(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
2020 年 10 月 8 日 (08.10.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/199143 A1

(51) 国际专利分类号:
H04Q 11/00 (2006.01)

(21) 国际申请号: PCT/CN2019/081161

(22) 国际申请日: 2019 年 4 月 3 日 (03.04.2019)

(25) 申请语言: 中文

(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)。

(72) 发明人: 沈胜宇 (**SHEN, Shengyu**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 吴聿旻 (**WU, Yumin**); 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, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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,

(54) **Title:** AI TRAINING NETWORK AND METHOD

(54) 发明名称: AI训练网络及方法

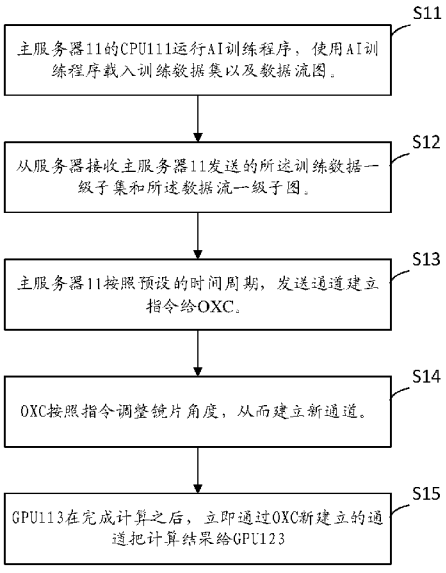

图 3

S11 A CPU 111 of a main server 11 running an AI training program so that the AI training program is loaded into a training data set and a data flow diagram
S12 Receive from a server a primary subset of the training data and a primary subgraph of the data flow which are sent by the main server 11
S13 The main server 11 transmitting a channel establishing instruction to an OXC according to a preset time period
S14 The OXC adjusting the angle of a lens so as to establish a new channel
S15 Transmit a computing result to a GPU 123 by means of the channel established by the OXC after a GPU 113 completes the computation

(57) **Abstract:** Artificial intelligence training technology, which is applied to an artificial intelligence AI training network: pre-emptively beginning to establish an optical channel for communications before graphic processing units in different servers need to communicate, and transmitting a computing result to the graphic processing unit of the next server without waiting or with only a small amount of waiting time immediately after the graphic processing unit of the previous sever completes the computation thereof, thereby reducing the time consumption of AI training.

(57) 摘要: 一种人工智能训练技术，应用于人工智能AI训练网络，在位于不同服务器的图形处理单元需要通信之前，提前开始建立通信用的光通道，一旦前一个服务器的图形处理单元完成自身的计算后，无需等待或者仅等待少量时间即可立刻把计算结果发送给下一个服务器的图形处理单元，从而节约了AI训练的时间消耗。

WO 2020/199143 A1

[见续页]

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))。

AI 训练网络及方法

技术领域

本申请涉及人工智能领域，尤其涉及一种 AI 训练网络及方法。

背景技术

在人工智能领域的 AI 训练中，使用大量加速器（加速器例如可以是 GPU、CPU，可以提供算力）进行计算，计算一个神经网络的最优结构参数，使得该网络能完成特定的工作。所谓“AI 训练”，就是给机器“喂”大量的数据，让它慢慢学会识别和区分对象。例如 ImageNet1K 分类是一种常见的场景，在该场景中可以给定 128 万张图片，其中包含 1000 个不同的对象。同时每张照片已经给出了正确的 label，即给出了该图片中的对象类别。则 AI 训练的任务在于找到一个合适的神经网络架构(如 Alexnet)和其中每个参数的赋值，使得该网络能够尽可能正确的识别图片中的对象。

在具体实现中，多个加速器使用训练算法进行分别进行计算，并把各自的学习结果合并在一起，并在此分发给每个加速器，然后进入下一次迭代。如此经过多轮迭代运算后，机器就能习得更多的关键细节，从而显得更加智能。相较于中央处理器（CPU）而言，图形处理单元（graphics processing unit，GPU）更适合这类迭代运算，因此 GPU 更普遍的应用于 AI 训练。

随着应用场景需求的提升，神经网络规模和数据集规模急剧增长，例如 Nvidia DGX-2 和 google TPU 这样的大规模加速器服务器集群应运而生，以便提高更强的算力。随着高算力加速器集群的规模越来越大，在 GPU 芯片之间传递数据变得更加频繁，这导致了 GPU 芯片之间传递数据的快慢对整个训练过程的耗时所造成的影响越来越明显。因此，如何降低建立光通道 GPU 芯片之间传递数据所耗费的时间，是目前亟需解决的问题。

发明内容

第一方面，提供一种 AI 训练方法，应用于人工智能 AI 训练网络，所述 AI 训练网络包括第一服务器、第二服务器和光交叉连接 OXC 连接，其中所述第一服务器包括第一图形处理单元，所述第二服务器包括第二图形处理单元，所述第一服务器和所述第二服务器分别与所述光交叉连接 OXC 连接，所述方法包括：第一图形处理单元按照第一数据流图对第一数据集进行 AI 训练计算；在所述第一图像处理单元完成对第一数据集的 AI 训练计算之前，触发所述光交叉连接 OXC 开始进行通道切换，通道切换完成后，所述第一图形处理单元与第二图形处理单元之间的光通道建立成功；所述第一图形单元完成计算后，通过已建立完成的所述光通道发送计算结果给所述第二图形单元；所述第二图形单元使用第二数据流图对所述计算结果进行 AI 训练计算。

由于现有技术中占有在应用该方法，第一图形处理单元完成自身的计算之后（也就是在有数据需要传输之后）才开始启动通道的建立，因此不得不等待整个通道的建立时间。而在本实施例中，在有数据需要传输之前即开始了通道的建立，一旦位于第一服务器的第一图形处理单元完成自身的计算后，即可立刻把计算结果发送给下一个服务器的图形处理单元。无需等待同道的建立或者仅等待少量时间来等待通道的建立，从而节约了 AI 训练的时间消耗。

第一方面的第一种可能实现方式中，所述 AI 训练网络还包括主服务器。其中，所述 OXC 进行通道切换具体包括：所述 OXC 接收主服务器的通道建立指令，所述通道建立指令中想携带调整参数；所述 OXC 按照所述调整参数对光通道进行切换。

该方案提供了一种调整 OXC 的具体解决方案。

基于第一方面的第一种可能实现方式中，可选的， 所述主服务器周期性发送所述通道建立指令给所述 OXC。例如，主服务器根据所述第一图形处理单元发送数据给所述第二图形处理单元的的时间周期，以及所述 OXC 的通道切换时间，获得所述通道建立指令的发送周期。

该方案提供了一种根据两个图形处理单元之间发送数据的规律性，周期性指令 OXC 进行通道切换的方案。

第一方面的第二种可能实现方式中，OXC 是微机电系统 MEMS 或者是硅光 SiP。

第二方面，提供一种 AI 训练网络，对应于上面的 AI 训练方法，并具有相应的有益效果。

第三方面，提供一种光交叉连接管理方法，光交叉连接 OXC 分别连接 AI 训练网络中的第一服务器和 AI 训练网络中的第二服务器，其中所述第一服务器包括第一图形处理单元，所述第二服务器包括第二图形处理单元，包括：根据所述第一图形处理单元发送数据给所述第二图形处理单元的的时间周期，以及所述 OXC 的通道切换时间，获得通道建立指令的发送周期；按照所述发送周期，周期性的发送通道切换指令给所述 OXC，指示所述 OXC 进行建立所述第一图像处理单元和所述第二图像处理单元之间的通道。

该方案介绍了如何周期性的建立光交叉 OXC 中的光通道，以便及时的转发第一图形处理单元需要发送给第二图像处理单元的数据。位于第一服务器的第一图形处理单元完成自身的计算后，无需等待或者仅等待少量时间通道即可立刻把计算结果发送给下一个服务器的图形处理单元，从而节约了 AI 训练的时间消耗。

第三方面的第一种可能实现方式中，所述 OXC 进行通道切换具体包括：所述 OXC 接收主服务器的通道建立指令，所述通道建立指令中想携带调整参数；所述 OXC 按照所述调整参数对光通道进行调整。

可选的，基于第三方面的第一种可能实现方式：所述主服务器周期性发送所述通道建立指令给所述 OXC。此外，在第一图形处理单元按照第一数据流图对第一数据集进行 AI 训练计算之前，还可以包括：根据所述第一图形处理单元发送数据给所述第二图形处理单元的的时间周期，以及所述 OXC 的通道切换时间，获得所述通道建立指令的发送周期。

第三方面的第二种可能实现方式中，在第一图形单元完成计算之前，所述通道切换完成。

第三方面的第三种可能实现方式中，微机电系统 MEMS 和硅光 SiP 中的一种。

第四方面，提供一种光交叉连接管理服务器，例如前述的主服务器，可以执行光交叉连接管理方法，并具有相应的技术效果。

附图说明

图 1 提供了一种 AI 训练网络实施例的架构图；

图 2 提供了一种图像处理单元之间数据流量测量图；

图 3 提供了一种 AI 训练实施例流程图；

图 4 提供了微机电系统中进行切换光通道的实施例示意图。

具体实施方式

人工智能（artificial intelligence，AI）训练网络由多台服务器组成服务器阵列，通过运行 AI 程序执行 AI 训练。图 1 提供了一种 AI 训练网络的架构，如图所示，阵列包括服务器 11、服务器 12、服务器 13 和服务器 14，所述阵列还包括光交叉连接 15、光交叉连接 16 和光交叉连接 17。本发明实施例中的服务器可以是专用服务器、通用服务器、工作站、笔记本电脑等具有运算能力

的设备。服务器之间可以通过数据交换网络 18 通信，数据交换网络 18 例如是以太网或者光纤通道（fibre channel，FC）。在这些通过数据交换网络互连的服务器中，可以由图中的某一个服务器作为主服务器，而余下的服务器作为从服务器，主服务器可以把命令通过数据交换网络 18 发送给其他服务器。此外，主服务器可以通过所述数据交换网络 18 从阵列外部接收 AI 训练的指令和原始数据。主服务器可以由服务器之间通过程序选举产生也可以由第三方指定，为了方便介绍，把服务器 11 定义为主服务器，其余服务器是从服务器。需要说明的是，在实际的阵列中还有更多的器件和设备，例如服务器的网络接口卡、内存 RAM、输入输出设备，数据交换网络 18 中的以太网交换机、路由器等，为简介起见，图 1 中没有示出。

一个完整的 AI 训练是迭代的运行以下步骤直到计算结果收敛到足够的精度：（1）前向传播：TF 将输入数据从上图的左侧输入神经网络，按照算子依赖关系，顺序运行各个算子，直到在上图的右端得到结果；（2）计算损失：将步骤（1）中的到的结果，与正确答案之间的差，作为损失；（3）后向传播：按照链式求导规则，将步骤（2）中的损失，逐级反向传播，得到所有参数的梯度；（4）当每一次迭代的损失值趋于平坦，不再有剧烈的下降，可以进行收敛。从上述步骤（1）、（2）和（3）的迭代运行的特征可知，AI 训练的计算量和通讯特征都是迭代重复的。在经过几个迭代之后，就能比较精确的预测在什么时候，会发送多大尺寸的报文，从哪个 GPU 发送到那个 GPU。

服务器中包含中央处理器（CPU）和图形处理单元（GPU），以服务器 11 为例，其内部包含了：CPU111、CPU112 和 GPU113、GPU114。CPU111 和 CPU112 可以通过总线（例如快速路径互连 QPI 总线或者超传输 HT 总线）通信或者节点控制器（node controller，NC）通信。CPU111 和 GPU113 之间，以及 CPU112 和 GPU114 之间，均可以通过快捷外围部件互连标准（peripheral component interconnect express，PCIe）总线通信。除了 PCIe 之外，ISA、PCI、AGP、AGI 与 AGU 也是可行的 GPU 接口标准。CPU 下发计算命令给 GPU，GPU 完成 CPU 下发的计算命令。

每个服务器可以运行一个操作系统（OS），OS 之上可以运行 AI 训练程序，AI 训练程序例如：谷张量流（tensorflow，TF）程序、CNTK、Caffe 和 MXNET。AI 训练软件 TF 需要用户首先给出一个神经网络的结构，称为数据流图，数据流图包括多个算子，算子可以是矩阵乘法、求平均值、求最大值、sigmoid 激活函数等。一些算子之间有依赖关系，也就是说：通过一个算子计算得出的输出结果作为另一个算子的输入数据。阵列拥有大量的 GPU，为了提高计算效率，需要把算子分散到多个 GPU，以便由多个 GPU 共同完成数据流图的计算。由于 GPU 分配到的算子之间的依赖关系，导致了 GPU 之间也产生了依赖关系，即：前一个 GPU 的输出结果作为下一个 GPU 的输入数据。由于依赖关系的存在，就需要在 GPU 之间的通信，为此，这两个 GPU 所分配到的算子除了包含计算算子（计算算子用于进行函数计算）之外，还有通讯算子（通讯算子用于 GPU 之间通信）。

发生通信的两个 GPU 可能属于同一个服务器，也可能属于不同的服务器。当发生通信的 2 个 GPU 属于相同的服务器时，可以通过服务器内部的总线进行通信。而当发生通信的 2 个 GPU 属于不同的服务器时，需要依靠服务器外部的通信渠道——也就是使用图 1 中的光交叉连接（OXC）进行通信，例如，GPU113 的发出的数据，在依次通过 OXC15、OXC17 和 OXC17 之后，可以到达 GPU144。OXC 也通过以太网或 FC 连接到数据交换网络 18 上，以便从以太网接受来自服务器 CPU 的命令，按照命令调整光开关输入和输出之间的连接关系。光交叉连接（optical cross-connect，OXC）器件包括但不限于微机电系统(micro electro mechanical system，MEMS)和硅光(silicon

photonics，SiP)。MEMS 是微米大小的机械系统，其加工技术由半导体加工技术改造而来，操作范围在微米范围内。本申请中提到的 MEMS 光开关，是由 MEMS 工艺制造的，能够按照外部指令进行偏转的反射镜组成的阵列，用于把入射的光束，反射到特定的方向。光束的传播可以是在自由空间中进行。MEMS 的缺点在于通道切换（从原有通道切换到新建立的通道）的速度很慢，大约 10ms 左右，比电交换的 ns 级相差了 6 个数量级。硅光是使用硅片作为光传导介质的光系统，与 MEMS 不同的是，硅片依靠波导通道同时完成光束的传播和方向保持。硅光能够提供比 MEMS 更快的通道切换速度。

然而，不论是 MEMS 还是硅光，从原通道切换到新通道的切换时间对 AI 训练所耗费的时间而言始终是不可忽略的，因此缩小这个切换时间从而整体上提高 AI 训练的耗时，是一个需要解决的问题。GPU 芯片之间传递数据所耗费的时间包括两部分：切换通道的数据和实际传输数据的时间。本发明实施例可以在需要传输数据之前，提前切换好通道，当需要传输数据时可以直接使用现成的通道，从而减少了切换过程对计算时间的 AI 训练影响。

图 2 是在真实 AI 训练过程中，通过软件界面所截取的两个 GPU 间数据流量图，图中横坐标是时间（单位：秒），纵坐标是数据大小（单位：兆字节），每一个标记点代表一次数据传输。从图中可以看出，数据传输具有明显的时间周期性：每间隔大概 200ms 的空闲期会发生频繁的数据传输，这样的频繁传输大概持续 500ms 后结束。数据传输的大小集中在 5MB 以下，5MB~10MB 也有较多分布，10MB~20MB 和 30~40M 有少量分布，根据统计，每一次周期传输的报文总量在 GB 的数量级。在其他实施场景中，可能偶有不符合周期性规律的情况，但是大多数情况下仍然是周期性的，因此仍然可以使用本发明实施例提供的方案获得收益。

因此，本发明实施例中通过发挥 AI 训练流量的高度重复特定和可预测性，提前发送通道切换指令给 OXC，以便触发通道的切换。第一种通道切换时机：可以在前一个 GPU 前一个 GPU 计算完成之前指令 OXC 建立传输通道，并且在前一个 GPU 计算完成之前通道完成切换，通道切换完成后可以直接把数据发送给后一个 GPU。从而避免有数据传输时才临时建立通道，规避 OXC 的低切换速度引起的高延时。根据图 2 的统计可以知道数据的产生是周期性的，主服务器根据过去时刻数据的产生时间，可以预计到后续的数据生成的时刻（也就是流量发生的时刻）、以及通道切换所需要花费的时间，可以计算出触发 OXC 开始通道建立的最晚时刻。只要 OXC 在等于或者略微早于这个最晚时刻开始通道的切换，那么新的通道可以在待传输的数据生成之前建立完成。

第二种通道切换时机：在前一个 GPU 计算完成之前开始（不要求前一个 GPU 计算完成之前通道切换完成）建立传输通道。这种情况下，时机更加灵活，可以在前一个 GPU 计算完成之前完成通道的切换，也可以在数据生成完成之后才完成通道的切换，因此第二种通道切换时机涵盖了第一种通道切换时机。如果使用第二种通道切换时机进行对 OXC 进行通道切换的触发，有可能在通道切换尚未切换完成的时候，前一个 GPU 已经计算完成。由于只有通道切换完成后数据才能被传输，因此前一个 GPU 需要等待一段时间才能发送数据给后一个 GPU，但是相比于现有技术（在有数据需要传输时才开始触发通道的切换）而言，由于通道切换的开始时间得到了提前，因此仍然节约了时间。

下面参照图 3，对本发明 AI 训练实施例流程进行更详细的介绍。

步骤 S11，主服务器 11 的 CPU111 运行 AI 训练程序，使用 AI 训练程序载入训练数据集以及数据流图。主服务器 11 把训练数据集和所述数据流图拆分成几个部分，通过数据交换网络 18，分别发往所述从服务器 12、所述从服务器 13 以及所述从服务器 14，以使得每个服务器分担一部

分训练任务。其中，每个从服务器收到的那部分数据流图用于计算收到的那部分数据集，也就是说收到的那部分数据流图和和收到的那部分数据集之间有对应关系。如果把所有服务器的训练任务合起来，就可以组成所述训练数据集以及所述数据流图。

主服务器 11 在执行执行调度的功能之外，还可以承担一部分所述训练数据集和一部分所述数据流图的计算；主服务器 11 也可以不承担计算任务，仅仅执行调度的功能。主服务器 11 拥有处理器和与接口，接口用于和 OXC 通信，如果主服务器承担计算任务，那么还可以进一步包含图形处理单元。本实施例中共有 4 个服务器，假设计算任务在它们之间平均分配，那么每个服务器处理：1/4 的训练数据集和与 1/4 的训练数据对应的 1/4 的数据流图。为了方便后续的介绍，由单个服务器负责的部分训练数据集称为数据一级子集，由单个服务器负责的部分数据流图称为数据流一级子图。

步骤 S12，从服务器接收主服务器 11 发送的所述训练数据一级子集和所述数据流一级子图。从服务器的 CPU 按照 GPU 的数量把训练数据一级子集和数据流一级子图再次拆分，一个数据一级子集拆分成多个数据二级子集、一个数据流一级子图拆分成多个数据流二级子图。然后把数据二级子集和数据流二级子图发送给对应的 GPU，指令所述 GPU 对接收到的数据二级子集按照接收到的数据流二级子图进行计算。

各个服务器按照自己的数据流一级子图开始计算自己负责的数据一级子集。具体的计算操作由 GPU 执行，以服务器 12 为例，在通过数据交换网络 18 收到需要自己计算的 1/4 的训练数据集和 1/4 的所述数据流图后，服务器 12 的 CPU（CPU121 和/或 CPU122）把计算任务分配给归属的 GPU，例如 GPU123 和 GPU124 分别承担 1/8 的训练数据集和 1/8 的所述数据流图。

步骤 S13，主服务器 11（例如 CPU111 或者 CPU112）按照预设的时间周期，发送通道建立指令给 OXC。如前所述，GPU 之间可能存在依赖关系,由于依赖关系，GPU 之间会周期性的发送大量数据，以图 2 为例，主服务器 11 周期性的发送指令给 OXC，在图 2 所示的例子中，发送持续时间大致是 0.5s（500ms），发送后的间隔时间大致是 0.2s（200ms），因此这个时间周期可以大致是 200ms+500ms=700ms。因此，每间隔 700ms 建立一次对应的通道即可。

通道建立指令中包括调整参数，用于指令按照 OXC 按照所述调整参数对光通道进行调整。本实施例中，调整参数中包括需要调整的镜片的编号以及需要调整的角度。参见图 4，本实施例假设 GPU123（第二图形处理单元）的输入依赖于 GPU113（第一图形处理单元）的输出，那么需要被调整的镜片是位于 GPU123 和 GPU113 之间的 OXC，也就是 MEMS15，MEMS15 包括微机电控制器 150 以及 2 个反射镜片阵列，每个镜片阵列包括多个镜片，镜片的偏转角度在物理上是可调的。GPU113 发出的电信号转换成光信号之后，通过光纤通道 151、反射镜 152、反射镜 153、光纤通道 154 到达 GPU124。如图 4 所示，在调整之前，反射镜提高的偏转角度是 45 °，光信号的反射路径是 155-156-158，此时如果 GPU113 发出的数据，那么数据会到达 GPU124。调整镜片 152 和/或镜片 153 的角度均可以达到修改反射路径的目的，反射路径一旦修改完成也就意味着新的通道建立成功。本实施例中，调整的是反射镜 153，当把反射镜 15 的提高的反射角度调整为 30° 之后，GPU113 和 GPU123 之间的通道建立成功。主服务器 11 发送给 OXC15 的通道建立指令包含的调整参数例如是：{反射镜 15，反射镜角度 30° }。

为了方便介绍，也可以把 GPU113（第一图形处理单元）需要承担的训练数据集和数据流图分别称为第一训练数据集、第一数据流图；把 GPU123（第二图形处理单元）需要承担的训练数据集（GPU113 的计算结果）和数据流图分别称为第二训练数据集、第二数据流图。

需要说明的是，本实施例中，主服务器 11（例如 CPU111 或者 CPU112）按照预设的时间周期，发送通道建立指令给 OXC 的时机可以早于 GPU113 发送数据给 GPU123 的时机，以便提前触发通道的建立。GPU113 一旦完成计算，就可以立即通过这个通道把信号发送给 GPU123。因此本实施例可以增加有这样的限制：GPU113 完成自身分配的训练数据集的计算之前，先要完成 GPU113 与 GPU123 之间通道的建立。例如：GPU113 发送数据给 GPU123 的时间周期是 2 秒，具体而言，分别需要在 10 秒、12 秒、14 秒……的时刻发送数据给 GPU123，建立通道需要花费 0.4 秒，那么主服务器 11 可以通知 OXC 在 9.6 秒、11.6 秒、13.6 秒……之前开始建立 GPU113 和 GPU123 之间的通道。这个例子中，和现有技术相比节约了 0.4 秒的通道建立时间。

需要说明的是，在其他实施例中，“GPU113 完成自身分配的训练数据集的计算之前，先要完成 GPU113 与 GPU123 之间通道的建立”，这一限制并不是必须的。在其他实施例中，可以不限于在 GPU113 完成计算之前完成通道的建立，只要在 GPU113 完成计算之前启动通道的建立即可。例如：GPU113 分别需要在 10 秒、12 秒、14 秒……发送数据给 GPU123，通道建立需要是 0.4 秒，那么主服务器 11 可以通知 OXC 在 9.7 秒、11. 7 秒、13. 7 秒……的时刻开始建立 GPU113 和 GPU123 之间的通道。在这个例子中， GPU113 在完成计算之后，需要等待 0.1 秒之后才能把数据通过通道发送给 GPU123，和现有技术相比节约了 0.3 秒的通道建立时间。或者，，那么主服务器 11 可以通知 OXC 在 9 秒、11 秒、13 秒……的时刻开始建立 GPU113 和 GPU123 之间的通道。在这个例子中，通道建立完成之后的 0.2 秒 GPU113 才完成计算，和现有技术相比节约了 0.4 秒的通道建立时间。

需要特别说明的是，在步骤 S13-S15 中主服务器 11 所执行的功能，例如发送通道建立指令以及接受通道建立指令的响应，并不限于由主服务器 11 执行，也可以改由集群中其他服务器或者第三方设备执行。

需要说明的是，步骤 S12 和步骤 S13 之间没有依赖关系。二者可以并行执行，也可以任意一个先执行。

步骤 S14，MEMS15 接收包含调整参数{反射镜 15，反射镜角度 30° }的通道建立指令，MEMS 控制器 150 按照指令把反射镜 15 的角度进行调整，调整后反射的光线角度为 30° 。

经过调整之后，光信号的反射路径 155-156-157 建立完成，也就是说 GPU11 和 GPU123 之间的通道切换完成。微机电系统 15 发送通道建立成功的响应消息给主服务器 11，以便告知主服务器 11：GPU113 和 GPU123 之间的通道成功建立。主服务器 11 收到该响应消息之后，主服务器 11 的 CPU 把通道建立成功的消息通知给 GPU113。

步骤 S15，GPU113 收到主服务器 11 发送的通知之后。如果已经完成计算，则可以立即通过光路径 155-156-157 把计算结果给 GPU123，GPU123 使用收到的数据进行后续计算。如果未完成计算，则等待完成计算之后，可以立即通过光路径 155-156-157 发送计算结果给 GPU123。GPU123 收到 GPU113 的计算结果后，按照自己的数据流子图进行下一步计算。

由以上步骤可以看出，GPU113 一旦完成计算，在 MEMS15 中已经有现成的通道供其使用，因此可以立即使用 MEMS15 把数据发给 GPU123，由于不用耗费时间等待通道的建立，因此节约了时间。对于 MEMS 而言，每一次跨服务器的 GPU 通信可以节约左右 10ms，而在一次 AI 训练中，服务器阵列需要频繁的在不同服务器的 GPU 之间传输信号，因此应用该实施例可大量节约时间。

在步骤 13 中提到“预设的时间周期”， OXC 中的镜片基于该时间周期提前翻转，下面对如

何获得这个周期进行示例性的说明。需要强调的是，该时间周期的获取还可以有更多的办法，此处提供两种以加深本领域人员对实施例的理解。

方法一：由管理员设置。对于同一种类型的 AI 训练，时间周期是相似性，因此管理员可以根据自身经验掌握这个时间周期，通过人工在软件中设置这个周期的数值。

方法二：由服务器阵列自行获得。对于需要进行 GPU 之间通信的 GPU 而言，其收到的子数据流图中除了包含计算算子之外，还可以包含通讯算子，通讯算子可以对 GPU 之间的依赖关系进行描述。需要发送数据的 GPU 的子数据流图中拥有“发送”算子，需要接收数据的 GPU 收到的子数据流图中拥有“接收”算子。当 GPU 使用通讯算子数据传输时，可以记录这个数据传输的相关信息：源 GPU、目的 GPU、传输数据量大小、发生的时间点等信息。通过这些信息（这些相关信息可以由源 GPU 记录或者目的 GPU 记录或者共同记录）即可以获得如图 2 所示的 GPU 之间的流量图，从而掌握 GPU 之间传输数据的规律性。这些信息可以存放在 GPU 所在的服务器的存储器中，也可以汇总到一个统一的存放位置，例如汇总到主服务器的存储器中，或者汇总到服务器阵列之外的第三方设备。在持续记录一定时间之后，软件可以掌握时间周期，并把时间周期保存到可被读取的存储位置。

本发明还提供一种程序产品的实施例，运行在主服务器中，程序产品包括程序代码，主服务器运行所述程序代码可以对 OXC 进行管理，例如：根据所述第一图形处理单元发送数据给所述第二图形处理单元的的时间周期，以及所述 OXC 的通道切换时间，获得通道建立指令的发送周期；按照所述发送周期，周期性的发送通道切换指令给所述 OXC，指示所述 OXC 进行建立所述第一图像处理单元和所述第二图像处理单元之间的通道。

权 利 要 求 书

1、一种 AI 训练方法，应用于人工智能 AI 训练网络，所述 AI 训练网络包括第一服务器、第二服务器和光交叉连接 OXC 连接，其中所述第一服务器包括第一图形处理单元，所述第二服务器包括第二图形处理单元，所述第一服务器和所述第二服务器分别与所述光交叉连接 OXC 连接，所述方法包括：

第一图形处理单元按照第一数据流图对第一数据集进行 AI 训练计算；在所述第一图形单元完成对所述第一数据集的 AI 训练计算之前，触发所述 OXC 开始进行通道切换，通道切换完成后，所述第一图形处理单元与第二图形处理单元之间的光通道建立成功；

所述第一图形单元完成计算后，通过已建立完成的所述光通道发送计算结果给所述第二图形单元；

所述第二图形单元使用第二数据流图对所述计算结果进行 AI 训练计算。

2、根据权利要求 1 所述的 AI 训练方法方法，其中，所述 AI 训练网络还包括主服务器，所述 OXC 进行通道切换具体包括：

所述 OXC 接收主服务器的通道建立指令，所述通道建立指令中携带调整参数；

所述 OXC 按照所述调整参数对光通道进行切换。

3、根据权利要求 2 所述的 AI 训练方法方法，其中，所述方法还包括：

所述主服务器周期性发送所述通道建立指令给所述 OXC。

4、根据权利要求 3 所述的 AI 训练方法方法，其中，所述方法还包括：在第一图形处理单元按照第一数据流图对第一数据集进行 AI 训练计算之前，还包括：

根据所述第一图形处理单元发送数据给所述第二图形处理单元的的时间周期，以及所述 OXC 的通道切换时间，获得所述通道建立指令的发送周期。

5、根据权利要求 1-4 任一项所述的 AI 训练方法方法，其中，所述述通道切换完成的时间是：

在第一图形单元完成计算之前 。

6、根据权利要求 1-4 任一项所述的 AI 训练方法方法，其中，所述 OXC 是：

微机电系统 MEMS 和硅光 SiP 中的一种。

7、一种 AI 训练网络，所述 AI 训练网络包括第一服务器、第二服务器和光交叉连接 OXC 连接，其中所述第一服务器包括第一图形处理单元，所述第二服务器包括第二图形处理单元，所述第一服务器和所述第二服务器分别与所述光交叉连接 OXC 连接，其中：

所述第一图形处理单元用于：按照第一数据流图对第一数据集进行 AI 训练计算，以及通过已建立完成的所述光通道发送计算结果给所述第二图形单元；

所述光交叉连接 OXC 用于：在所述第一图像处理单元完成对第一数据集的 AI 训练计算之前，开始进行通道切换，其中，通道切换完成后，所述第一图形处理单元与第二图形处理单元之间的光通道建立成功；

所述第二图形单元用于：使用第二数据流图对所述计算结果进行 AI 训练计算。

8、根据权利要求 7 所述的 AI 训练网络，其中，所述 AI 训练网络还包括主服务器，所述包括主服务器用于：

发送通道建立指令给所述 OXC，所述通道建立指令中想携带调整参数；

所述 OXC 按照所述调整参数对光通道进行通道切换。

9、根据权利要求 8 所述的 AI 训练网络，其中，所述主服务器还用于：

周期性发送所述通道建立指令给所述 OXC。

10、根据权利要求 9 所述的 AI 训练网络，其中，所述所述主服务器还用于：

根据所述第一图形处理单元发送数据给所述第二图形处理单元的的时间周期，以及所述 OXC 的通道切换时间，获得所述通道建立指令的发送周期。

11、根据权利要求 7-10 任一项所述的 AI 训练网络，其中，所述光交叉连接 OXC 进一步用于：

在所述第一图像处理单元完成对第一数据集的 AI 训练计算之前，完成所述通道切换。

12、根据权利要求 7-10 任一项所述的 AI 训练网络，其中，所述 OXC 是：

微机电系统 MEMS 和硅光 SiP 中的一种。

13、一种光交叉连接管理方法，光交叉连接 OXC 分别连接 AI 训练网络中的第一服务器和 AI 训练网络中的第二服务器，其中所述第一服务器包括第一图形处理单元，所述第二服务器包括第二图形处理单元，包括：

根据所述第一图形处理单元发送数据给所述第二图形处理单元的的时间周期，以及所述 OXC 的通道切换时间，获得通道建立指令的发送周期；

按照所述发送周期，周期性的发送通道切换指令给所述 OXC，指示所述 OXC 在所述第一图像处理单元完成对第一数据集的 AI 训练计算之前，建立所述第一图像处理单元和所述第二图像处理单元之间的通道。

14、一种光交叉连接管理服务器，光交叉连接管理服务器和光交叉连接 OXC 通信，所述 OXC 与 AI 训练网络中的第一服务器和 AI 训练网络中的第二服务器通信，其中所述第一服务器包括第一图形处理单元，所述第二服务器包括第二图形处理单元，所述光交叉连接管理服务器包括处理器，所述处理器用于：

根据所述第一图形处理单元发送数据给所述第二图形处理单元的的时间周期，以及所述 OXC 的通道切换时间，获得通道建立指令的发送周期；

按照所述发送周期，周期性的发送通道切换指令给所述 OXC，在所述第一图像处理单元完成对第一数据集的 AI 训练计算之前，指示所述 OXC 进行建立所述第一图像处理单元和所述第二图像处理单元之间的通道。

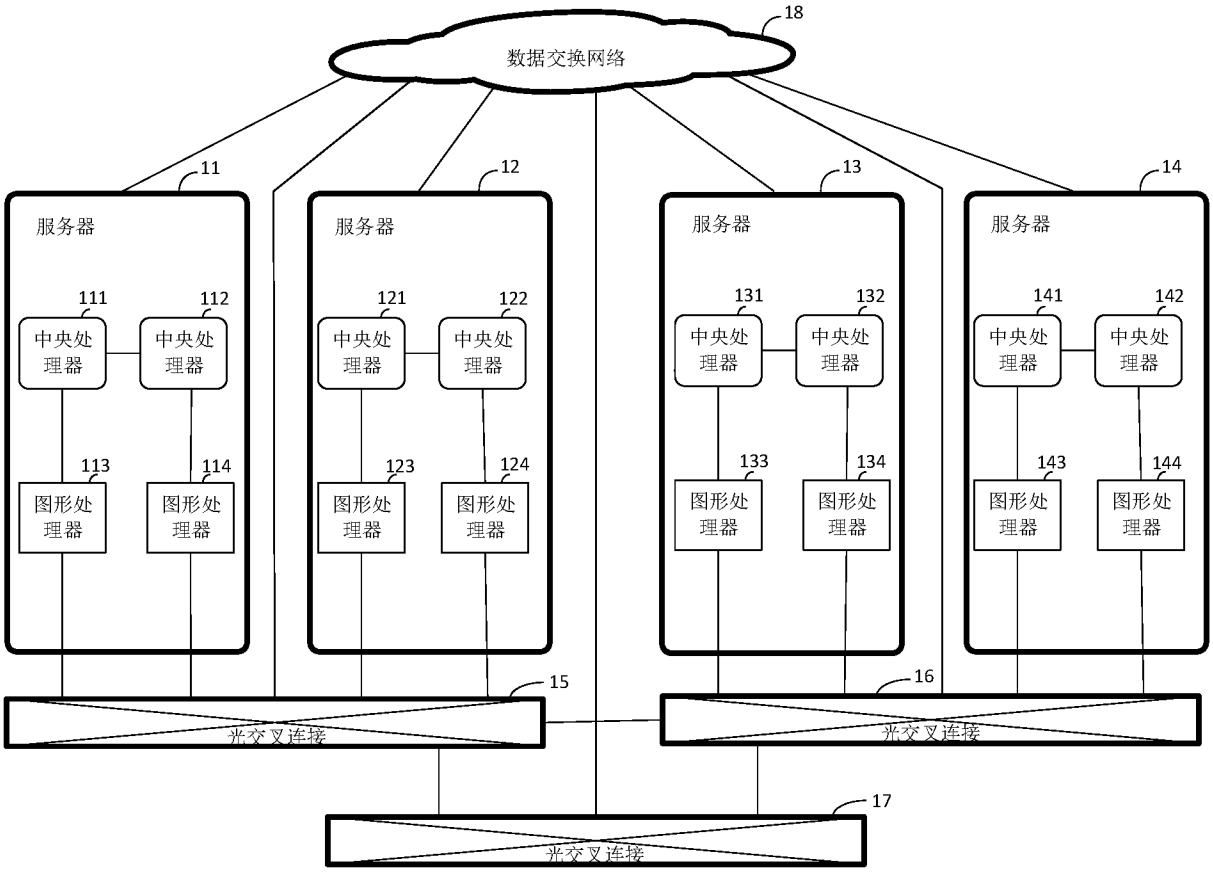
18
数据交换网络
11
12
13
14
服务器
服务器
服务器
服务器
111
112
121
122
131
132
141
142
中央处理器
中央处理器
中央处理器
中央处理器
中央处理器
中央处理器
中央处理器
中央处理器
113
114
123
124
133
134
143
144
图形处理器
图形处理器
图形处理器
图形处理器
图形处理器
图形处理器
图形处理器
图形处理器
15
16
光交叉连接
光交叉连接
17
光交叉连接

图 1

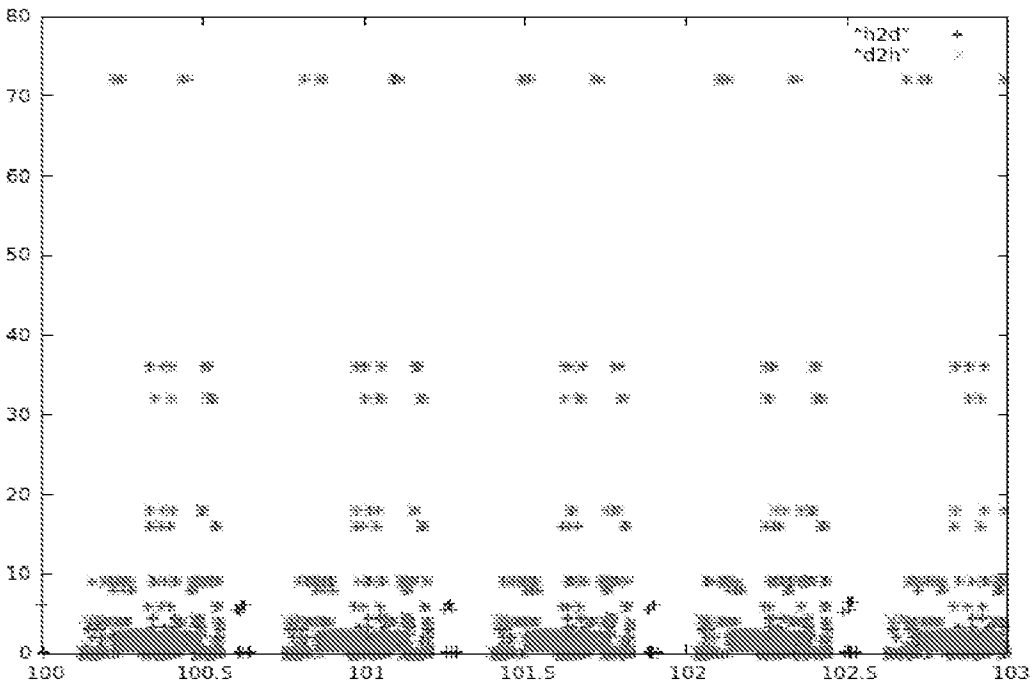
80
70
60
50
40
30
20
10
0
100
100.5
101
101.5
102
102.5
103

图 2

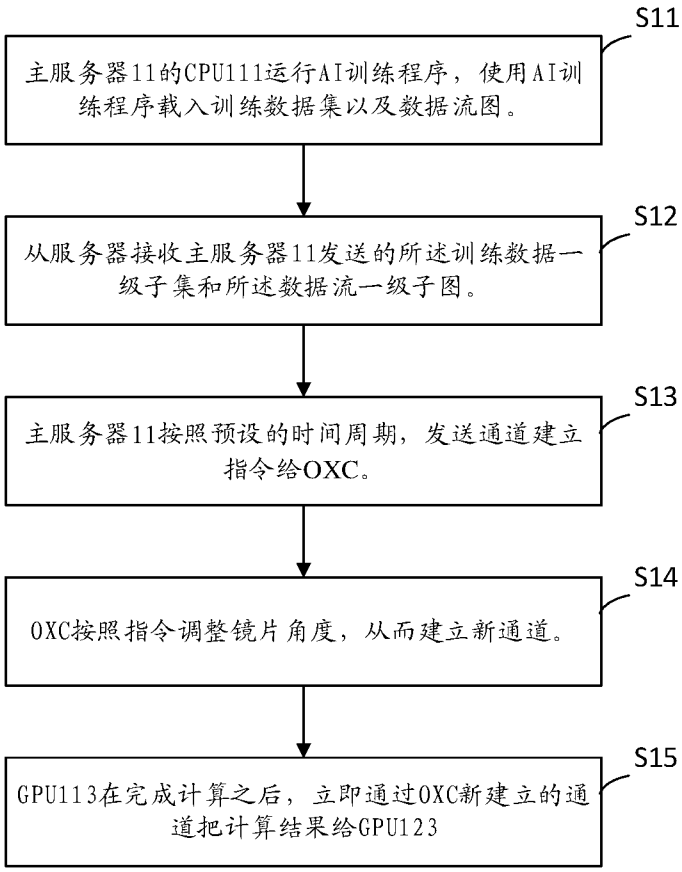
S11
主服务器11的CPU111运行AI训练程序，使用AI训练程序载入训练数据集以及数据流图。
S12
从服务器接收主服务器11发送的所述训练数据一级子集和所述数据流一级子图。
S13
主服务器11按照预设的时间周期，发送通道建立指令给OXC。
S14
OXC按照指令调整镜片角度，从而建立新通道。
S15
GPU113在完成计算之后，立即通过OXC新建立的通道把计算结果给GPU123

图 3

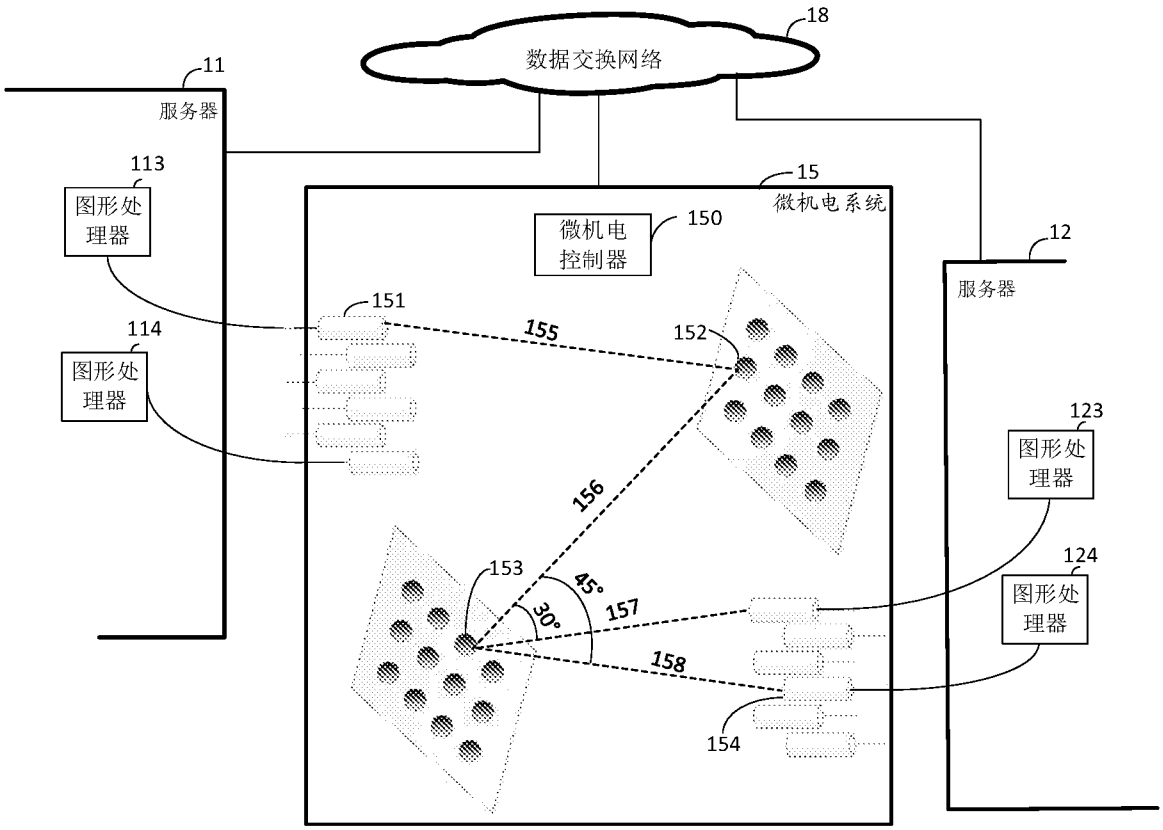
18
数据交换网络
11
服务器
113
图形处理器
114
图形处理器
15
微机电系统
微机电控制器
150
151
155
152
156
153
45°
30°
157
158
154
12
服务器
123
图形处理器
124
图形处理器

图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/081161

A. CLASSIFICATION OF SUBJECT MATTER

H04Q 11/00(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)

H04Q

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)

CNABS; CNTXT; CNKI; VEN; USTXT; EPTXT; WOTXT: 人工智能, 训练, 光交叉连接, 图形处理单元, 集群, 提前, 通道, 通路, 切换, 减少, 降低, 延时, AI, OXC, GPU, path, switch, delay

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 108353217 A (EQUINIX, INC.) 31 July 2018 (2018-07-31) description, paragraphs [0025]-[0056], and figures 1 and 2A-2B	1-14
Y	CN 102546749 A (CHINA TELECOM CORPORATION LIMITED) 04 July 2012 (2012-07-04) description, paragraph [0039]	1-14
A	CN 1984006 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 20 June 2007 (2007-06-20) entire document	1-14
A	CN 106941633 A (WUHAN RESEARCH INSTITUTE OF POSTS AND TELECOMMUNICATIONS) 11 July 2017 (2017-07-11) entire document	1-14
A	CN 107885762 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 06 April 2018 (2018-04-06) entire document	1-14
A	US 2003026524 A1 (KAKIZAKI, S. et al.) 06 February 2003 (2003-02-06) 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	Date of mailing of the international search report
09 December 2019	**06 January 2020**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration** **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/CN2019/081161

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	108353217	A	31 July 2018	US	9736556	B2	15 August 2017
				US	10206017	B2	12 February 2019
				US	2017078771	A1	16 March 2017
				AU	2016321133	A1	19 April 2018
				US	2017366882	A1	21 December 2017
				CA	2998117	C	15 January 2019
				BR	112018004710	A2	25 September 2018
				JP	6557408	B2	07 August 2019
				EP	3326381	A1	30 May 2018
				AU	2016321133	B2	31 January 2019
				CN	108353217	B	15 March 2019
				WO	2017044405	A1	16 March 2017
				CA	2998117	A1	16 March 2017
				JP	2018528697	A	27 September 2018
				HK	1256203	A0	13 September 2019
				SG	11201801605	A1	28 March 2018
CN	102546749	A	04 July 2012	CN	102546749	B	29 July 2015
CN	1984006	A	20 June 2007	CN	100452739	C	14 January 2009
CN	106941633	A	11 July 2017		None		
CN	107885762	A	06 April 2018	US	2019087383	A1	21 March 2019
US	2003026524	A1	06 February 2003	JP	4676099	B2	27 April 2011
				JP	2003051785	A	21 February 2003
				US	2007223918	A1	27 September 2007
				US	7542672	B2	02 June 2009

国际检索报告

国际申请号
PCT/CN2019/081161

A. 主题的分类

H04Q 11/00(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)

H04Q

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））

CNABS;CNTXT;CNKI;VEN;USTXT;EPTXT;WOTXT：人工智能，训练，光交叉连接，图形处理单元，集群，提前，通道，通路，切换，减少，降低，延时，AI，OXC，GPU，path，switch，delay

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
Y	CN 108353217 A (环球互连及数据中心公司) 2018年 7月 31日 (2018-07-31) 说明书第[0025]-[0056]段，图1、2A-2B	1-14
Y	CN 102546749 A (中国电信股份有限公司) 2012年 7月 4日 (2012-07-04) 说明书第[0039]段	1-14
A	CN 1984006 A (北京邮电大学) 2007年 6月 20日 (2007-06-20) 全文	1-14
A	CN 106941633 A (武汉邮电科学研究院) 2017年 7月 11日 (2017-07-11) 全文	1-14
A	CN 107885762 A (北京百度网讯科技有限公司) 2018年 4月 6日 (2018-04-06) 全文	1-14
A	US 2003026524 A1 (KAKIZAKI S等) 2003年 2月 6日 (2003-02-06) 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	国际检索报告邮寄日期
2019年 12月 9日	2020年 1月 6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 张婧 电话号码 86-(20)-28950441

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/081161

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	108353217	A	2018年 7月 31日	US	9736556	B2	2017年 8月 15日
				US	10206017	B2	2019年 2月 12日
				US	2017078771	A1	2017年 3月 16日
				AU	2016321133	A1	2018年 4月 19日
				US	2017366882	A1	2017年 12月 21日
				CA	2998117	C	2019年 1月 15日
				BR	112018004710	A2	2018年 9月 25日
				JP	6557408	B2	2019年 8月 7日
				EP	3326381	A1	2018年 5月 30日
				AU	2016321133	B2	2019年 1月 31日
				CN	108353217	B	2019年 3月 15日
				WO	2017044405	A1	2017年 3月 16日
				CA	2998117	A1	2017年 3月 16日
				JP	2018528697	A	2018年 9月 27日
				HK	1256203	A0	2019年 9月 13日
				SG	11201801605	A1	2018年 3月 28日
CN	102546749	A	2012年 7月 4日	CN	102546749	B	2015年 7月 29日
CN	1984006	A	2007年 6月 20日	CN	100452739	C	2009年 1月 14日
CN	106941633	A	2017年 7月 11日		无		
CN	107885762	A	2018年 4月 6日	US	2019087383	A1	2019年 3月 21日
US	2003026524	A1	2003年 2月 6日	JP	4676099	B2	2011年 4月 27日
				JP	2003051785	A	2003年 2月 21日
				US	2007223918	A1	2007年 9月 27日
				US	7542672	B2	2009年 6月 2日