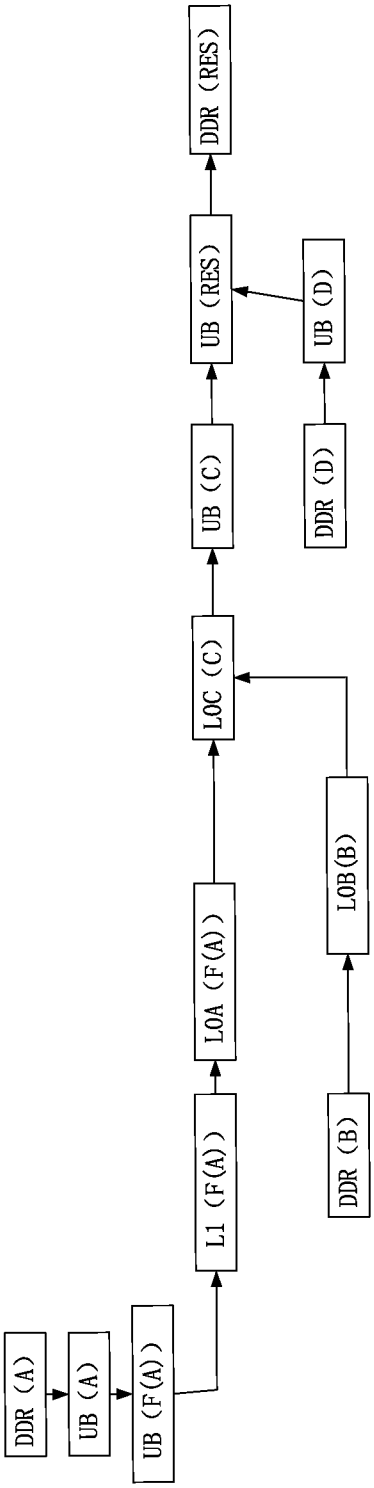

DDR (A)
UB (A)
UB (F(A))
L1 (F(A))
LOA (F(A))
LOC (C)
UB (C)
UB (RES)
DDR (RES)
DDR (B)
LOB(B)
DDR (D)
UB (D)

图8

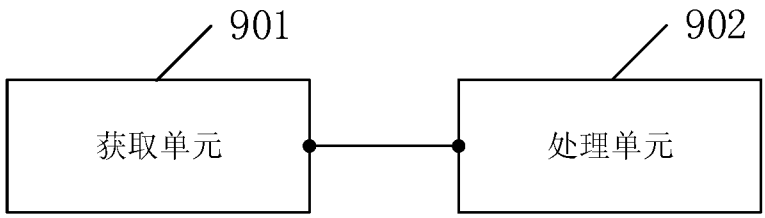


图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2020/115759

A. CLASSIFICATION OF SUBJECT MATTER

G06F 8/41(2018.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, EPODOC, WPI, CNKI: 中间表达, 中间表示, 算子, 计算, 运算, 编译, 语句, 图, 树, 代码, 迁移, 存储, 位置, 路径, 数据, 张量, 切分, 分片, 分配, IR, intermediate representation, operator, compiler, address list, graph, tree, coder, circle, migration, storage, location, data, tensor, DSL, split, distribution

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2018136912 A1 (THE MATHWORKS, INC.) 17 May 2018 (2018-05-17) description, paragraphs [0019]-[0022], [0061]-[0104], [0138]-[0142] and [0204]-[0211], and figures 1-10	1-6, 9-17, 20-24
A	CN 108292374 A (GOOGLE INC.) 17 July 2018 (2018-07-17) entire document	1-24
A	CN 108292241 A (GOOGLE INC.) 17 July 2018 (2018-07-17) entire document	1-24
A	CN 108345937 A (GOOGLE INC.) 31 July 2018 (2018-07-31) entire document	1-24
A	WO 2018217222 A1 (THE CHARLES STARK DRAPER LABORATORY, INC.) 29 November 2018 (2018-11-29) entire document	1-24

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
02 December 2020	**18 December 2020**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2020/115759

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	2018136912	A1	17 May 2018	EP	3403221	A1	21 November 2018
				WO	2018094099	A1	24 May 2018
CN	108292374	A	17 July 2018	US	2017132513	A1	11 May 2017
				EP	3374933	A2	19 September 2018
				WO	2017083399	A2	18 May 2017
CN	108292241	A	17 July 2018	JP	2018538607	A	27 December 2018
				US	2017124452	A1	04 May 2017
				KR	20180077218	A	06 July 2018
				WO	2017075438	A1	04 May 2017
				US	2018247197	A1	30 August 2018
				KR	20200015829	A	12 February 2020
				EP	3353656	A1	01 August 2018
				EP	3705994	A1	09 September 2020
				JP	2020129404	A	27 August 2020
CN	108345937	A	31 July 2018	DE	102018100239	A1	12 July 2018
				SG	10201800155P	A	30 August 2018
				US	9798527	B1	24 October 2017
				GB	2560410	A	12 September 2018
				GB	201800179	D0	21 February 2018
				WO	2018129327	A1	12 July 2018
				HK	1259126	A1	22 November 2019
WO	2018217222	A1	29 November 2018	US	2018341493	A1	29 November 2018

国际检索报告

国际申请号
PCT/CN2020/115759

A. 主题的分类

G06F 8/41(2018.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT,EPODOC,WPI,CNKI： 中间表达，中间表示，算子，计算，运算，编译，语句，图，树,代码，迁移，存储，位置，路径，数据，张量，切分，分片，分配，IR, intermediate representation, operator, compiler, address list, graph, tree, coder, circle, migration, storage, location, data, tensor, DSL, split, distribution

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 2018136912 A1 (MATHWORKS, INC.) 2018年 5月 17日 (2018-05-17) 说明书第[0019]-[0022]、[0061]-[0104]、[0138]-[0142]、[0204]-[0211]段、图1-10	1-6，9-17，20-24
A	CN 108292374 A (谷歌有限责任公司) 2018年 7月 17日 (2018-07-17) 全文	1-24
A	CN 108292241 A (谷歌有限责任公司) 2018年 7月 17日 (2018-07-17) 全文	1-24
A	CN 108345937 A (谷歌有限责任公司) 2018年 7月 31日 (2018-07-31) 全文	1-24
A	WO 2018217222 A1 (THE CHARLES STARK DRAPER LABORATORY, INC.) 2018年 11月 29日 (2018-11-29) 全文	1-24

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2020年 12月 2日	2020年 12月 18日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 孙娟 电话号码 (86-10)53961383

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2020/115759

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	2018136912	A1	2018年 5月 17日	EP	3403221	A1	2018年 11月 21日
				WO	2018094099	A1	2018年 5月 24日
CN	108292374	A	2018年 7月 17日	US	2017132513	A1	2017年 5月 11日
				EP	3374933	A2	2018年 9月 19日
				WO	2017083399	A2	2017年 5月 18日
CN	108292241	A	2018年 7月 17日	JP	2018538607	A	2018年 12月 27日
				US	2017124452	A1	2017年 5月 4日
				KR	20180077218	A	2018年 7月 6日
				WO	2017075438	A1	2017年 5月 4日
				US	2018247197	A1	2018年 8月 30日
				KR	20200015829	A	2020年 2月 12日
				EP	3353656	A1	2018年 8月 1日
				EP	3705994	A1	2020年 9月 9日
				JP	2020129404	A	2020年 8月 27日
CN	108345937	A	2018年 7月 31日	DE	102018100239	A1	2018年 7月 12日
				SG	10201800155P	A	2018年 8月 30日
				US	9798527	B1	2017年 10月 24日
				GB	2560410	A	2018年 9月 12日
				GB	201800179	D0	2018年 2月 21日
				WO	2018129327	A1	2018年 7月 12日
				HK	1259126	A1	2019年 11月 22日
WO	2018217222	A1	2018年 11月 29日	US	2018341493	A1	2018年 11月 29日