

一种算子编译方法以及装置

5 本申请要求于 2023 年 5 月 31 日提交中国国家知识产权局、申请号为 202310639274.4、发明名称为“一种数据处理系统”的中国专利申请的优先权和要求于 2023 年 8 月 22 日提交中国国家知识产权局、申请号为 202311070464.5、发明名称为“一种算子编译方法以及装置”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

10 本申请实施例涉及计算机领域，尤其涉及一种算子编译方法以及装置。

背景技术

近年来，随着大数据等新兴应用的快速发展，信息量的爆炸式增长带来计算规模飞速扩大，计算形式日趋多元化，多元的应用场景对硬件计算系统的计算性能和存储性能均提出了更高的要求。例如，在面向大数据的应用场景中，当前主流硬件中的存算分离架构严重制约了性能和能效提升，由于处理器运算时都需要把数据从存储器经过数据总线搬运到处理器中，数据处理完之后再搬运回存储器中。随着数据量的扩大，这一过程的延时也随之增大，甚至超过存储器和处理器本身的读写和处理延时，另外大量数据在存储器和处理器间的频繁搬运也带来了严重的传输功耗问题。

为了解决存算分离架构下的数据搬运能效问题，存算一体架构应运而生，存算一体架构通过在存储器芯片内部实现嵌入式计算功能来大幅减少存储器和处理器间的数据搬运规模和频次，以显著提高计算系统的算力和能效。但是，目前应用于存算一体架构中的指令集的大都为指令集都为标量指令集，存储器利用标量指令集处理计算任务时，需要对逐个标量进行处理，导致计算任务的处理并行度低，无法充分发挥存算一体架构优势。

25 发明内容

申请实施例提供了一种算子编译方法，该方法中存算一体系统能够利用算子编译方法编译得到张量指令集，并利用张量指令集执行存算一体系统的计算任务，从而提升计算任务的处理并行度。本申请实施例还提供了算子编译方法对应的计算设备、计算设备集群、计算机可读存储介质以及计算机程序产品。

30 第一方面，本申请实施例提供了一种算子编译方法，该方法可以由存算一体系统执行，也可以由存算一体系统的部件，例如存算一体系统的处理器、芯片或芯片系统等执行，还可以由能实现全部或部分存算一体系统功能的逻辑模块或软件实现。第一方面提供的法包括：存算一体系统获取编译算子，编译算子用于指示对源代码的编译规则。存算一体系统基于编译算子对源代码进行编译，得到一条或多条张量指令，张量指令用于执行存算一体系统中计算任务，张量指令包括以下一项或多项指令段：操作码、张量地址、数据类型、数据位宽和张量长度。存算一体系统将张量指令存储至指令存储列表，指令存储列表用于存储编译完成的张量指令。

35 本申请实施例中存算一体系统能够基于编译算子对源代码进行编译得到张量指令，并通过执行张量指令实现对计算任务中张量的处理，相较于目前的存算一体系统利用现有的标量指令处理计算任务，本申请实施例提供的算子编译方法以及张量指令集提升了存算一体系统处理计算任务的处理并行度，进而提升存算一体系统的计算性能。

40 在一些可能的实现方式中，存算一体系统基于编译算子对源代码进行编译之前，存算一体系统创建编译算子表，编译算子表用于存储一个或多个编译算子对应的编译信息，编译信息包括以下一项或多项：算子编号、行地址、列地址、输出地址、数据类型、数据位宽和张量长度。存算一体系统基于编译算子对源代码进行编译的过程中，存算一体系统直接在编译算子表中查询编译算子和编译信息，根据编译算子和编译信息生成源代码对应的张量指令。

45 本申请实施中存算一体系统能够将源代码中编译信息预先存入编译算子表，从而在进行源代码编译的过程中，存算一体系统能够基于编译算子表快速编译生成源代码信息对应的张量指令，从而提升了张

量指令的编译效率。

在一些可能的实施方式中，存算一体系统基于所述算子编译表中存储的编译信息生成一条或多条张量指令的过程中，存算一体系统根据编译算子表中的算子编号生成张量指令的操作码，根据编译算子表中的行地址、列地址和输出地址生成张量指令的张量地址，根据编译算子表中的数据类型、数据位宽和
5 张量长度分别生成张量指令的数据类型、数据位宽和张量长度。

本申请实施例中存算一体系统能够基于编译算子表快速编译生成源代码对应的张量指令，从而提升了张量指令的编译效率。

在一些可能的实现方式中，编译算子的类型包括筛选（filter）算子、排序（sort）算子和聚合（Aggregation）算子。筛选算子的算子结构包括操作码、行地址、列地址、输出地址、输入数据类型
10 和张量长度。排序算子的算子结构包括操作码、行地址、列地址、输入数据类型和张量长度。聚合算子的算子结构包括操作码、行地址、列地址、输入数据类型和张量长度。

本申请实施例中存算一体系统提供了多种编译算子，多种编译算子能够对不同的源代码进行编译得到张量指令，从而提升了张量指令集的丰富性。

在一些可能的实现方式中，指令类型包括以下一项或多项：仅地址参与指令、数值参与指令、阵列控制指令和近存电路控制指令。其中，仅地址参与指令是仅有张量地址参与的和指令类型。数值参与指令是存在数值的张量指令。阵列控制指令是指进行阵列操作的指令。近存电路控制指令是对阵列 ID 进行指定的近存电路操作的指令。
15

本申请实施例中存算一体系统能够编译得到不同指令类型的张量指令，从而利用不同张量类型的张量指令执行多种存算操作，从而提升了对不同存算操作的适用性。

在一些可能的实现方式中，不同类型的张量指令具有不同的指令段。仅地址参与指令的指令段包括操作码、张量地址、类型、位宽和张量长度。数值参与指令的指令段包括操作码、张量地址、数值、类型、位宽和张量长度。阵列控制指令的指令段包括操作码、张量地址、数值和张量长度。近存电路控制指令的指令段包括操作码、存储阵列标识、近存电路类型和近存电路状态。
20

本申请实施例中存算一体系统提供了不同指令类型的张量指令以及指令段格式，从而提升了方案的
25 可实现性。

在一些可能的实现方式中，存算一体系统根据指令关联矩阵和阵列状态列表确定指令存储列表中的张量指令的指令状态，指令关联矩阵用于存储张量指令之间的依赖关系，阵列状态列表用于指示阵列模组的状态，阵列模组用于执行张量指令，指令状态包括可执行状态和等待状态。

本申请实施例中存算一体系统能够基于指令关联矩阵和阵列状态列表确定张量指令的指令状态，进而确定张量指令的执行顺序，提升了存算一体系统的计算效率。
30

在一些可能的实现方式中，当张量指令的指令状态处于可执行状态时，存算一体系统向访存控制单元发送张量指令。存算一体系统根据张量指令确定执行操作，执行操作包括以下一项或多项：近存计算操作、存内计算操作和常规读写操作。

本申请实施例中存算一体系统生成的张量指令能够适用于多种的存算一体架构，从而提升了张量指令对存算一体系统的适用性。
35

在一些可能的实现方式中，张量指令集第一张量指令和第二张量指令，第一张量指令和第二张量指令在相同的阵列模组中执行，存算一体系统在阵列模组的不同执行单元中分别同时执行第一张量指令和第二张量指令，执行单元包括阵列和近存电路。

本申请实施例存算一体系统中阵列模组的不同执行单元能够同时执行不同的张量指令，从而提升张量指令的执行并行度和阵列模组的硬件利用效率，进一步提升了存算一体系统的计算效率。
40

在一些可能的实现方式中，张量指令之间的依赖关系包括以下一项或多项：读后写依赖、写后写依赖和写后读依赖。其中，读后写依赖是指在并发计算中一个指令的读操作在另一个指令写操作之后发生。写后写依赖是指在并发计算中一个指令的写操作在另一个指令的写操作之后发生。写后读依赖是指在并发计算中一个指令的写操作在另一个指令的读操作之后发生。

本申请实施例中指令关联矩阵能够存储多种张量指令之间的依赖关系，从而降低了张量指令执行过程中的指令冲突概率，提升了存算一体系统的计算效率。
45

在一些可能的实现方式中,阵列状态列表包括以下一项或多项:阵列模组空闲状态和近存电路状态。阵列模组空闲状态能够指示阵列模组空闲和忙碌状态,近存电路状态能够指示近存电路中不同运算的空闲状态。存算一体系统能够通过访存控制单元采集阵列模组空闲状态和近存电路状态,并将阵列模组空闲状态和近存电路状态存储至阵列状态列表。

5 本申请实施例中阵列状态列表中能够阵列模组空闲状态和近存电路状态,从而提升了存算一体系统执行张量指令的执行效率。

在一些可能的实现方式中,指令存储列表包括以下一项或多项:张量指令的指令段、张量指令编号、编译算子编号和阵列块标识。

10 本申请实施例中存算一体系统通过指令存储列表存储编译完成的张量指令,从而提升了张量指令的查询效率。

第二方面,本申请实施例提供了一种算子编译装置,该装置包括获取单元、编译单元和处理单元。获取单元用于获取编译算子,编译算子用于指示对源代码的编译规则。编译单元用于基于编译算子对源代码进行编译,得到一条或多条张量指令,张量指令用于执行存算一体系统中计算任务,张量指令包括以下一项或多项指令段:操作码、张量地址、数据类型、数据位宽和张量长度。处理单元用于将张量指令存储至指令存储列表,指令存储列表用于存储编译完成的张量指令。

15 在一些可能的实现方式中,处理单元还用于创建编译算子表,编译算子表用于存储一个或多个编译算子对应的编译信息,编译信息包括以下一项或多项:算子编号、行地址、列地址、输出地址、数据类型、数据位宽和张量长度;编译单元具体用于基于编译算子表中存储的编译信息生成一条或多条张量指令。

20 在一些可能的实现方式中,编译单元具体用于根据编译算子表中的算子编号生成张量指令的操作码,根据编译算子表中的行地址、列地址和输出地址生成张量指令的张量地址,根据编译算子表中的数据类型、数据位宽和张量长度分别生成张量指令的数据类型、数据位宽和张量长度。

在一些可能的实现方式中,指令类型包括以下一项或多项:仅地址参与指令、数值参与指令、阵列控制指令和近存电路控制指令。

25 在一些可能的实现方式中,编译单元还用于根据指令关联矩阵和阵列状态列表确定指令存储列表中的张量指令的指令状态,指令关联矩阵用于存储张量指令之间的依赖关系,阵列状态列表用于指示阵列模组的状态,阵列模组用于执行张量指令,指令状态包括可执行状态和等待状态。

30 在一些可能的实现方式中,处理单元还用于当张量指令的指令状态处于可执行状态时,向访存控制单元发送张量指令。根据张量指令确定执行操作,执行操作包括以下一项或多项:近存计算操作、存内计算操作和常规读写操作。

在一些可能的实现方式中,张量指令集第一张量指令和第二张量指令,第一张量指令和第二张量指令在相同的阵列模组中执行,处理单元还用于在阵列模组的不同执行单元中分别同时执行第一张量指令和第二张量指令,执行单元包括阵列和近存电路。

35 在一些可能的实现方式中,张量指令之间的依赖关系包括以下一项或多项:读后写依赖、写后写依赖和写后读依赖。

在一些可能的实现方式中,阵列状态列表包括以下一项或多项:阵列模组空闲状态和近存电路状态。

在一些可能的实现方式中,指令存储列表包括以下一项或多项:张量指令的指令段、张量指令编号、编译算子编号和阵列块标识。

40 第三方面,本申请实施例提供了一种计算设备,计算设备包括处理器,处理器与存储器耦合,处理器用于存储指令,当指令被处理器执行时,以使得计算设备执行上述第一方面或第一方面任意一种可能的实现方式所述的方法。

第四方面,本申请实施例提供了一种计算设备集群,计算设备集群包括一个或多个计算设备,计算设备包括处理器,处理器与存储器耦合,处理器用于存储指令,当指令被处理器执行时,以使得计算设备集群执行上述第一方面或第一方面任意一种可能的实现方式所述的方法。

45 第五方面,本申请实施例提供了一种计算机可读存储介质,其上存储有指令,指令被执行时,以使得计算机执行上述第一方面或第一方面任意一种可能的实现方式所述的方法。

第六方面，本申请实施例提供了一种计算机程序产品，计算机程序产品中包括指令，指令被执行时，以使得计算机实现上述第一方面或第一方面任意一种可能的实现方式所述的方法。

可以理解，上述提供的任何一种算子编译装置、计算设备、计算设备集群、计算机可读介质或计算机程序产品等所能达到的有益效果可参考对应的方法中的有益效果，此处不再赘述。

附图说明

- 图 1 为本申请实施例提供了一种存算一体系统的系统架构示意图；
- 图 2 为本申请实施例提供了一种算子编译方法的流程示意图；
- 图 3 为本申请实施例提供了一种张量指令执行流程示意图；
- 图 4 为本申请实施例提供了一种编译算子表的示意图；
- 图 5 为本申请实施例提供了一种基于编译算子生成张量指令的示意图；
- 图 6 为本申请实施例提供了一种张量指令的示意图；
- 图 7 为本申请实施例提供的另一种基于编译算子生成张量指令的示意图；
- 图 8 为本申请实施例提供了一种指令存储列表的示意图；
- 图 9 为本申请实施例提供了一种指令关联矩阵的示意图；
- 图 10 为本申请实施例提供了一种阵列状态列表的示意图；
- 图 11 为本申请实施例提供了一种执行张量指令的流程示意图；
- 图 12 为本申请实施例提供了一种存算一体模组内不同执行单元的执行状态示意图；
- 图 13 为本申请实施例提供了一种算子编译装置的结构示意图；
- 图 14 为本申请实施例提供了一种计算设备的结构示意图；
- 图 15 为本申请实施例提供了一种计算设备集群的结构示意图。

具体实施方式

本申请实施例提供了一种算子编译方法以及装置，用于提升存算一体系统的执行计算任务处理并行度。

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”、“第三”、“第四”等（如果存在）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数据在适当情况下可以互换，以便这里描述的实施例能够以除了在这里图示或描述的内容以外的顺序实施。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，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，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步骤或单元。

在本申请实施例中，“示例性的”或者“例如”等词用于表示作例子、例证或说明。本申请实施例中描述为“示例性的”或者“例如”的任何实施例或设计方案不应被解释为比其它实施例或设计方案更优选或更具优势。确切而言，使用“示例性的”或者“例如”等词旨在以具体方式呈现相关概念。

首先，介绍本申请实施例中涉及的部分术语，方便本领域技术人员理解技术方案。

存算一体架构是一种将存储功能和计算功能集成在一起的技术架构。存算一体架构通过在存储器芯片内部实现嵌入式计算功能来减少存储器和处理器间的数据搬运规模和频次，从而提高系统算力和能效。

近存计算是指将存储器和逻辑处理模块的物理距离拉近，使得数据传输的延时和功耗得以降低。

存内计算是指在存储器内部进行运算，从消除可能的数据传输，降低系统的延时和功耗。

乱序执行（out of order execution, OOE）是一种应用在高性能微处理器中来利用指令周期以避免特定类型的延迟消耗。处理器乱序执行中能够根据输入数据的可用性确定执行指令的顺序。在这种方式下，可以避免因为获取下一条程序指令所引起的处理器等待，可以处理下一条可以立即执行的指令。

为了使得本申请的技术方案更加清楚、易于理解，下面结合附图对本申请的系统架构进行介绍。

请参阅图 1，图 1 为本申请实施例提供了一种存算一体系统的系统架构示意图。在图 1 所示的示例中，存算一体系统 10 的为中央处理器 20 下发的存算一体任务，存算一体任务例如大数据分析任务、机

器学习任务、图像处理任务、多媒体编解码任务和数据库查询和存储任务等。存算一体系统 10 用于执行存算一体计算任务。在存算一体系统中，本申请实施例对存算一体任务和计算任务不做区分。

如图 1 所示，存算一体系统 10 包括编译器 101 和存算一体芯片 102，其中，存算一体芯片 102 包括调度器 1021、访存控制器单元 1022 和阵列模组 1023。下面对编译器 101 和存算一体芯片 102 分别进行详细介绍。

编译器 101 用于接收中央处理器 20 发送的存算一体任务，并将存算一体任务对应的源代码编译为指令，该指令的指令格式即为存算一体芯片 102 能够直接执行的指令格式。编译器 101 编译得到的一条或多条指令可以构成指令集，由于存算一体系统 10 能够利用该指令集在存算一体芯片 102 进行张量计算，该指令集也可以称为存算一体指令集或张量指令集。编译器 101 还用于将编译完成的张量指令集发送至存算一体芯片 102，由存算一体芯片 102 存储并执行该张量指令集。

存算一体芯片 102 用于在芯片内实现存算一体任务的处理。具体的，存算一体芯片 102 用于执行编译器 101 编译完成的张量指令集，通过执行张量指令集完成中央处理器 20 下发的存算一体任务。存算一体芯片 102 包括调度器 1021、访存控制器单元 1022 和阵列模组 1023，下面分别介绍这几个部分的具体功能。

其中，调度器 1021 用于接收编译器 101 编译完成的张量指令集，并将张量指令集存储至指令存储列表中。调度器 1021 还用于基于指令关联矩阵和阵列状态列表确定张量指令的指令状态，当指令状态为可执行状态时，调度器 1021 调度访存控制单元 1022 执行张量指令。

访存控制单元 1022 用于具体执行张量指令。具体的，访存控制单元 1022 用于根据张量指令生成阵列模组 1023 对应的电路控制信号，并利用该电路控制信号控制阵列模组 1023 执行存储和计算等操作。访存控制单元 1022 可以控制一个或多个阵列模组 1023 执行存储和计算等操作。

阵列模组 1023 用于执行存算一体任务。具体的，阵列模组 1023 包括近存电路和存储阵列。其中，近存电路用于执行近存计算操作，并将执行近存计算操作后的数据通过缓存存储到存储阵列。存储阵列用于进行数据读写操作，除此之外，存储阵列还用于进行存内计算操作，并将存内计算操作的结果通过缓存存储在存储阵列。

基于图 1 所示的存算一体系统 10，本申请还提供一种算子编译方法。下面结合实施例，对本申请的算子编译方法进行介绍。

请参阅图 2，图 2 为本申请实施例提供的一种算子编译方法的流程示意图。在图 2 所示的示例中，该方法包括如下步骤：

S201. 存算一体系统获取编译算子。

存算一体系统 10 接收中央处理器 20 发送的编译算子，编译算子用于指示编译器 101 对源代码的编译规则。具体的，编译器 101 可以基于编译算子对源代码进行词法分析、语法分析和代码优化等。存算一体系统 10 还可以接收中央处理器 20 发送的存算一体任务对应的源代码，编译器 101 能够基于该编译算子对源代码进行分析和转换，生成存算一体系统 10 能够执行的具体指令。

请参阅图 3，图 3 为本申请实施例提供的一种张量指令执行的流程示意图。在图 3 所示示例的步骤 1 中，存算一体系统 10 接收中央处理器 20 发送的编译算子，并基于编译算子对源代码进行编译。例如，存算一体系统 10 接收的编译算子为筛选算子，筛选算子可以根据特定的条件从数据集中选择出符合条件的元素，并将其返回为一个新的子集。

S202. 存算一体系统基于编译算子对源代码进行编译，得到张量指令，张量指令用于执行存算一体系统的计算任务。

存算一体系统 10 获取源代码之后，首先在内存中创建编译算子表，该编译算子表用于存储存储一个或多个编译算子对应的编译信息，编译信息包括以下一项或多项：算子编号、行地址、列地址、输出地址、数据类型、数据位宽和张量长度。存算一体系统 10 按照编译算子表的表结构将源代码中编译信息存储至编译算子表。

在上述编译算子表中，行地址为一条源代码中的条件语句后的张量的地址，列地址为一条源代码中涉及的张量的首地址，也称为待查询的首地址。算子编号是指不同编译算子对应的标识。位宽是指一条源代码占用的地址位数。下面举例介绍本申请实施例提供的编译算子表。

请参阅图 4，图 4 为本申请实施例提供的一种基于编译算子表进行源代码编译的示意图。在图 4 所示的示例中，表 (a) 为存算一体系统 10 创建的编译算子表，编译算子表的表项包括算子编号、行地址、列地址、输出地址、数据类型、数据位宽和张量长度。

其中，行地址为一条源代码中的条件语句后的张量的地址。例如，对于源代码 “A(A==B)→C”，表示判断 A 是否等于 B，若 A 等于 B，则将比较的结果传递给 C。其中条件语句为 “判断 A 是否等于 B”，因此，条件语句后的张量地址为张量 B 的地址，因此，存算一体系统 10 将张量 B 的地址作为行地址存入编译算子表中。

在图 4 所示的示例中，列地址为一条源代码中涉及的张量的首地址，也称为待查询的首地址。再例如，在上述源代码 “A(A==B)→C” 中，源代码中涉及的张量的首地址为张量 A 的地址，因此，存算一体系统 10 将张量 A 的地址作为列地址存入编译算子表中。

在图 4 所示的示例中，算子编号是指不同编译算子对应的编号。例如，在上述源代码为 “A(A==B)→C” 中，源代码为比较运算，其对应的编译算子为筛选 (filter) 算子，筛选算子对应的算子编号为 “tab1”，因此，存算一体系统 10 将 “tab1” 存入编译算子表中。

在图 4 所示的示例中，位宽是指算一条源代码所占用的地址位数。例如，在上述源代码 “A(A==B)→C” 占用的地址位数为 100 位，因此，存算一体系统 10 将 “100” 存入编译算子表。

可以理解的是，存算一体系统 10 创建的编译算子表可以存储源代码涉及的一个或多个编译算子，以及编译算子对应的编译信息。因此，存算一体系统 10 可以建立算子编号和编译信息之间的映射的关系，从而在源代码编译过程中快速查询编译算子表，确定张量指令的指令段的内容。

下面继续介绍存算一体系统 10 创建编译算子表后，基于编译算子表对源代码进行编译的过程。

存算一体系统 10 基于编译算子和编译算子表生成源代码对应的张量指令的过程中，首先，存算一体系统 10 基于编译算子的算子结构确定张量指令的各个指令段，并基于算子结构对应的编译信息确定各个指令段的具体内容。具体的，基于编译算子的算子编号查询编译算子表中的编译信息，确定相应指令段中要填入的编译信息，从而生成张量指令的每段指令段的内容，得到张量指令。下面举例介绍源代码的编译过程。

请继续参阅图 4，在图 4 所示的示例中，存算一体系统 10 基于编译算子的算子结构确定张量指令的各个指令段，其中不同的编译算子具有不同的算子结构。例如，筛选算子的算子结构包括操作码、行地址、列地址、输出地址、输入数据类型和张量长度。存算一体系统 10 基于筛选算子确定张量指令的各个指令段，筛选算子对应的张量指令的指令段包括操作码、变量 A 地址、变量 B 地址、变量 C 地址、类型、位宽和张量长度。其中，变量 A 地址对应算子结构的行地址，变量 B 地址对应算子结构的列地址，变量 C 地址对应算子结构的输出地址，类型对应算子结构的输入数据类型。

在 4 所示的示例中，存算一体系统 10 基于编译算子生成张量指令的过程中，基于筛选算子的算子编号在编译算子表中查询对应的行地址、列地址、输出地址、数据类型和张量长度，并将查询到的行地址、列地址、输出地址、数据类型和张量长度作为张量指令相应指令段的内容。例如，筛选算子的算子编号为 “tab1”，存算一体系统 10 基于 “tab1” 查询编译算子表的中行地址为 “row1”，列地址为 “0x000010”，存算一体系统 10 即可以得到张量指令的指令段中变量 A 地址和变量 B 地址。存算一体系统 10 可以进一步基于筛选算子的其他算子结构对源代码进行编译，得到完整的张量指令。

本申请实施例中编译算子的算子结构包括以下一项或多项：操作码、行地址、列地址、输出地址、输入数据类型和张量长度。其中，不同的编译算子类型对应不同的算子结构，编译算子的类型例如筛选 (filter) 算子、排序 (sort) 算子和聚合 (Aggregation) 算子。筛选算子的算子结构包括操作码、行地址、列地址、输出地址、输入数据类型和张量长度。排序算子的算子结构包括操作码、行地址、列地址、输入数据类型和张量长度。聚合算子的算子结构包括操作码、行地址、列地址、输入数据类型和张量长度。

本申请实施例中编译算子的算子结构与张量指令的指令段存在映射关系。例如，张量指令的指令段包括一项或多项：操作码、变量 A 地址、变量 B 地址、变量 C 地址、类型、位宽和张量长度。算子结构中的操作码、行地址、列地址、输出地址、输入数据类型和张量长度分别对应指令段中的操作码、变量 A 地址、变量 B 地址、变量 C 地址、类型和张量长度。

请参阅图 5，图 5 本申请实施例提供的一种基于编译算子的算子结构确定张量指令的指令段的示意图。在图 5 所示的示例中，存算一体系统 10 基于筛选算子、排序算子和聚合算子的算子结构生成对应的张量指令。

在图 5 所示的示例中，以筛选算子为例，筛选算子的算子结构包括操作码、行地址、列地址、输出地址、输入数据类型和张量长度。存算一体系统 10 基于筛选算子的算子结构，通过查询编译算子表确定张量地址，并进一步编译得到张量指令的指令段的操作码、变量 A 地址、变量 B 地址、变量 C 地址、类型、位宽和张量长度。

例如，张量指令的操作码可基于源代码“A(A=B)→C”确定为比较操作，对应的操作码为“00101110”，变量 A 地址可以通过查询编译算子表的列地址确定为“0x1111”。变量 B 地址可以通过查询编译算子表的行地址确定为“0x1001”。变量 C 地址、类型、位宽和张量长度则可以基于源代码和编译算子的算子结构分别确定为“0x3442”、“0”、“01”和“000000101”。

本申请实施例中一个编译算子可以对应一种或中不同的指令操作码。例如，在图 4 所示的示例中，筛选算子包括比较（compare，cmp）、大于等于（greater than or equal，gte）、大于（greater than，gt）、小于等于（less than or equal，lt）、小于（Less Than，lt）等操作码，排序算子包括按升序排序（ascending）和按降序排序（descending）等操作码，聚合算子包括求和（sum）、求平均（avg）和计数（count）等操作码。

本申请实施例中不同类型的张量指令具有不同的指令段。其中，张量指令的指令类型包括下一项或多项类型：仅地址参与指令、数值参与指令、阵列控制指令和近存电路控制指令。下面具体介绍这些指令类型已经对应的指令段。

其中，仅地址参与指令是仅有张量地址参与的和指令类型。仅地址参与指令例如从张量 A 地址和张量 B 地址分别读取 A 和 B 进行指定的操作后将结果 C 写入张量 C 地址。仅地址参与指令的指令段包括以下一项或多项：操作码、张量地址、类型、位宽和张量长度。其中张量地址可以是多个张量的地址，张量长度是指张量中包含元素的数量。

数值参与指令是存在数值的张量指令。数值参与指令例如从张量 A 地址读取 A，结合数值 C 进行指定操作后将结果 B 写入张量 B 地址。数值参与指令的指令段包括以下一项或多项：操作码、张量地址、数值、类型、位宽和张量长度。

阵列控制指令是指进行阵列操作的指令。阵列控制指令例如从张量 A 地址读取 A，结合数值 C 进行指定操作对应的张量 A 地址的阵列进行控制操作。阵列控制指令的指令段包括以下一项或多项：操作码、张量地址、数值和张量长度。

近存电路控制指令是对阵列 ID 进行指定的近存电路操作的指令。近存电路控制指令具体由指令中的近存电路类型和近存电路状态决定。近存电路控制指令的指令段包括以下一项或多项：操作码、存储阵列 ID、近存电路类型和近存电路状态。

请参阅图 6，图 6 为本申请实施例提供的一种不同指令类型的张量指令的指令段示意图。在图 6 所示的示例中，仅地址参与指令的指令段包括操作码、张量 A 地址、张量 B 地址、张量 C 地址、类型、位宽和张量长度。数值参与指令的指令段包括操作码、张量 A 地址、张量 B 地址、数值 C、类型、位宽和张量长度。阵列控制指令的指令段包括操作码、张量 A 地址、数值 C 和张量长度。近存电路控制指令的指令段包括操作码、存储阵列 ID、近存电路类型和近存电路状态。

从图 6 所示的示例中可以看出，当指令对应的类型不需要指令段中的部分分段时，可以将那部分指令段对应的地址或是数值置 0 以表示跳过。在上述指令段中，除了张量地址和数值之外，还有类型、位宽、张量长度的指令段。其中，类型包括整型数和浮点数，整型数包括 1 至 16bit 的数，浮点数包括 BF16、FP16、FP32 和 FP64 等。张量长度用于标识张量地址的拓展地址长度，当张量长度为 1 时所表示的张量数据就等于标量数据。

需要说明的是，存算一体系统 10 基于编译算子编译源代码时，基于不同的编译算子，可以对一条破源代码编译得到一条或多条张量指令。下面结合附图举例说明。

请参阅图 7，图 7 为本申请实施例提供的一种基于编译算子编译得到张量指令的示意图。在图 7 所示的示例中，存算一体系统 10 基于单一编译算子编译得到一条或多条张量指令，例如，存算一体系统

10 基于筛选算子对“select * from t where t.a>10”进行编译得到 1 条张量指令，即张量指令 0。
再例如，存算一体系统 10 基于投影（Projection）算子对“select a+b*c from t”进行编译得到 2 条张量指令，即张量指令 1 和张量指令 2。

在图 7 所示的示例中，单条源代码命令可以被编译为单条或多条存张量指令的组合。例如，表中的
5 三条算子对应不同的源代码命令，筛选算子用于按顺序是取大于 10 数，投影算子用于计算 $|a+b*c|$ ，升序排序算子用于升序排序。筛选算子对应命令 1，投影算子对应命令 1 和命令 2，升序排序算子对应命令 3。其中，命令 0 和命令 2 可以使用单条指令实现。而命令 1 和命令 3 都需要多条指令实现。

S203. 存算一体系统将张量指令存储至指令存储列表。

存算一体系统 10 编译得到张量之后，将张量指令存储至指令存储列表。其中，指令存储列表包括
10 以下一项或多项：张量指令的指令段、张量指令编号、编译算子编号和阵列块标识。下面具体介绍指令存储列表。

请参阅图 8，图 8 为本申请实施例提供的一种指令存储列表的示意图。在图 8 所示的示例中，存算
一体系统 10 生成张量指令的先后顺序将张量指令存储至指令存储列表，先存入指令存储列表的指令将
15 会优先执行。指令存储列表中除了存储张量指令的指令段之外，还存储了张量指令的张量指令编号、编译算子编号、读写地址、阵列块标识和读写坐标位置。

例如，指令存储列表中存储了指令标号为“0”、编译算子编号为“0”的张量指令的指令段，该张量
指令的读写地址分别“[0x0011, 0x0104]”和“[0x0111]”，张量指令进行操作的存储阵列的阵列块标识
为“0”，张量指令进行操作的存储阵列读写坐标位置分别为“[(1, 4, 1, 8), (2, 3, 4, 7)]”和
“[(0, -1, 0, -1)]”。

20 请继续参阅图 3，在图 3 所示示例的步骤 2 中，存算一体系统 10 基于编译算子对源代码编译生成张量指令之后，将张量指令存储至指令存储列表。

S204. 存算一体系统基于指令关联矩阵和阵列状态列表执行指令存储列表中的张量指令。

存算一体系统 10 基于指令关联矩阵和阵列状态列表执行指令存储列表中的张量指令。具体的，存
算一体系统 10 根据指令关联矩阵和阵列状态列表确定指令存储列表中的张量指令的指令状态，指令状
25 态包括可执行状态和等待状态。其中，指令关联矩阵用于存储张量指令之间的依赖关系，阵列状态列表
用于指示阵列模组的状态，阵列模组用于执行张量指令。

请继续参阅图 3，在图 3 所示示例的步骤 2 至步骤 7 中，存算一体系统 10 基于得到编译完成的张量
指令之后，对张量指令进行解码，并将解码后的张量指令存储至指令存储列表，存算一体系统 10 基于
指令关联矩阵和阵列状态列表确定张量指令是否可执行，若张量指令可以执行，则向访存控制单元发送
30 张量指令，若张量指令不可执行，则在指令存储列表中继续等待执行。

下面具体介绍指令关联矩阵和阵列状态列表。

本申请实施例中指令关联矩阵用于存储指令存储列表中各指令之间的关联关系。指令关联矩阵为一
个长宽都与指令存储列表中张量指令数相同的矩阵，矩阵元素表示指令依赖关系。张量指令之间的依赖
关系包括以下一项或多项：读后写依赖、写后写依赖和写后读依赖。

35 其中，读后写依赖指的是在并发计算中一个指令的读操作在另一个指令写操作之后发生，即一个指
令要在另一个指令完成对某个共享数据的写入后，才能读取该共享数据。写后写依赖指的是在并发计算
中一个指令的写操作在另一个指令的写操作之后发生，即一个指令要在另一个指令完成对某个共享数据
的写入后，才能对该共享数据进行修改。写后读依赖指的是在并发计算中一个指令的写操作在另一个指
令的读操作之后发生，即一个指令要在另一个指令完成对某个共享数据的读取后，才能对该共享数据进
40 行修改。

请参阅图 9，图 9 为本申请实施例提供的一种指令关联矩阵的示意图。在图 9 所示的示例中，指令
关联矩阵中的矩阵元素中 (i, j) 的值代表第 i 个指令对第 j 个指令的冲突关系，其中，0 表示两者之
间没有地址冲突，1 表示写后读 (write after read, WAR) 依赖关系，2 表示写后写 (write after write,
WAW) 依赖关系，3 表示读后写 (read after write, RAW) 依赖关系。例如，指令 1 读后写依赖指令 0，
45 指令 2 写后写依赖指令 1，指令 3 写后读依赖指令 1 并且指令 3 写后写依赖指令 2。

本申请实施例中阵列状态列表用于查询存储阵列的状态以及近存电路的状态，从而确定张量指令是

否可以发送到相应的存储阵列或者近存电路中执行。阵列状态列表包括以下一项或多项：阵列模组空闲状态和近存电路状态。

请参阅图 10，图 10 为本申请实施例提供的一种阵列状态列表的示意图。在图 10 所示的阵列状态列表中，阵列状态列表显示编号分别为“CB1”、“CB3”、“CB3”和“CB4”共 4 个存储阵列的状态，其中，
5 阵列状态为“1”表示阵列处于忙碌状态，阵列状态为“0”表示阵列处于空闲状态。阵列状态列表还可以记录近存电路的状态，例如，阵列状态列表还能够记录近存电路中加法运算器的状态或者近存电路中异或运算器的状态。

存算一体系统 10 根据指令关联矩阵和阵列状态列表确定指令存储列表中的张量指令的指令状态，具体的，当待执行的张量指令满足指令关联矩阵的依赖关系，并且，张量指令进行操作的存储阵列或者
10 近存电路处于空闲状态，则该张量指令的指令状态为可执行状态。本申请实施例中存算一体系统 10 基于张量指令的指令状态向访存控制单元发送张量指令的方式也称作乱序执行。

当张量指令的指令状态处于可执行状态时，向访存控制单元发送张量指令。存算一体系统 10 根据张量指令确定执行操作，执行操作包括以下一项或多项：近存计算操作、存内计算操作和常规读写操作。其中，近存计算操作是指经由阵列或缓存读取所需数据，再调用对应近存电路完成计算的操作。存内计
15 算操作是指直接利用存储阵列进行运算。常规读写操作是指对存储阵列进行读写操作。上述不同的存算类型中张量指令的计算并行度不同。

下面结合示例介绍存算一体系统 10 执行张量指令的几种操作。

请参阅图 11，图 11 为本申请实施例提供的一种访存控制单元执行存算一体操作的示意图。在图 11 所示的示例中，访存控制单元 1022 接收到张量指令之后，生成存算一体阵列模型组 1023 对应的电路控
20 制信号，存算一体阵列模型组 1023 基于该控制电路信号确定存算操作的类型。

在图 11 所示的示例中，当存算操作的类型为近存计算操作时，阵列模组 1023 经由阵列或缓存读取所需数据，再调用对应近存电路完成计算的操作。在近存计算过程中，存算一体芯片 102 中阵列模组 1023 进行单行阵列读取，然后经由缓存传输至近存电路。

在图 11 所示的示例中，当存算操作的类型为存内计算操作时，阵列模组 1023 直接利用存储阵列执行计算操作，例如，在存储阵列内进行多行阵列累乘、累加操作，并经由缓存将计算结果写入存储阵列。
25 当存算操作的类型为常规读写操作时，阵列模组 1023 对存储阵列每个数据单元中的数据进行读写操作，并经由缓存传输数据。

本申请实施中在阵列模组 1023 的不同执行单元中分别同时执行第一张量指令和第二张量指令，不同执行单元包括阵列和近存电路，第一张量指令和第二张量指令为不同的张量指令。下面结合示例进行
30 介绍。

请参阅图 12，图 12 为本申请实施例提供的一种不同张量指令在存算一体阵列模型执行的示意图。在图 12 所示的示例中，其中图（a）为 4 条不同张量指令示意图，图（b）为不同张量指令的在不同时间的执行状态示意图。在图（a）所示的 4 条张量指令中，指令 0、指令 1 和指令 2 都分布在一个阵列模组当中，不能并行读写。指令 3 分布在另一个阵列模组当中，其处理过程和指令 0、指令 1 和指令 2 是
35 可以并行的。同时，指令 1 和指令 0 之间存在读后写冲突，因此，指令 1 需要等待指令 0 写入完成后才可以进行读出。

从图（b）所示的示例中可以看出，在指令执行过程中，首先指令 0 和指令 1 并行进入调度，阵列模组 1023 识别两者之间的读后写 RAW 依赖关系，因此，指令 1 开始等待指令 0 的完成。在这期间，指令 2 和指令 3 也输入调度。指令 3 由于在另一个阵列模组上，不存在和其他指令的冲突，因此可以不受
40 限制地执行。

从图（b）所示的示例中可以看出，阵列模组 1023 的阵列和近存电路可以分别同时执行指令 0 和指令 2。例如，阵列模组 1023 的近存电路计算时，存储阵列处于空闲状态，因此，在指令 0 进行近存电路计算操作时，指令 2 可以同时进行读取操作。

另外，从图（b）所示的示例中还可以看出，由于指令 2 作为 max 指令，因此，读取部分分为两步，
45 分别为识别最大值和读出最大值。在完成第一步读取之后，指令 0 完成对指令 2 的等待，指令 0 占用了阵列完成写入操作，之后指令 2 继续进行读取操作。因此，指令 1 在指令 2 完成写入后才可以空

阵列上完成写操作。

从以上实施例中可以看出，本申请实施例中存算一体系统能够基于编译算子生成张量指令集，并通过张量指令集能够对张量对象进行处理，从而实现对存算一体系统的存储阵列级操作，提升了存算一体系统处理计算任务的处理并行度，进而提升存算一体系统的计算性能。

5 基于上述方法实施例，本申请实施例还提供一种算子编译装置，下面具体介绍本申请实施例提供的算子编译装置。

请参阅图 13，图 13 为本申请实施例提供的一种算子编译装置的结构示意图。在图 13 所示的示例中，算子编译装置 1300 用于实现上述各实施例中存算一体系统所执行的各个步骤，该算子编译装置 1300 包括获取单元 1301、编译单元 1302 和处理单元 1303。

10 获取单元 1301 用于获取编译算子，编译算子用于指示对源代码的编译规则。编译单元 1302 用于基于编译算子对源代码进行编译，得到一条或多条张量指令，张量指令用于执行存算一体系统中计算任务，张量指令包括以下一项或多项指令段：操作码、张量地址、数据类型、数据位宽和张量长度。处理单元 1303 用于将张量指令存储至指令存储列表，指令存储列表用于存储编译完成的张量指令。

15 在一些可能的实现方式中，处理单元 1303 还用于创建编译算子表，编译算子表用于存储一个或多个编译算子对应的编译信息，编译信息包括以下一项或多项：算子编号、行地址、列地址、输出地址、数据类型、数据位宽和张量长度。编译单元 1302 具体用于基于编译算子表中存储的编译信息生成一条或多条张量指令。

20 在一些可能的实现方式中，编译单元 1302 具体用于根据编译算子表中的算子编号生成张量指令的操作码，根据编译算子表中的行地址、列地址和输出地址生成张量指令的张量地址，根据编译算子表中的数据类型、数据位宽和张量长度分别生成张量指令的数据类型、数据位宽和张量长度。

在一些可能的实现方式中，指令类型包括以下一项或多项：仅地址参与指令、数值参与指令、阵列控制指令和近存电路控制指令。

25 在一些可能的实现方式中，编译单元 1302 还用于根据指令关联矩阵和阵列状态列表确定指令存储列表中的张量指令的指令状态，指令关联矩阵用于存储张量指令之间的依赖关系，阵列状态列表用于指示阵列模组的状态，阵列模组用于执行张量指令，指令状态包括可执行状态和等待状态。

在一些可能的实现方式中，处理单元 1303 还用于当张量指令的指令状态处于可执行状态时，向访存控制单元发送张量指令。根据张量指令确定执行操作，执行操作包括以下一项或多项：近存计算操作、存内计算操作和常规读写操作。

30 在一些可能的实现方式中，张量指令集第一张量指令和第二张量指令，第一张量指令和第二张量指令在相同的阵列模组中执行，处理单元 1303 还用于在阵列模组的不同执行单元中分别同时执行第一张量指令和第二张量指令，执行单元包括阵列和近存电路。

在一些可能的实现方式中，张量指令之间的依赖关系包括以下一项或多项：读后写依赖、写后写依赖和写后读依赖。

在一些可能的实现方式中，阵列状态列表包括以下一项或多项：阵列模组空闲状态和近存电路状态。

35 在一些可能的实现方式中，指令存储列表包括以下一项或多项：张量指令的指令段、张量指令编号、编译算子编号和阵列块标识。

40 应理解以上装置中单元的划分仅仅是一种逻辑功能的划分，实际实现时可以全部或部分集成到一个物理实体上，也可以物理上分开。且装置中的单元可以全部以软件通过处理元件调用的形式实现；也可以全部以硬件的形式实现；还可以部分单元以软件通过处理元件调用的形式实现，部分单元以硬件的形式实现。例如，各个单元可以为单独设立的处理元件，也可以集成在装置的某一个芯片中实现，此外，也可以以程序的形式存储于存储器中，由装置的某一个处理元件调用并执行该单元的功能。此外这些单元全部或部分可以集成在一起，也可以独立实现。这里所述的处理元件又可以成为处理器，可以是一种具有信号的处理能力的集成电路。在实现过程中，上述方法的各步骤或以上各个单元可以通过处理器元件中的硬件的集成逻辑电路实现或者以软件通过处理元件调用的形式实现。

45 值得说明的是，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，本申请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申请所必须的。

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，也属于本申请的保护范围内。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申请所必须的。

5 请参阅图 14，图 14 为本申请实施例提供的一种计算设备的结构示意图。如图 14 所示，该计算设备 1400 包括：处理器 1401、存储器 1402、通信接口 1403 和总线 1404，处理器 1401、存储器 1402 与通信接口 1403 通过总线（图中未标注）耦合。存储器 1402 存储有指令，当存储器 1402 中的执行指令被执行时，计算设备 1400 执行上述方法实施例中存算一体系统所执行的方法。

10 计算设备 1400 可以是配置成实施以上方法的一个或多个集成电路，例如：一个或多个特定集成电路（application specific integrated circuit, ASIC），或，一个或多个微处理器（digital signal processor, DSP），或，一个或者多个现场可编程门阵列（field programmable gate array, FPGA），或这些集成电路形式中至少两种的组合。再如，当装置中的单元可以通过处理元件调度程序的形式实现时，该处理元件可以是通用处理器，例如中央处理器（central processing unit, CPU）或其它可以调用程序的处理器。再如，这些单元可以集成在一起，以片上系统（system-on-a-chip, SOC）的形式实现。

15 处理器 1401 可以是中央处理单元（central processing unit, CPU），还可以是其它通用处理器、数字信号处理器（digital signal processor, DSP）、专用集成电路（application specific integrated circuit, ASIC）、现场可编程门阵列（field programmable gate array, FPGA）或者其它可编程逻辑器件、晶体管逻辑器件，硬件部件或者其任意组合。通用处理器可以是微处理器，也可以是任何常规的处理器。

20 存储器 1402 可以是易失性存储器或非易失性存储器，或可包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器（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。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（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）。

30 存储器 1402 中存储有可执行的程序代码，处理器 1401 执行该可执行的程序代码以分别实现前述获取单元 1301、处理单元 1302 和处理单元 1303 的功能，从而实现上述路由表的压缩方法。也即，存储器 1402 上存有用于执行上述路由表的压缩方法的指令。

通信接口 1403 使用例如但不限于网络接口卡、收发器一类的收发模块，来实现计算设备 1400 与其他设备或通信网络之间的通信。

35 总线 1404 除包括数据总线之外，还可以包括电源总线、控制总线和状态信号总线等。总线可以是快捷外围部件互连标准（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。总线可以分为地址总线、数据总线、控制总线等。

40 请参阅图 15，图 15 为本申请实施例提供的一种计算设备集群的示意图。如图 15 所示，该计算设备集群 1500 包括至少一台计算设备 1400。

如图 15 所示，所述计算设备集群 1500 包括至少一个计算设备 1400。计算设备集群 1500 中的一个或多个计算设备 1400 中的存储器 1402 中可以存有相同的用于执行上述路由表的压缩方法的指令。

45 在一些可能的实现方式中，该计算设备集群 1500 中的一个或多个计算设备 1400 的存储器 1402 中也可以分别存有用于执行上述路由表的压缩方法的部分指令。换言之，一个或多个计算设备 1400 的组合可以共同执行用于执行上述路由表的压缩方法的指令。

需要说明的是，计算设备集群 1500 中的不同的计算设备 1400 中的存储器 1402 可以存储不同的指令，分别用于执行上述算子编译装置的部分功能。也即，不同的计算设备 1400 中的存储器 1402 存储的指令可以实现获取单元 1301、编译单元 1302 和处理单元 1303 中的一个或多个模块的功能。

在一些可能的实现方式中，计算设备集群 1500 中的一个或多个计算设备 1400 可以通过网络连接。
5 其中，所述网络可以是广域网或局域网等等。

在本申请的另一个实施例中，还提供一种计算机可读存储介质，计算机可读存储介质中存储有计算机执行指令，当设备的处理器执行该计算机执行指令时，设备执行上述方法实施例中存算一体系统所执行的方法。

在本申请的另一个实施例中，还提供一种计算机程序产品，该计算机程序产品包括计算机执行指令，
10 该计算机执行指令存储在计算机可读存储介质中。当设备的处理器执行该计算机执行指令时，设备执行上述方法实施例中存算一体系统所执行的方法。

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

在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式
15 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。

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
20 者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

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

所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个
25 计算机可读取存储介质中。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或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（ROM，
30 read-only memory）、随机存取存储器（RAM，random access memory）、磁碟或者光盘等各种可以存储程序代码的介质。

权 利 要 求 书

1. 一种算子编译方法，其特征在于，应用于存算一体系统，所述方法包括：
获取编译算子，所述编译算子用于指示源代码的编译规则；
基于所述编译算子对所述源代码进行编译，得到一条或多条张量指令，所述张量指令用于执行所述
5 存算一体系统中计算任务，所述张量指令包括以下一项或多项指令段：操作码、张量地址、数据类型、
数据位宽和张量长度；
将所述张量指令存储至指令存储列表，所述指令存储列表用于存储编译完成的张量指令。
2. 根据权利要求 1 所述的方法，其特征在于，所述基于所述编译算子对所述源代码进行编译之前，
所述方法还包括：
10 创建编译算子表，所述编译算子表用于存储一个或多个所述编译算子对应的编译信息，所述编译信息
包括以下一项或多项：算子编号、行地址、列地址、输出地址、数据类型、数据位宽和张量长度；
所述基于所述编译算子对所述源代码进行编译包括：
基于所述编译算子表中存储的编译信息生成一条或多条张量指令。
3. 根据权利要求 2 所述的方法，其特征在于，所述基于所述算子编译表中存储的编译信息生成一条
15 或多条张量指令包括：
根据所述编译算子表中的算子编号生成所述张量指令的操作码；
根据所述编译算子表中的行地址、列地址和输出地址生成所述张量指令的张量地址；
根据所述编译算子表中的数据类型、数据位宽和张量长度分别生成所述张量指令的数据类型、数据
位宽和张量长度。
4. 根据权利要求 1 至 3 中任一项所述的方法，其特征在于，所述指令类型包括以下一项或多项：仅
20 地址参与指令、数值参与指令、阵列控制指令和近存电路控制指令。
5. 根据权利要求 1 至 4 中任一项所述的方法，其特征在于，所述方法还包括：
根据指令关联矩阵和阵列状态列表确定所述指令存储列表中的张量指令的指令状态，所述指令关联
矩阵用于存储所述张量指令之间的依赖关系，所述阵列状态列表用于指示阵列模组的状态，所述阵列模
25 组用于执行所述张量指令，所述指令状态包括可执行状态和等待状态。
6. 根据权利要求 5 所述的方法，其特征在于，所述方法还包括：
当所述张量指令的指令状态处于所述可执行状态时，向访存控制单元发送所述张量指令；
根据所述张量指令确定执行操作，所述执行操作包括以下一项或多项：近存计算操作、存内计算操
作和常规读写操作。
7. 根据权利要求 1 至 6 中任一项所述的方法，其特征在于，所述张量指令集第一张量指令和第二张
30 量指令，所述第一张量指令和所述第二张量指令在相同的阵列模组中执行，所述方法还包括：
在所述阵列模组的不同执行单元中分别同时执行所述第一张量指令和所述第二张量指令，所述执行
单元包括阵列和近存电路。
8. 根据权利要求 5 至 7 中任一项所述的方法，其特征在于，所述张量指令之间的依赖关系包括以下
35 一项或多项：读后写依赖、写后写依赖和写后读依赖。
9. 根据权利要求 5 至 8 中任一项所述的方法，其特征在于，所述阵列状态列表包括以下一项或多项：
阵列模组空闲状态和近存电路状态。
10. 根据权利要求 5 至 9 中任一项所述的方法，其特征在于，所述指令存储列表包括以下一项或多
项：所述张量指令的指令段、张量指令编号、编译算子编号和阵列块标识。
11. 一种算子编译装置，其特征在于，所述装置包括：
40 获取单元，用于获取编译算子，所述编译算子用于指示对源代码的编译规则；
编译单元，用于基于所述编译算子对所述源代码进行编译，得到一条或多条张量指令，所述张量指
令用于执行所述存算一体系统中计算任务，所述张量指令包括以下一项或多项指令段：操作码、张量地
址、数据类型、数据位宽和张量长度；
45 处理单元，用于将所述张量指令存储至指令存储列表，所述指令存储列表用于存储编译完成的张量

指令。

12. 根据权利要求 11 所述的装置，其特征在于，所述处理单元还用于：

创建编译算子表，所述编译算子表用于存储一个或多个所述编译算子对应的编译信息，所述编译信息包括以下一项或多项：算子编号、行地址、列地址、输出地址、数据类型、数据位宽和张量长度；

5 所述编译单元具体用于：

基于所述编译算子表中存储的编译信息生成一条或多条张量指令。

13. 根据权利要求 12 所述的装置，其特征在于，所述编译单元具体用于：

根据所述编译算子表中的算子编号生成所述张量指令的操作码；

根据所述编译算子表中的行地址、列地址和输出地址生成所述张量指令的张量地址；

10 根据所述编译算子表中的数据类型、数据位宽和张量长度分别生成所述张量指令的数据类型、数据位宽和张量长度。

14. 根据权利要求 11 至 13 中任一项所述的装置，其特征在于，所述指令类型包括以下一项或多项：仅地址参与指令、数值参与指令、阵列控制指令和近存电路控制指令。

15. 根据权利要求 11 至 14 中任一项所述的装置，其特征在于，所述编译单元还用于：

15 根据指令关联矩阵和阵列状态列表确定所述指令存储列表中的张量指令的指令状态，所述指令关联矩阵用于存储所述张量指令之间的依赖关系，所述阵列状态列表用于指示阵列模组的状态，所述阵列模组用于执行所述张量指令，所述指令状态包括可执行状态和等待状态。

16. 根据权利要求 15 所述的装置，其特征在于，所述处理单元还用于：

当所述张量指令的指令状态处于所述可执行状态时，向访存控制单元发送所述张量指令；

20 根据所述张量指令确定执行操作，所述执行操作包括以下一项或多项：近存计算操作、存内计算操作和常规读写操作。

17. 根据权利要求 11 至 16 中任一项所述的装置，其特征在于，所述张量指令集第一张量指令和第二张量指令，所述第一张量指令和所述第二张量指令在相同的阵列模组中执行，所述处理单元还用于：

25 在所述阵列模组的不同执行单元中分别同时执行所述第一张量指令和所述第二张量指令，所述执行单元包括阵列和近存电路。

18. 根据权利要求 15 至 17 中任一项所述的装置，其特征在于，所述张量指令之间的依赖关系包括以下一项或多项：读后写依赖、写后写依赖和写后读依赖。

19. 根据权利要求 15 至 18 中任一项所述的装置，其特征在于，所述阵列状态列表包括以下一项或多项：阵列模组空闲状态和近存电路状态。

30 20. 根据权利要求 15 至 19 中任一项所述的装置，其特征在于，所述指令存储列表包括以下一项或多项：所述张量指令的指令段、张量指令编号、编译算子编号和阵列块标识。

21. 一种计算设备，其特征在于，包括处理器，所述处理器与存储器耦合，所述处理器用于存储指令，当所述指令被所述处理器执行时，以使得所述电子设备执行权利要求 1 至 10 中任一项所述的方法。

35 22. 一种计算设备集群，其特征在于，包括至少一个计算设备，所述计算设备包括处理器，所述处理器与存储器耦合，所述处理器用于存储指令，当所述指令被所述处理器执行时，以使得所述计算设备集群执行权利要求 1 至 10 中任一项所述的方法。

23. 一种计算机可读存储介质，其上存储有指令，其特征在于，所述指令被执行时，以使得计算机执行权利要求 1 至 10 中任一项所述的方法。

40 24. 一种计算机程序产品，所述计算机程序产品中包括指令，其特征在于，所述指令被执行时，以使得计算机实现权利要求 1 至 10 中任一项所述的方法。