(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
2019 年 9 月 12 日 (12.09.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/169986 A1

(51) 国际专利分类号:
H04L 12/733 (2013.01)

(21) 国际申请号: PCT/CN2019/074052

(22) 国际申请日: 2019 年 1 月 31 日 (31.01.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810178287.5 2018年3月5日 (05.03.2018) 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)。

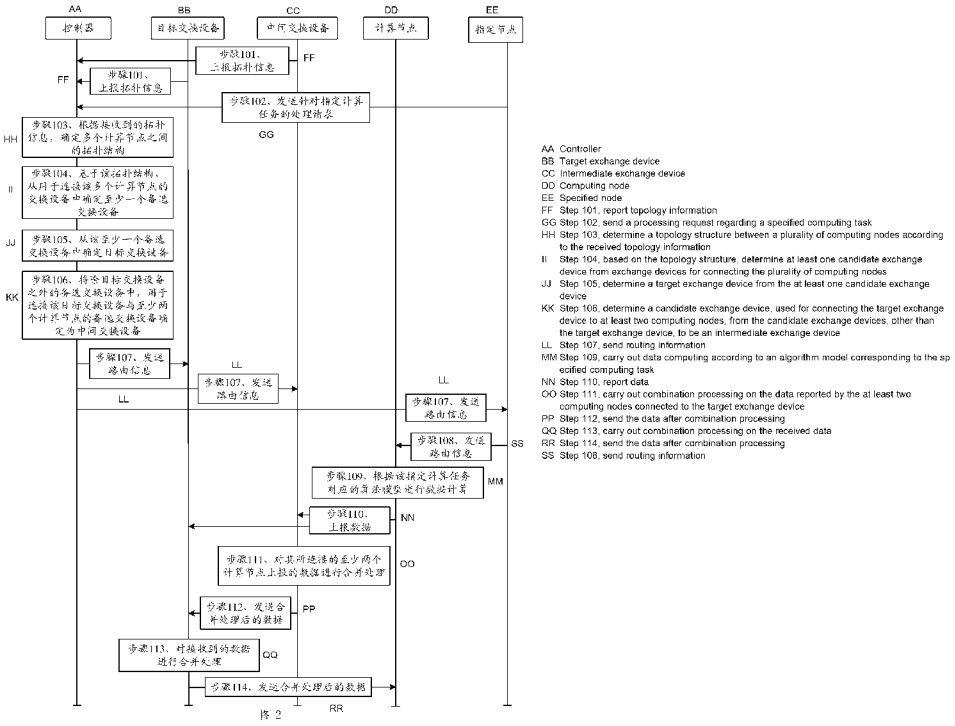
(72) 发明人: **黄伊 (HUANG, Yi)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **夏寅贲 (XIA, Yinben)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **刘孟竹 (LIU, Mengzhu)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(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,

WO 2019/169986 A1

(54) **Title:** DATA PROCESSING METHOD, APPARATUS AND SYSTEM

(54) 发明名称: 数据处理方法、装置及系统



(57) **Abstract:** Provided are a data processing method, apparatus and system, relating to the field of distributed computing. After a controller receives a processing request, carrying identifiers of a plurality of computing nodes for executing a specified computing task, sent by a specified node, a target exchange device can be determined from exchange devices for connecting the plurality of computing nodes; and routing information for indicating a data forwarding path between the plurality of computing nodes and the target exchange device is respectively sent to the target exchange device and the specified node, wherein the target exchange device is used for carrying

[见续页]

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, 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))。

out, according to the routing information, combination processing on data reported by the plurality of computing nodes and then sending the data to each computing node, the specified node is used for sending the routing information to each computing node, and each computing node can report the data to the target exchange device according to the routing information. The method provided in the present application can reduce the probability of network congestion and improve the execution efficiency of a computing task.

(57) 摘要：本申请提供了一种数据处理方法、装置及系统，涉及分布式计算领域，控制器接收到指定节点发送的携带有用于执行指定计算任务的多个计算节点的标识的处理请求后，可以从用于连接该多个计算节点的交换设备中确定目标交换设备；并分别向该目标交换设备以及该指定节点发送用于指示该多个计算节点与该目标交换设备之间的数据转发路径的路由信息；其中，目标交换设备用于根据该路由信息对该多个计算节点上报的数据进行合并处理后发送至每个计算节点，指定节点用于将该路由信息发送至每个计算节点，每个计算节点可以根据该路由信息向目标交换设备上报数据。本申请提供的方法可以降低网络拥塞的概率，提高计算任务的执行算效率。

数据处理方法、装置及系统

本申请要求了2018年3月5日提交的，申请号为CN 201810178287.5 发明名称为“数据处理方法、装置及系统”的中国申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及分布式计算领域，特别涉及一种数据处理方法、装置及系统。

背景技术

分布式机器学习一般采用数据并行的方式进行模型训练。在采用该数据并行的方式进行模型训练时，各个计算节点（也称为worker）中均存储有算法模型，且每个节点可以分别获取到部分样本数据，并能对获取到的样本数据进行训练得到模型参数。各个计算节点需要将计算得到的模型参数发送至参数服务器（parameter server，ps），该参数服务器用于对各个计算节点上报的模型参数进行汇聚更新，并将更新后的模型参数再发送至各个计算节点。

相关技术中，通常采用高性能计算（High Performance Computing，HPC）数据中心网络来实现分布式机器学习。具体的，可以选取一台服务器作为参数服务器，选取另一台服务器作为主节点，并可以选取多台其他服务器作为计算节点。其中，该主节点用于向该多个计算节点下发该参数服务器的网络地址，以及驱动该多个计算节点执行分布式机器学习任务。在该HPC数据中心网络中，该参数服务器与各个计算节点之间可以通过交换设备交互数据，以实现模型参数的上报，以及更新后的模型参数的下发。

但是，在该分布式机器学习的过程，数据中心网络中的数据传输量较大，可能会出现网络拥塞，导致计算节点与参数服务器之间的数据传输时延较大，影响分布式机器学习的效率。

发明内容

本申请提供了一种数据处理方法、装置及系统，可以解决相关技术中的数据中心网络在实现分布式计算时，网络中数据传输量较大，可能会出现网络拥塞，影响计算效率的问题。技术方案如下：

第一方面，提供了一种数据处理方法，应用于数据中心网络的控制器，该方法可以包括：

接收指定节点发送的针对指定计算任务的处理请求，该处理请求中包括用于执行该指定计算任务的多个计算节点的标识，之后控制器可以从用于连接该多个计算节点的交换设备中确定目标交换设备，并分别向目标交换设备以及指定节点发送指定计算任务对应的路由信息，该路由信息用于指示该多个计算节点与目标交换设备之间的数据转发路径。

其中，该路由信息用于在目标交换设备对多个计算节点上报的数据进行合并处理后根据该路由信息将该合并处理后的数据发送至每个计算节点。也即是，目标交换设备可以根据该路由信息对该多个计算节点上报的数据进行合并处理后发送至每个计算节点。此外，指定节点接收到路由信息后，可以将该路由信息发送至该多个计算节点中除该指定节点之外的每个计算节点，每个计算节点可以根据该路由信息向该目标交换设备上报数据。

本申请提供的方法，由于控制器可以选取目标交换设备对多个计算节点上报的数据进行合并处理，因此各计算节点无需再通过交换设备向参数服务器发送数据，参数服务器也无需再通过交换设备将合并处理后的结果反馈至各计算节点，有效减小了数据中心网络中的数据

传输量，降低了网络拥塞的概率以及数据传输的时延，提高了计算任务的执行算效率。

可选的，该多个计算节点与该目标交换设备之间的数据转发路径上可以包括至少一个交换设备，该方法还可以包括：

将该数据转发路径上包括的至少一个交换设备中，与该多个计算节点中的至少两个计算节点连接的交换设备确定为中间交换设备；并向中间交换设备发送路由信息，该路由信息用于该中间交换设备将与其连接的至少两个计算节点上报的数据进行合并处理后根据该路由信息将合并处理后的数据发送至该目标交换设备。

各计算节点在向目标交换设备上报数据的过程中，通过中间交换设备对至少两个计算节点上报的数据进行合并处理后再发出，相比于中间交换设备直接转发数据，可以进一步减小网络中的数据传输量，进而可以进一步降低网络拥塞的概率。

可选的，控制器分别向目标交换设备以及该指定节点发送指定计算任务对应的路由信息的过程可以包括：

向目标交换设备发送包括该目标交换设备的直连设备的标识的路由信息，该目标交换设备的直连设备为计算节点或者中间交换设备；

向指定节点发送包括每个计算节点的直连设备的标识的路由信息，每个计算节点的直连设备为目标交换设备或者中间交换设备，该指定节点用于将每个计算节点的直连设备的标识发送至对应的计算节点；

相应的，控制器向中间交换设备发送路由信息的过程可以包括：

向中间交换设备发送包括该中间交换设备的直连设备的标识的路由信息，中间交换设备的直连设备为计算节点、该目标交换设备或其他中间交换设备。

其中每个设备的标识可以为设备的 IP 地址。

控制器向每个设备发送的路由信息可以仅包括该设备的直连设备的标识，从而可以在保证数据正常转发的基础上，进一步降低路由信息的数据量，有效提高路由信息的传输效率。

可选的，控制器从用于连接该多个计算节点的交换设备中确定目标交换设备的过程可以包括：

分别计算用于连接该多个计算节点的交换设备中，每个交换设备与各个计算节点之间的路由跳数之和；将路由跳数之和最少的交换设备确定为目标交换设备。

在本发明实施例中，选取路由跳数之和最少的交换设备作为目标交换设备，可以保证选取出的目标交换设备与各个计算节点之间的总路径较短，可以有效降低网络中的数据传输量较少，进而可以降低网络拥塞的概率。

作为一种可选的实现方式，控制器将路由跳数之和最少的交换设备确定为目标交换设备的过程可以包括：

当路由跳数之和最少的交换设备包括多个时，分别确定每个路由跳数之和最少的交换设备的性能参数，该性能参数包括可用带宽、吞吐量、计算负载以及被选为目标交换设备的次数中的至少一种；将多个路由跳数之和最少的交换设备中，性能参数满足预设条件的交换设备确定为目标交换设备。

根据交换设备的性能参数选取目标交换设备，可以保证选取出的目标交换设备的性能较好，能够保证较高的计算效率。

作为另一种可选的实现方式，控制器将路由跳数之和最少的交换设备确定为目标交换设备的过程可以包括：

当路由跳数之和最少的交换设备包括多个时，分别确定每个路由跳数之和最少的交换设备与各个所述计算节点之间的路由跳数的均衡程度；将多个路由跳数之和最少的交换设备中，路由跳数的均衡程度最高的交换设备确定为目标交换设备。

根据路由跳数的均衡程度选取目标交换设备，可以保证选取出的目标交换设备与各个计算节点之间的路径长度较为均衡，进而可以保证各个计算节点上报数据时所需的时长较为接近，使得目标交换设备可以在较短的时间内接收到所有计算节点上报的数据，并进行合并处理，进一步提高了计算任务的执行效率。

可选的，控制器在计算每个交换设备与各个计算节点之间的路由跳数之和之前，还可以先检测该多个计算节点是否均直接连接至同一个交换设备；当该多个计算节点均直接连接至同一个交换设备时，控制器可以直接将该多个计算节点直接连接的交换设备确定为目标交换设备，而无需再计算交换设备与各个计算节点之间的路由跳数之和，可以提高目标交换设备的确定效率；当该多个计算节点直接连接至不同的交换设备时，控制器再计算每个交换设备与各个计算节点之间的路由跳数之和。

可选的，该方法还可以包括：

接收该数据中心网络中每个交换设备上报的拓扑信息；根据接收到的拓扑信息，确定该多个计算节点之间的拓扑结构；相应的，控制器在确定与该各个计算节点均具有连接关系的交换设备时，可以基于该拓扑结构确定。

可选的，控制器从与多个计算节点均具有连接关系的交换设备中确定目标交换设备时，还可以先从用于连接该多个计算节点的交换设备中确定至少一个备选交换设备，每个备选交换设备能够通过下行路径与至少两个计算节点连接；之后，控制器可以从该至少一个备选交换设备中确定该目标交换设备。

可选的，控制器在确定目标交换设备时，还可以先从用于连接该多个计算节点的交换设备中确定至少一个备选交换设备，其中每个备选交换设备可以通过下行路径与该多个计算节点中的至少两个计算节点连接；之后，控制器可以再从该至少一个备选交换设备中确定目标交换设备。

可选的，指定节点发送的处理请求还可以包括：该指定计算任务对应的合并处理类型；相应的，该方法还可以包括：

向目标交换设备发送该指定计算任务对应的合并处理类型，该目标交换设备用于按照该合并处理类型对该多个计算节点上报的数据进行合并处理。

由于不同的计算任务对应的合并处理类型可能不同，按照指定计算任务对应的合并处理类型对接收到的数据进行合并处理，可以保证数据处理的精度。

第二方面，提高了另一种数据处理方法，应用于数据中心网络的交换设备，该方法可以包括：接收控制器发送的指定计算任务对应的路由信息，该路由信息用于指示多个计算节点与目标交换设备之间的数据转发路径，该多个计算节点用于执行该指定计算任务；进一步的，交换设备可以对该多个计算节点上报的数据进行合并处理，并可以根据该路由信息，发送合并处理后的数据。其中，该路由信息为控制器接收到指定节点发送的针对该指定计算任务的处理请求后，从用于连接该多个计算节点的交换设备中确定目标交换设备后发送的。

本申请提供的方法，由于交换设备可以对多个计算节点上报的数据进行合并处理后再发出，因此各计算节点无需再通过交换设备向参数服务器发送数据，参数服务器也无需再通过

交换设备将合并处理后的结果反馈至各计算节点，有效减小了数据中心网络中的数据传输量，降低了网络拥塞的概率以及数据传输的时延，提高了计算任务的执行算效率。

可选的，交换设备在对该多个计算节点上报的数据进行合并处理之前，还可以接收该控制器发送的该指定计算任务对应的合并处理类型；相应的，交换设备对该多个计算节点上报的数据进行合并处理的过程可以包括：按照该合并处理类型，对该多个计算节点上报的数据进行合并处理。

可选的，该交换设备可以为目标交换设备；此时，该目标交换设备根据该路由信息，发送合并处理后的数据的过程可以包括：

根据该路由信息，向每个计算节点发送合并处理后的数据。

可选的，该交换设备可以为用于连接该目标交换设备和至少两个该计算节点的中间交换设备；此时，该中间交换设备对该多个计算节点上报的数据进行合并处理的过程可以包括：对至少两个该计算节点上报的数据进行合并处理；

相应的，中间交换设备根据该路由信息，发送合并处理后的数据的过程可以包括：根据该路由信息，向该目标交换设备发送合并处理后的数据。

第三方面，提供了一种数据处理装置，应用于数据中心网络的控制器，该装置可以包括至少一个模块，该至少一个模块用于实现上述第一方面所提供的数据处理方法。

第四方面，提供了一种数据处理装置，应用于数据中心网络的交换设备，该装置可以包括至少一个模块，该至少一个模块用于实现上述第二方面所提供的数据处理方法。

第五方面，提供了一种控制器，该控制器可以包括处理器、存储器以及通信接口；该存储器中存储有供该处理器运行的计算机程序，该处理器、存储器以及该通信接口可以用于实现上述第一方面所提供的数据处理方法。

第六方面，提供了一种交换设备，该交换设备包括交换芯片、处理器以及存储器，该交换芯片、处理器以及存储器可以用于实现上述第二方面所提供的数据处理方法。

第七方面，提供了一种数据处理系统，该系统可以包括：控制器、多个计算节点以及至少一个交换设备；该控制器可以包括第三方面所示的数据处理装置，或者可以为第五方面所示的控制器；每个交换设备可以包括第四方面所示的数据处理装置，或者可以为第七方面所示的交换设备。

第八方面，提供了一种计算机可读存储介质，该计算机可读存储介质中存储有指令，当该计算机可读存储介质在计算机上运行时，使得计算机执行上述第一方面或第二方面所提供的数据处理方法。

第九方面，提供了一种包含指令的计算机程序产品，当该计算机程序产品在计算机上运行时，可以使得计算机执行上述第一方面或第二方面所提供的数据处理方法。

综上所述，本申请提供了一种数据处理方法、装置及系统，指定节点向控制器发送的针对指定计算任务的处理请求中包括用于执行该指定计算任务的多个计算节点的标识，控制器接收到该处理请求后，可以从用于连接该多个计算节点的交换设备中确定目标交换设备，并分别向该目标交换设备以及该指定节点发送用于指示该多个计算节点与该目标交换设备之间的数据转发路径的路由信息，以使得每个计算节点可以根据该路由信息向该目标交换设备上报数据，目标交换设备可以根据该路由信息对该多个计算节点上报的数据进行合并处理后再发送至每个计算节点。因此通过本申请提供的方法，各计算节点无需再通过交换设备向参数服务器发送数据，参数服务器也无需再通过交换设备将合并处理后的结果反馈至各计算节点，有效减小了数据中心网络中的数据传输量，降低了网络拥塞的概率以及数据传输的时延，提高了计算任务的执行算效率。

附图说明

图 1A 是本发明实施例提供的数据处理方法所涉及的数据中心网络的架构图；

图 1B 是本发明实施例提供的一种交换设备的架构图；

图 1C 是本发明实施例提供的另一种交换设备的架构图；

图 1D 是本发明实施例提供的一种数据中心网络中的控制器的架构图；

图 2 是本发明实施例提供的一种数据处理方法的流程图；

图 3 是本发明实施例提供的另一种数据中心网络的架构图；

图 4 是本发明实施例提供的一种多个计算节点之间的拓扑结构示意图；

图 5 是本发明实施例提供的一种确定目标交换设备的方法流程图；

图 6 是本发明实施例提供的另一种多个计算节点之间的拓扑结构示意图；

图 7 是本发明实施例提供的一种数据处理装置的结构示意图；

图 8 是本发明实施例提供的一种发送模块的结构示意图；

图 9 是本发明实施例提供的一种确定模块的结构示意图；

图 10 是本发明实施例提供的另一种数据处理装置的结构示意图；

图 11 是本发明实施例提供的又一种数据处理装置的结构示意图。

具体实施方式

图 1A 是本发明实施例提供的数据处理方法所涉及的数据中心网络的架构图，如图 1A 所示，该数据中心网络可以包括控制器 01、多个计算节点 02 以及用于连接该多个计算节点 02 的至少一个交换设备 03。其中，该控制器 01 以及每个计算节点 02 均可以部署在服务器中，该交换设备 03 可以为具备数据转发以及数据处理功能的交换机。参考图 1A 可以看出，该控制器 01 与每个交换设备 03 均建立有通信连接，任意两个计算节点 02 之间可以通过交换设备 03 建立通信连接。

在本发明实施例中，该多个计算节点 02 可以用于实现分布式机器学习等分布式计算任务，例如该多个计算节点 02 可以实现基于深度神经网络（Deep Neural Networks，DNN）人工智能（Artificial Intelligence，AI）模型训练。在通过该多个计算节点 02 中的若干计算节点实现某个分布式计算任务时，可以在该若干计算节点中的每个计算节点中均部署该分布式计算任务的算法模型，并且可以在该若干计算节点中选取一个指定节点，或者也可以在

其他计算节点中选取一个指定节点。该指定节点中部署有用于驱动该分布式计算任务的驱动程序，该若干计算节点可以在该指定节点的驱动下，并行执行分布式计算任务。

随着计算硬件的快速发展，以张量处理单元（Tensor Processing Unit，TPU）和图形处理单元（Graphics Processing Unit，GPU）为核心的计算节点的计算性能得到了大幅度的提升，这使得各个计算节点执行分布式计算任务时的计算时间大幅缩短，因此对各计算节点与参数服务器之间的通信时间也提出了较高要求，一般需要将该通信时间限制在毫秒级时间周期内。

为了缩短各个计算节点与参数服务器之间的通信时间，本发明实施例提供的数据中心网络中，可以将参数服务器的功能卸载（offload）至交换设备中，即可以由该交换设备对各个计算节点上报的数据进行合并处理后，再反馈至各个计算节点，从而可以有效缩短数据的通信时间，提高分布式计算任务的执行效率。

图 1B 是本发明实施例提供的一种交换设备的架构图，如图 1B 所示，数据中心网络中的每个交换设备 02 可以包括交换功能组件 021、网络计算组件 022 以及网络管理组件 023。其中，交换功能组件 021 用于实现传统交换设备的数据转发功能；该网络计算组件 022 用于对多个计算节点 02 上报的数据进行合并处理；网络管理组件 023 用于感知网络拓扑、存储不同分布式计算任务所对应的路由信息，以及根据该路由信息指导交换功能组件 021 转发网络计算组件 022 合并处理后的数据。

图 1C 是本发明实施例提供的另一种交换设备的架构图，如图 1C 所示，该交换设备 02 的硬件部分主要可以包括交换芯片 02a、中央处理器（Central Processing Unit，CPU）02b 以及存储器 02c。软件部分可以包括至少一个容器 02d、部署在每个容器 02d 内的参数服务器 02e 以及网络管理组件 023。该部署在容器 02d 内的参数服务器 02e 可以是指能够实现参数服务器功能的应用程序。其中，交换芯片 02a 可以为交换设备 02 的交换功能组件，用于实现二层报文或者三层报文的转发。CPU 02b 以及该参数服务器 02e 可以为交换设备 02 的网络计算组件；该 CPU 可以为基于 x86 指令集（或者其他类型的指令集）的 CPU，其具有较高的计算性能，用于提供常规虚拟化容器等软件的处理需求以及用于支撑参数服务器的数据计算功能；该参数服务器 02e 运行在 CPU 02b 上，具备分布式计算所需的数据合并处理（也可以称为汇聚处理）功能。

此外，参考图 1C，该交换芯片 02a 与 CPU 02b 之间可以通过高速互连接口 b1 连接，该高速互连接口 b1 可以为网络适配器（Network Interface Card，NIC）接口，能够满足分布式计算数据传输的高带宽和低时延要求。其中，网络适配器一般也称为网卡。该高速互连接口 b1 的带宽速率可以为交换设备的对外接口 a1 的带宽速率的多倍，例如该高速互连接口 b1 的单向带宽速率可以大于 40Gbps（吉比特每秒）。该高速互连接口 b1 可以有效降低多个计算节点或者交换设备同时向一个交换设备上报数据（也称为多打一）时出现导致网络拥塞的概率。

图 1D 是本发明实施例提供的一种数据中心网络中的控制器的架构图，如图 1D 所示，该控制器 01 可以为基于软件定义网络（Software-defined Network，SDN）架构的控制器，该 SDN 架构可以包括应用层、控制层和转发层。其中，控制器 01 的应用层包括分布式计算加速 SDN 应用（简称加速应用）011，控制层包括分布式计算加速 SDN 控制器（简称加速控制器）012，转发层包括分布式计算加速 SDN 数据通道（简称数据通道）013。

其中，该加速应用 011 主要用于通过网络服务接口（例如 Restful 接口）与指定节点交

互。例如，该加速应用可以接收指定节点发送的处理请求，并可以将控制器确定的路由信息（该路由信息中可以包括用于实现参数服务器功能的交换设备的标识）反馈给该指定节点。此外，该加速应用 011 还可以与加速控制器 012 交互，可以向该加速控制器 012 提供指定计算任务对应的计算节点的标识，以及合并处理类型等信息，并可以接收该加速控制器 012 反馈的路由信息。

该加速控制器 012 可以为控制器 01 中用于实现分布式计算加速的功能主体，该加速控制器 012 中保存了数据中心网络的物理拓扑，并可以根据该物理拓扑，确定用于加速指定分布式计算任务的路由信息。此外，加速控制器 012 还可以统一获取数据中心网络中各个交换设备的性能参数，该性能参数可以包括可用带宽、吞吐量和计算负载等。

该数据通道 013 可以为逻辑上的数据转发通道，构成了控制器与指定节点之间的数据转发路径，以及控制器与各交换设备之间的数据转发路径。

图 2 是本发明实施例提供的一种数据处理方法的流程图，该方法可以应用于图 1A 所示的数据中心网络中，参考图 2，该方法可以包括：

步骤 101、每个交换设备向控制器上报拓扑信息。

其中，每个交换设备上报的拓扑信息可以包括该交换设备的标识，以及该交换设备所连接的设备（例如计算节点或者其他交换设备）的标识，该设备的标识可以为设备的网络地址，例如互联网协议（Internet Protocol，IP）地址。在本发明实施例中，数据中心网络中的每个交换设备均具备拓扑感知功能，每个交换设备可以在数据中心网络的拓扑稳定后，获取其所连接的设备的标识并上报至控制器。图 2 中仅示出了从该数据中心网络所包括的多个交换设备中选取出的目标交换设备和中间交换设备，实际上该数据中心网络中的每个交换设备均可以向控制器上报拓扑信息。控制器获取到各个交换设备的上报的拓扑信息后，即可确定出数据中心网络的整体拓扑结构。

示例的，图 3 是本发明实施例提供的另一种数据中心网络的架构图，如图 3 所示，假设该数据中心网络中包括控制器 01，计算节点 V1 至计算节点 V8，以及交换设备 SW1 至 SW6。其中，交换设备 SW2 所连接的设备包括：计算节点 V1、计算节点 V2、交换设备 SW1 以及交换设备 SW6，则该交换设备 SW2 可以通过其网络管理组件获取到拓扑信息，并通过其交换功能组件向控制器 01 上报拓扑信息。该交换设备 SW2 向控制器 01 上报的拓扑信息中可以包括交换设备 SW2 的 IP 地址、计算节点 V1 的 IP 地址、计算节点 V2 的 IP 地址、交换设备 SW1 的 IP 地址以及交换设备 SW6 的 IP 地址。

控制器 01 根据交换设备 SW1 至交换设备 SW6 上报的拓扑信息，可以确定该数据中心网络的拓扑结构为二层的叶脊（leaf-spine）拓扑结构。其中，交换设备 SW2 至交换设备 SW5 为叶（leaf）交换设备（即第一层交换设备），交换设备 SW1 和交换设备 SW6 为脊（spine）交换设备（即第二层交换设备），并且每个 leaf 交换设备连接有两个计算节点。

步骤 102、指定节点向控制器发送针对指定计算任务的处理请求。

该处理请求可以包括用于执行该指定计算任务的多个计算节点的标识，该指定节点为预先从数据中心网络所包括的计算节点中选定的用于驱动执行指定计算任务的计算节点。该指定节点中可以部署有用于驱动该多个计算节点执行该指定计算任务的分布式驱动程序。一个具体的实施例中，该指定节点以及用于执行该指定计算任务的多个计算节点均为开发人员预先选定的，并且该指定节点可以为该多个计算节点中的一个节点，或者，该指定节点也可以

为单独设置的一个计算节点，本发明实施例对此不做限定。

示例的，假设该指定计算任务为分布式AI训练任务，用于执行该分布式AI训练任务的计算节点包括计算节点V1、计算节点V2和计算节点V7，且指定节点为计算节点V1。参考图3中编号为1的虚线，该计算节点V1可以通过控制器01提供的接口（例如Restful接口）向控制器01发送处理请求，该处理请求中可以包括该分布式AI训练任务的标识，以及计算节点列表，该计算节点列表中记录有计算节点V1的IP地址、计算节点V2的IP地址及计算节点V7的IP地址。

步骤103、控制器根据接收到的拓扑信息，确定多个计算节点之间的拓扑结构。

控制器01接收到指定节点发送的处理请求后，可以根据预先确定的数据中心网络的拓扑结构，确定用于执行该指定计算任务的多个计算节点之间的拓扑结构。该多个计算节点之间的拓扑结构可以包括该多个计算节点，以及用于连接该多个计算节点的交换设备。

一般的，对于leaf-spine拓扑结构的数据中心网络，考虑到spine交换设备所具备的负载均衡的特性，即每个spine交换设备均可以与数据中心网络中的所有leaf交换设备连接，因此可以将数据中心网络中的所有spine交换设备均纳入至该多个计算节点之间的拓扑结构中。

示例的，控制器01可以根据预先确定的数据中心网络的拓扑结构，以及该计算节点V1、计算节点V2和计算节点V7中每个计算节点的IP地址，确定该三个计算节点之间的拓扑结构。该三个计算节点之间的拓扑结构可以如图4所示。从图4可以看出，该拓扑结构包括该三个计算节点，以及用于连接该三个计算节点的所有交换设备。用于连接该三个计算节点的交换设备的集合为{SW2，SW5，SW1，SW6}。

步骤104、控制器基于该拓扑结构，从用于连接该多个计算节点的交换设备中确定至少一个备选交换设备。

其中，每个备选交换设备可以为能够通过下行路径与该多个计算节点中的至少两个计算节点连接的交换设备，且该下行路径上可以包括其他交换设备，也可以不包括其他交换设备。

示例的，如图4所示，假设用于连接计算节点V1、计算节点V2和计算节点V7的交换设备的集合为{SW2，SW5，SW1，SW6}。由于其中spine交换设备SW1可以通过下行路径分别与计算节点V1、计算节点V2和计算节点V7连通，spine交换设备SW6也可以通过下行路径分别与计算节点V1、计算节点V2和计算节点V7连通；leaf交换设备SW2通过下行路径能够与计算节点V1和计算节点V2连通，而leaf交换设备SW5通过下行路径只能与计算节点V7连通，因此控制器01可以将交换设备SW1、交换设备SW2以及交换设备SW6确定为备选交换设备。

由于本发明实施例提供的方法中，每个交换设备均具备对数据进行合并处理的能力，而备选交换设备可以通过下行路径与至少两个计算节点连接，因此能够对其所连接的至少两个计算节点上报的数据进行合并处理后再发出，从该备选交换设备中确定目标交换设备以及中间交换设备，可以保证进行分布式计算时，网络中的数据传输量较小。

步骤105、控制器从该至少一个备选交换设备中确定目标交换设备。

图5是本发明实施例提供的一种确定目标交换设备的方法流程图，参考图5，该确定目标交换设备的过程可以包括：

步骤1051、检测多个计算节点是否均直接连接至同一个备选交换设备。

当该多个计算节点均直接连接至同一个备选交换设备时，控制器可以执行步骤1052；当

该多个计算节点直接连接至不同的备选交换设备时，控制器可以执行步骤 1053。

示例的，参考图 3，假设用于执行某个指定计算任务的计算节点包括计算节点 V3 和计算节点 V4，则由于该两个计算节点均直接连接至备选交换设备 SW3，因此控制器 01 可以执行步骤 1052。

或者，如图 4 所示，假设用于执行某个指定计算任务的计算节点包括计算节点 V1、计算节点 V2 和计算节点 V7，由于计算节点 V1 和计算节点 V2 均直接连接至备选交换设备 SW2，而计算节点 V7 通过交换设备 SW5 连接至备选交换设备 SW1，该三个计算节点直接连接的备选交换设备不为同一个备选交换设备，因此控制器 01 可以执行步骤 1053。

步骤 1052、将该多个计算节点直接连接的备选交换设备确定为目标交换设备。

当该多个计算节点均直接连接至同一个备选交换设备时，控制器可以直接将该多个计算节点直接连接的备选交换设备确定为目标交换设备。参考图 3，如上述计算节点包括计算节点 V3 和计算节点 V4 示例中，控制器 01 可以直接将备选交换设备 SW3 确定为目标交换设备。

在本发明实施例中，该多个计算节点在执行指定计算任务时，可以将计算得到的数据上报至目标交换设备，目标交换设备可以对该各个计算节点上报的数据进行合并处理，并将合并处理后的数据再分别发送至每个计算节点。由于该目标交换设备即可实现参数服务器的功能，因此该各个计算节点无需再通过交换设备向参数服务器上报数据，参数服务器也无需再通过交换设备下发合并处理后的数据，因此有效减少了数据中心网络中的数据传输量，降低了网络拥塞的概率和数据传输的时延，进而可以有效提高指定计算任务的执行算效率。

步骤 1053、计算每个备选交换设备与各个计算节点之间的路由跳数之和。

当该多个计算节点直接连接至不同的备选交换设备时，控制器可以基于该多个计算节点之间的拓扑结构，计算每个备选交换设备与各个计算节点之间的路由跳数之和，并将路由跳数之和最少的备选交换设备确定目标交换设备。

在本发明实施例中，在统计第一备选交换设备与各个计算节点之间的路由跳数之和时，控制器可以将该第一备选交换设备与各个计算节点之间的拓扑结构中，每相邻两个设备之间的路径记为一跳。

示例的，对于图 4 所示的拓扑结构，由于备选交换设备 SW1 与该三个计算节点之间的拓扑结构中，计算节点 V1 与备选交换设备 SW2 之间的路径可以记为一跳，计算节点 V2 与备选交换设备 SW2 之间的路径可以记为一跳，备选交换设备 SW2 与备选交换设备 SW1 之间的路径可以记为一跳，计算节点 V7 与交换设备 SW5 之间的路径记为一跳，交换设备 SW5 与备选交换设备 SW1 之间的路径可以记为一跳，因此控制器 01 可以确定该备选交换设备 SW1 与该三个计算节点之间的路由跳数之和为 5。同样的，控制器 01 可以确定备选交换设备 SW6 与该三个计算节点之间的路由跳数之和也为 5，备选交换设备 SW2 与该三个计算节点之间的路由跳数之和也为 5。

步骤 1054、当路由跳数之和最少的备选交换设备包括一个时，将该路由跳数之和最少的备选交换设备确定为目标交换设备。

若控制器能确定出一个路由跳数之和最少的备选交换设备，则可以直接将该路由跳数之和最少的备选交换设备确定为目标交换设备。由于在本发明实施例提供的方法中，每个交换设备均具备对数据进行合并处理的能力，当某个交换设备接收到多个其他设备（例如计算节点或交换设备）上报的数据后，可以将接收到的数据进行合并处理后再发送至下一跳交换设备，因此在各个计算节点每次上报数据的过程中，每相邻两个设备之间的一跳路径可以仅用

于传输一份数据。

根据上述分析可知，通过路由跳数之和能够直观反映出数据传输过程中，每个备选交换设备与各计算节点之间的数据传输量，选择路由跳数之和最少的备选交换设备作为目标交换设备，使得目标交换设备与各个计算节点之间数据传输量均较少，可以有效降低数据传输时延以及网络拥塞的概率，进而可以有效提高计算任务的执行效率。

步骤 1055、当路由跳数之和最少的备选交换设备包括多个时，分别确定每个路由跳数之和最少的备选交换设备的性能参数。

若路由跳数之和最少的备选交换设备包括多个，控制器则可以根据各个路由跳数之和最少的备选交换设备的性能参数，确定该目标交换设备。其中，每个路由跳数之和最少的备选交换设备的性能参数可以包括可用带宽、计算负载、吞吐量以及被选为目标交换设备的次数中的至少一种。其中，计算负载可以是指交换设备对数据进行合并处理时的负载。在本发明实施例中，控制器可以实时或者周期性的获取数据中心网络中每个交换设备的性能参数，例如，该控制器 01 可以通过其加速控制器 012 周期性的获取每个交换设备的性能参数。

步骤 1056、将多个路由跳数之和最少的备选交换设备中，性能参数满足预设条件的备选交换设备确定为目标交换设备。

具体的实施例中，根据控制器所获取到的性能参数所包括的参数类型的不同，该预设条件也有所不同。例如：当该性能参数包括可用带宽时，该预设条件可以为：交换设备的可用带宽最高；当该性能参数包括吞吐量时，该预设条件可以为：交换设备的吞吐量最低；当该性能参数包括计算负载时，该预设条件可以为：交换设备的计算负载最低；当该性能参数包括被选为目标交换设备的次数时，该预设条件可以为：被选为目标交换设备的次数最少。此外，若该性能参数包括多种类型的参数，控制器可以根据预设的参数优先级，以优先级较高的参数为基准依次进行判断。

例如，假设该预设的参数优先级为：可用带宽、计算负载、吞吐量和被选为目标交换设备的次数，则控制器在确定目标交换设备时，可以先对比各个备选交换设备的可用带宽，并选择可用带宽最高的备选交换设备作为目标交换设备；若可用带宽最高的备选交换设备包括多个，则控制器可以继续对比该多个可用带宽最高的备选交换设备的计算负载，若该多个可用带宽最高的备选交换设备中，计算负载最低的备选交换设备包括多个，则控制器可以继续对比各个备选交换设备的吞吐量，直至确定出满足该预设条件的目标交换设备。此外，若控制器通过上述判断过程，确定出的性能参数满足该预设条件的备选交换设备包括多个，则控制器可以从该性能参数满足预设条件的多个备选交换设备中任意确定一个备选交换设备作为该目标交换设备。

示例的，由于在图 4 所示的拓扑结构中，备选交换设备 SW1、备选交换设备 SW2 和备选交换设备 SW6 对应的路由跳数之和均为 5，则控制器可以对比该三个备选交换设备的性能参数。假设控制器获取到的性能参数为计算负载，且备选交换设备 SW1 的计算负载最低，则控制器可以将该备选交换设备 SW1 确定为目标交换设备。

在本发明实施例中，通过路由跳数之和，以及交换设备的性能参数选取目标交换设备，可以保证选取出的目标交换设备与各个计算节点之间的数据传输量较少，且目标交换设备的性能较好，能够保证较高的计算效率。

在本发明实施例中，控制器除了可以基于性能参数确定目标交换设备，还可以分别确定每个路由跳数之和最少的备选交换设备与各个计算节点之间的路由跳数的均衡程度，并将路

由跳数的均衡程度最高的备选交换设备确定为目标交换设备。当然，控制器也可以在上述步骤 1056 中，确定出多个性能参数均满足该预设条件的备选交换设备后，再基于该路由跳数的均衡程度确定目标交换设备。由于目标交换设备需要获取到用于执行指定计算任务的所有计算节点上报的数据后，才能进行合并处理，因此选取路由跳数的均衡程度最高的交换设备作为目标交换设备，可以保证该各个计算节点上报数据时所需的时长较为接近，使得目标交换设备可以在较短的时间内接收到所有计算节点上报的数据，并进行合并处理，降低了目标交换设备的等待时长，进一步提高了计算任务的执行效率。

其中，路由跳数的均衡程度可以由路由跳数的方差、均方差或平均差等参数确定。且该均衡程度的高低与上述任一参数的参数值大小负相关，即参数值越小表明均衡程度越高。示例的，对于多个路由跳数之和最少的备选交换设备中的每个备选交换设备，控制器可以分别统计该备选交换设备与每个计算节点之间的路由跳数，并计算该备选交换设备与各个计算节点之间的路由跳数的方差，之后可以将方差最小的备选交换设备确定为目标交换设备。

需要说明的是，在本发明实施例中，当控制器在上述步骤 104 中确定出多个备选交换设备时，还可以将该多个备选交换设备中的任意一个确定为目标交换设备。或者，控制器还可以进一步从该多个备选交换设备中，确定出能够通过下行路径与该多个计算节点均连接的候选交换设备，进而再从该候选交换设备中确定目标交换设备。例如，对于图 4 所示的拓扑结构，备选交换设备为 SW1、SW2 和 SW6，由于其中备选交换设备 SW1 和 SW6 能够通过下行路径与该三个计算节点建立连接，因此控制器可以将该备选交换设备 SW1 和 SW6 作为候选交换设备，并从该两个候选交换设备中确定目标交换设备。

步骤 106、控制器将除目标交换设备之外的备选交换设备中，用于连接该目标交换设备与至少两个计算节点的备选交换设备确定为中间交换设备。

在本发明实施例中，若控制器在上述步骤 104 中确定出了多个备选交换设备，则在确定出目标交换设备之后，还可以将剩余的备选交换设备中，用于连接该目标交换设备与该多个计算节点中的至少两个计算节点的备选交换设备确定为中间交换设备。由于在执行指定计算任务的过程中，该中间交换设备可以将其所连接的至少两个计算节点上报的数据进行合并处理后发送至目标交换设备，由此可以进一步减小网络中的数据传输量。

示例的，参考图 4，假设备选交换设备为 SW1、SW2 和 SW6，则当控制器 01 确定目标交换设备为 SW1 之后，由于剩余的两个备选交换设备中，备选交换设备 SW2 能够连接该目标交换设备 SW1 和两个计算节点（V1 和 V2），因此可以将该备选交换设备 SW2 确定为中间交换设备。

或者，参考图 6，假设用于执行指定计算任务的计算节点包括计算节点 V1、V2、V3 和 V7，根据图 6 所示的拓扑结构可知，能够通过下行路径与至少两个计算节点连接的备选交换设备包括：SW21、SW23、SW1 和 SW6。若最终确定的目标交换设备为 SW1，剩余的三个备选交换设备中，由于备选交换设备 SW21 和 SW23，能够连接该目标交换设备 SW1 和两个计算节点（V1 和 V2），因此控制器 01 可以将该备选交换设备 SW21 和 SW23 均确定为中间交换设备。

步骤 107、控制器分别向该目标交换设备、中间交换设备以及该指定节点发送路由信息。

该路由信息可以用于指示该多个计算节点与该目标交换设备之间的数据转发路径。例如，该路由信息中可以包括该多个计算节点的标识以及该目标交换设备的标识。若该多个计算节点与目标交换设备之间的数据转发路径上还包括中间交换设备，则该路由信息中还可以包括该中间交换设备的标识。

在本发明实施例中，为了降低路由信息的数据量，提高路由信息的发送效率，控制器向

每个设备发送的该指定计算任务对应的路由信息中可以仅包括该设备在该数据转发路径中的直连设备的标识。其中，直连设备可以包括该多个计算节点、中间交换设备和目标交换设备，而对于该数据转发路径上未被选取为中间交换设备的其他交换设备，则不在该路由信息的统计范围内。

例如，该控制器向该目标交换设备发送的路由信息可以仅包括该目标交换设备的直连设备的标识，该目标交换设备的直连设备可以为计算节点或者中间交换设备。控制器向该指定节点发送的路由信息可以仅包括每个计算节点直接连接的交换设备的标识，即该控制器向指定节点发送的路由信息可以包括参数服务器列表，该参数服务器列表中记录有用于实现参数服务器功能的交换设备的标识；该指定节点用于将每个计算节点直接连接的中间交换设备或者目标交换设备的标识发送至对应的计算节点。控制器向每个中间交换设备发送的路由信息可以包括该中间交换设备的直连设备的标识，每个中间交换设备的直连设备为计算节点、该目标交换设备或者其他中间交换设备。

示例的，结合图 4，假设用于执行分布式 AI 训练任务的计算节点为计算节点 V1、V2 和 V7，目标交换设备为 SW1，中间交换设备为 SW2。其中，目标交换设备 SW1 的直连设备为中间交换设备 SW2 和计算节点 V7；中间交换设备 SW2 的直连设备为目标交换设备 SW1，以及计算节点 V1 和 V2；计算节点 V1 和 V2 的直连设备均为中间交换设备 SW2，计算节点 V7 的直连设备为目标交换设备 SW1。则控制器 01 向该目标交换设备 SW1 发送的路由信息可以仅包括中间交换设备 SW2 的 IP 地址，以及计算节点 V7 的 IP 地址；控制器 01 向中间交换设备 SW2 发送的路由信息可以包括目标交换设备 SW1 的 IP 地址，计算节点 V1 的 IP 地址，以及计算节点 V2 的 IP 地址。控制器 01 向指定节点 V1 发送的路由信息可以包括中间交换设备 SW2 的 IP 地址，以及目标交换设备 SW1 的 IP 地址。该控制器 01 向各个设备发送路由信息的过程可以如图 3 中编号为 2 的虚线所示。

需要说明的是，在上述步骤 102 中，控制器接收到指定节点发送的针对指定计算任务的处理请求后，还可以为该指定计算任务生成一个任务标识（taskID），例如，控制器为分布式 AI 训练任务生成的 taskID 可以为 1；或者，该控制器也可以直接将该处理请求中携带的该指定计算任务的标识确定为该任务标识。

相应的，控制器向各个设备发送该指定计算任务对应的路由信息时，还可以在该路由信息中携带该任务标识，以便各个设备可以基于该任务标识，存储不同计算任务所对应的路由信息。

示例的，目标交换设备 SW1 存储的路由信息可以如表 1 所示，从表 1 可以看出，taskID 为 1 的计算任务所对应的路由信息中包括 IP1 和 IP2 共两个 IP 地址，其中 IP1 可以为中间交换设备 SW2 的 IP 地址，IP2 为计算节点 V7 的 IP 地址；taskID 为 2 的计算任务所对应的路由信息中则可以包括 IP3 至 IP5 共三个 IP 地址。

表 1

taskID	路由信息	合并处理类型
1	IP1、IP2	计算加权平均值
2	IP3、IP4、IP5	求和

还需要说明的是，在上述步骤 102 中，指定节点向控制器发送的处理请求中还可以包括

该指定计算任务对应的合并处理类型。因此，控制器还可以向该目标交换设备以及每个中间交换设备发送该指定计算任务对应的合并处理类型，以便该目标交换设备和每个中间交换设备可以按照该合并处理类型，对多个计算节点上报的数据进行合并处理。由于不同的计算任务对应的合并处理类型可能不同，按照指定计算任务对应的合并处理类型对接收到的数据进行合并处理，可以保证数据处理的精度。

示例性的，该合并处理类型可以包括：计算平均值、计算加权平均值、求和、计算最大值和计算最小值中的任一种。

此外，控制器可以在向该目标交换设备以及每个中间交换设备发送该指定计算任务对应的路由信息的同时，发送该指定计算任务对应的合并处理类型；或者，控制器也可以单独发送该指定计算任务对应的合并处理类型，本发明实施例对此不做限定。

示例的，假设该 taskID 为 1 的分布式 AI 训练任务对应的合并处理类型为计算加权平均值，则该控制器向目标交换设备 SW1，以及中间交换设备 SW2 发送分布式 AI 训练任务对应的路由信息时，可以在该路由信息中声明该分布式 AI 训练任务对应的合并处理类型为计算加权平均值，以便各个交换设备可以存储该分布式 AI 训练任务对应的合并处理类型。例如，参考表 1，目标交换设备 SW1 可以存储 taskID 为 1 的计算任务对应的合并处理类型为计算加权平均值；taskID 为 2 的计算任务对应的合并处理类型为求和。

步骤 108、指定节点向每个计算节点发送路由信息。

指定节点接收到控制器发送的该指定计算任务对应的路由信息后，可以将该路由信息转发至各个计算节点，以便各个计算节点在完成数据计算后，可以根据接收到的路由信息上报数据。

进一步的，由于控制器向该指定节点发送的路由信息中可以仅包括每个计算节点的直连设备的标识，其中每个计算节点之间连接的设备为用于实现参数服务器功能的中间交换设备或目标交换设备。因此该指定节点向每个计算节点发送的路由信息中，也可以仅包括该计算节点所直接连接的用于实现参数服务器功能的交换设备的标识。

示例的，该指定节点 V1 向计算节点 V2 发送的路由信息中可以仅包括该计算节点 V2 所直接连接的中间交换设备 SW2 的 IP 地址，该指定节点 V1 向计算节点 V7 发送的路由信息中可以仅包括该计算节点 V7 所直接连接的目标交换设备 SW1 的 IP 地址。该指定节点 V1 向各个计算节点发送路由信息的过程可以如图 4 中编号为 3 的虚线所示。

步骤 109、每个计算节点根据该指定计算任务对应的算法模型进行数据计算。

在本发明实施例中，用于执行该指定计算任务的每个计算节点中预先存储有该指定计算任务对应的算法模型，每个计算节点在接收到该指定节点下发的驱动指令后，即可根据该算法模型对获取到的输入数据进行数据计算。

示例的，假设该分布式 AI 训练任务为基于 DNN 的图像识别应用的训练任务。该训练任务可以包括多个相同计算集合的迭代，每次迭代过程中，可以分别向每个计算节点输入多张样本图片，每个计算节点可以根据预先存储的神经网络模型对输入的多张样本图片进行数据计算，得到图像识别应用所使用的神经网络模型的梯度（即误差修正数据）。

步骤 110、每个计算节点向对应的交换设备上报数据。

进一步的，每个计算节点完成数据计算后，即可根据接收到的路由信息，向对应的交换设备上报计算得到的数据。

示例的，参考图 4 中编号为 4 的虚线，计算节点 V1 和计算节点 V2 可以根据接收到的路

由信息，将计算得到的梯度发送至中间交换设备 SW2；参考图 4 中编号为 5 的虚线，计算节点 V7 可以根据接收到的路由信息，将计算得到的梯度直接发送至目标交换设备 SW1。参考图 4 还可以看出，该计算节点 V7 向目标交换设备 SW1 上报梯度时，需要通过交换设备 SW5 进行数据的透传，即该交换设备 SW5 仅转发数据，而不会对数据进行处理。

步骤 111、中间交换设备对其所连接的至少两个计算节点上报的数据进行合并处理。

在本发明实施例中，参考图 1C，每个交换设备中可以配置有用于对数据进行合并处理的参数服务器。每个中间交换设备在接收到控制器发送的路由信息后，可以配置和启动本地参数服务器实例，并可以在接收到其所连接的至少两个计算节点上报的数据后，基于该参数服务器实例，对接收到的数据进行合并处理。

示例的，中间交换设备 SW2 接收到计算节点 V1 和计算节点 V2 上报的梯度后，可以对该两个计算节点上报的梯度进行合并处理。

进一步的，由于控制器还可以向每个中间交换设备发送该指定计算任务对应的合并处理类型；因此相应的，每个中间交换设备在接收到用于执行该指定计算任务的至少两个计算节点上报的数据后，还可以按照该指定计算任务对应的合并处理类型，对该至少两个计算节点上报的数据进行合并处理。

示例的，假设控制器 01 向中间交换设备 SW2 发送路由信息时，还声明了分布式 AI 训练任务对应的合并处理类型为计算加权平均值。则中间交换设备 SW2 接收到计算节点 V1 和计算节点 V2 上报的梯度后，可以计算该两个计算节点上报的梯度的加权平均值。其中，每个计算节点上报梯度时，还可以上报该梯度所对应的权重，因此中间交换设备 SW2 可以根据各个计算节点上报的权重，计算梯度的加权平均值。

步骤 112、中间交换设备向目标交换设备发送合并处理后的数据。

每个中间交换设备对其所连接的至少两个计算节点上报的数据进行合并处理后，即可根据接收到的路由信息，向目标交换设备发送合并处理后的数据。由于该中间交换设备能够对至少两个计算节点上报的数据进行合并处理后再发出，相比于交换设备分别转发两个计算节点上报的数据，本发明实施例提供的方法中，中间交换设备仅需向目标交换设备上报一路数据，从而能够有效减少数据中心网络中的数据传输量，降低网络拥塞的概率。

示例的，参考图 4 中编号为 6 的虚线，中间交换设备 SW2 可以根据接收到的路由信息中，目标交换设备 SW1 的 IP 地址，将其计算得到的加权平均值发送至该目标交换设备 SW1。

步骤 113、目标交换设备对接收到的数据进行合并处理。

在本发明实施例中，参考图 1C，目标交换设备中可以配置有用于对数据进行合并处理的参数服务器。目标交换设备在接收到控制器发送的路由信息后，可以配置和启动本地参数服务器实例，并可以在接收到计算节点和/或中间交换设备上报的数据后，基于该参数服务器实例，对接收到的数据进行合并处理。

示例的，目标交换设备 SW2 接收到计算节点 V7 上报的梯度，以及中间交换设备 SW2 上报的加权平均值之后，可以对该梯度以及加权平均值进行合并处理。

进一步的，由于控制器还可以向目标交换设备发送该指定计算任务对应的合并处理类型；因此相应的，目标交换设备接收到该指定计算任务对应的计算节点和/或中间交换设备上报的数据后，还可以按照该指定计算任务对应的合并处理类型，对接收到的数据进行合并处理。

示例的，假设目标交换设备存储有表 1 所示的对应关系，则目标交换设备 SW1 接收到 IP 地址为 IP2 的计算节点 V7 上报的梯度，以及 IP 地址为 IP1 的中间交换设备 SW2 上报的加权

平均值之后，可以计算该梯度和该加权平均值的加权平均值。其中，计算节点 V7 上报梯度时，还可以上报该梯度所对应的权重，因此目标交换设备 SW1 可以根据该计算节点 V7 上报的权重，计算计算节点 V7 上报的梯度与中间交换设备 SW2 上报的加权平均值的加权平均值。

步骤 114、目标交换设备向每个计算节点发送合并处理后的数据。

最后，目标交换设备即可根据路由信息，将合并处理后的数据分别发送至每个计算节点，以便该各个计算节点根据合并处理后的数据继续执行该指定计算任务。并且，该目标交换设备向第一计算节点发送合并处理后的数据时的数据转发路径，与该第一计算节点上报数据时的数据转发路径可以相同，也可以不同，本发明实施例对此不做限定。该第一计算节点可以为该多个计算节点中的任一节点。

可选的，在本发明实施例中，由于分布式计算任务一般包括多个迭代的计算过程，因此每个计算节点可以在下一个迭代开始前，向该目标交换设备发送获取请求，目标交换设备可以在接收到该获取请求后，将合并处理后的数据发送至各个计算节点。

示例的，参考图 4 中编号为 7 的虚线，目标交换设备 SW1 可以将计算得到的加权平均值分别发送至计算节点 V1、计算节点 V2 和计算节点 V7。例如，该目标交换设备 SW1 可以通过中间交换设备 SW2 将加权平均值分别转发至计算节点 V1 和计算节点 V2，并可以通过交换设备 SW5 将加权平均值转发至计算节点 V7。该计算节点 V1、计算节点 V2 和计算节点 V7 用于根据该加权平均值继续进行图像识别应用的模型训练。

需要说明的是，本发明实施例提供的数据处理方法的步骤的先后顺序可以进行适当调整，步骤也可以根据情况进行相应增减。例如，步骤 104 可以根据情况进行删除，则在上述步骤 105 中，控制器可以直接从用于连接该多个计算节点的交换设备中确定目标交换设备；相应的，在上述步骤后 106 中，控制器可以将该目标交换设备与各个计算节点之间的数据转发路径上，用于连接该目标交换设备和至少两个计算节点的交换设备确定为目标交换设备。或者，上述步骤 106 和步骤 111 也可以根据情况进行删除，即控制器可以仅确定一个目标交换设备，由该目标交换设备对各个计算节点上报的数据进行合并处理。又或者，上述步骤 1051 和步骤 1052 也可以根据情况进行删除，即控制器在确定各个计算节点之间的拓扑结构之后，可以直接根据路由跳数之和确定目标交换设备。又或者，上述步骤 1053 和步骤 1054 也可以根据情况进行删除，即控制器可以直接基于各交换设备的性能参数（或路由跳数的均衡程度）确定目标交换设备。任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到变化的方法，都应涵盖在本发明的保护范围之内，因此不再赘述。

综上所述，本发明实施例提供了一种数据处理方法，指定节点向控制器发送的针对指定计算任务的处理请求中包括用于执行该指定计算任务的多个计算节点的标识，控制器接收到该处理请求后，可以从用于连接该多个计算节点的交换设备中确定目标交换设备，并分别向该目标交换设备以及该指定节点发送用于指示该多个计算节点与该目标交换设备之间的数据转发路径的路由信息，以使得每个计算节点可以根据该路由信息向该目标交换设备上报数据，目标交换设备可以根据该路由信息对该多个计算节点上报的数据进行合并处理后再发送至每个计算节点。由于目标交换设备可以对多个计算节点上报的数据进行合并处理，因此各计算节点无需再通过交换设备向参数服务器发送数据，参数服务器也无需再通过交换设备将合并处理后的结果反馈至各计算节点，有效减小了数据中心网络中的数据传输量，降低了网络拥塞的概率以及数据传输的时延，提高了计算任务的执行算效率。

可选的，本发明实施例提供的方法还可以应用于HPC数据中心网络中，该HPC数据中心网络可以使用消息传输接口（MPI，Message Passing Interface）作为分布式信息交互的编程接口，并且可以采用可扩展分层聚合协议（Scalable Hierarchical Aggregation and Reduction Protocol，SHARP）技术将MPI的集合操作中的部分操作（例如Reduction操作和Aggregation操作）卸载（offload）到交换设备上，即由交换设备执行该部分操作。也即是，该数据中心网络即可以支持SHARP技术，也可以支持本申请所提供的数据处理方法，具体的，数据中心网络中的各个交换设备可以在管理服务器的控制下实现该SHARP技术，并可以在控制器的控制下实现本发明实施例提供的数据处理方法。由于SHARP技术受限于MPI集合操作，需要计算节点采用特定的MPI函数库，其应用灵活性较低；又由于MPI集合操作中并未设定根汇聚节点，导致管理服务器选择用于执行该部分操作的交换设备时的计算复杂度较高，因此仅采用该SHARP技术难以支持较大规模的数据中心网络。而该HPC数据中心网络在采用本发明实施例提供的数据处理方法后，由于交换设备可以实现参数服务器的功能，因此各个交换设备不再受限于MPI集合操作，有效提高了数据处理的灵活性；并且，由于控制器仅需选取用于实现参数服务器功能的目标交换设备和中间交换设备，因此其选取过程的计算复杂度较低，能够支持较大规模的数据中心网络。

图7是本发明实施例提供的一种数据处理装置的结构示意图，该装置可以应用于图1A所示的数据中心网络的控制器01中，参考图7，该装置可以包括：

接收模块201，用于接收指定节点发送的针对指定计算任务的处理请求，该处理请求包括用于执行该指定计算任务的多个计算节点的标识。接收模块201的具体实现可以参考上述图2所示实施例中的步骤102中的详细描述，此处不再予以赘述。

确定模块202，用于从用于连接该多个计算节点的交换设备中确定目标交换设备。确定模块202的具体实现可以参考上述图2所示实施例中的步骤104和步骤105中的详细描述，此处不再予以赘述。

发送模块203，用于分别向该目标交换设备以及该指定节点发送该指定计算任务对应的路由信息，该路由信息用于指示该多个计算节点与该目标交换设备之间的数据转发路径。发送模块203的具体实现可以参考上述图2所示实施例中的步骤107中的详细描述，此处不再予以赘述。

其中，该路由信息用于在目标交换设备对该多个计算节点上报的数据进行合并处理后根据该路由信息将合并处理后的数据发送至每个计算节点。即该目标交换设备可以根据该路由信息对该多个计算节点上报的数据进行合并处理后发送至每个计算节点。此外，该指定节点可以在接收到该路由信息后，将该路由信息发送至每个计算节点，每个计算节点用于根据该路由信息向该目标交换设备上报数据。

此外，该接收模块201和发送模块203的功能可以与图1D所示架构中，加速应用011和数据通道013的功能类似；确定模块202的功能可以与图1D所示架构中加速控制器012的功能类似。

可选的，该多个计算节点与该目标交换设备之间的数据转发路径上可以包括至少一个交换设备；

相应的，该确定模块202，还可以用于：从该至少一个交换设备中确定至少一个中间交换设备，每个中间交换设备与至少两个计算节点连接。具体实现可以参考上述图2所示实施

例中的步骤 106 中的详细描述，此处不再予以赘述。

该发送模块 203，还可以用于向每个中间交换设备发送该路由信息，每个中间交换设备用于根据该路由信息，将该中间交换设备连接的至少两个计算节点上报的数据进行合并处理后发送至该目标交换设备。具体实现可以参考上述图 2 所示实施例中的步骤 107 中的详细描述，此处不再予以赘述。

可选的，该路由信息可以包括：每个计算节点的标识、目标交换设备的标识，以及中间设备的标识；图 8 是本发明实施例提供的一种发送模块的结构示意图，参考图 8，该发送模块 203 可以包括：

第一发送子模块 2031，用于向该目标交换设备发送该目标交换设备的直连设备的标识，该目标交换设备的直连设备为计算节点或者中间交换设备。

第二发送子模块 2032，用于向该指定节点发送每个计算节点的直连设备的标识，每个计算节点的直连设备为目标交换设备或者中间交换设备，该指定节点用于将每个计算节点的直连设备的标识发送至对应的计算节点。

第三发送子模块 2033，用于向中间交换设备发送该中间交换设备的直连设备的标识，中间交换设备的直连设备为计算节点、该目标交换设备或者其他中间交换设备。

以上各发送子模块的具体实现可以参考上述图 2 所示实施例中的步骤 107 中的详细描述，此处不再予以赘述。

图 9 是本发明实施例提供的一种确定模块的结构示意图，参考图 9，该确定模块 202 可以包括：

计算子模块 2021，用于计算用于连接该多个计算节点的交换设备中，每个交换设备与各个计算节点之间的路由跳数之和。该计算子模块 2021 的具体实现可以参考上述图 5 所示实施例中的步骤 1053 中的详细描述，此处不再予以赘述。

第一确定子模块 2022，用于将路由跳数之和最少的交换设备确定为目标交换设备。

该第一确定子模块 2022 的具体实现可以参考上述图 5 所示实施例中的步骤 1054 至步骤 1056 中的详细描述，此处不再予以赘述。

可选的，如图 9 所示，该确定模块 202 还可以包括：

检测子模块 2023，用于实现上述图 5 所示实施例中步骤 1051 所示的方法。

第二确定子模块 2024，用于实现上述图 5 所示实施例中步骤 1052 所示的方法。

相应的，计算子模块 2021 可以用于实现上述图 5 所示实施例中步骤 1053 所示的方法。

可选的，该确定模块 202 还可以用于实现上述图 2 所示实施例中步骤 103 至步骤 105 所示的方法。

可选的，该处理请求还可以包括：该指定计算任务对应的合并处理类型；

相应的，该发送模块 203，还可以用于向该目标交换设备发送该指定计算任务对应的合并处理类型，该目标交换设备用于按照该合并处理类型对该多个计算节点上报的数据进行合并处理。其具体实现可以参考上述图 2 所示实施例中的步骤 107 中的详细描述，此处不再予以赘述。

综上所述，本发明实施例提供了一种数据处理装置，该装置接收到的针对指定计算任务的处理请求中包括用于执行该指定计算任务的多个计算节点的标识，该装置可以从用于连接该多个计算节点的交换设备中确定目标交换设备，并分别向该目标交换设备以及该指定节点发送用于指示该多个计算节点与该目标交换设备之间的数据转发路径的路由信息，以使得每

个计算节点可以根据该路由信息向该目标交换设备上报数据，目标交换设备可以根据该路由信息对该多个计算节点上报的数据进行合并处理后再发送至每个计算节点。由于目标交换设备可以对多个计算节点上报的数据进行合并处理，因此各计算节点无需再通过交换设备向参数服务器发送数据，参数服务器也无需再通过交换设备将合并处理后的结果反馈至各计算节点，有效减小了数据中心网络中的数据传输量，降低了网络拥塞的概率以及数据传输的时延，提高了计算任务的执行算效率。

图 10 是本发明实施例提供的另一种数据处理装置的结构示意图，该装置可以应用于图 1A 所示的数据中心网络的交换设备 03 中，参考图 10，该装置可以包括：

接收模块 301，用于接收控制器发送的指定计算任务对应的路由信息，该路由信息用于指示多个计算节点与目标交换设备之间的数据转发路径，该多个计算节点用于执行该指定计算任务。接收模块 301 的具体实现可以参考上述图 2 所示实施例中的步骤 107 中的详细描述，此处不再予以赘述。

处理模块 302，用于对该多个计算节点上报的数据进行合并处理。处理模块 302 的具体实现可以参考上述图 2 所示实施例中的步骤 111 或者步骤 113 中的详细描述，此处不再予以赘述。

发送模块 303，用于根据该路由信息，发送合并处理后的数据。发送模块 303 的具体实现可以参考上述图 2 所示实施例中的步骤 112 或者步骤 114 中的详细描述，此处不再予以赘述。

其中，该路由信息为该控制器接收到指定节点发送的针对该指定计算任务的处理请求后，从用于连接该多个计算节点的交换设备中确定目标交换设备后发送的。

此外，该接收模块 301 和发送模块 303 的功能可以与图 1B 所示架构中交换功能组件 021 的功能类似；该处理模块 302 的功能可以与图 1B 所示架构中网络计算组件 022 的功能类似。

可选的，该接收模块 301，还用于在对该多个计算节点上报的数据进行合并处理之前，接收该控制器发送的该指定计算任务对应的合并处理类型。其具体实现可以参考上述图 2 所示实施例中步骤 107 中的详细描述，此处不再予以赘述。

相应的，该处理模块 302，可以用于：按照该合并处理类型，对该多个计算节点上报的数据进行合并处理。其具体实现可以参考上述图 2 所示实施例中步骤 111 或者步骤 113 中的详细描述，此处不再予以赘述。

可选的，当该交换设备可以为目标交换设备时，该发送模块 303 可以用于实现上述图 2 所示实施例中步骤 114 所示的方法。

可选的，当该交换设备为用于连接该目标交换设备和至少两个该计算节点的中间交换设备时，该处理模块 302，可以用于实现上述图 2 所示实施例中步骤 111 所示的方法。

该发送模块 303，可以用于实现上述图 2 所示实施例中步骤 112 所示的方法。

此外，在本发明实施例中，该数据处理装置还可以包括拓扑感知模块，该拓扑感知模块用于在数据中心网络的拓扑稳定后，获取交换设备所连接的其他设备的标识并上报至控制器。该拓扑感知模块的作用可以与图 1B 所示架构中网络管理组件 023 的功能类似。

综上所述，本发明实施例提供了一种数据处理装置，该装置可以根据控制器发送的指定计算任务所对应的路由信息，对用于执行该指定计算任务的多个计算节点上报的数据进行合并处理后再发送至每个计算节点。因此各计算节点无需再通过交换设备向参数服务器发送数

据，参数服务器也无需再通过交换设备将合并处理后的结果反馈至各计算节点，有效减小了数据中心网络中的数据传输量，降低了网络拥塞的概率以及数据传输的时延，提高了计算任务的执行算效率。

关于上述实施例中的装置，其中各个模块执行操作的实现方式已经在有关该方法的实施例中进行了详细描述，故此处不再阐述说明。

请参考图 11，其示出了本申请实施例提供的一种数据处理装置 600 的结构示意图，该数据处理装置可以配置于图 1A 所示的控制器 01 中，参见图 11，该数据处理装置 600 可以包括：处理器 610、通信接口 620 和存储器 630，通信接口 620 和存储器 630 分别与处理器 610 相连，示例地，如图 11 所示，通信接口 620 和存储器 630 通过总线 640 与处理器 610 相连。

其中，处理器 610 可以为中央处理器（CPU），处理器 610 包括一个或者一个以上处理核心。处理器 610 通过运行计算机程序，从而执行各种功能应用以及数据处理。该处理器 610 的具体实现可以参考上述图 2 所示实施例中步骤 103 至步骤 106 中的详细描述，以及图 5 所示实施例中的详细描述，此处不再予以赘述。

其中，通信接口 620 可以为多个，该通信接口 620 用于数据处理装置 600 与外部设备进行通信，该外部设备例如显示器、第三方设备（例如，存储设备、移动终端和交换设备等）等。该通信接口 620 的具体实现可以参考上述图 2 所示实施例中步骤 101、步骤 102 以及步骤 107 中的详细描述，此处不再予以赘述。

其中，存储器 630 可以包括但不限于：随机存取存储器（RAM）、只读存储器（ROM）、可擦除可编程只读存储器（EPROM）、快闪存储器、光存储器。该存储器 630 负责信息存储，例如，该存储器 630 用于存储供处理器 610 运行的计算机程序。

可选地，该数据处理装置 600 还可以包括：输入/输出（I/O）接口（图 11 中未示出）。I/O 接口与处理器 610、通信接口 620 以及存储器 630 连接。I/O 接口例如可以为通用串行总线（USB）。

本发明实施例还提供了一种交换设备，如图 1C 所示，该交换设备 02 可以包括交换芯片 02a、CPU 02b 以及存储器 02c。其中存储器 02c 中可以存储有计算机程序，该 CPU 02b 可以通过执行该计算机程序，实现上述图 2 所示实施例中步骤 111 或者步骤 113 所示的方法，其具体实现过程此处不再予以赘述。该交换芯片 02a 可以用于实现上述图 2 所示实施例中步骤 101、步骤 112 和步骤 114 所示的方法，其具体实现过程此处不再予以赘述。

本发明实施例还提供了一种数据处理系统，参考图 1A，该系统可以包括：控制器 01、多个计算节点 02 以及至少一个交换设备 03。

该控制器 01 可以包括如图 7 或图 11 所示的数据处理装置，该数据处理装置可以包括图 8 所示的发送模块和图 9 所示的确定模块；或者，该控制器 01 可以为图 1D 所示的控制器。每个交换设备可以包括如图 10 所示的数据处理装置，或者，每个交换设备可以为如图 1B 或图 1C 所示的交换设备。

本发明实施例提供了一种计算机可读存储介质，该计算机可读存储介质中存储有指令，当该计算机可读存储介质在计算机上运行时，使得计算机执行上述方法实施例提供的数据处理方法。

本发明实施例还提供了一种包含指令的计算机程序产品，当该计算机程序产品在计算机上运行时，使得计算机执行上述方法实施例所提供的数据处理方法。

权 利 要 求

1、一种数据处理方法，其特征在于，应用于数据中心网络的控制器，所述方法包括：

接收指定节点发送的针对指定计算任务的处理请求，所述处理请求包括用于执行所述指定计算任务的多个计算节点的标识，所述多个计算节点包括所述指定节点；

从用于连接所述多个计算节点的交换设备中确定目标交换设备；

分别向所述目标交换设备以及所述指定节点发送所述指定计算任务对应的路由信息，所述路由信息用于指示所述多个计算节点与所述目标交换设备之间的数据转发路径；

其中，所述路由信息用于在所述目标交换设备对所述多个计算节点上报的数据进行合并处理后根据所述路由信息将所述合并处理后的数据发送至每个所述计算节点。

2、根据权利要求 1 所述的方法，其特征在于，所述多个计算节点与所述目标交换设备之间的数据转发路径上包括至少一个交换设备，所述方法还包括：

将所述数据转发路径上包括的至少一个交换设备中，与所述多个计算节点中的至少两个计算节点连接的交换设备确定为中间交换设备；

向所述中间交换设备发送所述路由信息，所述路由信息用于所述中间交换设备将与所述中间交换设备连接的至少两个计算节点上报的数据进行合并处理后根据所述路由信息将所述合并处理后的数据发送至所述目标交换设备。

3、根据权利要求 2 所述的方法，其特征在于，所述分别向所述目标交换设备以及所述指定节点发送所述指定计算任务对应的路由信息，包括：

向所述目标交换设备发送包括所述目标交换设备的直连设备的标识的路由信息，所述目标交换设备的直连设备为计算节点或者中间交换设备；

向所述指定节点发送包括每个所述计算节点的直连设备的标识的路由信息，每个所述计算节点的直连设备为目标交换设备或者中间交换设备，所述指定节点用于将每个所述计算节点的直连设备的标识发送至对应的计算节点；

所述向所述中间交换设备发送所述路由信息，包括：

向所述中间交换设备发送包括所述中间交换设备的直连设备的标识的路由信息，所述中间交换设备的直连设备为计算节点、所述目标交换设备或其他中间交换设备。

4、根据权利要求 1 至 3 任一所述的方法，其特征在于，所述从用于连接所述多个计算节点的交换设备中确定目标交换设备，包括：

分别计算用于连接所述多个计算节点的交换设备中，每个交换设备与各个所述计算节点之间的路由跳数之和；

将路由跳数之和最少的交换设备确定为目标交换设备。

5、根据权利要求 4 所述的方法，其特征在于，所述将路由跳数之和最少的交换设备确定为目标交换设备，包括：

当路由跳数之和最少的交换设备包括多个时，分别确定每个所述路由跳数之和最少的交换设备的性能参数，所述性能参数包括可用带宽、计算负载、吞吐量以及被选为目标交换设

备的次数中的至少一种；

将多个所述路由跳数之和最少的交换设备中，性能参数满足预设条件的交换设备确定为目标交换设备。

6、根据权利要求 4 所述的方法，其特征在于，所述将路由跳数之和最少的交换设备确定为目标交换设备，包括：

当路由跳数之和最少的交换设备包括多个时，分别确定每个所述路由跳数之和最少的交换设备与各个所述计算节点之间的路由跳数的均衡程度；

将多个所述路由跳数之和最少的交换设备中，路由跳数的均衡程度最高的交换设备确定为目标交换设备。

7、根据权利要求 4 所述的方法，其特征在于，在计算每个交换设备与各个所述计算节点之间的路由跳数之和之前，所述方法还包括：

检测所述多个计算节点是否均直接连接至同一个交换设备；

当所述多个计算节点均直接连接至同一个交换设备时，将所述多个计算节点直接连接的交换设备确定为所述目标交换设备；

所述计算用于连接所述多个计算节点的交换设备中，每个交换设备与各个所述计算节点之间的路由跳数之和，包括：

当所述多个计算节点直接连接至不同的交换设备时，计算用于连接所述多个计算节点的交换设备中，每个交换设备与各个所述计算节点之间的路由跳数之和。

8、根据权利要求 1 至 3 任一所述的方法，其特征在于，所述从用于连接所述多个计算节点的交换设备中确定目标交换设备，包括：

从用于连接所述多个计算节点的交换设备中确定至少一个备选交换设备，每个所述备选交换设备通过下行路径与所述多个计算节点中的至少两个计算节点连接；

从所述至少一个备选交换设备中确定所述目标交换设备。

9、根据权利要求 1 至 3 任一所述的方法，其特征在于，所述处理请求还包括：所述指定计算任务对应的合并处理类型；所述方法还包括：

向所述目标交换设备发送所述指定计算任务对应的合并处理类型，所述目标交换设备用于按照所述合并处理类型对所述多个计算节点上报的数据进行合并处理。

10、一种数据处理方法，其特征在于，应用于数据中心网络的交换设备，所述方法包括：

接收控制器发送的指定计算任务对应的路由信息，所述路由信息用于指示多个计算节点与目标交换设备之间的数据转发路径，所述多个计算节点用于执行所述指定计算任务；

对所述多个计算节点上报的数据进行合并处理；

根据所述路由信息，发送合并处理后的数据；

其中，所述路由信息为所述控制器接收到指定节点发送的针对所述指定计算任务的处理请求后，从用于连接所述多个计算节点的交换设备中确定目标交换设备后发送的。

11、根据权利要求 10 所述的方法，其特征在于，在对所述多个计算节点上报的数据进行合并处理之前，所述方法还包括：

接收所述控制器发送的所述指定计算任务对应的合并处理类型；

所述对所述多个计算节点上报的数据进行合并处理，包括：

按照所述合并处理类型，对所述多个计算节点上报的数据进行合并处理。

12、根据权利要求 10 或 11 所述的方法，其特征在于，所述交换设备为所述目标交换设备；所述根据所述路由信息，发送合并处理后的数据，包括：

根据所述路由信息，向每个所述计算节点发送合并处理后的数据。

13、根据权利要求 10 或 11 所述的方法，其特征在于，所述交换设备为用于连接所述目标交换设备和至少两个所述计算节点的中间交换设备；

所述对所述多个计算节点上报的数据进行合并处理，包括：

对至少两个所述计算节点上报的数据进行合并处理；

所述根据所述路由信息，发送合并处理后的数据，包括：

根据所述路由信息，向所述目标交换设备发送合并处理后的数据。

14、一种数据处理装置，其特征在于，应用于数据中心网络的控制器，所述装置包括：

接收模块，用于接收指定节点发送的针对指定计算任务的处理请求，所述处理请求包括用于执行所述指定计算任务的多个计算节点的标识，所述多个计算节点包括所述指定节点；

确定模块，用于从用于连接所述多个计算节点的交换设备中确定目标交换设备；

发送模块，用于分别向所述目标交换设备以及所述指定节点发送所述指定计算任务对应的路由信息，所述路由信息用于指示所述多个计算节点与所述目标交换设备之间的数据转发路径；

其中，所述路由信息用于在所述目标交换设备对所述多个计算节点上报的数据进行合并处理后根据所述路由信息将所述合并处理后的数据发送至每个所述计算节点。

15、根据权利要求 14 所述的装置，其特征在于，所述多个计算节点与所述目标交换设备之间的数据转发路径上包括至少一个交换设备；

所述确定模块，还用于将所述数据转发路径上包括的至少一个交换设备中，与所述多个计算节点中的至少两个计算节点连接的交换设备确定为中间交换设备；

所述发送模块，还用于向所述中间交换设备发送所述路由信息，所述路由信息用于将与所述中间交换设备连接的至少两个计算节点上报的数据进行合并处理后根据所述路由信息将所述合并处理后的数据发送至所述目标交换设备。

16、根据权利要求 15 所述的装置，其特征在于，所述发送模块，包括：

第一发送子模块，用于向所述目标交换设备发送包括所述目标交换设备的直连设备的标识的路由信息，所述目标交换设备的直连设备为计算节点或者中间交换设备；

第二发送子模块，用于向所述指定节点发送包括每个所述计算节点的直连设备的标识的路由信息，每个所述计算节点的直连设备为目标交换设备或者中间交换设备，所述指定节点

用于将每个所述计算节点的直连设备的标识发送至对应的计算节点；

第三发送子模块，用于向所述中间交换设备发送包括所述中间交换设备的直连设备的标识的路由信息，所述中间交换设备的直连设备为计算节点、所述目标交换设备或其他中间交换设备。

17、根据权利要求 14 至 16 任一所述的装置，其特征在于，所述确定模块，包括：

计算子模块，用于计算用于连接所述多个计算节点的交换设备中，每个交换设备与各个所述计算节点之间的路由跳数之和；

第一确定子模块，用于将路由跳数之和最少的交换设备确定为目标交换设备。

18、根据权利要求 17 所述的装置，其特征在于，所述第一确定子模块，用于：

当路由跳数之和最少的交换设备包括多个时，分别确定每个所述路由跳数之和最少的交换设备的性能参数，所述性能参数包括可用带宽、计算负载、吞吐量以及被选为目标交换设备的次数中的至少一种；

将多个所述路由跳数之和最少的交换设备中，性能参数满足预设条件的交换设备确定为目标交换设备。

19、根据权利要求 17 所述的装置，其特征在于，所述第一确定子模块，用于：

当路由跳数之和最少的交换设备包括多个时，分别确定每个所述路由跳数之和最少的交换设备与各个计算节点所述之间的路由跳数的均衡程度；

将多个所述路由跳数之和最少的交换设备中，路由跳数的均衡程度最高的交换设备确定为目标交换设备。

20、根据权利要求 17 所述的装置，其特征在于，所述确定模块，还包括：

检测子模块，用于检测所述多个计算节点是否均直接连接至同一个交换设备；

第二确定子模块，用于当所述多个计算节点均直接连接至同一个交换设备时，将所述多个计算节点直接连接的交换设备确定为所述目标交换设备；

所述计算子模块，用于：当所述多个计算节点直接连接至不同的交换设备时，计算用于连接所述多个计算节点的交换设备中，每个交换设备与各个所述计算节点之间的路由跳数之和。

21、根据权利要求 14 至 16 任一所述的装置，其特征在于，所述确定模块，用于：

从用于连接所述多个计算节点的交换设备中确定至少一个备选交换设备，每个所述备选交换设备通过下行路径与至少两个计算节点连接；

从所述至少一个备选交换设备中确定所述目标交换设备。

22、根据权利要求 14 至 16 任一所述的装置，其特征在于，所述处理请求还包括：所述指定计算任务对应的合并处理类型；

所述发送模块，还用于向所述目标交换设备发送所述指定计算任务对应的合并处理类型，所述目标交换设备用于按照所述合并处理类型对所述多个计算节点上报的数据进行合并

处理。

23、一种数据处理装置，其特征在于，应用于数据中心网络的交换设备，所述装置包括：

接收模块，用于接收控制器发送的指定计算任务对应的路由信息，所述路由信息用于指示多个计算节点与目标交换设备之间的数据转发路径，所述多个计算节点用于执行所述指定计算任务；

处理模块，用于对所述多个计算节点上报的数据进行合并处理；

发送模块，用于根据所述路由信息，发送合并处理后的数据；

其中，所述路由信息为所述控制器接收到指定节点发送的针对所述指定计算任务的处理请求后，从用于连接所述多个计算节点的交换设备中确定目标交换设备后发送的。

24、根据权利要求 23 所述的装置，其特征在于，

所述接收模块，还用于在对所述多个计算节点上报的数据进行合并处理之前，接收所述控制器发送的所述指定计算任务对应的合并处理类型；

所述处理模块，用于：

按照所述合并处理类型，对所述多个计算节点上报的数据进行合并处理。

25、根据权利要求 23 或 24 所述的装置，其特征在于，所述交换设备为所述目标交换设备；所述发送模块，用于：

根据所述路由信息，向每个所述计算节点发送合并处理后的数据。

26、根据权利要求 23 或 24 所述的装置，其特征在于，所述交换设备为用于连接所述目标交换设备和至少两个所述计算节点的中间交换设备；

所述处理模块，用于：对至少两个所述计算节点上报的数据进行合并处理；

所述发送模块，用于：根据所述路由信息，向所述目标交换设备发送合并处理后的数据。

27、一种数据处理系统，其特征在于，所述系统包括：控制器、多个计算节点以及至少一个交换设备；

所述控制器包括如权利要求 14 至 22 任一所述的数据处理装置；

每个所述交换设备包括如权利要求 23 至 26 任一所述的数据处理装置。

28、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储有指令，当所述计算机可读存储介质在计算机上运行时，使得计算机执行权利要求 1 至 13 任一所述的数据处理方法。

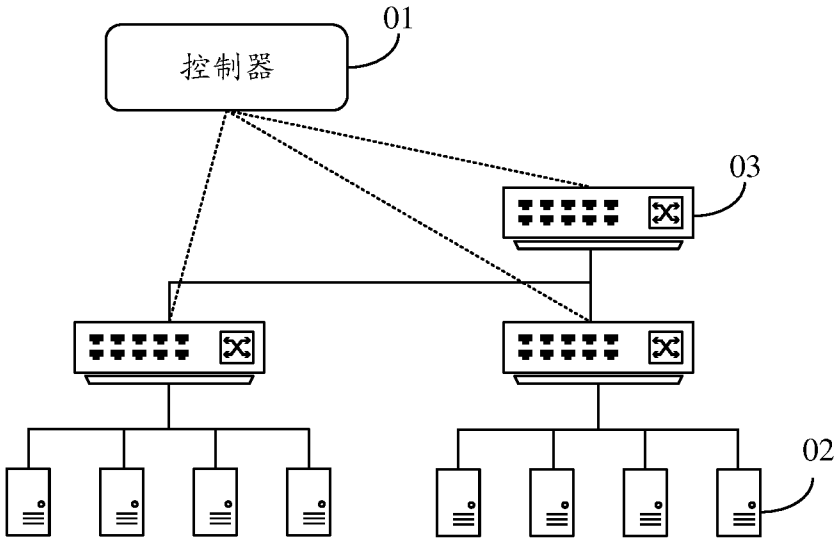


图 1A

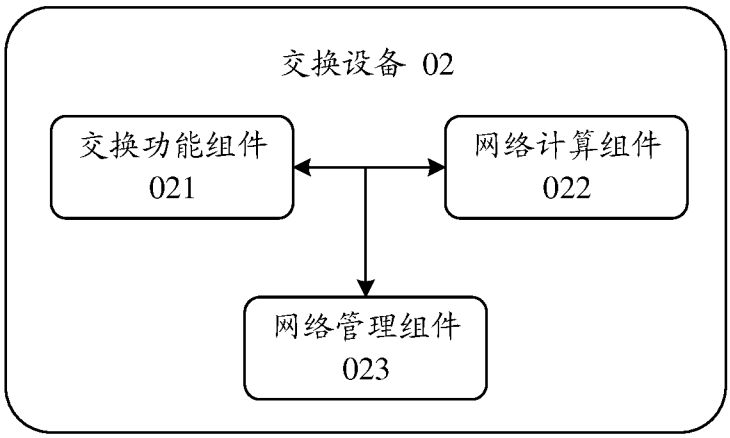


图 1B

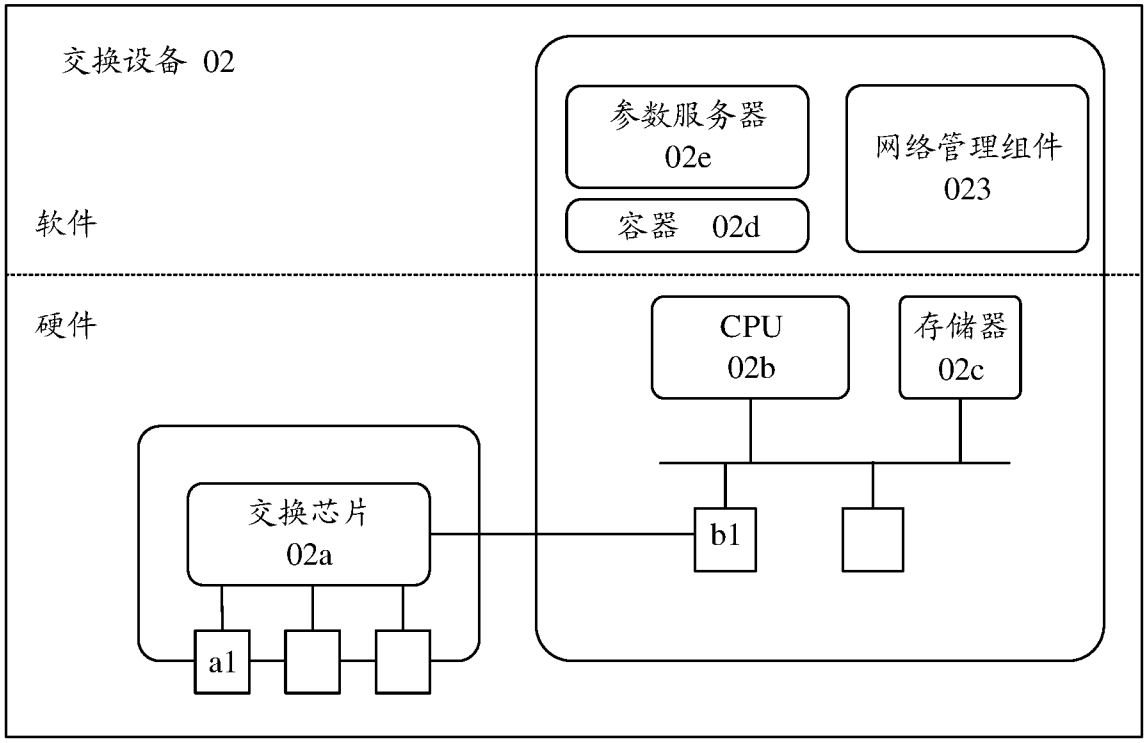

交换设备 02
参数服务器
02e
网络管理组件
023
容器 02d
软件
硬件
CPU
02b
存储器
02c
交换芯片
02a
b1
a1

图 1C

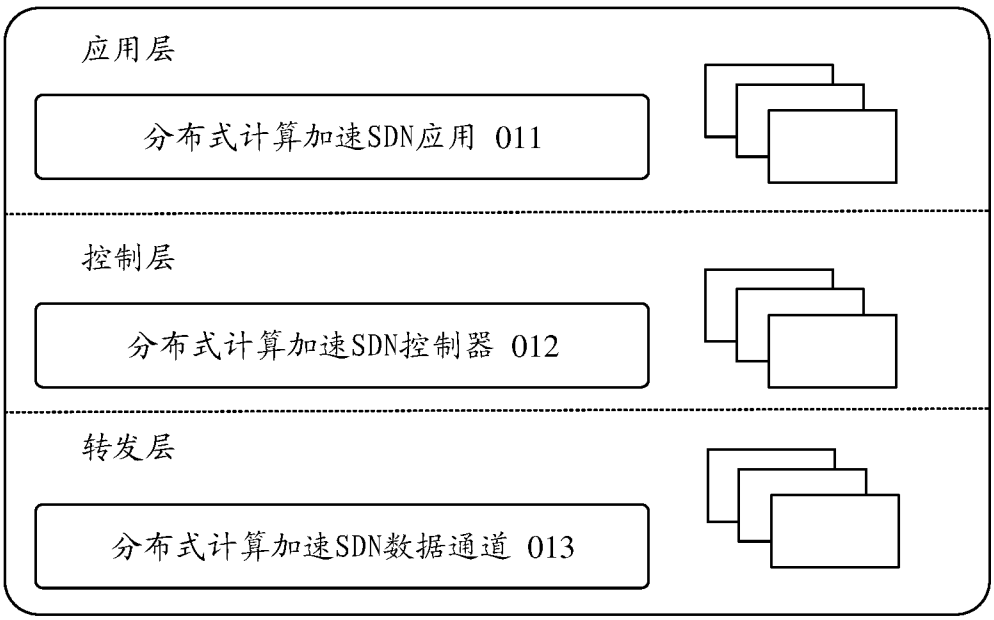

应用层
分布式计算加速SDN应用 011
控制层
分布式计算加速SDN控制器 012
转发层
分布式计算加速SDN数据通道 013

图 1D

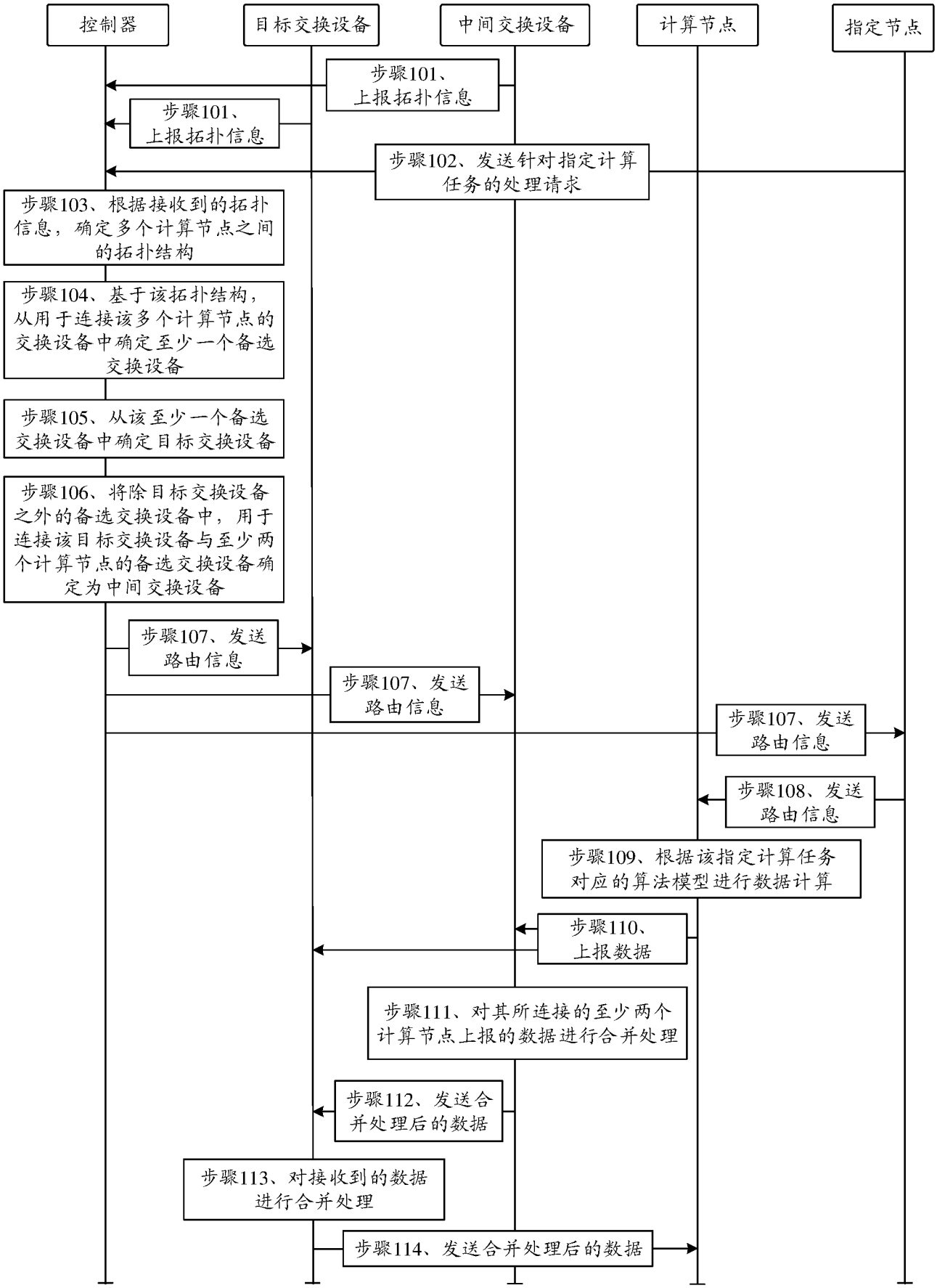

控制器
目标交换设备
中间交换设备
计算节点
指定节点
步骤101、上报拓扑信息
步骤101、上报拓扑信息
步骤102、发送针对指定计算任务的处理请求
步骤103、根据接收到的拓扑信息，确定多个计算节点之间的拓扑结构
步骤104、基于该拓扑结构，从用于连接该多个计算节点的交换设备中确定至少一个备选交换设备
步骤105、从该至少一个备选交换设备中确定目标交换设备
步骤106、将除目标交换设备之外的备选交换设备中，用于连接该目标交换设备与至少两个计算节点的备选交换设备确定为中间交换设备
步骤107、发送路由信息
步骤107、发送路由信息
步骤107、发送路由信息
步骤108、发送路由信息
步骤109、根据该指定计算任务对应的算法模型进行数据计算
步骤110、上报数据
步骤111、对其所连接的至少两个计算节点上报的数据进行合并处理
步骤112、发送合并处理后的数据
步骤113、对接收到的数据进行合并处理
步骤114、发送合并处理后的数据

图 2

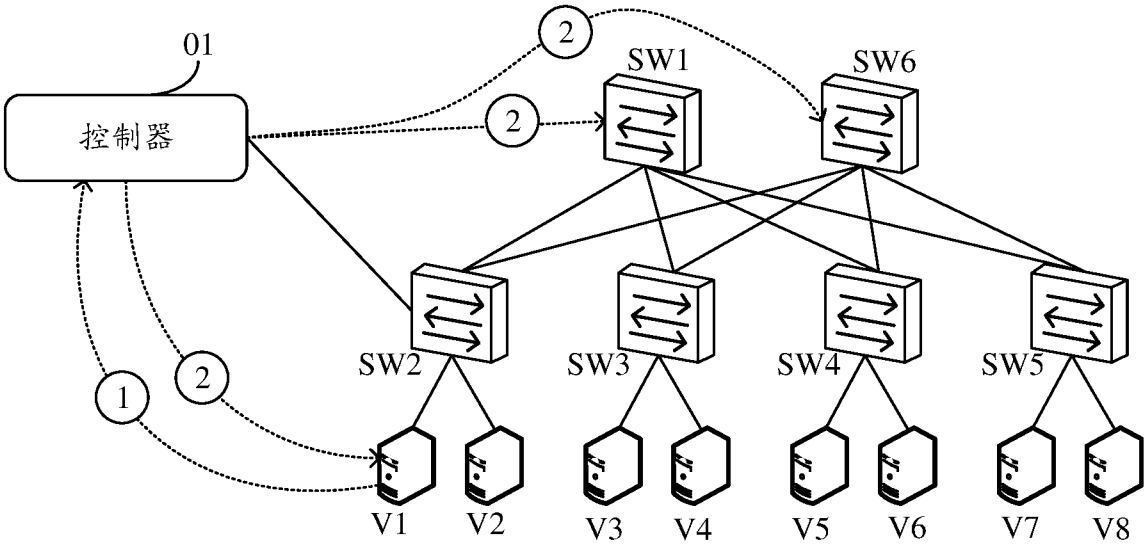

01
控制器
2
2
SW1
SW6
SW2
SW3
SW4
SW5
1
2
V1
V2
V3
V4
V5
V6
V7
V8

图 3

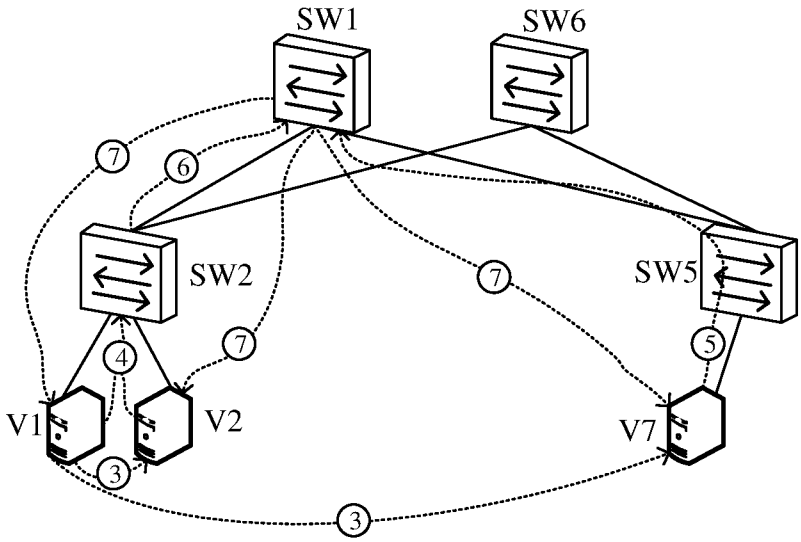

SW1
SW6
7
6
SW2
7
SW5
4
7
5
V1
V2
V7
3
3

图 4

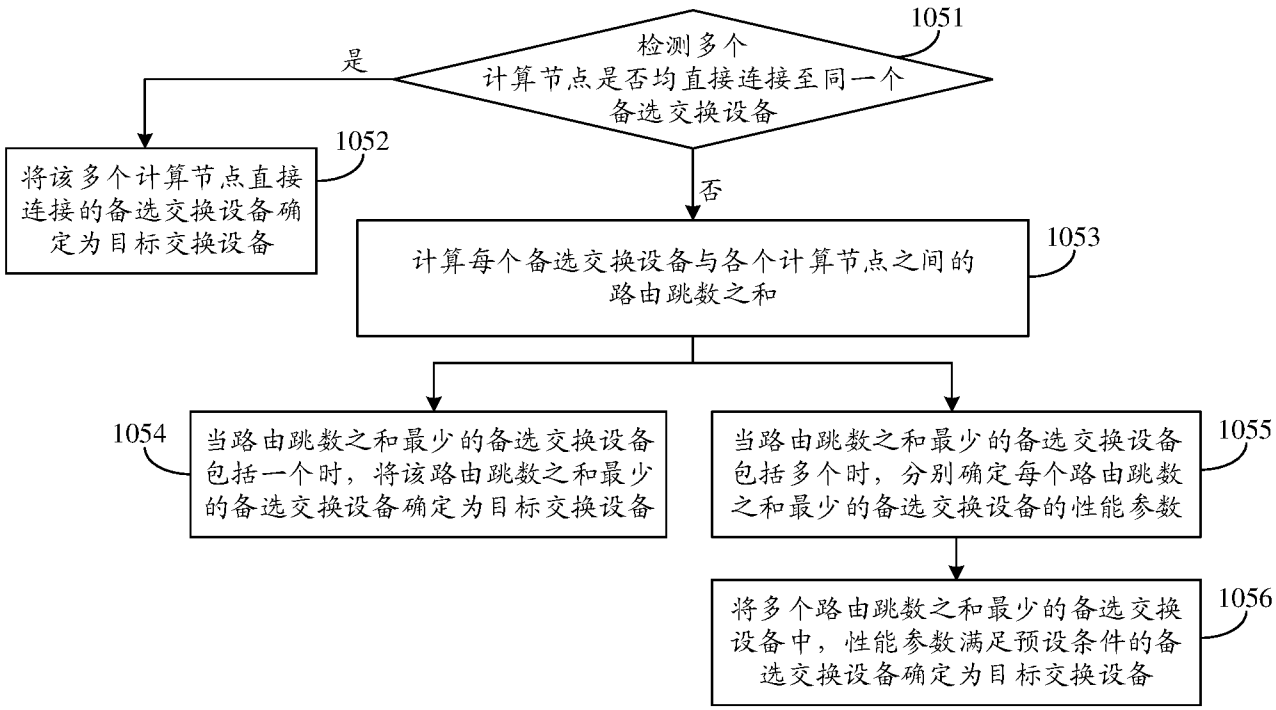

检测多个
计算节点是否均直接连接至同一个
备选交换设备
1051
是
否
将该多个计算节点直接
连接的备选交换设备确
定为目标交换设备
1052
计算每个备选交换设备与各个计算节点之间的
路由跳数之和
1053
当路由跳数之和最少的备选交换设备
包括一个时，将该路由跳数之和最少
的备选交换设备确定为目标交换设备
1054
当路由跳数之和最少的备选交换设备
包括多个时，分别确定每个路由跳数
之和最少的备选交换设备的性能参数
1055
将多个路由跳数之和最少的备选交换
设备中，性能参数满足预设条件的备
选交换设备确定为目标交换设备
1056

图 5

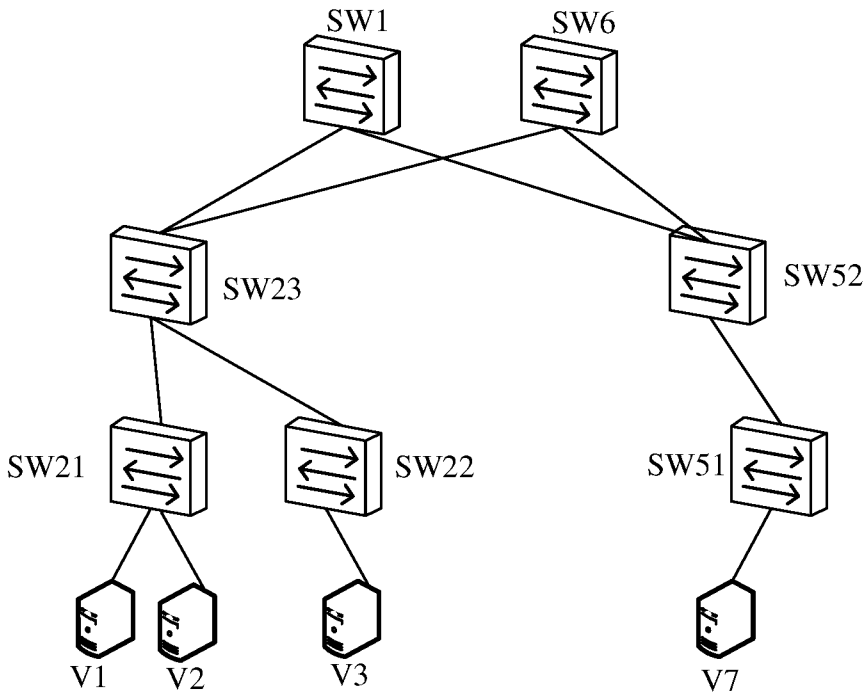

SW1
SW6
SW23
SW52
SW21
SW22
SW51
V1
V2
V3
V7

图 6

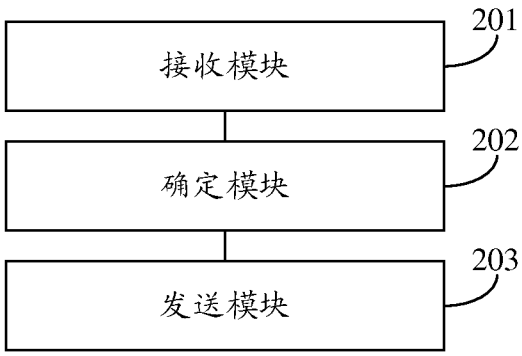

201
接收模块
202
确定模块
203
发送模块

图 7

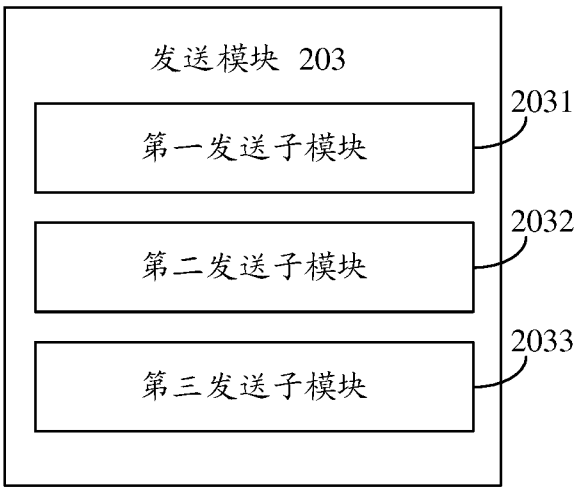

发送模块 203
2031
第一发送子模块
2032
第二发送子模块
2033
第三发送子模块

图 8

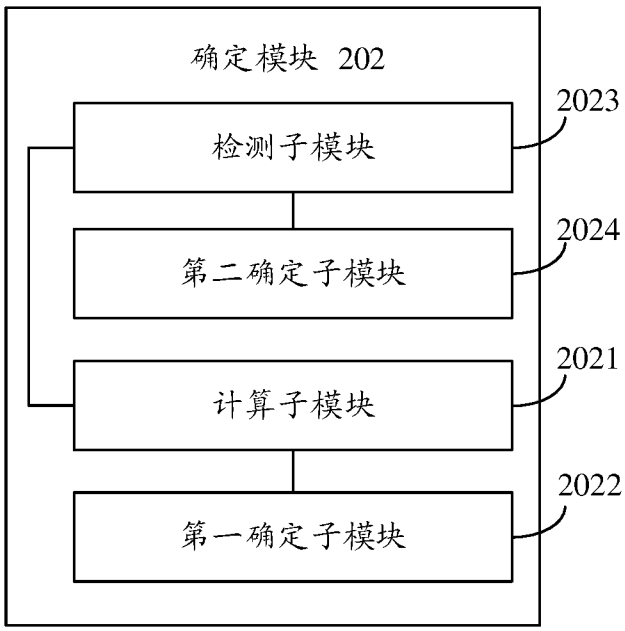

确定模块 202
2023
检测子模块
2024
第二确定子模块
2021
计算子模块
2022
第一确定子模块

图 9

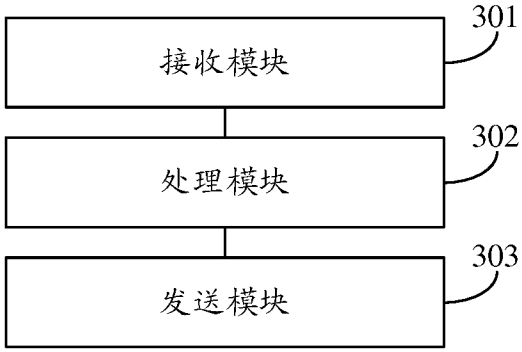

301
接收模块
302
处理模块
303
发送模块

图 10

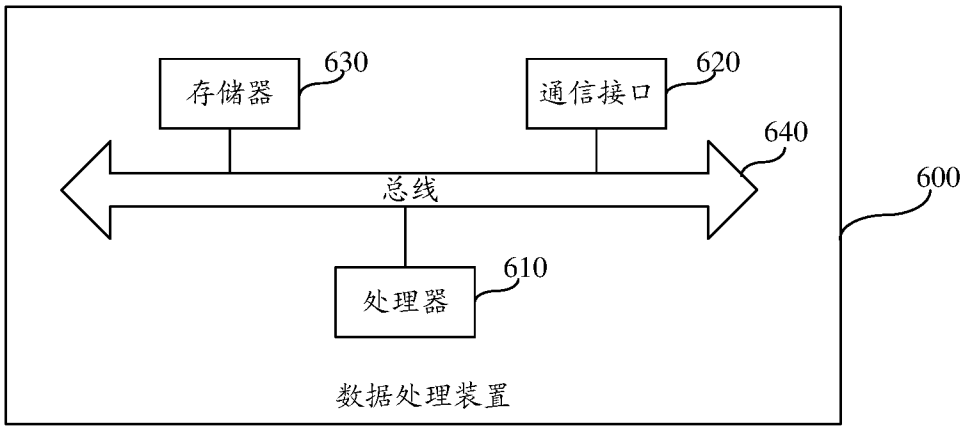

630
存储器
620
通信接口
640
总线
600
610
处理器
数据处理装置

图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/074052

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/733(2013.01)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)

H04L; H04W

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; CPRSABS; CNKI; 3GPP; DWPI; VEN: 分布式, 学习, 交换设备, 交换机, 数据, 合并, 路由, 交换, 控制器, 节点, 参数服务器, 拥塞, distribut+, learn+, switch+, data, combin+, router?, exchang+, controller?, node?, parameter server?, congest+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 105357124 A (HUAZHONG UNIVERSITY OF SCIENCE & TECHNOLOGY) 24 February 2016 (2016-02-24) description, paragraphs [0002]-[0027], and figures 1-4	1-28
A	CN 105634938 A (SHENYANG INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES) 01 June 2016 (2016-06-01) entire document	1-28
A	US 2014289745 A1 (SEVEN NETWORKS, INC.) 25 September 2014 (2014-09-25) entire document	1-28

☐ 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
27 March 2019	**07 May 2019**

Name and mailing address of the ISA/CN	Authorized officer
National Intellectual Property Administration, PRC (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/074052

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	105357124	A	24 February 2016	CN	105357124	B	17 August 2018
CN	105634938	A	01 June 2016	CN	105634938	B	06 June 2017
US	2014289745	A1	25 September 2014	US	9516127	B2	06 December 2016
				US	10178199	B1	08 January 2019

国际检索报告

国际申请号 PCT/CN2019/074052

A. 主题的分类

H04L 12/733(2013.01)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)

H04L; H04W

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;CPRSABS;CNKI;3GPP;DWPI;VEN:分布式，学习，交换设备，交换机，数据，合并，路由，交换，控制器，节点，参数服务器，拥塞，distribut+，learn+， switch+，data， combin+，router?， exchang+， controller?, node?, parameter server?, congest+

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 105357124 A (华中科技大学) 2016年 2月 24日 (2016-02-24) 说明书第0002-0027段，图1-4	1-28
A	CN 105634938 A (中国科学院沈阳自动化研究所) 2016年 6月 1日 (2016-06-01) 全文	1-28
A	US 2014289745 A1 (SEVEN NETWORKS INC) 2014年 9月 25日 (2014-09-25) 全文	1-28

☐ 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 3月 27日	2019年 5月 7日

ISA/CN的名称和邮寄地址	受权官员
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	刘永喆
传真号 (86-10)62019451	电话号码 86-(010)-62089976

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/074052

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	105357124	A	2016年 2月 24日	CN	105357124	B	2018年 8月 17日
CN	105634938	A	2016年 6月 1日	CN	105634938	B	2017年 6月 6日
US	2014289745	A1	2014年 9月 25日	US	9516127	B2	2016年 12月 6日
				US	10178199	B1	2019年 1月 8日