

(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



(10) 国际公布号

WO 2022/165718 A1

(43) 国际公布日

2022 年 8 月 11 日 (11.08.2022)

(51) 国际专利分类号:

G06F 9/38 (2006.01)

中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(21) 国际申请号:

PCT/CN2021/075307

(74) 代理人: 北京中博世达专利商标代理有限公司

(22) 国际申请日:

2021 年 2 月 4 日 (04.02.2021)

(BEIJING ZBSD PATENT & TRADEMARK AGENT LTD.); 中国北京市海淀区交大东路 31 号 11

(25) 申请语言:

中文

号楼 8 层, Beijing 100044 (CN)。

(26) 公布语言:

中文

(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)。

(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, 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,

(72) 发明人: 熊旭红 (XIONG, Xuhong); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 石洁珂 (SHI, Jieke);

(54) Title: INTERFACE CONTROLLER, DATA TRANSMISSION METHOD, AND SYSTEM ON CHIP

(54) 发明名称: 一种接口控制器、数据传输方法及片上系统

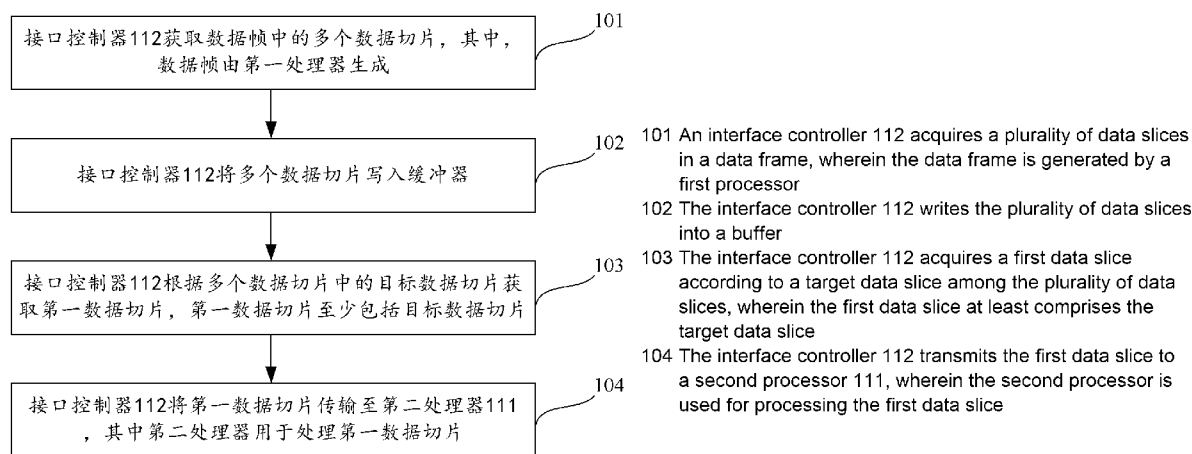


图 4b

(57) Abstract: The embodiments of the present application relate to the technical field of interfaces. Provided are an interface controller, a data transmission method, and a system on chip, which can improve the overall performance of a system. The data transmission method comprises: acquiring a plurality of data slices in a data frame, wherein the data frame is generated by a first processor; writing the plurality of data slices into a buffer; acquiring a first data slice according to a target data slice among the plurality of data slices, wherein the first data slice at least comprises the target data slice; and transmitting the first data slice to a second processor, wherein the second processor is used for processing the first data slice.

(57) 摘要: 本申请的实施例提供一种接口控制器、数据传输方法及片上系统, 涉及接口技术领域, 能够提升系统整体性能。该数据传输方法包括: 获取数据帧中的多个数据切片, 其中, 数据帧由第一处理器生成; 将多个数据切片写入缓冲器; 根据多个数据切片中的目标数据切片获取第一数据切片, 第一数据切片至少包括目标数据切片; 将第一数据切片传输至第二处理器, 其中第二处理器用于处理第一数据切片。

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))。

一种接口控制器、数据传输方法及片上系统

技术领域

5 本申请涉及接口技术领域，尤其涉及一种接口控制器、数据传输方法及片上系统。

背景技术

在片上系统（system on chip, SoC，或者称作系统级芯片）中，对数据帧实现某种算法的处理的一般过程是：将数据帧分别通过不同的处理器（通常，处理器可以包含一个或多个知识产权保护核（intellectual property core, IP 核），其中一个 IP
10 核即一个逻辑单元）的联合处理形成算法链路流水线（pipeline）。其中，算法链路用到的处理器可以是纯硬件的处理器或者可软件编程的处理器，例如：中央处理器（central processing unit, CPU）、数字信号处理器（digital signal processor, DSP）、图像信号处理器（image signal processor, ISP）、视频编码器（video codec）、神经网络处理器（neural processing unit, NPU）、图形处理器（graphics processing
15 unit, GPU）、或显示子系统（display sub system, DSS）等。

一般，处理器之间的数据交互通常都是通过内存共享实现。例如，参照图 1 所示，前一级处理器对当前数据帧处理完后，放入共享内存（缓存器或 DDR 中至少一个），并通知后一级处理器继续处理，后一级处理器从共享内存读取该数据帧，并进行后续处理。可见，在上述过程中算法链路在前一级处理器和后一级处理器之间是串行执行的，因此对于周期性实时处理，且延时敏感的算法链路（例如 30FPS（30 帧每秒）的游戏超分），要求一帧数据的处理要在单位周期内完成，如图 2 所示。前一级处理器
20 处理并输出一帧数据会占据单位周期中一帧的处理时间 T1，此时后一级处理器都处在空闲状态。当一帧数据处理完时，单位周期剩下的时间段 T2 让后一级处理器处理。由于剩下的时间段 T2 可能会很短，因此这种方法对后一级处理器的性能提出了很高的要求。并且，通常当数据帧太大时，在使用共享内存时，无法放到片内的缓存器（SoC
25 cache），因此需要放在外部存储器（例如，双倍速率动态随机存储器（double data rate dynamic random access memory, DDR）上，从而会导致存储器的带宽和功耗规格限制引起的整体的性能。总之，目前的 SoC 架构下处理器协同执行算法时，数据传输的方式会影响整体的性能提升。

发明内容

30 本申请提供一种接口控制器、数据传输方法及片上系统，能够提升系统整体性能。

第一方面，提供一种数据传输方法。该数据传输方法应用于接口控制器，接口控制器连接第一处理器以及第二处理器。该数据传输方法包括：获取数据帧中的多个数据切片，其中，数据帧由第一处理器生成；将多个数据切片写入缓冲器（buffer）；
35 根据多个数据切片中的目标数据切片获取第一数据切片，第一数据切片至少包括目标数据切片；将第一数据切片传输至第二处理器，其中第二处理器用于处理第一数据切片。在该方案中，接口控制器可以将第一处理器处理的数据帧中的多个数据切片在缓冲器中进行存储，并能够根据多个数据切片中的目标数据切片获取第一数据切片，其中第一数据切片至少包括该目标数据切片，然后将第一数据切片发送至第二处理器，

并由第二处理器处理对应的第一数据切片，上述过程中由于将数据帧分为了多个数据切片，因此相对于数据帧减小了每次传输数据的大小，并通过接口控制器的缓冲器向第二处理器传输，因此可以避免采用共享内存的方式在第一处理器和第二处理器之间传输数据，并且由于接口控制器每获取一个第一数据切片即可向第二处理器传输并通知第二处理器进行处理，这样在第二处理器处理第一数据切片时，接口控制器进一步用于将其他数据切片传输至所述第二处理器，这样可以有效的形成计算链路流水，并且由于将数据帧划分成了数据切片，则在每个数据切片的数据在第一处理器处理完成后就可以通过接口控制器传输给第二处理器，由第二处理器开始处理，从而为第二处理器争取到了更多的处理时间，相对于现有技术中算法数据帧在各个处理器中串行执行，该方案中，只要第一处理器处理完一个数据切片的数据，则可以通过接口控制器传输给第二处理器进行处理，因此对第二处理器的性能要求较低，并且能够减少整个计算链路的延时；从而提升系统整体性能。

在一种可能的实现方式中，将数据帧划分为多个数据切片可以由第一处理器或者接口控制器执行。例如，第一处理器为类似于 GPU 的处理器时，由于 GPU 通常是按照数据切片对图像进行渲染，因此，第一处理器将数据帧划分为多个数据切片，获取数据帧中的多个数据切片；具体包括：接收第一处理器传输的多个数据切片。当然对于对数据帧不进行数据切片划分的第一处理器，将数据帧划分为多个数据切片可以由接口控制器执行，例如：在获取数据帧中的多个数据切片之前还包括：接收第一处理器传输的数据帧；获取数据帧中的多个数据切片，包括：将数据帧划分为多个数据切片。

在一种可能的实现方式中，根据多个数据切片中的目标数据切片获取第一数据切片，包括：在缓冲器读取目标数据切片，以及目标数据切片的边缘相邻的数据切片；根据目标数据切片，以及目标数据切片的边缘相邻的数据切片生成第一数据切片，其中第一数据切片覆盖目标数据切片，并且第一数据切片与目标数据切片的边缘相邻的数据切片具有重叠区域。示例性的，第二处理器可以为神经网络处理器 NPU；在 GPU 和 NPU 之间传输的数据为图像，此外，由于 NPU 中的神经网络在对数据切片进行计算时，在 NPU 输入侧的数据切片的宽度、高度会被依据神经网络的层数、步长 (stride) 收缩后输出，因此通常，NPU 输出侧的数据切片的宽度 (高度) 会小于输入侧的数据切片的宽度 (高度)，这样为了确保 NPU 输出侧的数据切片能够完全覆盖整个数据帧，则需要输入至 NPU 的数据切片的具有更大的宽度和高度。因此为了确保对第二处理器对第一数据切片处理后的数据切片与目标数据切片的尺寸相同，因此第一数据切片需要具有比目标数据切片更大的尺寸，即第一数据切片除了覆盖目标数据切片，还需要与目标数据切片的边缘相邻的数据切片具有重叠区域，其中重叠区域的宽度由 NPU 的卷积网络的层数及步长确定。

在一种可能的实现方式中，根据多个数据切片中的目标数据切片获取第一数据切片之前，还包括：确定目标数据切片的边缘相邻的数据切片存储于缓冲器中。其中当需要根据目标数据切片以及目标数据切片的边缘相邻的数据切片计算第一数据切片时，需要保证目标数据切片的边缘相邻的数据切片存储于缓冲器中。

在一种可能的实现方式中，还包括：生成比特地图表，当确定获取所述数据切片后，将所述比特地图表中所述数据切片对应位置的比特置为 1；确定目标数据切片的

边缘相邻的数据切片存储于缓冲器中，包括：确定所述比特地图表中所述目标数据切片对应位置周围的比特均为 1。其中，通过该比特地图表 bitmap 可以记录各个 bit 相应的数据切片已经存储于缓冲器中。

在一种可能的实现方式中，目标数据切片包括 $2n$ 个数据切片，其中， n 为正整数。

5 如果，目标数据切片(slice)的粒度是 1×1 ，即只包含一个数据切片。则对每个目标数据切片计算一次，需要至少获取目标数据切片的边缘相邻的八个其他数据切片（其中位于数据帧边缘的目标数据切片除外）；则第一数据切片与这些其他数据切片重叠区域的数据会多次被重复传输。为降低对重叠区域的数据重复传输，本申请的实施例提供的目标数据切片可以包括 $2n$ 个数据切片。例如，对 2×2 粒度的目标数据切片计算一次，最多需要目标数据切片的边缘相邻的 12 个数据切片，这样 2×2 粒度的目标数据切片中针对每个数据切片相当于平均满足提供其边缘相邻的三个数据切片即可进行第一数据切片的计算，降低了对重叠区域的数据重复传输。

15 在一种可能的实现方式中，接口控制器为了将数据帧划分为多个数据切片，或者为了确定数据切片在数据帧中的位置需要获取数据帧的信息。因此还包括：获取数据帧的信息，数据帧的信息包括基地址、宽度、高度及通道数；根据所述数据帧的信息将数据帧划分为多个数据切片。

20 在一种可能的实现方式中，接口控制器还与片上系统的缓存器连接；当确定缓冲器存满时，将缓冲器中的数据切片存入缓存器。这样，当确定缓冲器存满时，接口控制器可以将缓冲器中的数据切片存入系统的缓存器，此外，接口控制器还可以将缓冲器中的数据切片存入 DDR；当缓冲器有剩余空间时，再从片上系统的缓存器或 DDR 读出。这样如果缓冲器有足够的空间，数据的传输只在缓冲器中进行，如果缓冲器的空间不足，为避免造成数据传输中断，可以将数据切片片上系统的缓存器或 DDR。

25 第二方面，提供一种接口控制器，所述接口控制器连接第一处理器以及第二处理器，所述接口控制器包括：获取单元，用于获取数据帧中的多个数据切片，其中，所述数据帧由第一处理器生成；处理单元，用于将所述获取单元获取的所述多个数据切片写入缓冲器；所述处理单元，用于根据所述缓冲器中的所述多个数据切片中的目标数据切片获取第一数据切片，所述第一数据切片至少包括所述目标数据切片；发送单元，用于将所述处理单元获取的所述第一数据切片传输至所述第二处理器，其中所述第二处理器用于处理所述第一数据切片。

30 在一种可能的实现方式中，所述获取单元具体用于接收所述第一处理器传输的所述多个数据切片，其中所述第一处理器用于将所述数据帧划分为所述多个数据切片。

在一种可能的实现方式中，所述获取单元还用于接收所述第一处理器传输的所述数据帧；所述获取单元具体用于将所述数据帧划分为所述多个数据切片。

35 在一种可能的实现方式中，所述处理单元，具体用于在所述缓冲器读取目标数据切片，以及所述目标数据切片的边缘相邻的数据切片；根据所述目标数据切片，以及所述目标数据切片的边缘相邻的数据切片生成所述第一数据切片，其中所述第一数据切片覆盖所述目标数据切片，并且所述第一数据切片与所述目标数据切片的边缘相邻的数据切片具有重叠区域。

在一种可能的实现方式中，第二处理器为神经网络处理器 NPU。

在一种可能的实现方式中，所述处理单元还用于确定目标数据切片的边缘相邻的数据切片存储于缓冲器中。

在一种可能的实现方式中，所述处理单元还用于生成比特地图表，当确定获取所述数据切片后，将所述比特地图表中所述数据切片对应位置的比特置为 1；所述处理
5 器具体用于确定所述比特地图表中所述目标数据切片对应位置周围的比特均为 1。

在一种可能的实现方式中，所述目标数据切片包括 $2n$ 个数据切片其中 n 为正整数。

在一种可能的实现方式中，还包括：所述获取单元，还用于获取数据帧的信息，数据帧的信息包括基地址、宽度、高度及通道数；处理单元具体用于根据数据帧的信息将数据帧划分为多个数据切片。
10

在一种可能的实现方式中，其特征在于，接口控制器还与片上系统的缓存器连接；处理单元具体用于当确定缓冲器存满时，通过发送单元将缓冲器中的数据切片存入缓存器。

第三方面，提供一种接口控制器，包括缓冲器和一个或多个处理器，处理器用于
15 调用计算机指令，以执行如上述的方法。

第四方面，提供一种计算机可读存储介质，该计算机可读存储介质上存储有计算机程序，当该计算机程序在计算机上运行时，使得该计算机执行如上述的方法。

第五方面，提供一种片上系统，包括第一处理器、第二处理器以及如上述的接口
20 控制器，其中，所述接口控制器与所述第一处理器及所述第二处理器连接，接口控制器用于执行上述的数据传输方法。

在一种可能的实现方式中，在第二处理器处理第一数据切片时，接口控制器进一步用于将其他数据切片传输至所述第二处理器。可选地，第一处理器进一步在第二处理器处理第一数据切片时用于处理其他数据，以形成流水。

其中，第二方面至第五方面中任一种可能实现方式中所带来的技术效果可参见上
25 述第一方面不同的实现方式所带来的技术效果，此处不再赘述。

附图说明

图 1 为本申请的实施例提供的一种片上系统的结构示意图；

图 2 为本申请的实施例提供的一种处理器之间的算法链路流水线示意图；

图 3 为本申请的另一实施例提供的一种片上系统的结构示意图；

30 图 4a 为本申请的又一实施例提供的一种片上系统的结构示意图；

图 4b 为本申请的实施例提供的一种数据传输方法的流程示意图；

图 5 为本申请的实施例提供的一种数据帧的数据切片阵列的结构示意图；

图 6 为本申请的实施例提供的一种数据切片的结构示意图；

图 7 为本申请的实施例提供的一种数据帧的数据切片的处理顺序示意图；

35 图 8 为本申请的实施例提供的一种数据帧的数据切片的划分顺序示意图；

图 9 为本申请的又一实施例提供的一种片上系统的结构示意图；

图 10 为本申请的实施例提供的 NPU 输出侧的数据切片和 NPU 输入侧的数据切片的尺寸示意图；

图 11 为本申请的另一实施例提供的一种数据帧的数据切片阵列的结构示意图；

图 12 为本申请的实施例提供的一种 bitmap 的结构示意图；

图 13 为本申请另一实施例的提供的一种 bitmap 的结构示意图；

图 14 为本申请的又一实施例提供的一种 bitmap 的结构示意图；

图 15 为本申请的再一实施例提供的一种 bitmap 的结构示意图；

5 图 16 为本申请的另一实施例提供的一种 bitmap 的结构示意图；

图 17 为本申请的又一实施例提供的一种 bitmap 的结构示意图；

图 18a 为本申请的再一实施例提供的一种 bitmap 的结构示意图；

图 18b 为本申请的另一实施例提供的一种 bitmap 的结构示意图；

图 19 为本申请的另一实施例提供的一种处理器之间的算法链路流水线示意图；

10 图 20 为本申请的另一实施例提供的一种数据传输方法的流程示意图；

图 21 为本申请的实施例提供的一种接口控制器的结构示意图。

具体实施方式

在本实施例中，术语“第一”、“第二”等仅用于描述方便，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、
15 “第二”等的特征可以明示或者隐含地包括一个或者更多个该特征。在本申请的描述中，除非另有说明，“多个”的含义是两个或两个以上。例如，多个处理单元是指两个或两个以上的处理单元。

此外，本申请实施例中，“上”、“下”、“左”以及“右”不限于相对附图中的部件示意置放的方位来定义的，应当理解到，这些方向性术语可以是相对的概念，
20 它们用于相对于的描述和澄清，其可以根据附图中部件附图所放置的方位的变化而相应地发生变化。在附图中，为了清楚起见，夸大了层和区域的厚度，图示中的各部分之间的尺寸比例关系并不反映实际的尺寸比例关系。

本申请实施例中，除非另有明确的规定和限定，术语“连接”应做广义理解，例如，“连接”可以是固定连接，也可以是可拆卸连接，或成一体；可以是直接相连，
25 也可以通过中间媒介间接相连。

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“例如”等词旨在以具体方式呈现相关概念。

30 本申请的实施例提到的人工神经网络 (artificial neural network, ANN) 简称为神经网络 (neural network, NN) 或类神经网络，在机器学习和认知科学领域，是一种模仿生物神经网络（如动物的中枢神经系统，特别是大脑）的结构和功能的数学模型或计算模型，用于对函数进行估计或近似。人工神经网络可以包括卷积神经网络 (convolutional neural network, CNN)、深度神经网络 (deep neural network, DNN)、
35 时延神经网络 (time delay neural network, TDNN) 和多层感知器 (multilayer perceptron, MLP) 等神经网络。

本申请提供的技术方案可以应用于包括各类芯片，例如片上系统，的电子设备，例如手机、移动终端、个人计算机 (personal computer, PC)、服务器、笔记本电脑、平板电脑、车载电脑、智能摄像头、智能手表、嵌入式设备等。本申请实施例对该电

子设备的具体形式不做特殊限制。

下面结合附图对本申请实施例提供的一种接口控制器、数据传输方法及片上系统进行详细描述。图3为本申请实施例提供的一种片上系统的结构示意图。需理解，后续仅以片上系统为例介绍，实际方案也可应用于其他类型的芯片或设备。如图3所示，片上系统100可以包括处理器110、处理器111、接口控制器112、存储器120、通信线路130、缓存器140、以及至少一个通信接口150及外围电路（图中未示出）。需要说明的是，除了图3所示的器件外，片上系统100还可以包括通信接口150。可以理解的是，本发明实施例示意的结构并不构成对片上系统100的具体限定。在本申请另一些实施例中，片上系统100可以包括比图示更多或更少的部件，或者可以组合某些部件，或者可以拆分某些部件，或者具有不同的部件布置。图示的部件可以以硬件或软件和硬件的组合实现。

处理器(processor)110是片上系统100的运算核心和控制核心(control unit)。其中，处理器110可以包括中央处理器(central processing unit, CPU)、数字信号处理器(digital signal processor, DSP)、图像信号处理器(image signal processor, ISP)、视频编码器(video codec)、神经网络处理器(neural processing unit, NPU)、图形处理器(graphics processing unit, GPU)、DSS，特定应用集成电路(application-specific integrated circuit, ASIC)。在本申请的实施例中，处理器的形式可以是纯硬件的或者能够加载软件程序的硬件。在具体实现中，作为一种实施例，处理器110可以包括一个或多个处理器核(core)，例如图3中的核0和核1。在具体实现中，作为一种实施例，片上系统100可以包括多个处理器，例如图1中的处理器110、处理器111和接口控制器112。这些处理器中的每一个可以是一个单核(single-core)处理器（即处理器包括一个core），也可以是一个多核(multi-core)处理器（即处理器包括多个core）。

存储器120可以是独立于处理器110存在，通过通信线路130与处理器110相连接。在一种示例中，存储器120可以用于存储指令和数据，包括用于执行本申请方案的指令，该存储器120例如是静态随机存取存储器(SRAM)。该指令由处理器110来控制执行。在本申请的实施例中，可以由接口控制器112执行存储器120中存储的指令，从而实现本申请下述实施例提供的数据传输方法。可选的，本申请实施例中的指令也可以称之为应用程序代码，本申请实施例对此不作具体限定。

通信线路130可包括一通路，在上述组件之间传送信息。通信接口150，使用任何收发器一类的装置，用于与其他设备或通信网络通信，如以太网，无线接入网(radio access network, RAN)，无线局域网(wireless local area networks, WLAN)等。

缓存器(cache)140是位于处理器110和存储器120之间的高速存储器；其容量比存储器120小，但是与处理器110的交换速度快。缓存器140可以用于保存处理器110刚用过或循环使用的指令或数据。如果处理器110需要再次使用该指令或数据，可从缓存器140中直接调用。例如，缓存器140保存的指令或数据是存储器120中的一小部分，但这一小部分指令或数据是短时间内处理器110即将访问的；当处理器110调用指令或数据时，可避开存储器120直接从缓存140中调用，从而加快读取速度。

如上所述，接口控制器112可以是一个单核或多核的处理器，此外，在本申请的

实施例中，接口控制器 112 中还设置有缓冲器。缓冲器也称作缓冲寄存器，当数据经过两个装置之间的缓冲器传输时，缓冲器主要用来弥补缓冲器两端的装置对数据处理速度的差距造成的数据累积。

基于上述的片上系统，本申请的实施例提供一种数据传输方法，应用于接口控制器 112。参照图 4a 所示，接口控制器 112 连接第一处理器 110 以及第二处理器 111，其中结合上述针对图 3 示出的片上系统的描述，接口控制器 112 与第一处理器 110 以及第二处理器 111 的连接方式可以通过通信线路连接。参照图 4b 所示该数据传输方法包括以下步骤：

101、接口控制器 112 获取数据帧中的多个数据切片，其中，数据帧由第一处理器生成。在步骤 101 中，数据帧的形式可以是图像、文本、音频或视频等。其中，由数据帧中的多个数据切片 (slice 或 tile) 可以是数据切片阵列的形式。以第一处理器为 GPU 为例，数据帧可以为一帧图像，图像的数据帧可以为包括各个通道上的像素数据的三维数据阵列 (3D data cube)。如图 5 所示，在 xyz 直角坐标系中，数据帧中 x 方向为图像帧的宽度 W (width，其中，该宽度可以为在 x 方向上一行像素的个数)，数据帧中 y 方向为图像帧的高度 H (height，其中，该高度可以为在 y 方向上一列像素的个数)，以 4K 图像为例，其分辨率为 4096×2160 ，则图像的数据帧的宽度 W 为 4096，高度 H 为 2160。数据帧中 z 方向为图像帧的通道 C (channel，例如，以 RGB (red、green、blue，红绿蓝) 为三原色的图像的数据帧包含三个通道：R 通道、G 通道、B 通道)。图像的数据帧可以在 xy 平面上被划分为多个数据切片，本申请的实施例还提供了—个数据切片的结构，如图 6 所示，对于图像的数据帧划分的数据切片，一个数据切片也包含了三个通道的数据，其与整帧数据的区别是一个数据切片的高度和宽度仅为整帧数据的分数倍。可以理解的是，每个数据切片的一个通道上包含若干行和若干列阵列分布的像素数据。此外，对于其他数据也有二维数据阵列的形式。当然也可以将彩色图像灰度化 (或二值化) 后，变换成单通道形式的灰阶图像，这样的灰阶图像的数据帧也可以是二维数据阵列的形式。

其中，在本申请的实施例中可以由第一处理器 110 将数据帧划分为多个数据切片或者由本申请的实施例提供的接口控制器 112 将数据帧划分为多个数据切片。例如，如果第一处理器 110 将数据帧划分为多个数据切片，则在步骤 101 中，接口控制器 112 可以直接接收第一处理器 110 传输的多个数据切片。其中，第一处理器 110 可以将数据帧划分为多个数据切片，并对每个数据切片进行处理，之后按照一定的顺序将数据切片依次发送给接口控制器 112，例如，参照图 7 所示，可以以 Z 字形扫描方式在将数据切片阵列中的数据切片处理后，依次发送给接口控制器 112；或者第一处理器 110 也可以乱序将处理后的数据切片发送给接口控制器 112。在另一种方式中，接口控制器 112 可以直接接收第一处理器 110 传输的数据帧；然后将数据帧划分为多个数据切片。需要说明的是，在 SoC 系统中，可以由 CPU 为第一处理器 110 以及接口控制器 112 配置寄存器的方式，通知是由第一处理器 110 或由接口控制器 112 将数据帧划分为多个数据切片。在一些示例中，接口控制器 112 为了将数据帧划分为多个数据切片，或者为了确定数据切片在数据帧中的位置需要获取数据帧的信息。该数据帧的信息包括数据帧的基地址、宽度、高度和通道数。其中，数据帧的信息可以由 CPU 为接口控

制器 112 配置寄存器的方式通知接口控制器 112。具体的，第一处理器 110 可以通过突发 (burst) 事件向接口控制器 112 发送数据帧的数据，其中一次突发 (burst) 事件可以发送一个数据帧、或者一个数据切片或者其他任意数据长度的数据。这样，若由接口控制器 112 将数据帧划分为多个数据切片，当一次突发 (burst) 事件发送任意长度的数据 (例如可以是一个数据帧或者其他任意长度的数据) 时，该突发 (burst) 事件可以同时携带数据的写地址、数据长度；则只要根据写地址和数据长度确定发送的数据长度超过一个数据切片，接口控制器 112 则根据数据帧的信息将接收到的数据划分为多个数据切片。例如，参照图 8 所示，对于按照 Z 字形扫描逐行输出数据帧的数据第一处理器 110，接口控制器 112 可以按照数据帧的基地址、宽度、高度和通道数依次将连续若干行数据划分为一个数据切片。若由第一处理器 110 将数据帧划分为多个数据切片，当一次突发 (burst) 事件发送一个数据切片时，该突发 (burst) 事件可以同时携带数据切片的写地址、数据长度；则接口控制器 112 可以根据数据切片的写地址、数据长度以及上述的数据帧的信息确定数据切片在数据帧中的位置。

102、接口控制器 112 将多个数据切片写入缓冲器。其中，当确定缓冲器存满时，接口控制器 112 可以将缓冲器中的数据切片存入系统的缓存器，此外，接口控制器 112 还可以将缓冲器中的数据切片存入 DDR；当缓冲器有剩余空间时，再从片上系统的缓存器或 DDR 读出。这样如果缓冲器有足够的空间，数据的传输只在缓冲器中进行，如果缓冲器的空间不足，为避免造成数据传输中断，可以将数据切片存入片上系统的缓存器或 DDR。

103、接口控制器 112 根据多个数据切片中的目标数据切片获取第一数据切片，第一数据切片至少包括目标数据切片。例如，第一数据切片包括目标数据切片，可以选择性进一步包括与该目标数据切片的边缘相邻的数据切片的至少一部分。在步骤 103 中，接口控制器 112 可以直接在缓冲器中读取任意目标数据切片并在后续过程中发送至第二处理器 111。或者按照一定的顺序 (例如 Z 字形) 读取目标数据切片在后续过程中发送至第二处理器 111。前提是，经步骤 101 和 102 的处理已经在缓冲器中写入该目标数据切片。

在本申请的实施例，接口控制器 112 在初次接收数据切片或者数据帧之前，可以预先生成一张比特地图表 (bitmap)；其中，该 bitmap 可以是默认预配置的，在初始状态 (例如在接口控制器 112 开始接收数据帧或数据切片之前) bitmap 的所有比特置为 0。按照上述的方式接收到第一处理器 110 发送的数据切片或者在接口控制器 112 本地将数据帧划分为数据切片后，将 bitmap 中数据切片对应位置的比特置为 1。当接口控制器将数据帧划分为数据切片时，可以直接根据数据切片在数据帧中的位置将各个数据切片分别影射至 bitmap 的一个 bit。而当数据切片是由第一处理器 110 发送时，则接口控制器 112 在获取数据切片的同时还可以接收第一处理器 110 发送的数据切片在数据帧中的位置信息 (例如可以是数据切片的坐标，或者数据切片的写地址和数据长度)，这样可以根据数据切片的位置信息将数据切片与 bitmap 中的 bit 一一对应。这样，bitmap 中被置为 1 的 bit 对应的数据切片是完全传输至接口控制器 112 的数据切片。如图 12 所示，bitmap 表中的“1”表示相应的数据切片已经准备好，“0”表示数据切片还没有完全从第一处理器 110 输出到接口控制器 112。

另外，参照图 9 所示，当第一处理器为 GPU，且第二处理器为 NPU 时，在 GPU 和 NPU 之间传输的数据为图像，此外，由于 NPU 中的神经网络在对数据切片进行计算时，在 NPU 输入侧的数据切片的宽度、高度会被依据神经网络的层数、步长 (stride) 收缩后输出，因此通常，NPU 输出侧的数据切片的宽度 (高度) 会小于输入侧的数据切片的宽度 (高度)，如图 10 所示，NPU 输入侧的数据切片的宽度 $W1 >$ 输出侧的数据切片的宽度 $W2$ ；输入侧的数据切片的高度 $H1 >$ 输出侧的数据切片的高度 $H2$ 。具体的，假定 W, H 方向 stride 都为 1，神经网络的层数一共为 N 层，每层是 3×3 卷积，那么 $W2 = W1 - N \times 1$ ， $H2 = H1 - N \times 1$ 。这样如果将图 5 示出的数据切片作为神经网络的输入，则经过 NPU 处理后输出的各个数据切片之间会存在间隔，影响输出的图像的效果。为了避免该问题，需要确保输出侧的数据切片能够完全覆盖整个数据帧，即需要保证 NPU 输出侧的数据切片的尺寸 ($W2$ 和 $H2$) 与图 5 中示出的各个数据切片的尺寸相同。则需要，输入神经网络的数据切片具有比图 5 示出的数据切片更大的宽度和高度。例如，步骤 103 具体可以是在确定目标数据切片的边缘相邻的数据切片存储于缓冲器中后 (结合上述的 bitmap，即确定比特地图表中目标数据切片对应位置周围的比特均为 1 后)，在缓冲器读取目标数据切片，以及目标数据切片的边缘相邻的数据切片；根据目标数据切片，以及目标数据切片的边缘相邻的数据切片生成第一数据切片，其中第一数据切片覆盖目标数据切片，并且第一数据切片与目标数据切片的边缘相邻的数据切片具有重叠区域。具体参见图 11，第一数据切片包括了目标数据切片以及与该目标数据切片的边缘相邻的多个数据切片的一部分，该部分即为所述重叠区域。其中重叠区域的宽度由 NPU 的卷积网络的层数及步长确定。结合图 11 所示，目标数据切片 (i, j) 对应的第一数据切片覆盖目标数据切片 (i, j)；并且第一数据切片与数据切片 ($i-1, j-1$)、($i-1, j$)、($i-1, j+1$)、($i, j-1$)、($i, j+1$)、($i+1, j-1$)、($i+1, j$) 和 ($i+1, j+1$) 具有重叠区域，即第一数据切片包括了数据切片 ($i-1, j-1$)、($i-1, j$)、($i-1, j+1$)、($i, j-1$)、($i, j+1$)、($i+1, j-1$)、($i+1, j$) 和 ($i+1, j+1$) 中的一部分，如图 11 所示。此外需要说明的是，对于位于数据帧边缘的数据切片，由于某些方向没有相邻的数据切片，例如数据切片 ($i-1, j-1$) 的左侧和上侧没有相邻的数据切片，则在计算数据切片 ($i-1, j-1$) 对应的第一数据切片时，可以通过填充 (padding) 的方式生成满足 NPU 输入侧的数据切片的尺寸的第一数据切片，例如，可以对数据切片 ($i-1, j-1$) 对应的第一数据切片在 ($i-1, j-1$) 左侧和上侧的部分填充 0 或者直接填充复制的数据切片 ($i-1, j-1$) 中的像素数据。

可替换地，图 11 中第一数据切片还可以包所述缘相邻的多个数据切片的全部。或者，第一数据切片还可以进一步包括的更多。例如，在图 11 中，该目标数据切片的边缘相邻的多个数据切片是在该边缘的每个方向上与该目标数据切片相邻的一个数据切片，但实际上，可替换地，在该边缘的每个方向上相邻的数据切片的个数可以是多个，这取决于高度 $H1, H2$ 之间的比例和宽度 $W1, W2$ 之间的比例。可以理解为，第一数据切片还可以在每个方向上包括多个与目标数据切片的边缘相邻的数据切片，即在目标数据切片的基础上向外扩展了多个数据切片以覆盖更大的范围，本实施例对第一数据切片在目标数据切片的基础上所扩展的程度不做限定。

由于第一处理器 110 输出数据切片时，数据切片可能是乱序的，而非一般的按 Z

字扫描的顺序输出。以下按乱序的情况进行说明（可以理解的是顺序输出可归类为乱序输出）。本申请的实施例可以采用贪婪算法对乱序输出的数据切片按序（例如 Z 字扫描顺序）处理可以处理的数据切片。例如，如果只有一个目标数据切片可以处理，即该目标数据切片以及目标数据切片的边缘相邻的数据切片均已存储至接口控制器 112 的缓冲器，则可以开始计算该目标数据切片对应的第一数据切片。如果同时有多个目标数据切片可以开始计算对应的第一数据切片，则按照 Z 字扫描顺序依次计算多个目标数据切片对应的第一数据切片。

具体的，参考图 12 提供的 bitmap，若一个目标数据切片 tile 的边缘相邻的数据切片 tile 都准备好（bit 为 1）就可以计算输入至 NPU 的第一数据切片，其中图 12 中虚线框为计算的第一数据切片。然后，就可以继续执行步骤 104，将第一数据切片传输至 NPU 由 NPU 对第一数据切片进行处理。而当同时有几个这类似的切片都满足上述条件时，可以采用贪婪算法按序（例如 Z 字扫描顺序）计算。

具体的，以 4K 图像的数据帧为例，数据切片的尺寸 256×256 ，则一个 bitmap 大小为 16×8 bits。参照图 13 所示，目标数据切片 (i, j) 及其边缘相邻的数据切片 $(i-1, j-1)$ 、 $(i-1, j)$ 、 $(i-1, j+1)$ 、 $(i, j-1)$ 、 $(i, j+1)$ 、 $(i+1, j-1)$ 、 $(i+1, j)$ 和 $(i+1, j+1)$ 均已存储至接口控制器的缓冲器。则触发第一数据切片计算的开始条件为： $Cur_stat[i, j] = flag[i-1, j-1] \& flag[i-1, j] \& flag[i-1, j+1] \& flag[i, j-1] \& flag[i, j+1] \& flag[i+1, j-1] \& flag[i+1, j] \& flag[i+1, j+1]$ ；其中， $flag[i-1, j-1]$ 表示坐标为 $(i-1, j-1)$ 的数据切片的 bit 位已经标记为 1， $Cur_stat[i, j]$ 表示坐标 (i, j) 、 $(i-1, j-1)$ 、 $(i-1, j)$ 、 $(i-1, j+1)$ 、 $(i, j-1)$ 、 $(i, j+1)$ 、 $(i+1, j-1)$ 、 $(i+1, j)$ 和 $(i+1, j+1)$ 的数据切片均已存储至接口控制器的缓冲器。

此外，参照图 14 所示，接口控制器 112 还可以维护一张已经计算过第一数据切片的目标数据切片的 bitmap，其中，该 bitmap 中的 bit 位为 1 表示针对对应的目标数据切片已经计算过第一数据切片。bit 位为 0 表示针对对应的目标数据切片还未计算过第一数据切片。在执行图 13 计算目标数据切片 (i, j) 对应的第一数据切片之前，已完成数据切片 $(i-2, j-2)$ 、 $(i-2, j-1)$ 、 $(i-2, j)$ 以及 $(i-1, j)$ 对应的第一数据切片的计算。本次执行完对目标数据切片 (i, j) 对应的第一数据切片的计算之后可以将图 14 的 bitmap 更新为图 15 所示的 bitmap。这样通过维护更新图 14、图 15 提供的 bitmap，可以判定出需要继续对当前图 13 中哪一个目标数据切片计算对应的第一数据切片。

以上计算的目标数据切片(slice)的粒度是 1×1 。这样对每个目标数据切片计算一次，需要至少获取与目标数据切片的边缘相邻的八个其他数据切片（其中位于数据帧的边缘的目标数据切片除外）；则第一数据切片与这些其他数据切片重叠区域的数据会多次被重复传输。为降低对重叠区域的数据重复传输，本申请的实施例提供的目标数据切片可以包括 $2n$ 个数据切片。例如，目标数据切片(slice)的粒度是 1×2 ， 2×2 ，... 等不同大小尺寸。

以下以目标数据切片(slice)的粒度是 2×2 为另一个实施例，参照图 16 所示，当包括数据切片 $(1, 1)$ 、 $(1, 2)$ 、 $(2, 1)$ 以及 $(2, 2)$ 的目标数据切片以及目标数据

切片周围的数据切片 (1,3)、(2,3)、(3,1)、(3,2) 以及 (3,3) 均已存储至接口控制器的缓冲器, 则对目标数据切片对应的第一数据切片进行计算。这样相较于图 13 提供的 1x1 粒度的目标数据切片需要同时完成目标数据切片的边缘相邻的八个相邻的其他数据切片才可以计算对应的第一数据切片, 2x2 粒度的目标数据切片最多需要

5 目标数据切片的边缘相邻的 12 个数据切片 (由于示例 2x2 粒度的目标数据切片是位于左上角, 因此图 13 中示出了目标数据切片的边缘相邻的 5 个数据切片), 这样 2x2 粒度的目标数据切片中针对数据切片相当于平均满足提供其边缘相邻的三个数据切片即可进行第一数据切片的计算, 降低了对重叠区域的数据重复传输。

在另一个示例中, 对于图 17 提供的 bitmap, 目标数据切片为 2x2 粒度, 具体包

10 括数据切片 (i, j)、(i+1, j)、(i, j+1) 以及 (i+1, j+1), 目标数据切片及其边缘相邻的其他数据切片 (i-1, j-1)、(i-1, j)、(i-1, j+1)、(i-1, j+2)、(i, j+1)、(i, j+2)、(i+1, j-1)、(i+1, j+2)、(i+2, j-1)、(i+2, j)、(i+2, j+1) 以及 (i+2, j+2) 均已存储至接口控制器的缓冲器。则触发第一数据切片计算的开始条件为: $Cur_stat[(i, j) \& (i+1, j) \& (i, j+1) \& (i+1, j+1)] = flag[i-1, j-1] \& flag[i-1, j] \& flag[i-1, j+1] \& flag[i-1, j+2] \& flag[i, j+1] \& flag[i, j+2] \& flag[i+1, j-1] \& flag[i+1, j+2] \& flag[i+2, j-1] \& flag[i+2, j] \& flag[i+2, j+1] \& flag[i+2, j+2]$; 其中, $flag[i-1, j-1]$ 表示坐标为 (i-1, j-1) 的数据切片的 bit 位已经标记为 1。

15

此外, 参照图 18a 所示, 接口控制器 112 还可以维护一张表示可以计算第一数据切片的目标数据切片的 bitmap, 其中, 该 bitmap 的 (x, y) 中, x 为 1 表示 2x2 粒度的目标数据切片及其边缘相邻的其他数据切片已经存储至接口控制器 112 的缓冲器, x 为 0 表示 2x2 粒度的目标数据切片及其边缘相邻的其他数据切片还未全部存储至接口控制器 112 的缓冲器; y 为 0 表示针对对应的目标数据切片还未计算过第一数据切片, y 为 1 表示针对对应的目标数据切片已经计算过第一数据切片。如上述图 17 提供的

20 的 bitmap, 在未对目标数据切片 (i, j)、(i+1, j)、(i, j+1) 以及 (i+1, j+1) 计算第一数据切片时, 目标数据切片在图 18a 中对应表示为 (1, 0), 本次执行完对目标数据切片对应的第一数据切片的计算之后可以将目标数据切片在图 18b 中对应更新为 (1, 1)。这样通过维护更新图 18a、图 18b 提供的 bitmap, 可以判定出对当前图 17 中哪一个目标数据切片计算对应的第一数据切片。

25

进一步地, 104、接口控制器 112 将第一数据切片传输至第二处理器 111, 其中第二处理器用于处理第一数据切片。具体的, 接口控制器 112 还可以向第二处理器 111 发送对应第一数据切片的第一处理指示, 其中, 第二处理器 111 根据第一处理指示处理第一数据切片。

30

在该方案中, 接口控制器 112 可以将第一处理器 110 处理的数据帧中的多个数据切片在缓冲器中进行存储, 并能够根据多个数据切片中的目标数据切片获取第一数据切片, 其中第一数据切片至少包括该目标数据切片, 然后将第一数据切片发送至第二处理器 111, 并由第二处理器 111 处理对应的第一数据切片, 上述过程中由于将数据

35

帧分为了多个数据切片，因此相对于数据帧减小了每次传输数据的大小，并通过接口控制器 112 的缓冲器向第二处理器 111 传输，因此可以避免采用共享内存的方式在第一处理器 110 和第二处理器 111 之间传输数据，并且由于接口控制器 112 每获取一个第一数据切片即可向第二处理器 111 传输并通知第二处理器 111 进行处理，这样在第二处理器处理第一数据切片时，接口控制器 112 进一步用于将其他数据切片传输至所述第二处理器 111，这样可以有效的形成计算链路流水，并且由于将数据帧划分成了数据切片，则在每个数据切片的数据在第一处理器 110 处理完成后就可以通过接口控制器 112 传输给第二处理器 111，由第二处理器 111 开始处理，从而为第二处理器 111 争取到了更多的处理时间，相对于现有技术中算法在各个处理器中串行执行，该方案中，对第二处理器 111 的性能要求较低；例如，第一处理器 110 对数据帧划分为多个数据切片的话，参照图 19 所示，在单位周期内第一处理器和第二处理器需要完成一帧数据的处理，在第一处理器的处理周期 T_1 内，当第一处理器 110 处理完数据切片 t_1 后，即可以通过接口控制器 112 将数据切片传输给第二处理器 111，接口控制器 112 每获取一个数据切片即可向第二处理器 111 传输并由第二处理器 111 进行处理，这样第二处理器 111 可以在第一处理器 110 处理完数据切片 t_1 之后开始对接口控制器 112 传输的数据切片进行处理（即第二处理器 111 在第一处理器的处理周期 T_1 内开启对数据帧的处理周期 T_2 ），为第二处理器 111 争取到了更多的处理时间，降低了对第二处理器 111 性能的要求，并且，第二处理器 111 处理 t_1 时，第一处理器 110 仍然可以继续处理其他数据，接口控制器 112 进一步用于将其他数据切片传输至所述第二处理器 111，这样可以有效的形成计算链路流水；另外，接口控制器 112 对数据帧划分为多个数据切片的话，如图 8 所示，只要第一处理器 110 对数据帧开始处理并向接口控制器 112 输出数据帧的数据，则接口控制器 112 就可以开始对数据帧进行切片，这样当接口控制器 112 获取到一个数据切片之后便可以将数据切片传输给第二处理器 111，并由第二处理器 111 进行处理，这样，第二处理器 111 也是可以在 t_1 之后开始对接口控制器 112 传输的数据切片进行处理，为第二处理器 111 争取到了更多的处理时间，降低了对第二处理器 111 性能的要求，并且能够减少整个计算链路的延时；从而提升系统整体性能。

以下以第一处理器为 GPU、第二处理器为 NPU 为例，参照图 20 所示，对本申请的实施例提供的数据传输方法进行说明如下。在初始状态，接口控制器将 bitmap 中的所有 bit 置为 0。其中该初始状态可以是上一次数据帧传输结束后的重置（reset）为初始状态，例如通常数据帧的帧尾携带数据结束指示，具体可以根据该数据结束指示将 bitmap 重置为初始状态。

在步骤 201 中，接口控制器确定 GPU 是否有数据输出。若是，则确定 GPU 输出的是数据切片还是数据帧。若确定 GPU 输出的是数据帧，则将该数据帧划分为多个数据切片，并执行步骤 202。若确定 GPU 输出的是数据切片则执行步骤 202。在步骤 202 中，接口控制器根据数据切片将 bitmap 中对应的 bit 置为 1。在步骤 203 中，接口控制器根据 bitmap 确定是否有目标数据切片以便可以计算对应的第一数据切片。若是，则计算第一数据切片，并执行步骤 204。在步骤 204 中，接口控制器将第一数据切片传输给 NPU，并向 NPU 发送第一处理指示，以通知 NPU 处理该第一数据切片。在步骤

205 中，确定是否所有的数据切片均以处理完毕，若是则结束，若否则执行步骤 203。以上步骤 201-205 介绍了本申请提供的一种采用接口控制器在 GPU 与 NPU 之间数据传输的方法的基本逻辑过程，其中每个过程中的具体方式可以参照上述步骤 101-106 中的描述。

5 上述主要对本申请实施例提供的方案进行了介绍。可以理解的是，上述接口控制器为了实现上述功能，其包含了执行各个功能相应的硬件结构和/或软件模块。本领域技术人员应该可以理解，结合本文中所公开的实施例描述的各示例的单元及算法步骤，本申请能够以硬件或硬件和计算机软件的结合形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行，取决于技术方案的特定应用和设计约束条件。专
10 业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

本申请实施例可以根据上述方法示例对接口控制器进行功能模块的划分，例如，可以对应各个功能划分各个功能模块，也可以将两个或两个以上的功能集成在一个处理模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块
15 的形式实现。需要说明的是，本申请实施例中对模块的划分是示意性的，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。下面以采用对应各个功能划分各个功能模块为例进行说明。

图 21 是本申请实施例提供的接口控制器 300 的逻辑结构示意图，接口控制器 300 能够实现本申请实施例提供的数据传输方法方法。接口控制器 300 可以是硬件结构、
20 软件模块、或硬件结构加软件模块。如图 21 所示，接口控制器 300 包括信息获取单元 301、处理单元 302 和发送单元 303 以及缓冲器。其中，获取单元 301 可以用于执行上述的步骤 101。处理单元 302 可以用于执行上述的步骤 102 和 103。发送单元 303 可以用于执行上述的步骤 104。其中，上述方法实施例涉及的各步骤的所有相关内容均可以援引到对应功能单元的功能描述，在此不再赘述。

25 本领域普通技术人员可知，上述方法中的全部或部分步骤可以通过程序指令相关的硬件完成，该程序可以存储于一计算机可读存储介质中，该计算机可读存储介质如 ROM、RAM 和光盘等。本申请实施例还提供一种计算机可读存储介质，该计算机可读存储介质可以包括存储器。上述提供的任一种接口控制器 300 中相关内容的解释及有益效果均可参考上文提供的对应的方法实施例，此处不再赘述。

30 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。当使用软件程序实现时，可以全部或部分地以计算机程序产品的形式来实现。该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、网络设备、用户设备、或者其他可编程装置。所
35 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个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、服务器或者数据中心通过有线（例如同轴电缆、光纤、数字用户线（digital subscriber line, DSL））或无线（例如红外、无线、微波等）方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能

够存取的任何可用介质或者是包含一个或多个可以用介质集成的服务器、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，数字视频光盘（digital video disc, DVD））、或者半导体介质（例如固态硬盘（solid state disk, SSD））等。

5 尽管在此结合各实施例对本申请进行了描述，然而，在实施所要求保护的本申请过程中，本领域技术人员通过查看所述附图、公开内容、以及所附权利要求书，可理解并实现所述公开实施例的其他变化。在权利要求中，“包括”（comprising）一词不排除其他组成部分或步骤，“一”或“一个”不排除多个的情况。单个处理器或其他单元可以实现权利要求中列举的若干项功能。相互不同的从属权利要求中记载了某些措施，但这并不表示这些措施不能组合起来产生良好的效果。

10 尽管结合具体特征及其实施例对本申请进行了描述，显而易见的，在不脱离本申请的精神和范围的情况下，可对其进行各种修改和组合。相应地，本说明书和附图仅仅是所附权利要求所界定的本申请的示例性说明，且视为已覆盖本申请范围内的任意和所有修改、变化、组合或等同物。显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

权 利 要 求 书

1. 一种数据传输方法，其特征在于，包括：

获取数据帧中的多个数据切片，其中，所述数据帧由第一处理器生成；

将所述多个数据切片写入缓冲器；

5 根据所述多个数据切片中的目标数据切片获取第一数据切片，所述第一数据切片至少包括所述目标数据切片；

将所述第一数据切片传输至第二处理器，其中所述第二处理器用于处理所述第一数据切片。

2. 根据权利要求1所述的数据传输方法，其特征在于，所述获取数据帧中的多个数据切片；包括：

接收所述第一处理器传输的所述多个数据切片，其中所述第一处理器用于将所述数据帧划分为所述多个数据切片。

3. 根据权利要求1所述的数据传输方法，其特征在于，在获取数据帧中的多个数据切片之前，所述方法还包括：接收所述第一处理器传输的所述数据帧；

15 所述获取数据帧中的多个数据切片，包括：将所述数据帧划分为所述多个数据切片。

4. 根据权利要求1-3任一项所述的数据传输方法，其特征在于，根据所述多个数据切片中的目标数据切片获取第一数据切片，包括：在所述缓冲器读取目标数据切片，以及所述目标数据切片的边缘相邻的数据切片；

20 根据所述目标数据切片，以及所述目标数据切片的边缘相邻的数据切片生成所述第一数据切片，其中所述第一数据切片覆盖所述目标数据切片，并且所述第一数据切片与所述目标数据切片的边缘相邻的数据切片具有重叠区域。

5. 根据权利要求1-4任一项所述的数据传输方法，其特征在于，所述第二处理器为神经网络处理器NPU。

25 6. 根据权利要求4所述的数据传输方法，其特征在于，
所述目标数据切片包括 $2n$ 个所述数据切片，其中， n 为正整数。

7. 一种接口控制器，其特征在于，所述接口控制器包括：

获取单元，用于获取数据帧中的多个数据切片，其中，所述数据帧由第一处理器生成；

30 处理单元，用于将所述获取单元获取的所述多个数据切片写入缓冲器；

所述处理单元，用于根据所述缓冲器中的所述多个数据切片中的目标数据切片获取第一数据切片，所述第一数据切片至少包括所述目标数据切片；

发送单元，用于将所述处理单元获取的所述第一数据切片传输至第二处理器，其中所述第二处理器用于处理所述第一数据切片。

35 8. 根据权利要求7所述的接口控制器，其特征在于，所述获取单元具体用于接收所述第一处理器传输的所述多个数据切片，其中所述第一处理器用于将所述数据帧划分为所述多个数据切片。

9. 根据权利要求7所述的接口控制器，其特征在于，所述获取单元还用于接收所述第一处理器传输的所述数据帧；所述获取单元具体用于将所述数据帧划分为所述多

个数据切片。

10. 根据权利要求 7-9 任一项所述的接口控制器，其特征在于，

5 所述处理单元，具体用于在所述缓冲器读取目标数据切片，以及所述目标数据切片的边缘相邻的数据切片；根据所述目标数据切片，以及所述目标数据切片的边缘相邻的数据切片生成所述第一数据切片，其中所述第一数据切片覆盖所述目标数据切片，并且所述第一数据切片与所述目标数据切片的边缘相邻的数据切片具有重叠区域。

11. 根据权利要求 7-10 任一项所述的接口控制器，其特征在于，所述第二处理器为神经网络处理器 NPU。

12. 根据权利要求 10 所述的接口控制器，其特征在于，

10 所述目标数据切片包括 $2n$ 个所述数据切片，其中， n 为正整数。

13. 一种片上系统，其特征在于，包括第一处理器、第二处理器以及接口控制器，其中，所述接口控制器与所述第一处理器及所述第二处理器连接，所述接口控制器用于执行如权利要求 1-6 任一项所述的数据传输方法。

15 14. 根据权利要求 13 所述的片上系统，其特征在于，在所述第二处理器处理所述第一数据切片时，所述接口控制器进一步用于将其他数据切片传输至所述第二处理器。

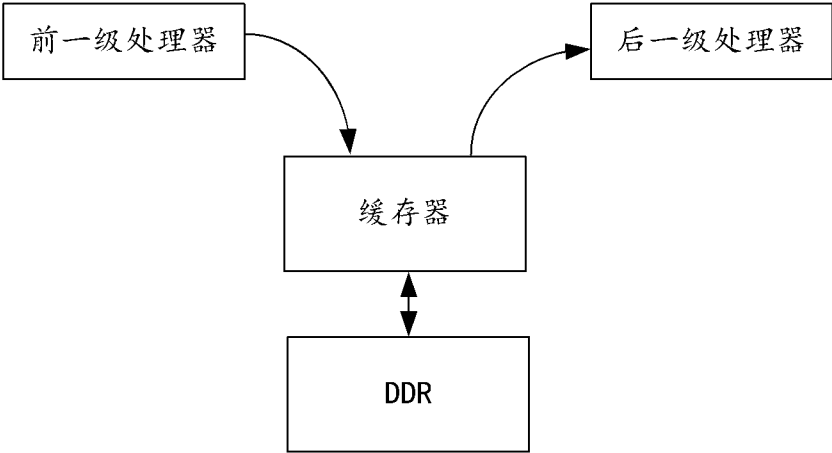


图 1

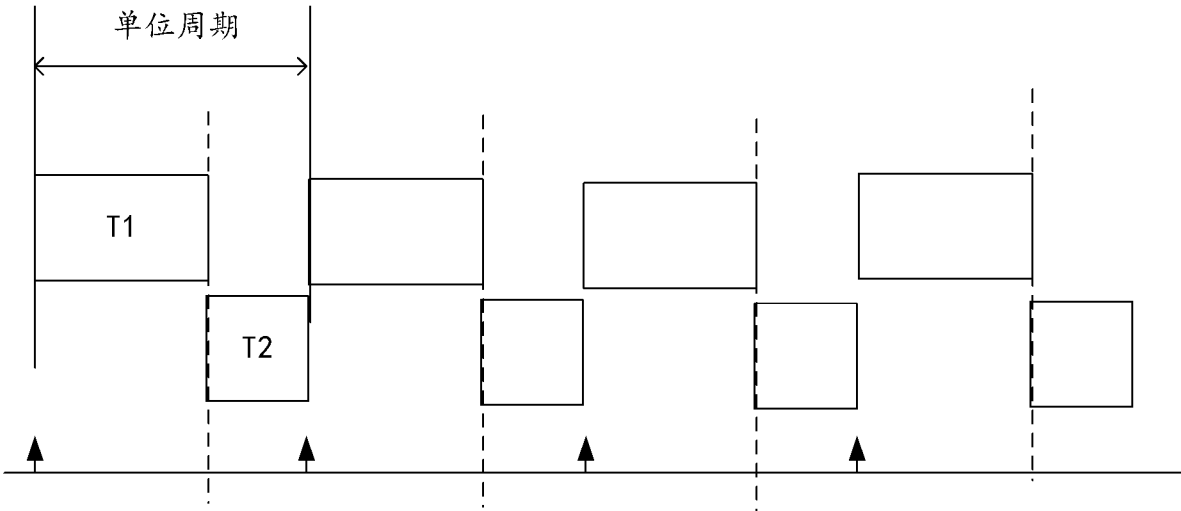


图 2

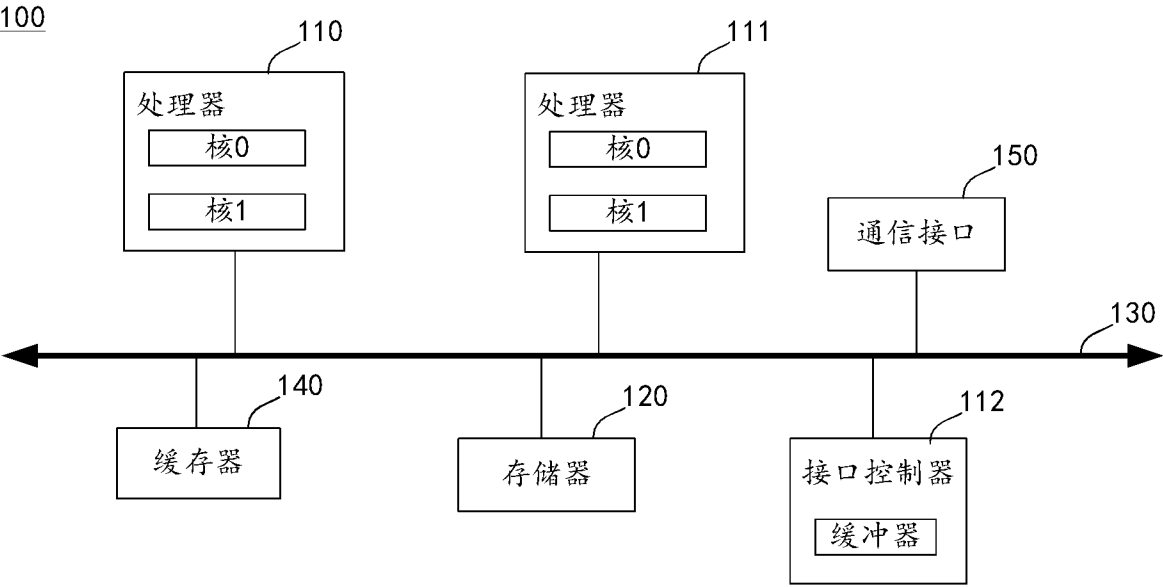


图 3

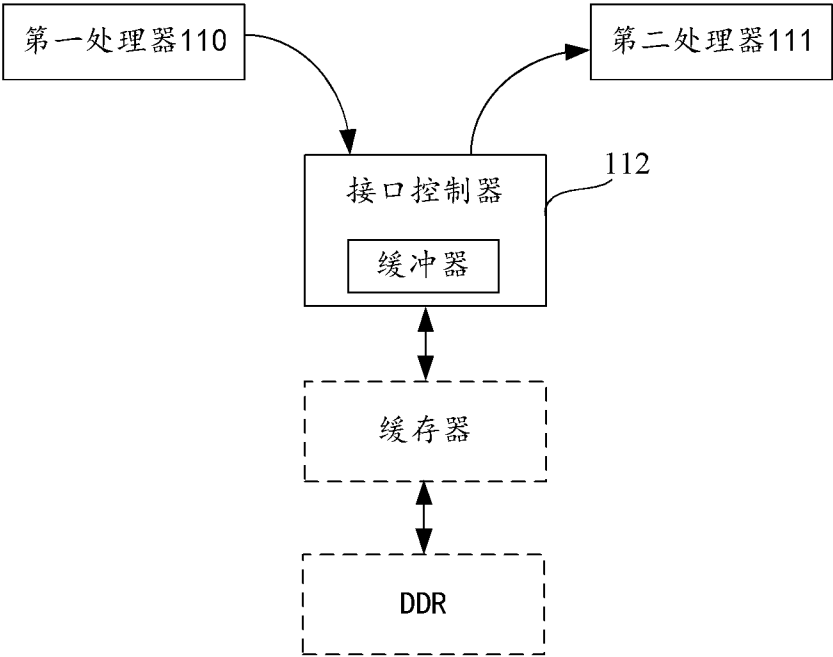


图 4a

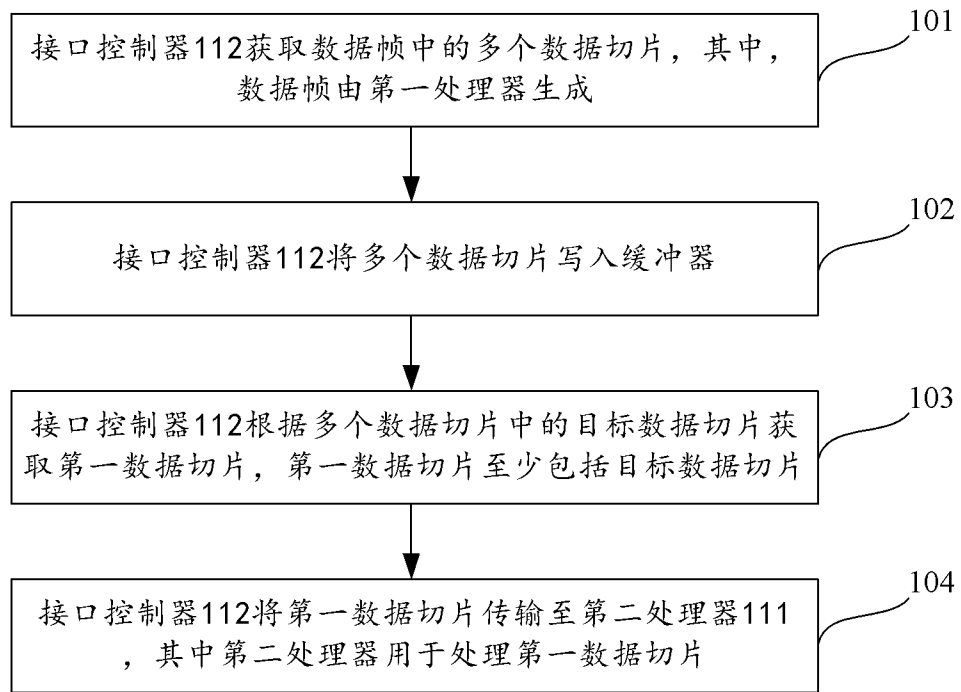


图 4b

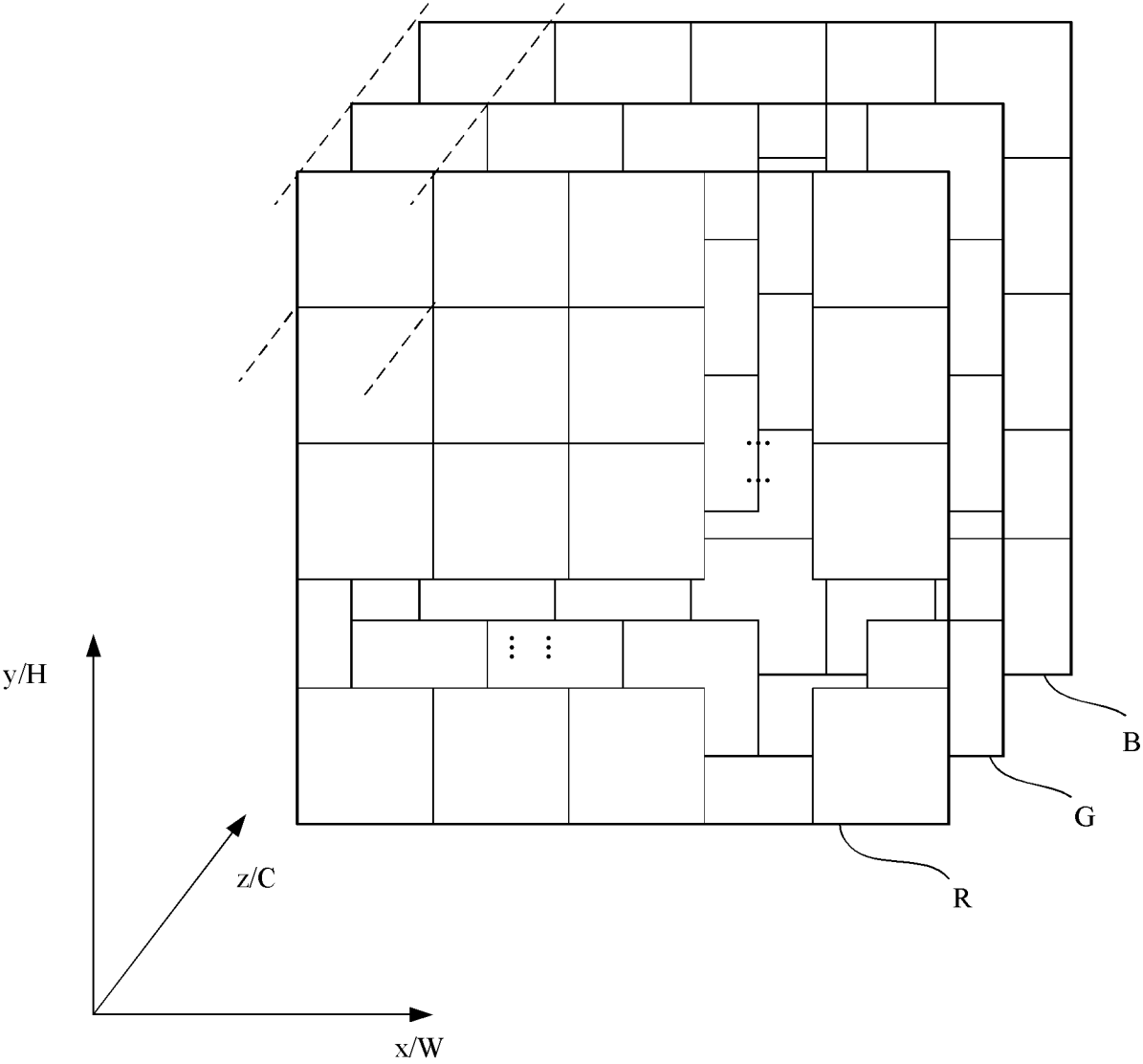


图 5

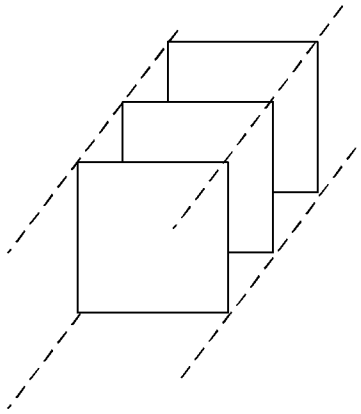


图 6

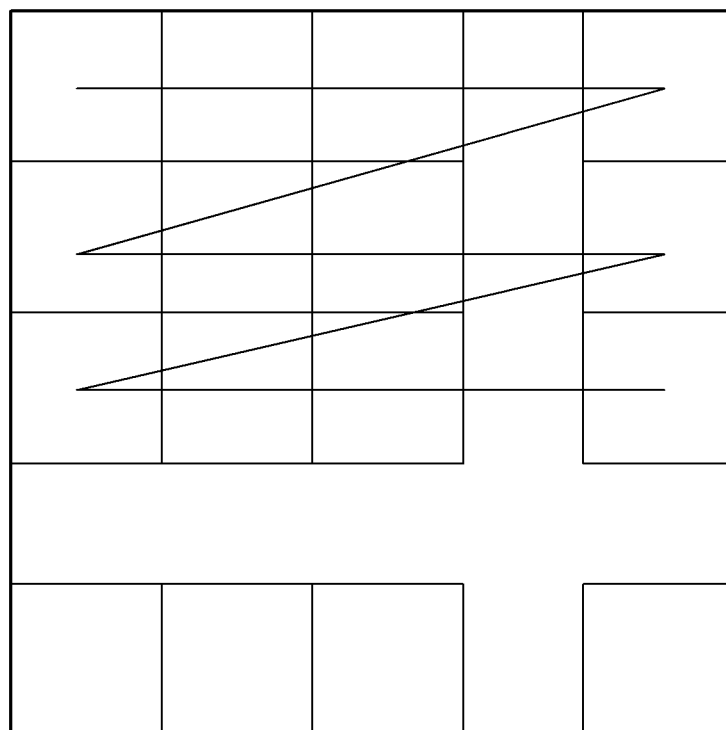


图 7

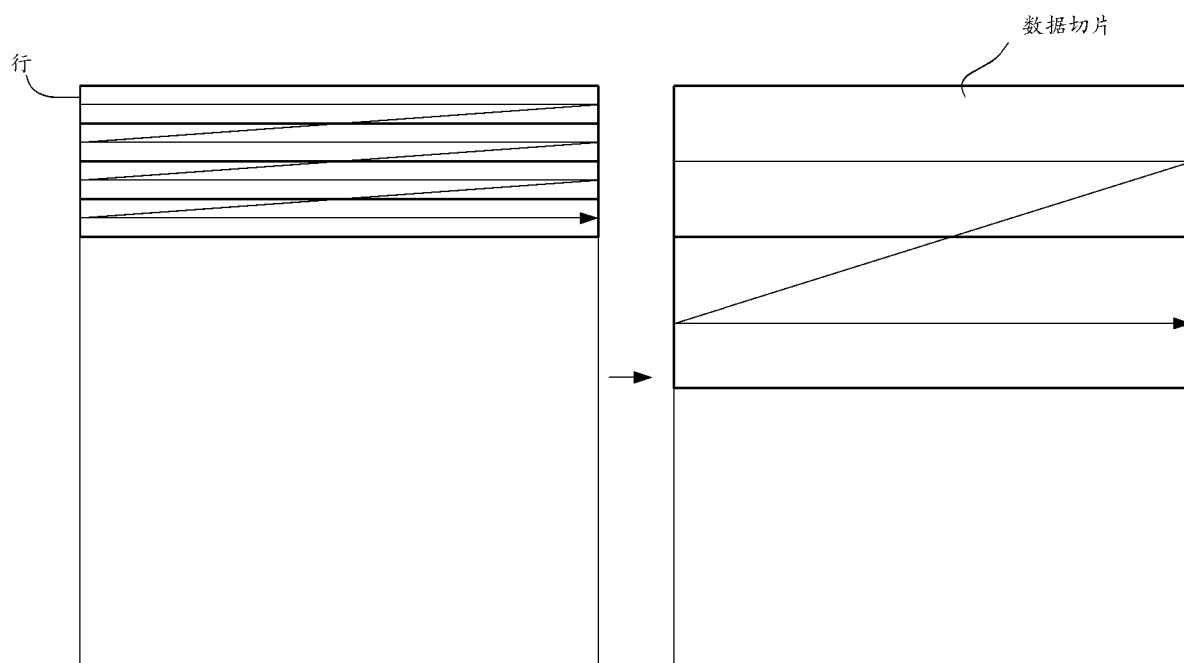


图 8

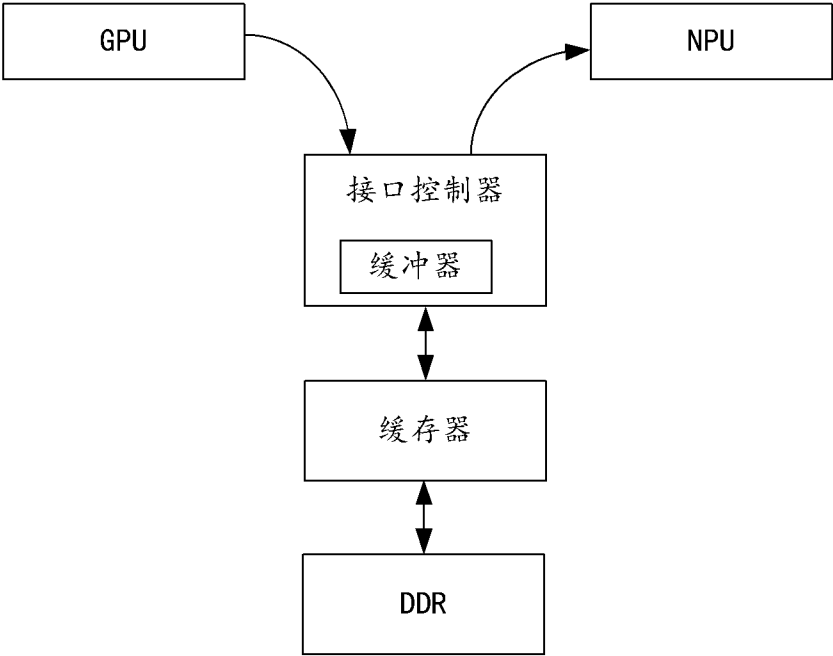


图 9

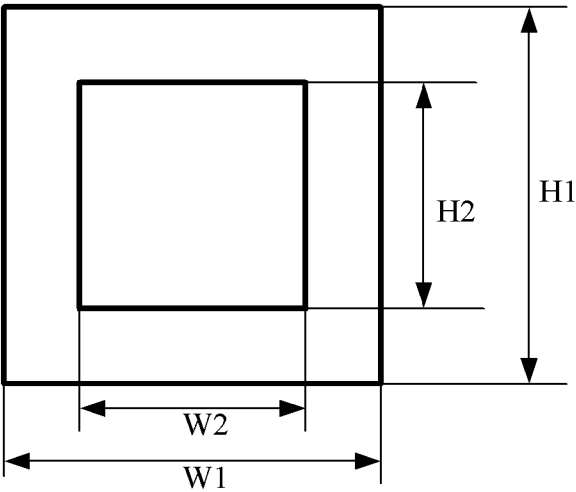


图 10

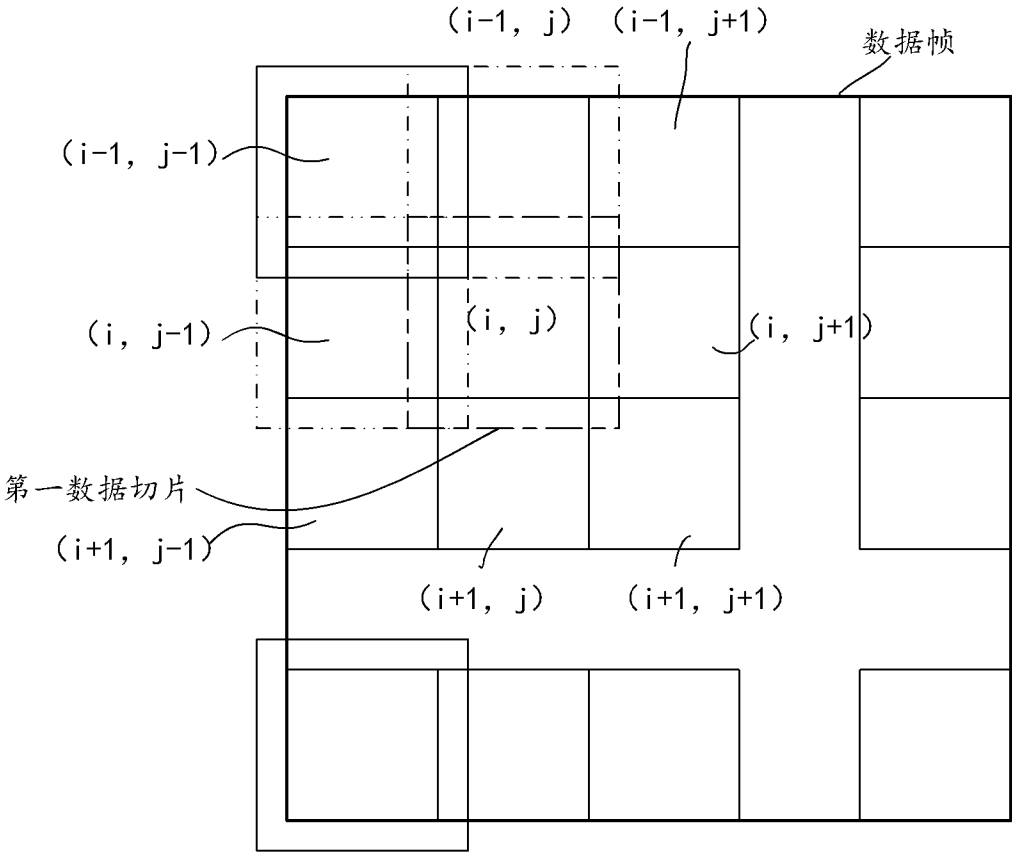


图 11

0	0	1	0	0	0	0
1	1	1	1	0	1	0
0	1	1	1	0	0	0
1	1	1	1	1	1	1
0	0	1	0	1	0	1

0	0
1	0
0	0
1	1
0	1

0	1	0	1	1
0	0	1	0	1

1	1
0	1

图 12

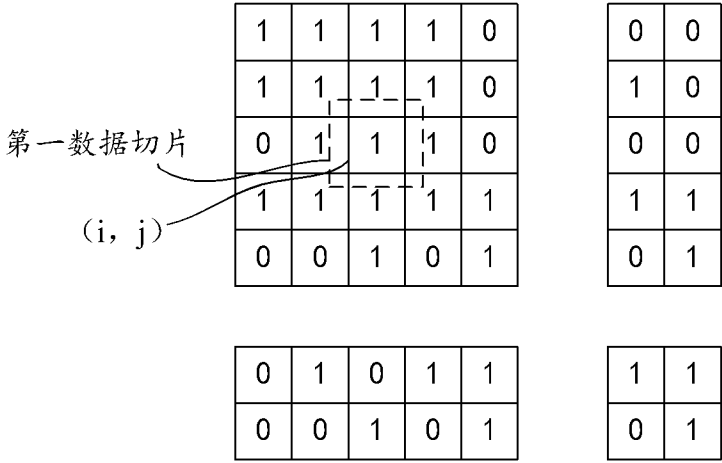


图 13

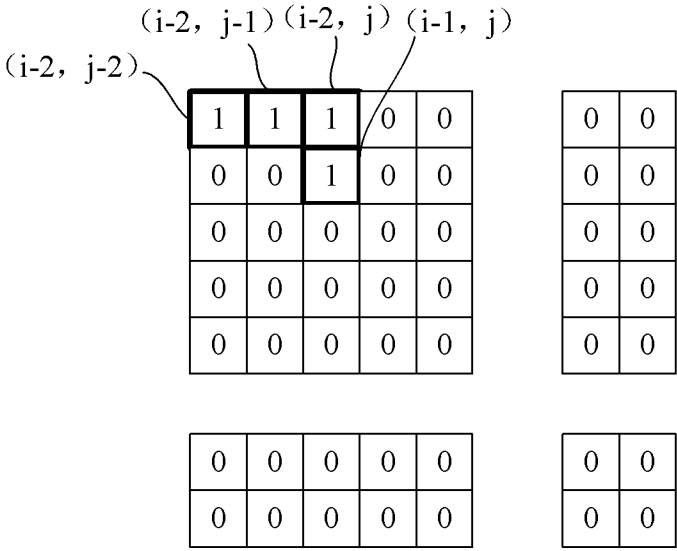


图 14

1	1	1	0	0	0	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

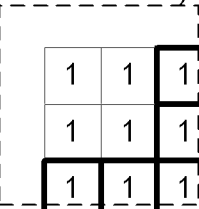
图 15

1	1	1	1	0	1	0	1
1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1
1	1	0	1	0	1	1	1
1	1	1	0	0	1	1	1
0	0	1	1	1	1	0	0
1	1	1	1	0	1	0	1
1	1	1	1	1	1	0	1
1	1	1	1	1	1	1	1
1	1	0	1	0	1	1	1
1	1	1	0	0	1	1	1
0	0	1	1	1	1	0	0

图 16

1	1	1	1	0	1	0	1
1	1	1	1	1	1	0	1
1	1	1	1	1	1	1	1
1	1	0	1	0	1	1	1
1	1	1	0	0	1	1	1
0	0	1	1	1	1	0	0

第一数据切片



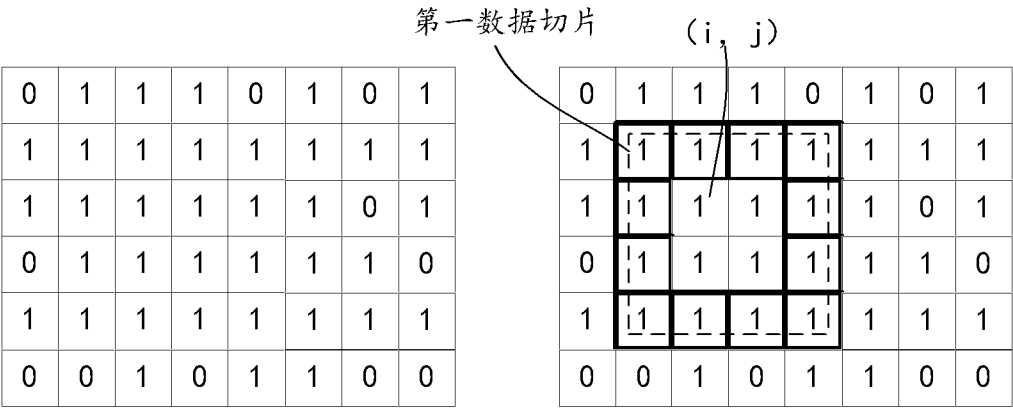


图 17

0, 0	0, 0	0, 0	0, 0
0, 0	1, 0	0, 0	0, 0
0, 0	0, 0	0, 0	0, 0

图 18a

0, 0	0, 0	0, 0	0, 0
0, 0	1, 1	0, 0	0, 0
0, 0	0, 0	0, 0	0, 0

图 18b

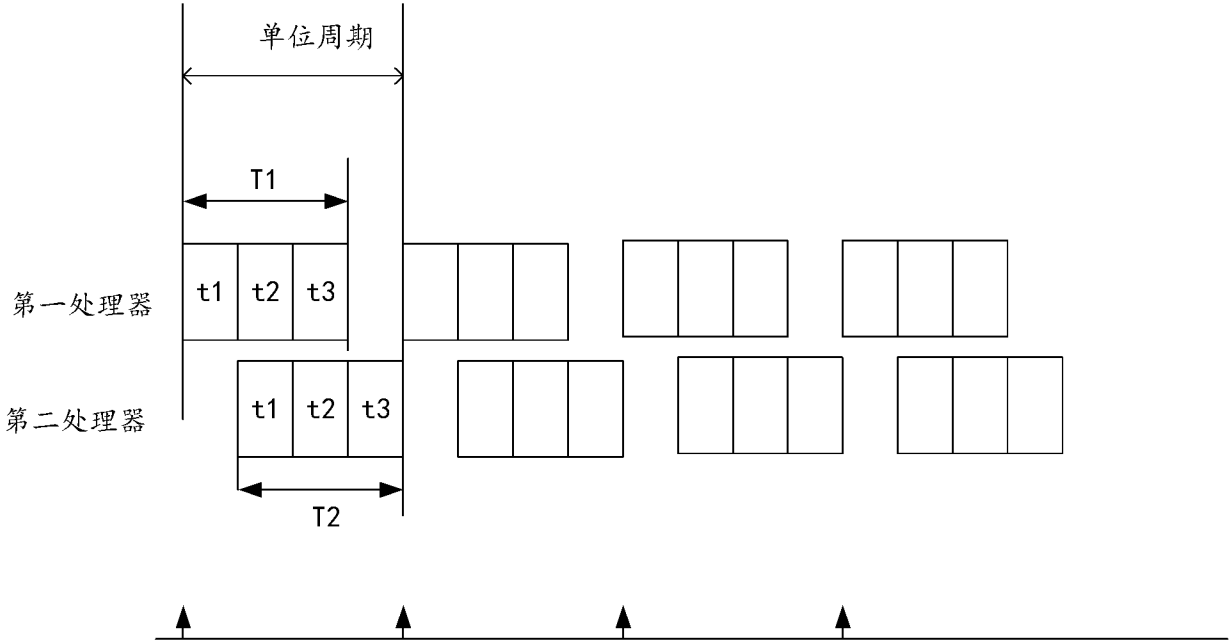


图 19

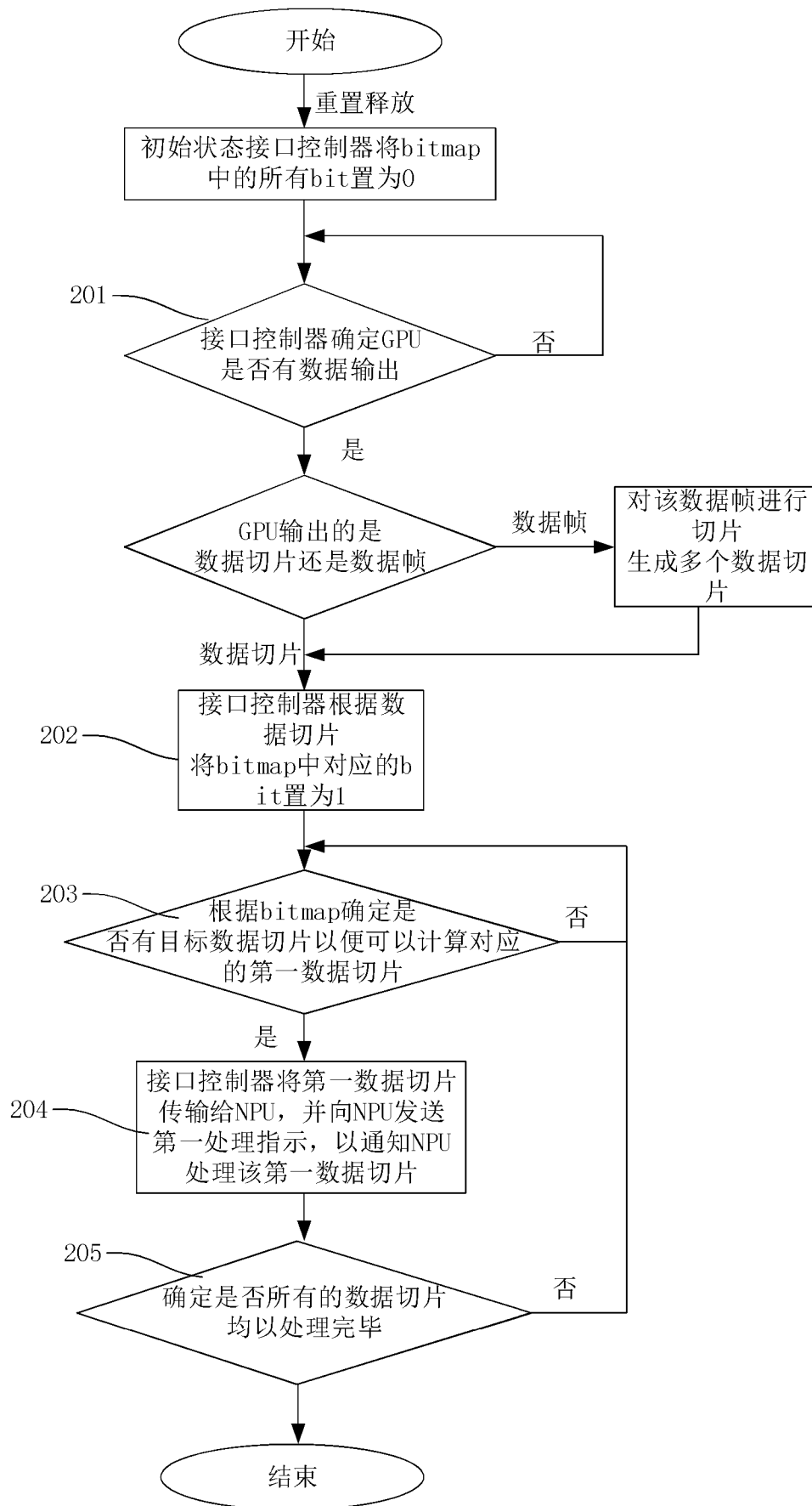


图 20

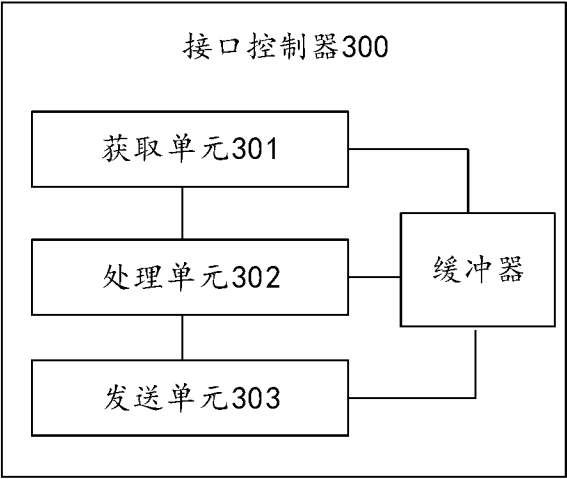


图 21

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/075307

A. CLASSIFICATION OF SUBJECT MATTER

G06F 9/38(2006.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F,H04N

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)

WPI, EPODOC, CNKI, CNPAT: 缓冲, 后一级, 独立, 依赖, NPU, GPU, 切片, 缓存, 处理器, 处理单元, 划分, 分割, 边界, 边缘, buffer, slice, divide, processor, subsequent, follow, prior, dependent, independent, border

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 111465919 A (SHENZHEN DAJIANG INNOVATION TECHNOLOGY CO., LTD.) 28 July 2020 (2020-07-28) description paragraphs [0054]-[0060], figure 6	1-14
A	US 2009125538 A1 (ELEMENTAL TECHNOLOGIES, INC.) 14 May 2009 (2009-05-14) entire document	1-14
A	US 2009310686 A1 (KIM Do-Hyoung et al.) 17 December 2009 (2009-12-17) entire document	1-14
A	WO 2017203096 A1 (PICTURALL OY) 30 November 2017 (2017-11-30) entire document	1-14



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

18 October 2021

Date of mailing of the international search report

01 November 2021

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/CN2021/075307

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	111465919	A	28 July 2020	WO	2019127244	A1	04 July 2019
				US	2020319818	A1	08 October 2020
				EP	3701364	A1	02 September 2020
US	2009125538	A1	14 May 2009	US	2012093234	A1	19 April 2012
				US	2017344513	A1	30 November 2017
US	2009310686	A1	17 December 2009	None			
WO	2017203096	A1	30 November 2017	US	2019189085	A1	20 June 2019
				EP	3465605	A1	10 April 2019

国际检索报告

国际申请号

PCT/CN2021/075307

A. 主题的分类

G06F 9/38(2006.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F, H04N

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))

WPI, EPDOC, CNKI, CNPAT:缓冲, 后一级, 独立, 依赖, NPU, GPU, 切片, 缓存, 处理器, 处理单元, 划分, 分割, 边界, 边缘, buffer, slice, divide, processor, subsequent, follow, prior, dependent, independent, border

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	CN 111465919 A (深圳市大疆创新科技有限公司) 2020年 7月 28日 (2020 - 07 - 28) 说明书第[0054]-[0060]段, 附图6	1-14
A	US 2009125538 A1 (ELEMENTAL TECHNOLOGIES, INC.) 2009年 5月 14日 (2009 - 05 - 14) 全文	1-14
A	US 2009310686 A1 (KIM Do-Hyoung等) 2009年 12月 17日 (2009 - 12 - 17) 全文	1-14
A	WO 2017203096 A1 (PICTURALL OY) 2017年 11月 30日 (2017 - 11 - 30) 全文	1-14

☐ 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

2021年 10月 18日

国际检索报告邮寄日期

2021年 11月 1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

授权官员

陈红英

电话号码 86-(10)-53961636

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/075307

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)
CN	111465919	A	2020年 7月 28日	WO	2019127244	A1	2019年 7月 4日
				US	2020319818	A1	2020年 10月 8日
				EP	3701364	A1	2020年 9月 2日
US	2009125538	A1	2009年 5月 14日	US	2012093234	A1	2012年 4月 19日
				US	2017344513	A1	2017年 11月 30日
US	2009310686	A1	2009年 12月 17日	无			
WO	2017203096	A1	2017年 11月 30日	US	2019189085	A1	2019年 6月 20日
				EP	3465605	A1	2019年 4月 10日