(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
2022 年 3 月 10 日 (10.03.2022)

WIPO | PCT

(10) 国际公布号
WO 2022/048328 A1

(51) 国际专利分类号:
H04L 1/00 (2006.01) *H04L 12/24* (2006.01)

(21) 国际申请号: PCT/CN2021/106804

(22) 国际申请日: 2021 年 7 月 16 日 (16.07.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010920720.5 2020年9月4日 (04.09.2020) 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: 端 启 航 (**DUAN, Qihang**); 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, 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,

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

(54) **发明名称**: 数据处理的方法、装置、设备及介质

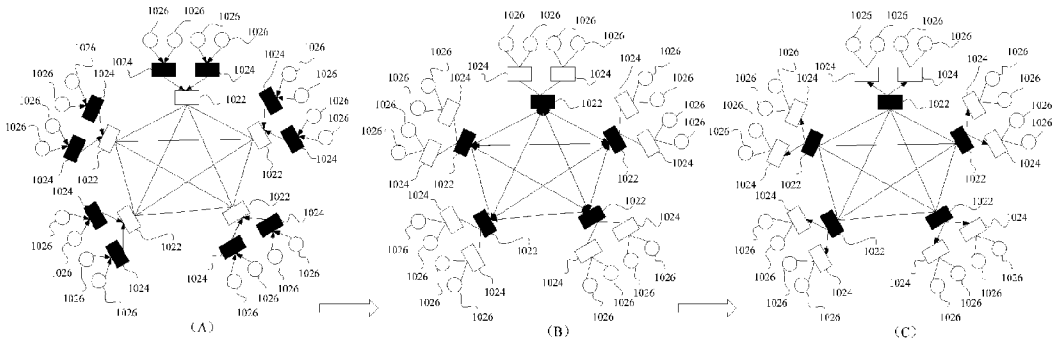

图5

(57) **Abstract:** The present application provides a data processing method, applied to a computing cluster. The computing cluster comprises a plurality of groups of a fully interconnected structure, and each group comprises an exchange node and a computing node. The method comprises: an exchange node, i.e., a first top node, used for communicating with a second top node in a second group, in a first group receives first data sent by a sub node; the first top node receives second data sent by the second top node; and the first top node then aggregates the first data and the second data to obtain third data. Top nodes in a plurality of groups participate in aggregation communication to aggregate data, so that the efficiency of aggregation is improved; moreover, the top nodes in the plurality of groups are used for aggregating the data, so that the utilization rate of the nodes is increased, node resources are prevented from being idle, and the resource utilization rate is increased.

(57) **摘要**: 本申请提供了一种数据处理的方法，应用于计算集群，计算集群包括多个全互连结构的组，每个组中包括交换节点和计算节点，该方法包括：第一组中用于与第二组中的第二顶节点通信的交换节点即第一顶节点接收子节点发送的第一数据，然后第一顶节点接收第二顶节点发送的第二数据，接着第一顶节点对第一数据和第二数据进行聚合，得到第三数据。由于多个组中的顶节点参与聚合通讯，对数据进行聚合，因而提高了聚合的效率，而且多个组中的顶节点被用于聚合数据，提高了节点的利用率，避免节点资源被闲置，提高了资源利用率。

WO 2022/048328 A1

[见续页]

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))。

数据处理的方法、装置、设备及介质

技术领域

本申请涉及计算技术领域，尤其涉及一种数据处理的方法、装置、设备以及计算机可读存储介质。

背景技术

随着高性能计算（high performance computing，HPC）以及（artificial intelligence，AI）技术的不断发展，许多新的应用应运而生。用户对各类应用场景的执行效率及性能也越来越追求极致。为此，越来越多的应用采用聚合通讯代替大量的点对点操作，以提升应用的运行性能。
86902977PCT02

聚合通讯，也称作集合通信，具体是指在一组进程之间完成特定的通信操作。该特定的通信操作可以包括广播（broadcast）、收集（gather）、归约（reduction）等操作。与点对点通讯相比，聚合通讯能够有效地提高消息传递的性能。

针对聚合通讯，业界还提出了一种可扩展分层聚合协议（scalable hierarchical aggregation and reduction protocol，SHARP），以提升聚合通讯的性能。具体地，SHARP 支持将聚合操作由计算节点卸载至交换网络如交换机上，无需在计算节点之间多次发送数据，如此减少了聚合操作整体在网络上的数据量，减少了聚合操作的时间。

然而，将聚合操作由计算节点卸载至交换网络时，通常是按树的方式在交换机上进行聚合，一方面聚合效率较低，另一方面未能充分利用带宽资源，资源利用率较低。

发明内容

本申请提供了一种数据处理的方法，该方法利用计算集群包括的多个全互连结构的组中的顶节点对数据进行聚合，而不局限于根节点，提高了聚合效率，并且该方法可以使得更多顶节点参与聚合，充分利用带宽资源，提高了资源利用率。本申请还提供了上述方法对应的装置、设备、计算机可读存储介质以及计算机程序产品。

第一方面，本申请提供了一种数据处理的方法。该方法应用于计算集群。该计算集群包括多个全互连结构的组。每个组实际是一种实现组内通信的子网络。所谓全互连是指每个组与该计算集群的其他组中的每一个组之间至少有一个直连链路。如此可以进一步缩短组间的链路长度，从而实现提供低延迟通信。

每个组中包括交换节点和计算节点。计算节点是指具有计算能力的计算机设备，例如可以是终端或服务器。其中，终端包括但不限于台式机、笔记本电脑、平板电脑或者智能手机。交换节点是指网络中具有数据转发功能的设备，例如可以是交换机或路由器。

每个组中用于与其他组进行通信的交换节点称为顶节点。每个组中的计算节点称为叶子节点。在一些可能的实现方式中，组中还可以包括用于组内通信的交换节点，该交换节点可以连接叶子节点和顶节点，实现组内通信。因此，上述用于组内通信的交换节点可以称为中间节点。一个组的中间节点具体是该组中除顶节点以外的交换节点。

顶节点、中间节点和叶子节点是基于节点在组中的层级对节点进行划分的，在一些可能的实现方式中，还可以通过父节点、子节点描述节点的层级关系。在一个组中，除了顶节点以外的节点均具有父节点，除了叶子节点以外的节点均具有子节点。

为了便于描述，本申请以计算集群包括第一组和第二组进行示例说明。其中，第一组是多个全互连结构的组中的任意一个组，第二组是多个全互连结构的组中除第一组以外的组。

第一组包括第一顶节点，第二组包括第二顶节点。第一顶节点是第一顶节点所在第一组中用于与第二组的第二顶节点进行通信的交换节点。第二顶节点是第二组中用于与第一组进行通信的交换节点。第一顶节点接收子节点发送的第一数据，以及接收第二顶节点发送的第二数据，然后第一顶节点对第一数据和第二数据进行聚合。

在该方法中，多个顶节点均参与聚合通讯，并进行聚合卸载，而不局限于根节点，如此可以提高聚合效率。并且，多个顶节点参与聚合通讯，可以避免资源闲置，充分利用网络资源，提高了资源利用率。

在一些可能的实现方式中，第一顶节点可以接受多个子节点发送的多个第一数据，对应地，第一顶节点可以先对子节点发送的多个第一数据进行聚合，然后第一顶节点将对所述多个第一数据进行聚合所得的数据与所述第二数据进行聚合，得到第三数据。

其中，第一顶节点对多个第一数据进行聚合的结果可以提供给第二顶节点，以便第二顶节点直接根据该结果以及第二数据进行聚合，得到第三数据。由此实现多个组中的顶节点对数据进行聚合，提高聚合效率，以及提供资源利用率。

在一些可能的实现方式中，第一顶节点可以向所述第二顶节点发送对所述多个第一数据进行聚合所得的数据。如此，第二顶节点无需再对多个第一数据进行聚合。由此，可以减少聚合次数，提高聚合效率，避免资源浪费。

在一些可能的实现方式中，第一顶节点可以通过以下方式中的任意一种向所述第二顶节点发送数据：全互连（all to all）、环形（ring）和递推倍增（recursive doubling）。

其中，全互连方式是指每个顶节点向与该顶节点直连的所有顶节点发送数据，具体可以是发送对多个第一数据进行聚合所得的数据。环形方式是指每个顶节点依次向相邻的顶节点发送数据。其中，每个顶节点发送的数据包括每个顶节点自身聚合的数据，以及接收到的相邻的顶节点聚合的数据。每个顶节点知晓需要聚合的数据的数量，当顶节点接收到所有的待聚合的数据时，该顶节点可以停止发送数据。递推倍增方式是指每两个节点相互发送数据，然后每个节点聚合数据。接着，每两组节点交互聚合后的数据，每组节点包括上述两个节点。以此类推，直至每个节点聚合所有数据为止。

无论是采用全互连、环形还是递推倍增，每个组的顶节点均可以获得所有的待聚合的数据，每个组的顶节点均可以对数据进行聚合，提高了聚合效率，并且提高了顶节点的参与率，从而提高了资源利用率。

在一些可能的实现方式中，所述子节点包括所述第一组中与所述第一顶节点直接相连的交换节点。对应地，所述第一数据包括部分聚合数据，所述第三数据为完全聚合数据。其中，部分聚合数据是指对部分待聚合的数据进行聚合所得的数据，完全聚合数据是指对所有待聚合的数据进行聚合所得的数据。

在一些可能的实现方式中，所述子节点包括所述第一组中与所述第一顶节点直接相连的计算节点，所述第一数据包括非聚合数据，所述第三数据为完全聚合数据。其中，非聚合数据是指待聚合的数据中尚未进行聚合的数据。非聚合的数据为叶子节点提供的数据。

在一些可能的实现方式中，第一顶节点在接收子节点发送的第一数据之前，还可以先加入通信域。通信域定义了一个聚合拓扑，该聚合拓扑描述了提供待聚合的数据的叶子节点以及对数据进行聚合的中间节点、顶节点。该聚合拓扑可以与物理拓扑相同，也可以与物理拓扑不同，叶子节点的数据可以按照聚合拓扑进行聚合。当聚合拓扑与物理拓扑相同时，节点利用率可以达到最大值，由此可以提升后续聚合过程中数据聚合的效率以及资源利用率。

在一些可能的实现方式中，聚合拓扑可以包括根节点，该根节点为控制节点，不同组中的顶节点为该根节点下的二级节点。第一顶节点可以向控制节点发送加入域请求，用于请求加入通信域。控制节点收到所有的加入域请求后，可以生成加入域响应。然后第一顶节点接收所述控制节点发送的加入域响应，该加入域响应用于表征所述第一顶节点成功加入通信域。如此，第一顶节点可以在业务流程中对数据进行聚合，提高聚合效率以及资源利用率。

在一些可能的实现方式中，所述控制节点可以为独立于所述计算节点和所述交换节点的节点，也可以为交换节点中的一个节点。当控制节点为独立的节点时，有利于提高计算集群的稳定性，当控制节点为交换节点中的一个节点时，可以充分利用已有资源，避免资源浪费。

在一些可能的实现方式中，聚合拓扑也可以没有根节点。所述第一顶节点可以接收第三顶节点发送的加入域请求，当加入域请求中的域标识包括于所述第一顶节点的域标识列表时，所述第一顶节点将第一顶节点的节点标识添加至对应通信域的节点标识列表。如此，第一顶节点可以在业务流程中对数据进行聚合，提高聚合效率以及资源利用率。

在一些可能的实现方式中，所述第三顶节点为所述第二顶节点中右连或左连所述第一顶节点的节点，或者是所述第二顶节点中与所述第一顶节点直接相连的节点。

具体地，每个顶节点可以接收左连（或右连）的顶节点发送的加入域请求，若加入域请求中的域标识在该顶节点的域标识列表中，则将该顶节点的节点标识添加至对应通信域的节点标识列表，然后转发加入域请求至下一顶节点，否则，直接转发该加入域请求。当某个顶节点接收到自己发送的加入域请求时，丢弃该加入域请求，并停止转发。

当然，顶节点也可以接收加入域请求，若加入域请求中的域标识在该顶节点的域标识列表中，则将该顶节点的节点标识添加至通信域的节点标识列表，否则将加入域请求丢弃。当顶节点收到加入域请求对应的子节点个数总和等于整个通信域中叶子节点个数时，通信域创建完毕，各叶子节点和各顶节点均加入通信域。

在一些可能的实现方式，所述计算集群的拓扑为蜻蜓网络拓扑。计算集群中的计算节点在进行聚合通讯时，可以根据蜻蜓网络拓扑构建聚合拓扑，按照聚合拓扑进行聚合，而不局限于在根节点进行最终的聚合，如此可以提高聚合效率，以及节点资源的利用率。

第二方面，本申请提供了一种数据处理的装置，该装置包括用于执行第一方面或第一方面任意一种可能实现方式中的数据处理的方法的各个模块。

第三方面，本申请提供一种电子设备，该电子设备包括处理器和存储器。所述处理器和所述存储器进行相互的通信。其中，所述存储器，用于存储计算机指令；所述处理器，用于根据所述计算机指令执行如本申请第一方面或第一方面的任意一种实现方式中的数据处理的方法。

第四方面，本申请提供一种计算机可读存储介质，所述计算机可读存储介质中存储有指令，当其在计算机上运行时，使得计算机执行上述第一方面或第一方面的任意一种实现方式中的数据处理的方法。

第五方面，本申请提供了一种包含指令的计算机程序产品，当其在计算机上运行时，使得计算机执行上述第一方面或第一方面的任意一种实现方式中的数据处理的方法。

本申请在上述各方面提供的实现方式的基础上，还可以进行进一步组合以提供更多实现方式。

附图说明

为了更清楚地说明本申请实施例的技术方法，下面将对实施例中所需使用的附图作以简单地介绍。

图 1 为本申请实施例提供的一种蜻蜓网络的系统架构图；

图 2 为本申请实施例提供的一种蜻蜓网络的系统架构图；

图 3 为本申请实施例提供的一种蜻蜓网络的系统架构图；

图 4 为本申请实施例提供的一种数据处理的方法的流程图；

图 5 为本申请实施例提供的一种数据处理的方法的流程示意图；

图 6 为本申请实施例提供的一种递推倍增方式发送数据的示意图；

图 7 为本申请实施例提供的一种蜻蜓网络的系统架构图；

图 8 为本申请实施例提供的一种分布式模型训练的场景示意图；

图 9 为本申请实施例提供的一种数据处理的装置的结构示意图；

图 10 为本申请实施例提供的一种电子设备的结构示意图。

具体实施方式

为了便于理解，首先对本申请实施例中所涉及到的一些技术术语进行介绍。

高性能计算是指利用大量处理单元的聚合计算能力进行计算，从而用于解决复杂问题，如天气预测、石油勘探、核爆模拟等问题。其中，大量处理单元的聚合计算能力可以是单个机器中多个处理器的聚合计算能力，也可以是集群中的多个计算机的聚合计算能力。

集群中多个计算机的聚合计算基于聚合通讯实现。所谓聚合通讯，也称作集合通信。在包括多个计算机的计算系统中，涉及一组处理器之间的全局数据迁移和全局控制的操作被称为聚合通讯。聚合通讯在并行分布式计算领域中有着大量而且重要的应用，在很多情况下聚合通讯比点对点通讯更为重要。与点对点通讯相比，聚合通讯能够很大程度地提高消息传递程序的性能。

业界常用的聚合通讯中间件包括消息传递接口（message passing interface，MPI）。MPI 定义了一种具有可移植性的编程接口以供用户调用从而实现对应的聚合通讯操作。聚合通讯

操作具体包括广播（broadcast）、栅栏同步（barrier）、归约（reduce）、分散（scatter）和收集（gather）等中的任意一种或多种。

根据聚合通讯的数据流的流向，可以将聚合通讯操作分为有根的通讯操作（rooted communication）和无根的通讯操作（non-rooted communication）。有根的通讯操作是指源于特定节点(即根节点)或者将消息注入根节点的操作，具体包括 broadcast、gather、scatter 和 reduce 等操作。无根的通讯操作是指集合通讯操作中除有根的通讯操作之外的通讯操作，具体包括全局收集（all gather），全局分散（all scatter），全局归约（all reduce）和 barrier 等。

在进行聚合通讯时，不同计算机之间的同步和通信可以通过互连网络实现。其中，互连网络可以是蜻蜓（dragonfly）网络。蜻蜓网络包括多个组，每个组实际是一种实现组内通信的子网络。每个组之间通过链路连接，并且组间链路相对较窄，类似于蜻艇的宽大身体和窄小翅膀，因此称为蜻艇网络。由于蜻蜓网络包括的各个组之间采用窄链路，能够很大程度减少全局链路的数量，降低组网成本。

进一步地，蜻蜓网络的各个组之间可以采用全互连（all to all）方式进行连接。所谓全互连是指每个组与该蜻蜓网络的其他组中的每一个组之间至少有一个直连链路。如此可以进一步缩短链路长度，从而实现提供低延迟通信。

接下来，以组间采用全互连方式进行连接的蜻蜓网络（也称作 dragonfly+网络）为例，对蜻蜓网络的架构进行详细说明。蜻蜓网络是一种网络拓扑结构。蜻蜓网络组规模较大，支持连接的组数也较多。但是，组间链路较窄，类似于蜻艇的宽大身体和窄小翅膀的结构，因此称为蜻艇网络。蜻蜓网络由于组间采用窄链路，能很大程度节省全局链路的数量，节约系统成本，而倍受青睐。蜻蜓网络为实现全局跳步为 1 的目标，组间采用全互连结构，并且为了降低网络直径。组内采用跳步数尽可能少的连接方式，通常为全互连或维扁平蝴蝶的结构。

参见图 1 所示的蜻蜓网络的架构示意图，该蜻蜓网络 100 包括多个全互连结构的组 102，每个组 102 中包括交换节点和计算节点。计算节点是指具有计算能力的计算机设备，例如可以是终端或服务器。其中，终端包括但不限于台式机、笔记本电脑、平板电脑或者智能手机。交换节点是指网络中具有数据转发功能的设备，例如可以是交换机或路由器。全互连结构的组是指不同组之间存在连接关系，使得各个组中交换节点或计算节点可以通过互连的网络结构进行通信。

每个组 102 中用于与其他组进行通信的交换节点称为顶节点 1022。每个组 102 中的计算节点称为叶子节点 1026。在一些可能的实现方式中，组 102 中还可以包括用于组内通信的交换节点，该交换节点可以连接叶子节点和顶节点，实现组内通信。因此，上述用于组内通信的交换节点可以称为中间节点 1024。一个组 102 的中间节点 1024 具体是该组 102 中除顶节点 1022 以外的交换节点。

上述顶节点 1022、中间节点 1024、叶子节点 1026 是基于节点在网络中的层级进行划分的。在一些可能的实现方式中，还可以通过父节点、子节点描述节点的层级关系。在一个组 102 中，除了顶节点 1022 以外的节点均具有父节点，除了叶子节点 1026 以外的节点均具有子节点。

一个顶节点 1022 的子节点为该顶节点 1022 所在组 102 中与该顶节点 1022 直接相连的节点。其中，一个组 102 中与顶节点 1022 直接相连的节点可以是中间节点 1024，也可以是叶子节点 1026。对应地，叶子节点 1026 的父节点可以是中间节点 1024 或者顶节点 1022。

一个中间节点 1024 的子节点为该中间节点 1024 所在组 102 中与该中间节点 1024 直接相连，并且除顶节点 1022 以外的节点。其中，一个组 102 中与中间节点 1024 直接相连并且除

顶节点 1022 以外的节点可以是另一中间节点 1024，也可以是叶子节点 1026。对应地，中间节点 1024 的父节点可以是另一中间节点 1024 或者是顶节点 1022。

图 1 仅仅是以一个组 102 中包括一个顶节点 1022，顶节点 1022 的子节点为多个中间节点 1024，中间节点 1024 的子节点为多个叶子节点 1026 进行示例说明。在本申请实施例其他可能的实现方式中，组 102 中可以包括多个顶节点 1022。

在一些可能的实现方式中，如图 2 所示，顶节点 1022 的子节点还可以包括叶子节点 1026。在另一些可能的实现方式中，如图 3 所示，中间节点 1024 的子节点还可以包括其他中间节点 1024，本申请实施例对此不作限定。

其中，交换节点（包括顶节点 1022 和中间节点 1024）不仅可以用于数据转发，还可以用于数据处理。例如，交换节点还可以用于多个计算节点进行聚合通讯时，对计算节点的数据进行聚合。

在一些计算任务中，需要将各计算节点上待聚合的数据进行聚合，并将最终聚合结果分发给各计算节点。其中，待聚合的数据具体可以是计算任务对应数据类型的数据，计算任务对应数据类型包括但不限于：整型、浮点型、布尔型，即，待聚合的数据可以为整型数据、浮点型数据或布尔型数据。

例如，在天气预测任务中，待聚合的数据可以是温度、湿度、风向、风速等中的至少一种。其中，温度、湿度、风速的数据类型可以是整型或浮点型。风向可以通过与指定方向如（北向）的夹角进行表征，因此，风向的数据类型可以是浮点型。又例如，在石油勘测任务中，待聚合的数据可以是不同地理位置的重力、磁力、电阻率等中的至少一个。其中，重力、磁力的数据类型可以是整型，电阻率的数据类型可以是浮点型。考虑到精确度，重力、磁力的数据类型还可以是浮点型。

聚合是指将各计算节点上的多个待聚合的数据进行合并处理得到一个数据的过程。对多个待聚合数据进行合并处理的过程，具体可以是对多个待聚合数据进行数学公式运算，例如，对多个待聚合的数据进行相加，获得的和为聚合结果。又例如，对多个待聚合数据进行相加，最后进行求平均值，获得的平均值为聚合结果。

如图 1 至图 3 所示，叶子节点 1026 在进行聚合通讯时还可以将聚合操作卸载至交换节点如中间节点 1024 和顶节点 1022 上。例如，多个叶子节点 1026 进行全局归约操作，如全局归约求和时，可以将求和操作卸载至对应的中间节点 1024 以及顶节点 1022。

然而，交换节点通常是按照树的方式进行聚合，具体是以其中一个顶节点 1022 作为根节点，将根节点作为聚合的最终节点进行逐层聚合。这种聚合方法效率低下，而且大量网络资源（如带宽资源）未被充分利用，降低了资源利用率，增加了聚合成本。

本申请实施例提供了一种数据处理的方法。该方法应用于计算集群。该计算集群的网络拓扑可以是蜻蜓网络拓扑结构，例如是如图 1 至图 3 所示的 dragonfly+网络拓扑结构。该计算集群包括多个全互连结构的组，每个组中包括交换节点和计算节点。为了便于描述，本申请以计算集群包括第一组和第二组进行示例说明。其中，第一组是多个全互连结构的组中的任意一个组，第二组是多个全互连结构的组中除第一组以外的组。

第一组包括第一顶节点，第二组包括第二顶节点。第一顶节点是第一顶节点所在第一组中用于与第二组的第二顶节点进行通信的交换节点。第二顶节点是第二组中用于与第一组进行通信的交换节点。第一顶节点接收子节点发送的第一数据，以及接收第二顶节点发送的第二数据，然后第一顶节点对第一数据和第二数据进行聚合。

在该方法中，多个顶节点 1022 均参与聚合通讯，并进行聚合卸载，而不局限于根节点，

如此可以提高聚合效率。并且，多个顶节点 1022 参与聚合通讯，可以避免资源闲置，充分利用网络资源，提高了资源利用率。

为了便于理解，下面结合图 1 至图 3 所示的蜻蜓网络中数据处理的方法对本申请的技术方案进行介绍。

参见图 1 至图 3 所示的蜻蜓网络的系统架构图，不同叶子节点 1026 进行聚合通讯时，与该叶子节点连接的中间节点 1024、顶节点 1022 也可以进行聚合计算，从而完成聚合通讯。具体地，叶子节点 1026、中间节点 1024、顶节点 1022 可以先加入通信域。

不同计算任务可以对应不同的通信域。通信域定义了一个聚合拓扑，该聚合拓扑描述了提供待聚合的数据的叶子节点 1026 以及对数据进行聚合的中间节点 1024、顶节点 1022。该聚合拓扑可以与物理拓扑相同，也可以与物理拓扑不同，叶子节点 1026 的数据可以按照聚合拓扑进行聚合。当聚合拓扑与物理拓扑相同时，节点利用率可以达到最大值，由此可以提升后续聚合过程中数据聚合的效率以及资源利用率。

参与聚合通讯的各节点加入通信域后，每个组 102 中的叶子节点 1026 可以向其父节点发送待聚合的数据，父节点可以对叶子节点 1026 发送的数据进行初步聚合。其中，叶子节点 1026 的父节点可以是中间节点 1024 或者是顶节点 1022。当叶子节点 1026 的父节点为中间节点 1024 时，该中间节点 1024 还可以向父节点发送初步聚合后的数据，以便父节点进行进一步聚合。中间节点 1024 的父节点可以是另一中间节点 1024 或者是顶节点 1022。中间节点 1024 的父节点为另一中间节点 1024 时，该中间节点 1024 可以向另一中间节点 1024 发送进一步聚合后的数据，以便另一中间节点 1024 对进一步聚合后的数据再进行聚合。

对应地，每个组 102 中的顶节点 1022 可以接收子节点发送的第一数据。当子节点为叶子节点 1026 时，第一数据可以包括叶子节点 1026 的数据，该数据为非聚合数据。其中，非聚合数据是指多个待聚合的数据中尚未进行聚合的数据。当子节点为中间节点 1024 时，第一数据可以包括对部分叶子节点 1026 的数据进行聚合的数据，该数据为部分聚合数据。

每个组 102 中的顶节点 1022 还可以接收其他组中的顶节点 1022 发送的第二数据。该第二数据是其他组 102 中的顶节点 1022 根据该组 102 中的叶子节点 1026 的数据获得的数据。当该组 102 中包括多个叶子节点时，第二数据可以是对该组中的叶子节点 1026 的数据进行聚合所得的数据。也即第二数据可以是部分聚合数据。在一些实施例中，一个组 102 仅包括一个叶子节点 1026 时，第二数据也可以是该叶子节点 1026 的数据，也即第二数据也可以是非聚合数据。

每个组 102 中的顶节点 1022 可以对第一数据和第二数据进行聚合，得到第三数据。其中，第三数据是对该计算任务对应的所有待聚合的数据进行聚合所得的数据，该第三数据也称作完全聚合数据。

在一些可能的实现方式中，如图 1 至图 3 所示，顶节点 1022 包括多个子节点，顶节点 1022 可以接收多个子节点分别发送的第一数据。对应地，顶节点 1022 在对第一数据和第二数据进行聚合时，可以先对多个子节点发送的多个第一数据进行聚合，然后将对多个第一数据进行聚合所得的数据与第二数据进行聚合，得到第三数据。

其中，每个顶节点 1022 可以向其他组的顶节点 1022 发送对第一数据进行聚合所得的数据，以便每个顶节点 1022 可以实现对所有待聚合的数据进行聚合。由此实现每个顶节点 1022 均参与数据聚合，提高聚合效率，以及资源利用率。

根据顶节点 1022 的子节点不同，上述第一数据可以分为以下三种情况：

第一种情况，如图 1 所示，顶节点 1022 的子节点均为中间节点 1024，对应地，第一数

据可以均为部分聚合数据；

第二种情况，如图 2 所示，顶节点 1022 的子节点包括中间节点 1024 和叶子节点 1026，对应地，第一数据可以包括部分聚合数据和非聚合数据；

第三种情况，顶节点 1022 的子节点均为叶子节点 1026，对应地，第一数据均为非聚合数据。

接下来，结合附图，从第一顶节点的角度对本申请实施例提供的数据处理的方法的具体实现进行详细说明。参见图 4 所示的数据处理的方法的流程图，该方法包括：

S402：第一顶节点接收子节点发送的第一数据。

在一些计算任务中，需要将各计算节点上待聚合的数据进行聚合，并将最终聚合结果分发给各计算节点。其中，待聚合的数据具体可以是计算任务对应数据类型的数据，计算任务对应数据类型包括但不限于：整型、浮点型、布尔型，即，待聚合的数据可以为整型数据、浮点型数据或布尔型数据。

聚合是指将各计算节点上的多个待聚合的数据进行合并处理得到一个数据的过程。对多个待聚合数据进行合并处理的过程，具体可以是对多个待聚合数据进行数学公式运算，例如，对多个待聚合的数据进行相加，获得的和为聚合结果。又例如，对多个待聚合数据进行相加，最后进行求平均值，获得的平均值为聚合结果。

在本实施例中，对计算节点上待聚合的数据进行聚合的过程被卸载至交换节点。以第一组对数据处理的过程进行示例说明。具体地，第一组中的第一顶节点具有子节点的，第一顶节点可以接收该子节点发送的第一数据。根据计算任务不同，第一数据可以是不同数据。例如在天气预测任务中，第一数据可以是温度、湿度、风向和/或风速。又例如，在石油勘测任务中，第一数据可以是重力、磁力、电阻率等中的一种或多种。

其中，第一顶节点的子节点包括中间节点时，该中间节点可以对叶子节点（计算节点）上待聚合的数据进行聚合，得到第一数据，该第一数据为部分聚合数据。第一顶节点的子节点包括叶子节点时，该叶子节点上待聚合的数据即为第一数据，该第一数据属于非聚合数据。第一顶节点的子节点可以向第一顶节点发送第一数据，例如发送上述部分聚合数据和/或非聚合数据，以便第一顶节点对第一数据进行进一步聚合得到最终聚合结果。

S404：第一顶节点接收所述第二顶节点发送的第二数据。

第二数据是第二组中的第二顶节点根据该第二顶节点的子节点发送的数据获得的数据。根据计算任务不同，该第二数据可以是不同数据。例如，在天气预测任务中，第二数据可以是温度、湿度、风向和/或风速。又例如，在石油勘测任务中，第二数据可以是重力、磁力、电阻率等中的一种或多种。其中，第一数据和第二数据是同一计算任务的数据类型对应的数据，例如，第一数据和第二数据可以是温度，或者是湿度等等。

该第二数据可以是非聚合数据，或者是部分聚合数据。当第二顶节点仅连接一个叶子节点，例如直接连接一个叶子节点，或者通过中间节点连接一个叶子节点，该第二数据为非聚合数据，具体是该叶子节点上的数据。当第二顶节点（直接或间接地）连接多个叶子节点时，该第二数据为部分聚合数据，该部分聚合数据是对第二顶节点连接的叶子节点上的数据进行聚合得到。

与第一顶节点类似，第二顶节点的子节点可以包括中间节点和/或叶子节点。当第二顶节点的子节点包括中间节点时，第二顶节点可以对中间节点发送的部分聚合数据进行进一步聚

合，得到第二数据。当第二顶节点的子节点包括叶子节点时，第二顶节点可以对叶子节点发送的非聚合数据进行聚合，得到第二数据。当第二顶节点的子节点包括中间节点和叶子节点时，第二顶节点可以对中间节点发送的部分聚合数据和叶子节点发送的非聚合数据进行进一步聚合，得到第二数据。

S406：第一顶节点对所述第一数据和所述第二数据进行聚合，得到第三数据。

当第一顶节点在聚合拓扑中仅包括一个子节点时，第一顶节点可以接收该子节点发送的一个第一数据。对应地，第一顶节点可以直接将第一数据和第二数据进行聚合，得到第三数据。

当第一顶节点在聚合拓扑中包括多个子节点时，第二顶节点可以接收多个子节点发送的多个第一数据。对应地，第一顶节点可以先对多个第一数据进行聚合，然后将对多个第一数据进行聚合所得的数据与第二数据进行聚合，从而得到第三数据。

进一步地，第一顶节点可以向第二顶节点发送数据，以便第二顶节点也对数据进行聚合。具体地，当第一顶节点仅接收一个第一数据时，第一顶节点可以直接发送该第一数据。当第一顶节点接收多个第一数据时，第一顶节点可以向第二顶节点发送对多个第一数据进行聚合所得的数据。如此，第二顶节点无需再对多个第一数据进行聚合。由此，可以减少聚合次数，提高聚合效率，避免资源浪费。

第一顶节点聚合得到第三数据（完全聚合数据）后，还可以向其子节点返回该第三数据，从而向对应的计算节点返回第三数据，以完成计算任务。

为了使得本申请的技术方案更加清楚，下面结合附图，从数据变化的角度对本申请实施例提供的数据处理方法的具体实现进行详细说明。

参见图 5 所示的数据处理的方法的流程示意图，圆圈代表计算节点，也即代表叶子节点 1026，方块代表交换节点，具体包括顶节点 1022 和中间节点 1024。其中，叶子节点提供待聚合的数据，交换节点如顶节点 1022 和中间节点 1024 对待聚合的数据进行聚合，然后转发最终聚合结果至叶子节点 1026。

每个组的中间节点 1024 可以接收该组的叶子节点 1026 发送的、待聚合的数据，该中间节点 1024 可以对叶子节点 1026 发送的数据进行聚合，得到部分聚合数据。如图 5 中（A）所示，填充黑色的方块表明该节点上具有经过聚合的数据，例如上述部分聚合数据。中间节点 1024 向顶节点 1022 发送上述部分聚合数据。接着，如图 5 中（B）所示，每个顶节点 1022 对该顶节点 1022 的子节点（具体为中间节点 1024）发送的部分聚合数据进行聚合，然后向其他顶节点 202 发送聚合结果，如此，每个顶节点 1022 可以将本节点的聚合结果和其他顶节点 1022 的聚合结果进行聚合，得到最终聚合结果。如图 5 中（C）所示，每个顶节点 1022 向该顶节点 1022 的子节点（具体为中间节点 1024）发送最终聚合结果。该中间节点 1024 可以向该中间节点 1024 的子节点（具体为叶子节点 1026）返回最终聚合结果。

如果按照树的方式进行聚合，则在多个顶节点 1022 中确定一个顶节点 1022 为根节点。在执行步骤（A）之后，每个顶节点 1022 将各自接收的部分聚合数据进行聚合，得到聚合结果。然后，多个顶节点 1022 中除根节点以外的其他顶节点 1022 向根节点发送上述聚合结果，根节点对该上述聚合结果进行聚合，得到最终聚合结果。根节点再将该最终聚合结果发送至多个顶节点 1022 中除根节点以外的其他顶节点 1022。每个顶节点 1022 再执行步骤（C），将最终聚合结果发送至叶子节点 1026。

由此可见，本申请实施例的聚合方法可以减少一轮聚合，由此可以提高聚合效率，而且多个顶节点 1022 均参与聚合，避免部分顶节点 1022 资源闲置，充分利用交换节点的资源，提高了资源利用率。

在图 5 所示实施例中，顶节点 1022 是通过全互连（all to all）方式向其他顶节点 1022 发送数据，如发送顶节点 1022 对接收的第一数据进行聚合所得的数据。其中，全交叉方式是指每个顶节点 1022 以点对点方式向其他顶节点 1022 发送数据。

在一些可能的实现方式中，顶节点 1022 还可以通过环形（ring）或者递推倍增（recursive doubling）方式向其他顶节点 1022 发送数据。下面对环形和递推倍增方式进行详细说明。

环形方式是指每个顶节点 1022 依次向相邻的顶节点发送数据。其中，每个顶节点 1022 发送的数据包括每个顶节点 1022 自身聚合的数据，以及接收到的相邻的顶节点 1022 聚合的数据。每个顶节点 1022 知晓需要聚合的数据的数量，当顶节点 1022 接收到所有的待聚合的数据时，该顶节点可以停止发送数据。

例如，多个顶节点 1022 分别为顶节点 1 至顶节点 5，各个顶节点 1022 聚合的来自于子节点的数据为数据 1 至数据 5，则顶节点 1 可以向顶节点 2 发送数据 1，顶节点 2 可以向顶节点 3 发送数据 1 和数据 2，顶节点 3 可以向顶节点 4 发送数据 1、数据 2 和数据 3，顶节点 4 可以向顶节点 5 发送数据 1、数据 2、数据 3 和数据 4。

顶节点 5 可以聚合上述数据 1、数据 2、数据 3、数据 4 以及数据 5。接着，顶节点 5 向顶节点 1 发送数据 2、数据 3、数据 4 和数据 5。如此，顶节点 1 可以聚合数据 1 至数据 5。顶节点 1 向顶节点 2 发送数据 3、数据 4 和数据 5，如此，顶节点 2 可以聚合数据 1 至数据 5。顶节点 2 向顶节点 3 发送数据 4 和数据 5，如此顶节点 3 可以聚合数据 1 至数据 5，顶节点 3 向顶节点 4 发送数据 5，如此顶节点 4 可以聚合数据 1 至数据 5。

递推倍增是指每两个节点相互发送数据，然后每个节点聚合数据。接着，每两组节点交互聚合后的数据，每组节点包括上述两个节点。以此类推，直至每个节点聚合所有数据为止。

为了便于理解，下面结合具体示例说明。参见图 6，计算集群包括 8 个顶节点，具体为 P0、P1、……P7，首先，P0、P1 相互发送数据，P2、P3 相互发送数据，P4、P5 相互发送数据，P6、P7 相互发送数据，然后，P0 至 P7 分别聚合数据。接着，P0、P1 这一组节点与 P2、P3 这一组节点交换聚合后的数据，并对交换的数据进行聚合。例如 P0、P2 相互交换聚合后的数据，并分别对交换的数据进行聚合，P1、P3 相互交换聚合的数据，并分别对交换的数据进行聚合，类似地，P4、P5 这一组节点与 P6、P7 这一组节点交换聚合后的数据，并分别对交换的数据进行聚合。接着，P0、P1、P2、P3 这一组节点与 P4、P5、P6、P7 这一组节点交换聚合后的数据，并分别对交换的数据进行聚合。例如 P0、P4 交换聚合后的数据，P1、P5 交换聚合后的数据，P2、P6 交换聚合后的数据，P3、P7 交换聚合后的数据，P0 至 P7 分别对交换后的数据进行聚合。如此，每个顶节点均可以对所有数据进行聚合，得到最终聚合结果。

图 4 所示实施例主要是从业务流程对数据聚合过程进行介绍。在进行数据聚合之前，还可以执行控制流程，将参与聚合的节点加入通信域。故第一顶节点在接收子节点发送的第一数据之前，第一顶节点还需要加入通信域。

其中，第一顶节点加入通信域可以有多种实现方式。下面分别针对有根节点和无根节点的情况对第一顶节点加入通信域的实现方式进行详细说明。

参见图 7，蜻蜓网络 100 中还包括控制节点 104，该控制节点 104 为独立于计算节点（即叶子节点 1026）和交换节点（包括顶节点 1022 和中间节点 1024）。控制节点 104 包括子网管理器（subnet manager）和聚合通讯控制器（collective controller）。聚合通讯控制器可以协同子网管理器监控整个聚合通讯的生命周期，负责拓扑变化感知、异常监控及资源分配和回收。聚合通讯控制器可以是软件模块，也可以是具有上述功能的硬件模块。交换节点（具体可以是交换机）包括聚合通讯代理（collective agent），collective agent 可以是软件模块，也可以是硬件模块，该代理用于实现业务流程和控制流程。计算节点包括参与聚合通讯的进程（具体可以是 MPI process）和聚合通讯库（collective library）。聚合通讯库提供有运行时的控制面及数据面接口，用于实现与聚合通讯代理和聚合通讯控制器的交互。

控制节点 104（具体是控制节点 104 中的聚合通讯控制器）可以读取配置文件，从而确定网络的物理拓扑，例如确定网络的物理拓扑为 dragonfly+。然后控制节点 104 根据上述物理拓扑生成一个聚合树的拓扑结构，用于数据聚合。其中，控制节点 104 为树的根节点，不同组中互连的顶节点为同一棵聚合树的二级节点。控制节点 104 根据该聚合树的拓扑结构，通知实际的交换机进行链接的建立，并通知顶节点为 dragonfly+拓扑。然后控制节点 104 通知上述顶节点建立全连接，具体是将顶节点间所有物理链路建立连接。

作为叶子节点 1026 的计算节点与控制节点 104 进行通信，得到其直连的交换机（具体是交换机中聚合通讯代理）的地址，然后计算节点向直连的交换机发送加入域请求。加入域请求用于请求加入通信域。加入域请求中包括唯一的域标识，进一步地，加入域请求中还可以包括域大小。域大小可以通过通信域中节点的数量表征。

中间节点 1024 接收到上述加入域请求，记录发送加入域请求的子节点的信息，如子节点的节点标识，如此，在业务流程中，中间节点 1024 可以对来自子节点（如叶子节点 1026）的数据进行聚合。然后，中间节点 1024 继续向父节点发送加入域请求。

当顶节点 1022 接收到加入域请求，记录发送加入域请求的子节点的信息，然后向控制节点 104 发送加入域请求。控制节点 104 接收到所有加入域请求后，可以获得所有顶节点的信息，然后控制节点 104 可以发送加入域响应，加入域响应用于表征顶节点 1022 成功加入通信域。顶节点 1022 接收到加入域响应，可以标记通信域通过全互连方式进行通信。然后，顶节点 1022 可以向顶节点 1022 的子节点回复加入域响应。以此类推，子节点继续向子节点的子节点回复加入域响应，直至叶子节点 1026 收到加入域响应。

在一些可能的实现方式中，还可以将交换节点作为控制节点 104。该控制节点 104 为虚拟的控制节点，即从交换节点中虚拟出具有相应功能的控制节点。具体实现时，可以将一个顶节点 1022 虚拟为控制节点 104。

在一些可能的实现方式中，顶节点 202 无需借助根节点也可以实现加入通信域。下面以第一顶节点加入通信域的过程进行示例说明。

具体地，第一顶节点接收第三顶节点发送的加入域请求，当加入域请求中的域标识包括于第一顶节点的域标识列表时，第一顶节点将该第一顶节点的节点标识添加至对应通信域的节点标识列表。

其中，第三顶节点可以是第二顶节点中右连或左连第一顶节点的节点。以第三顶节点为左连第一顶节点的节点为例进行说明。每个顶节点 1022 接收左连的顶节点 1022 发送的加入域请求，若加入域请求中的域标识在该顶节点的域标识列表中，则将该顶节点 1022 的节点标识添加至对应通信域的节点标识列表，然后转发加入域请求至下一顶节点 1022，否则，直接转发该加入域请求。当某个顶节点接收到自己发送的加入域请求时，丢弃该加入域请求，并停止转发。

在一些实施例中，第三顶节点也可以是第二顶节点中与所述第一顶节点直连的节点。顶节点 1022 接收加入域请求，若加入域请求中的域标识在该顶节点 1022 的域标识列表中，则将该顶节点 1022 的节点标识添加至通信域的节点标识列表，否则将加入域请求丢弃。当顶节点 1022 收到加入域请求对应的子节点个数总和等于整个通信域中叶子节点个数时，通信域创建完毕，各叶子节点 1026 和各顶节点 1022 均加入通信域。

接下来，将结合分布式模型训练场景对本申请实施例提供的数据处理的方法进行详细说明。

参见图 8 所示的分布式模型训练的场景示意图，如图 8 所示，在进行模型训练时，可以通过蜻蜓网络 100 进行分布式模型训练，一方面可以实现通过多个叶子节点 1026 进行并行训练，提高训练效率，另一方面，可以实现利用多个叶子节点 1026 各自的数据集进行训练，保护数据隐私。

具体地，蜻蜓网络 100 包括多个组 102，每个组 102 包括计算节点和交换节点。其中，每个组 102 中的计算节点为该组中的叶子节点 1026，每个组 102 中与其他组 102 的交换节点进行通信的交换节点为该组 102 的顶节点 1022，每个组 102 中除顶节点 1022 以外的交换节点为该组 102 的中间节点 1024。

每个叶子节点 1026 上包括一个相同的初始子模型，每个叶子节点 1026 分别采用不同的训练样本训练上述初始子模型，得到损失函数的梯度。各个叶子节点 1026 可以通过聚合通讯计算梯度的平均值，以更新初始子模型的权重，得到子模型。

其中，各个叶子节点 1026 通过聚合通讯中的全局归约（all reduce）接口函数计算梯度的平均值。具体地，各个叶子节点 1026 以及参与聚合通讯的顶节点 1022、中间节点 1024 加入通信域后，各个叶子节点 1026 分别向其连接的中间节点 1024 发送该叶子节点 1026 训练初始子模型得到的损失函数的梯度，中间节点 1024 对接收到的梯度进行初步聚合，向父节点（顶节点 1022 或另一中间节点 1024）发送聚合后的梯度。当顶节点 1022 接收到多个子节点发送的聚合后的梯度时，先对子节点发送的多个梯度进行聚合，然后每个顶节点 1022 将聚合后的梯度发送至其他顶节点 1022，如此，每个顶节点 1022 可以根据自身聚合后的梯度以及接收的其他顶节点 202 聚合后的梯度确定梯度和，进而可以确定梯度的平均值。各个顶节点 202 向其子节点返回梯度的平均值，子节点为中间节点 1024 时，中间节点 1024 返回梯度的平均值。各个叶子节点 1026 根据梯度的平均值更新初始子模型，得到子模型，由此实现分布式模型训练。

上述实施例是以计算节点包括一个参与聚合通讯的进程进行示例说明的。在一些可能的

实现方式中，至少一个计算节点可以包括多个参与聚合通讯的进程，基于此，计算节点还可以先在节点内对多个参与聚合通讯的进程的数据进行聚合，得到部分聚合数据。然后计算节点将该部分聚合数据发送至交换节点进行进一步聚合。

其中，计算节点在节点内对多个参与聚合通讯的进程的数据进行聚合，可以是通过计算节点自身的处理器如中央处理器(central processing unit，CPU)实现。在一些可能的实现方式中，计算节点的网卡也包括处理器，计算节点还可以对将多个聚合通讯的进程的数据进行聚合的过程卸载至网卡。

上文结合图 1 至图 8 对本申请实施例提供的数据处理的方法进行了详细介绍，下面将结合附图对本申请实施例提供的装置、设备进行介绍。

参见图 9 所示的数据处理的装置的结构示意图，该装置 900 应用于计算集群，所述计算集群包括多个全互连结构的组，每个组中包括交换节点和计算节点，所述装置 900 为第一组中用于与第二组的第二顶节点进行通信的交换节点，所述第二顶节点所述第二组中用于与所述第一组的所述装置进行通信的交换节点，所述装置 900 包括:

通信模块 902，用于接收子节点发送的第一数据，所述子节点用于指示所述第一组中与所述装置直接相连的节点;

所述通信模块 902，还用于接收所述第二顶节点发送的第二数据;

聚合模块 904，用于对所述第一数据和所述第二数据进行聚合，得到第三数据。

应理解的是，本申请实施例的装置 900 可以通过专用集成电路（application-specific integrated circuit，ASIC）实现，或可编程逻辑器件（programmable logic device，PLD）实现，上述 PLD 可以是复杂程序逻辑器件（complex programmable logical device，CPLD），现场可编程门阵列（field-programmable gate array，FPGA），通用阵列逻辑（generic array logic，GAL）或其任意组合。也可以通过软件实现图 4 所示的数据处理方法时，装置 900 及其各个模块也可以为软件模块。

在一些可能的实现方式中，所述聚合模块 904，还用于：对所述子节点发送的多个第一数据进行聚合；将对所述多个第一数据进行聚合所得的数据与所述第二数据进行聚合，得到第三数据。

在一些可能的实现方式中，所述通信模块 902 还用于：向所述第二顶节点发送对所述多个第一数据进行聚合所得的数据。

在一些可能的实现方式中，所述通信模块 902，还用于通过全互连（all to all）、环形（ring）和递推倍增（recursive doubling）中的任意一种向所述第二顶节点发送数据。

在一些可能的实现方式中，所述子节点包括所述第一组中与所述装置 900 直接相连的交换节点;

所述第一数据包括部分聚合数据，所述第三数据为完全聚合数据。

在一些可能的实现方式中，所述子节点包括所述第一组中与所述装置 900 直接相连的计算节点;

所述第一数据包括非聚合数据，所述第三数据为完全聚合数据。

在一些可能的实现方式中，所述装置 900 还包括:

控制模块，用于在接收子节点发送的第一数据之前，加入通信域。

在一些可能的实现方式中，所述控制模块具体用于:

向控制节点发送加入域请求;

接收所述控制节点发送的加入域响应，所述加入域响应用于表征所述装置成功加入通信域。

在一些可能的实现方式中，所述控制节点为以下任意一种：

独立于所述计算节点和所述交换节点的节点；或，

所述交换节点中的一个节点。

在一些可能的实现方式中，所述控制模块具体用于：

接收第三顶节点发送的加入域请求；

当加入域请求中的域标识包括于所述装置的域标识列表时，将所述装置的标识添加至对应通信域的节点标识列表。

在一些可能的实现方式中，所述第三顶节点为所述第二顶节点中右连或左连所述装置的节点，或者是所述第二顶节点中与所述装置直接相连的节点。

在一些可能的实现方式中，所述计算集群的拓扑为蜻蜓网络拓扑。

根据本申请实施例的数据处理的装置 900 可对应于执行本申请实施例中描述的方法，并且数据处理的装置900的各个模块/单元的上述和其它操作和/或功能分别为了实现图4所示实施例中的各个方法的相应流程，为了简洁，在此不再赘述。

本申请实施例还提供了一种电子设备 1000。该电子设备 1000 是具有数据转发功能的设备。具体地，该电子设备 1000 可以是交换机或者路由器等设备。该电子设备 1000 可以用于实现如图 9 所示实施例中数据处理的装置 900 的功能。

图 10 提供了一种电子设备 1000 的结构示意图，如图 10 所示，电子设备 100 包括总线 1001、处理器 1002、通信接口 1003 和存储器 1004。处理器 1002、存储器 1004 和通信接口 1003 之间通过总线 1001 通信。

总线 1001 可以是外设部件互连标准（peripheral component interconnect，PCI）总线或扩展工业标准结构（extended industry standard architecture，EISA）总线等。总线可以分为地址总线、数据总线、控制总线等。为便于表示，图 10 中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

处理器 1002 可以为中央处理器(central processing unit，CPU)。存储器 1004 可以包括易失性存储器(volatile memory)，例如随机存取存储器(random access memory，RAM)。存储器 1004 还可以包括非易失性存储器(non-volatile memory)，例如只读存储器(read-only memory，ROM)，快闪存储器，硬盘驱动器（hard disk drive，HDD）或固态驱动器（solid state drive，SSD）。

通信接口 1003 用于与外部通信。例如，接收子节点发送的第一数据，接收第二顶节点发送的第二数据，向第二顶节点发送对多个第一数据进行聚合所得的数据等等。

存储器 1004 中存储有可执行代码，处理器 1002 执行该可执行代码以执行前述数据处理的方法。

具体地，在实现图 9 所示实施例的情况下，且图 9 实施例中所描述的数据处理的装置 900 的各模块为通过软件实现的情况下，执行图 9 中的聚合模块 904 功能所需的软件或程序代码存储在存储器 1004 中。

通信模块 902 功能通过通信接口 1003 实现。通信接口 1003 接收子节点发送的第一数据，以及接收第二顶节点发送的第二数据，将第一数据、第二数据通过总线 1001 传输至处理器 1002，处理器 1002 执行存储器 1004 中存储的各模块对应的程序代码，如执行聚合模块 904 对应的程序代码，以执行对第一数据和第二数据进行聚合得到第三数据的步骤。

应理解，本申请实施例的电子设备1000可对应于本申请实施例中的图9所述的数据处理的装置900，电子设备1000用于实现上述图4所述方法中相应主体执行的方法的操作步骤，为了简洁，在此不再赘述。

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

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

以上所述，仅为本申请的具体实施方式。熟悉本技术领域的技术人员根据本申请提供的具体实施方式，可想到变化或替换，都应涵盖在本申请的保护范围之内。

权 利 要 求 书

1、一种数据处理的方法，其特征在于，应用于计算集群，所述计算集群包括多个全互连结构的组，每个组中包括交换节点和计算节点，所述方法包括：

第一顶节点接收子节点发送的第一数据，所述第一顶节点为所述第一顶节点所在第一组中用于与第二组的第二顶节点进行通信的交换节点，所述子节点用于指示所述第一组中与所述第一顶节点直接相连的节点，所述第二顶节点为所述第二组中用于与所述第一组的所述第一顶节点进行通信的交换节点；

所述第一顶节点接收所述第二顶节点发送的第二数据；

所述第一顶节点对所述第一数据和所述第二数据进行聚合，得到第三数据。

2、根据权利要求 1 所述的方法，其特征在于，所述第一顶节点对所述第一数据和所述第二数据进行聚合，得到第三数据，包括：

所述第一顶节点对所述子节点发送的多个第一数据进行聚合；

所述第一顶节点将对所述多个第一数据进行聚合所得的数据与所述第二数据进行聚合，得到第三数据。

3、根据权利要求 2 所述的方法，其特征在于，所述方法还包括：

所述第一顶节点向所述第二顶节点发送对所述多个第一数据进行聚合所得的数据。

4、根据权利要求 3 所述的方法，其特征在于，所述第一顶节点通过以下方式中的任意一种向所述第二顶节点发送数据：

全互连（all to all）、环形（ring）和递推倍增（recursive doubling）。

5、根据权利要求 1 至 4 任一项所述的方法，其特征在于，所述子节点包括所述第一组中与所述第一顶节点直接相连的交换节点；

所述第一数据包括部分聚合数据，所述第三数据为完全聚合数据。

6、根据权利要求 1 至 5 任一项所述的方法，其特征在于，所述子节点包括所述第一组中与所述第一顶节点直接相连的计算节点；

所述第一数据包括非聚合数据，所述第三数据为完全聚合数据。

7、根据权利要求 1 至 6 任一项所述的方法，其特征在于，在接收子节点发送的第一数据之前，所述方法还包括：

所述第一顶节点加入通信域。

8、根据权利要求 7 所述的方法，其特征在于，所述第一顶节点加入通信域，包括：

所述第一顶节点向控制节点发送加入域请求；

所述第一顶节点接收所述控制节点发送的加入域响应，所述加入域响应用于表征所述第一顶节点成功加入通信域。

9、根据权利要求 8 所述的方法，其特征在于，所述控制节点为以下任意一种：

独立于所述计算节点和所述交换节点的节点；或，

所述交换节点中的一个节点。

10、根据权利要求 7 所述的方法，其特征在于，所述第一顶节点加入通信域，包括：

所述第一顶节点接收第三顶节点发送的加入域请求；

当加入域请求中的域标识包括于所述第一顶节点的域标识列表时，所述第一顶节点将第一顶节点的节点标识添加至对应通信域的节点标识列表。

11、根据权利要求 10 所述的方法，其特征在于，所述第三顶节点为所述第二顶节点中右连或左连所述第一顶节点的节点，或者是所述第二顶节点中与所述第一顶节点直接相连的节点。

12、根据权利要求 1 至 11 任一项所述的方法，其特征在于，所述计算集群的拓扑为蜻蜓网络拓扑。

13、一种数据处理的装置，其特征在于，应用于计算集群，所述计算集群包括多个全互连结构的组，每个组中包括交换节点和计算节点，所述装置为第一组中用于与第二组的第二顶节点进行通信的交换节点，所述第二顶节点所述第二组中用于与所述第一组的所述装置进行通信的交换节点，所述装置包括：

通信模块，用于接收子节点发送的第一数据，所述子节点用于指示所述第一组中与所述装置直接相连的节点；

所述通信模块，还用于接收所述第二顶节点发送的第二数据；

聚合模块，用于对所述第一数据和所述第二数据进行聚合，得到第三数据。

14、根据权利要求 13 所述的装置，其特征在于，所述聚合模块具体用于：

对所述子节点发送的多个第一数据进行聚合；

将对所述多个第一数据进行聚合所得的数据与所述第二数据进行聚合，得到第三数据。

15、根据权利要求 14 所述的装置，其特征在于，所述通信模块还用于：

向所述第二顶节点发送对所述多个第一数据进行聚合所得的数据。

16、根据权利要求 15 所述的装置，其特征在于，所述通信模块具体用于:

通过全互连（all to all）、环形（ring）和递推倍增（recursive doubling）中的任意一种向所述第二顶节点发送数据。

17、根据权利要求 13 至 16 任一项所述的装置，其特征在于，所述子节点包括所述第一组中与所述装置直接相连的交换节点;

所述第一数据包括部分聚合数据，所述第三数据为完全聚合数据。

18、根据权利要求 13 至 17 任一项所述的装置，其特征在于，所述子节点包括所述第一组中与所述装置直接相连的计算节点;

所述第一数据包括非聚合数据，所述第三数据为完全聚合数据。

19、根据权利要求 13 至 18 任一项所述的装置，其特征在于，所述装置还包括:

控制模块，用于在接收子节点发送的第一数据之前，加入通信域。

20、根据权利要求 19 所述的装置，其特征在于，所述控制模块具体用于:

向控制节点发送加入域请求;

接收所述控制节点发送的加入域响应，所述加入域响应用于表征所述装置成功加入通信域。

21、根据权利要求 20 所述的装置，其特征在于，所述控制节点为以下任意一种:

独立于所述计算节点和所述交换节点的节点；或，

所述交换节点中的一个节点。

22、根据权利要求 19 所述的装置，其特征在于，所述控制模块具体用于:

接收第三顶节点发送的加入域请求;

当加入域请求中的域标识包括于所述装置的域标识列表时，将所述装置的标识添加至对应通信域的节点标识列表。

23、根据权利要求 22 所述的装置，其特征在于，所述第三顶节点为所述第二顶节点中右连或左连所述装置的节点，或者是所述第二顶节点中与所述装置直接相连的节点。

24、根据权利要求 13 至 23 任一项所述的装置，其特征在于，所述计算集群的拓扑为蜻蜓网络拓扑。

25、一种电子设备，其特征在于，所述电子设备包括处理器和存储器;

所述处理器用于执行所述存储器中存储的指令，以使得所述电子设备执行如权利要求 1 至 12 中任一项所述的方法。

26、一种计算机可读存储介质，其特征在于，包括指令，所述指令指示电子设备执行如权利要求 1 至 12 中任一项所述的方法。

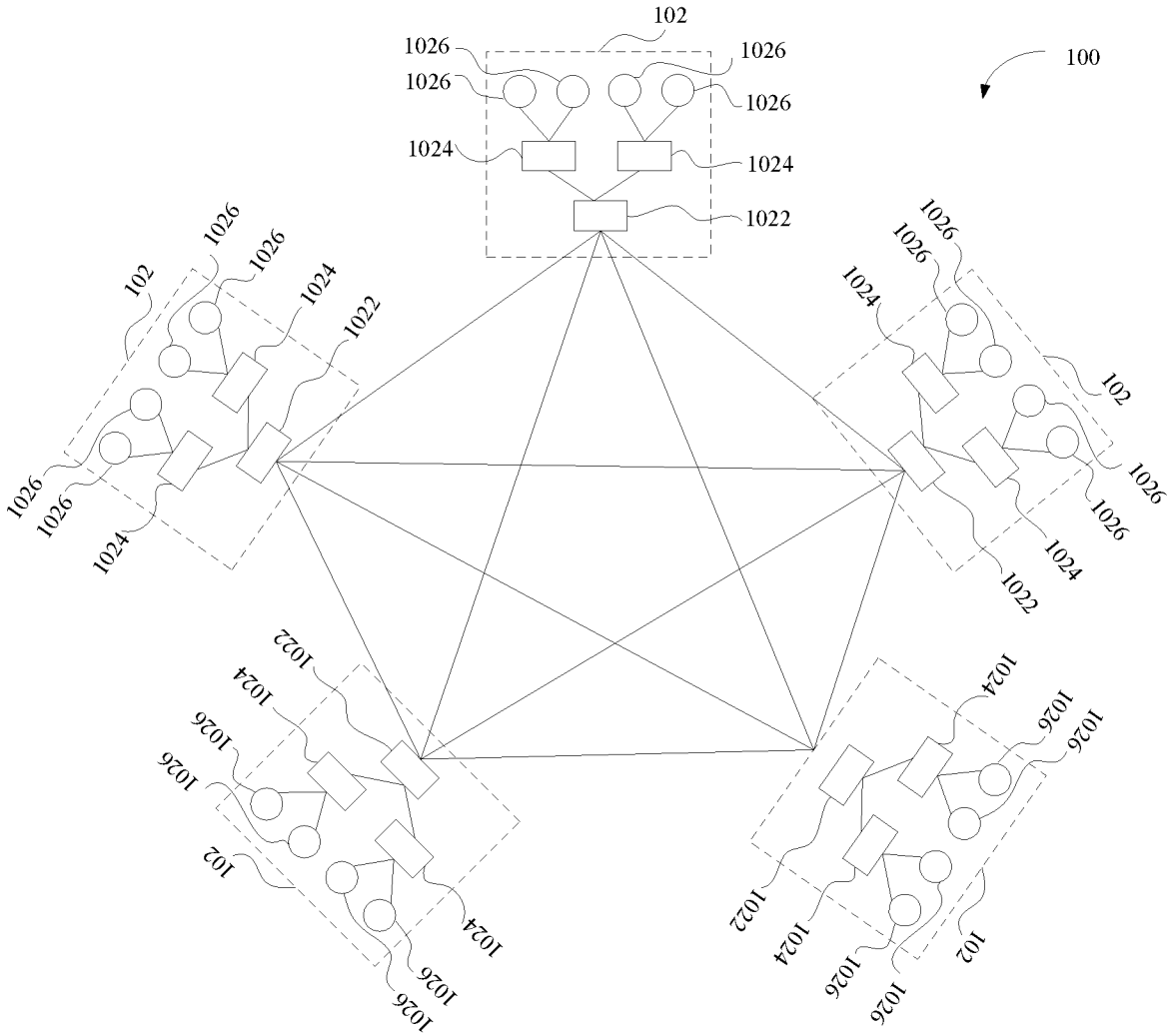

图 1

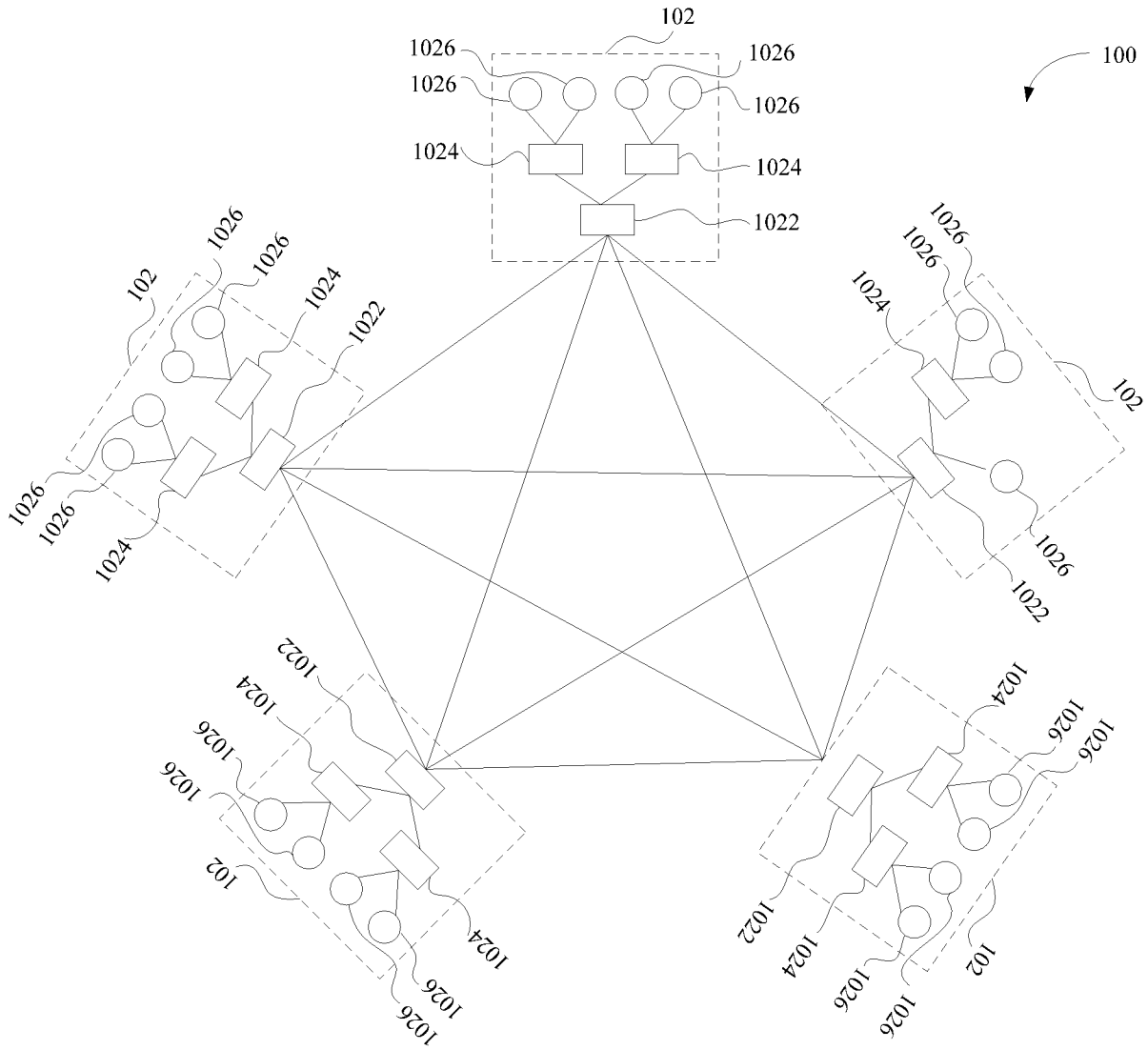
102
1026
1026
1026
1026
1024
1024
1022
100
102
1026
1026
1024
1022
1026
1026
1024
102
1026
1026
1024
1022
1026
1026
1024
102
1024
1026
1026
1022
1024
1026
1026
102

图 2

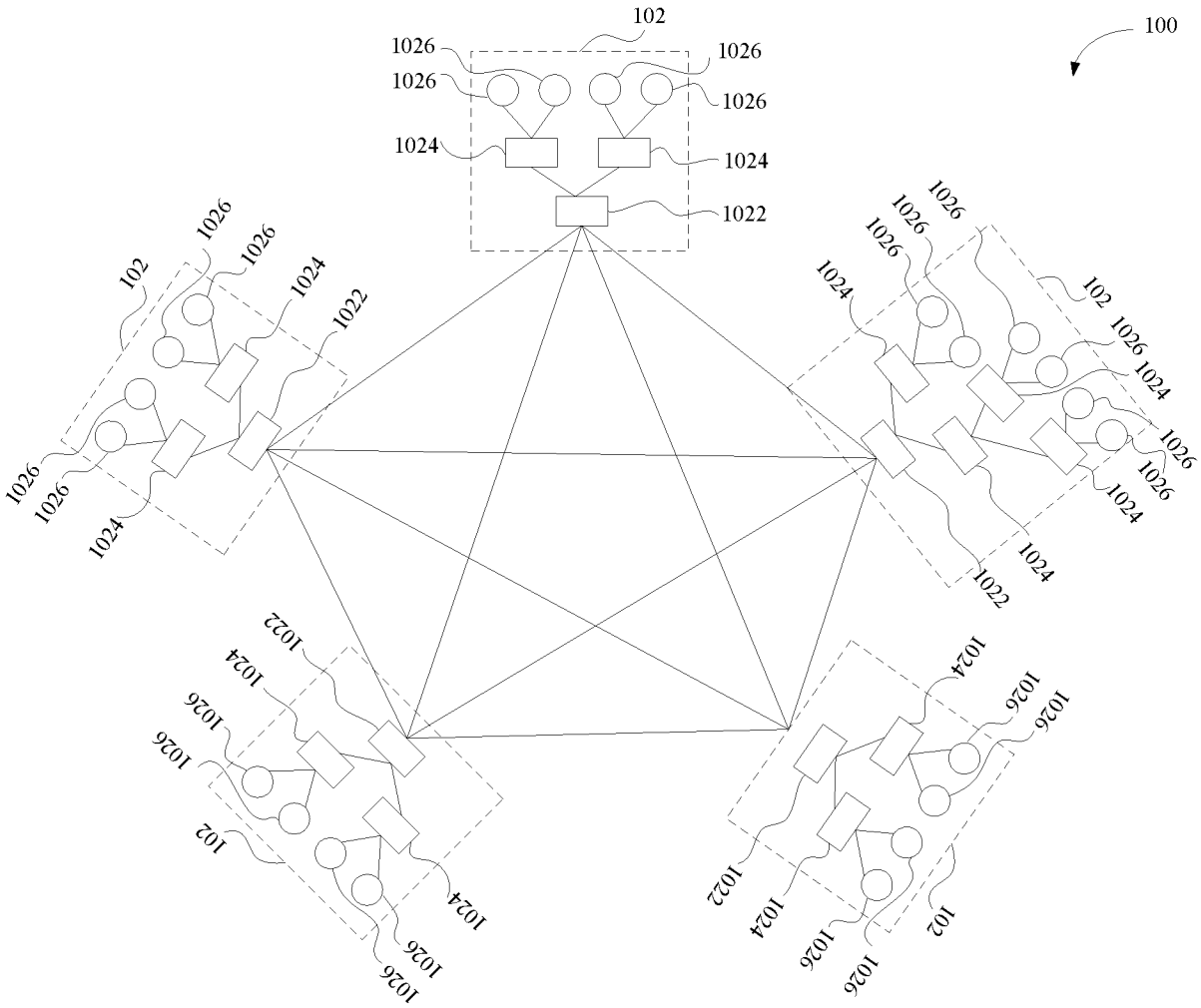
100
102
1026
1024
1022

图 3

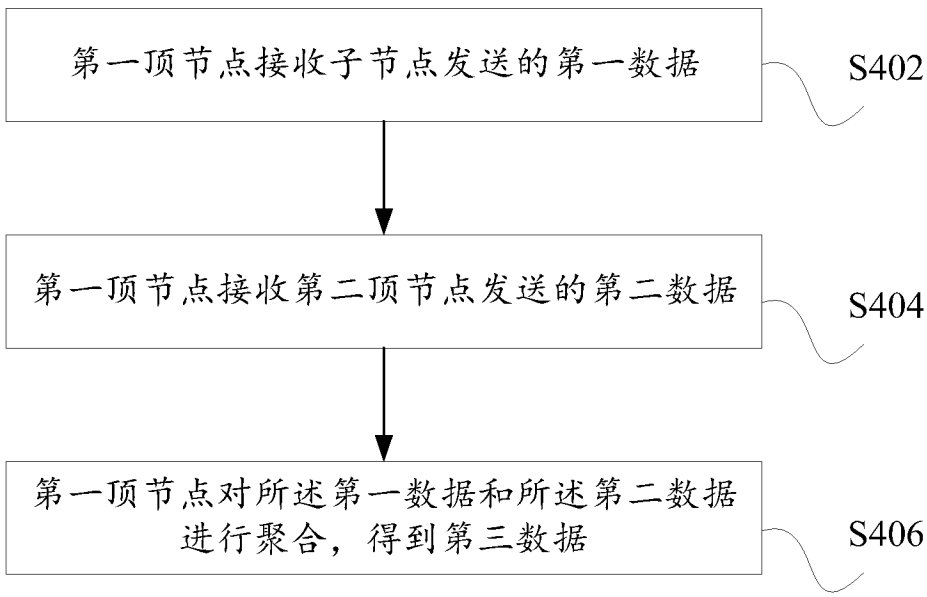
第一顶节点接收子节点发送的第一数据
S402
第一顶节点接收第二顶节点发送的第二数据
S404
第一顶节点对所述第一数据和所述第二数据进行聚合，得到第三数据
S406

图 4

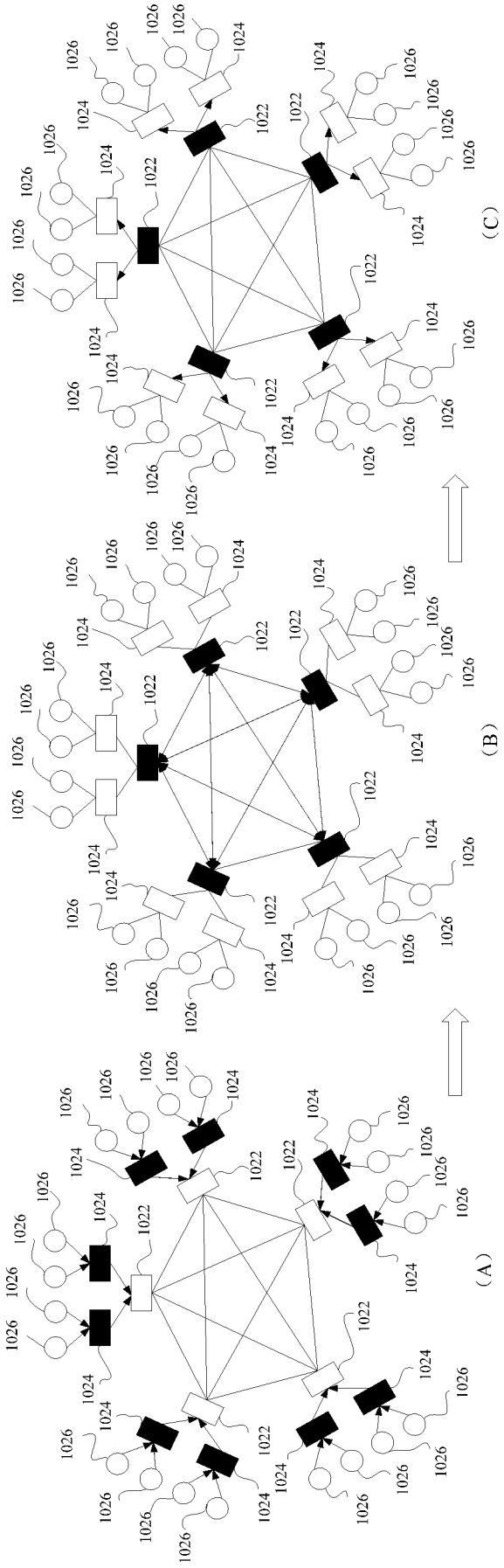

图5

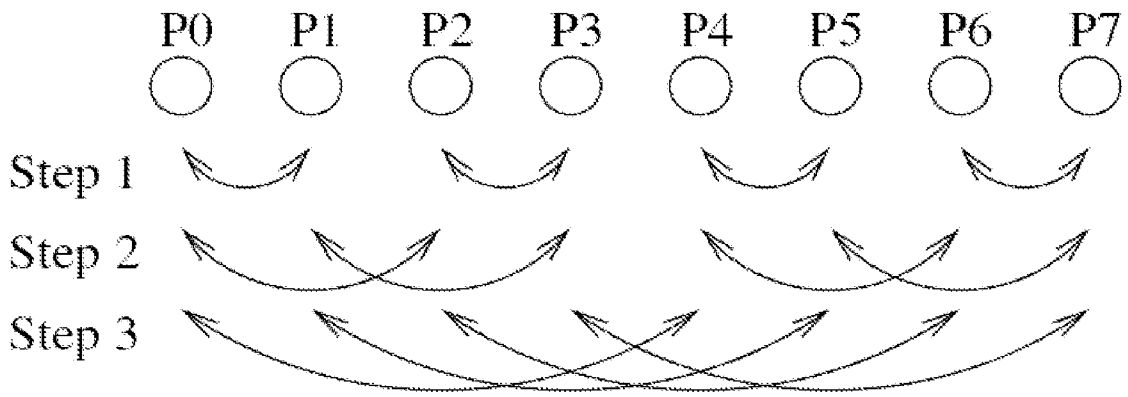

图 6

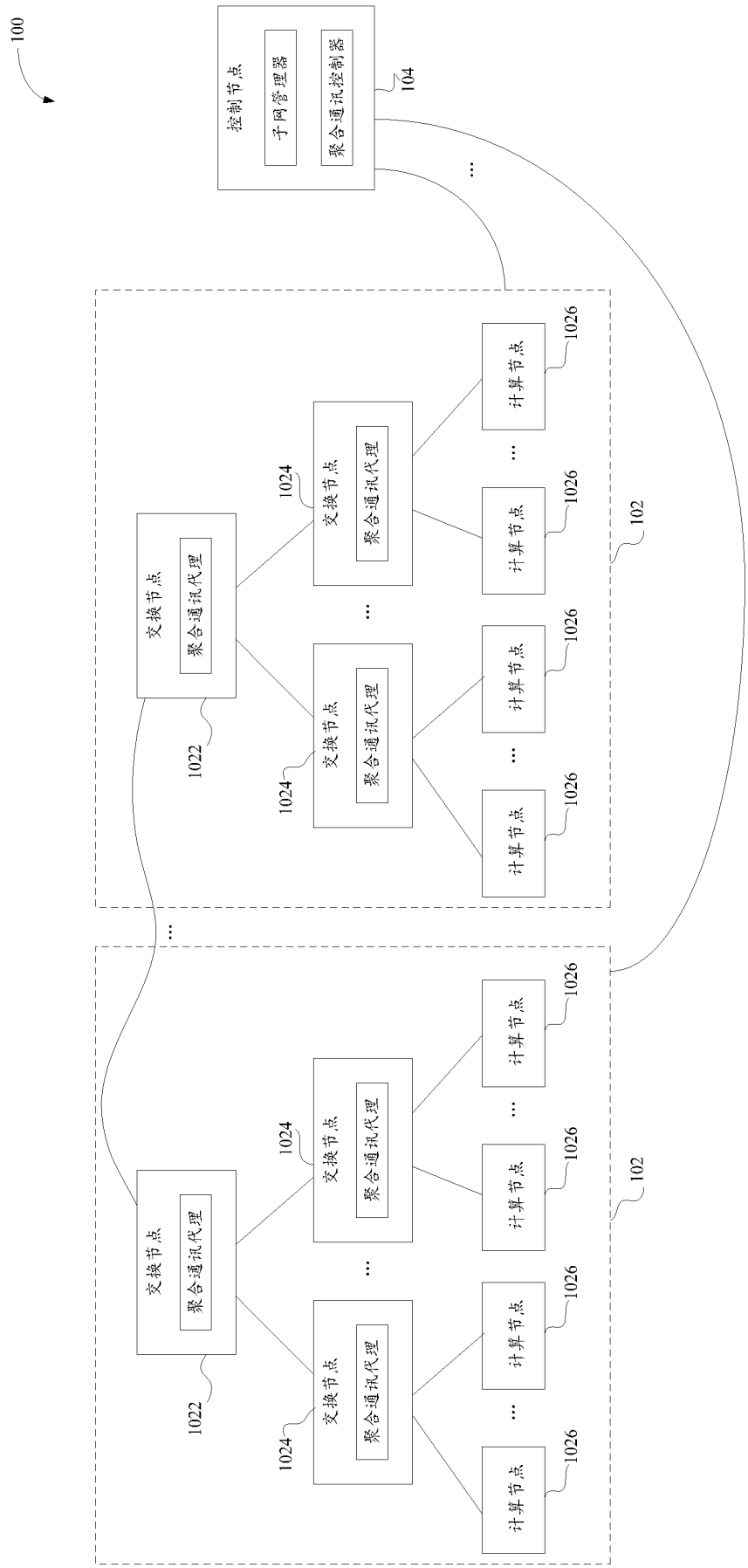
100
控制节点
子网管理器
聚合通讯控制器
104
102
交换节点
聚合通讯代理
1022
1024
1026
计算节点

图7

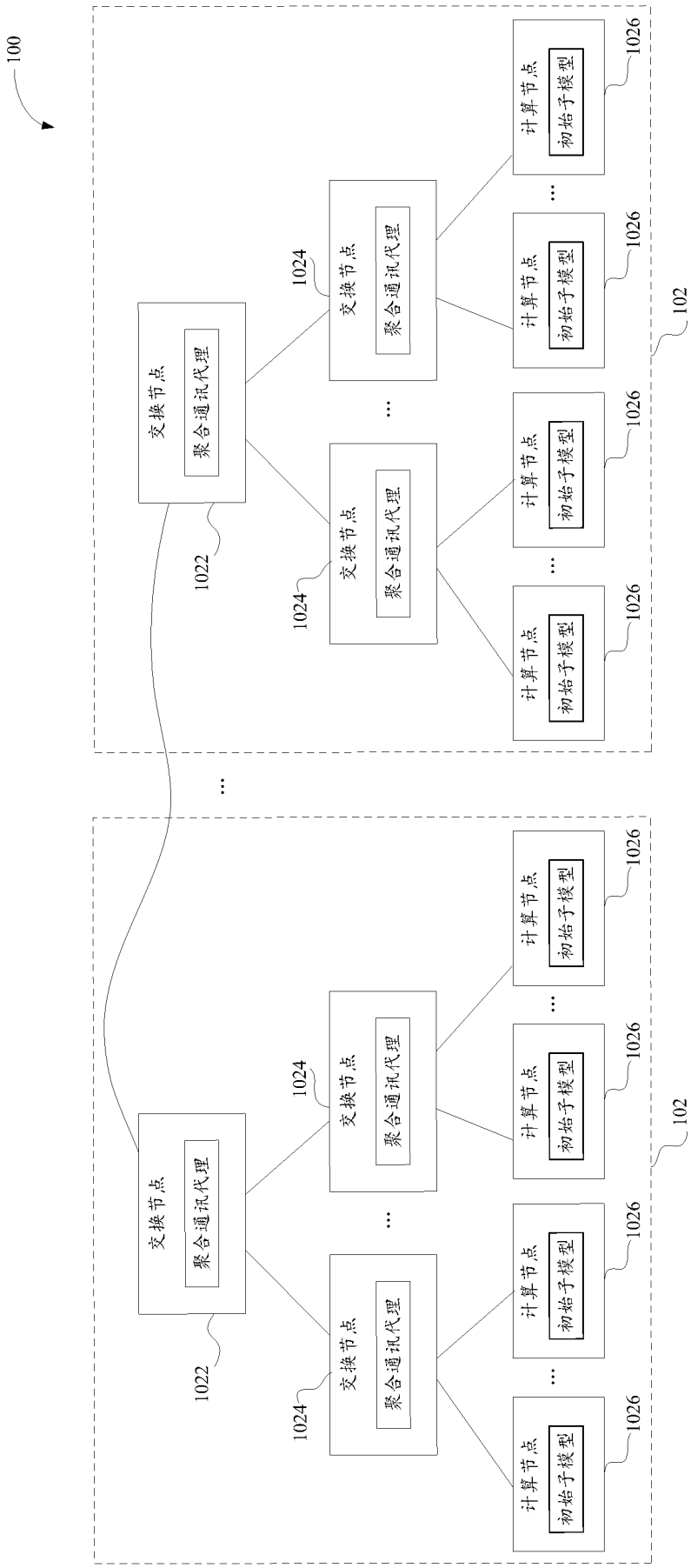

图8

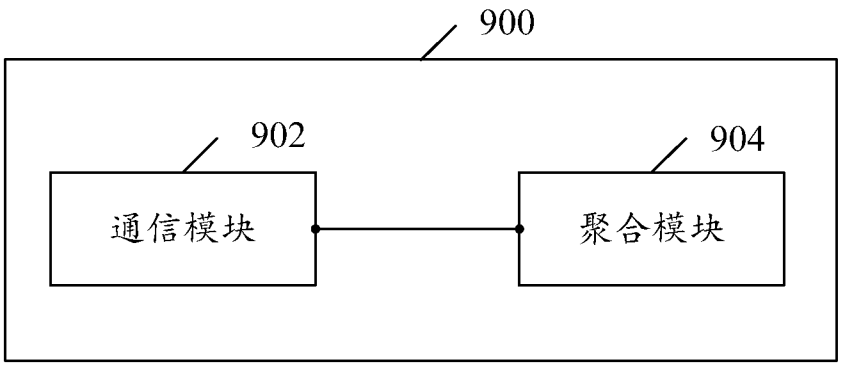
900
902
904
通信模块
聚合模块

图 9

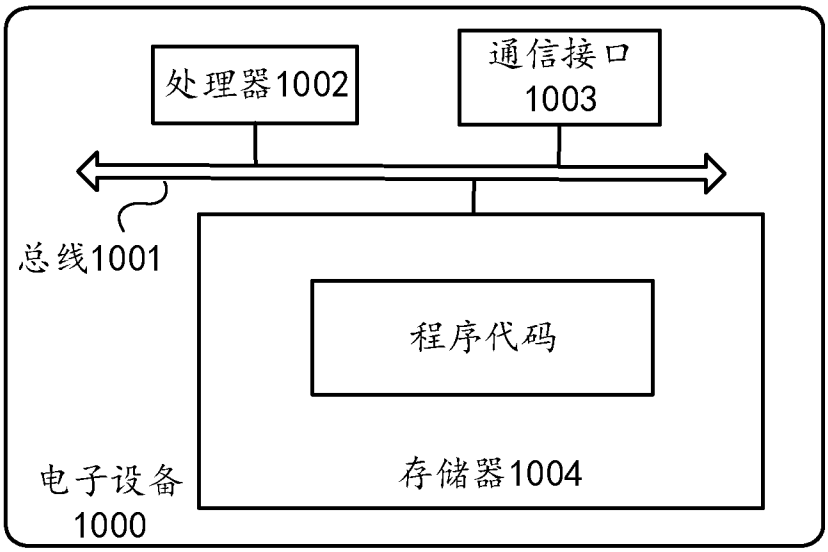
处理器1002
通信接口
1003
总线1001
程序代码
存储器1004
电子设备
1000

图 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/106804

A. CLASSIFICATION OF SUBJECT MATTER

H04L 1/00(2006.01)i; H04L 12/24(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)

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)

CNKI, CNPAT, WPI, EPODOC: 数据, 组, 全连通, 全互连, 聚合, 蜻蜓网络, data, group, all?to?all, fuse, dragonfly network

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 104009907 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) 27 August 2014 (2014-08-27) description, paragraphs [0012]-[0072], and figures 2-9	1-26
Y	CN 101035040 A (NANJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 12 September 2007 (2007-09-12) description, pages 10-11	1-26
A	CN 111245747 A (HUAWEI TECHNOLOGIES CO., LTD.) 05 June 2020 (2020-06-05) entire document	1-26
A	US 2016301602 A1 (AMAZON TECHNOLOGIES, INC.) 13 October 2016 (2016-10-13) entire document	1-26

☐ 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
29 September 2021	**14 October 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/CN2021/106804

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	104009907	A	27 August 2014	GB	201303181	D0	10 April 2013
				JP	2014164756	A	08 September 2014
				US	2014245324	A1	28 August 2014
				US	2016239356	A1	18 August 2016
				US	2017220398	A1	03 August 2017
				GB	2511089	A	27 August 2014
CN	101035040	A	12 September 2007	None			
CN	111245747	A	05 June 2020	EP	3573313	A1	27 November 2019
				US	2019230050	A1	25 July 2019
				JP	2018501703	A	18 January 2018
				WO	2017101114	A1	22 June 2017
				US	2018262446	A1	13 September 2018
				CN	107211036	A	26 September 2017
				EP	3211858	A1	30 August 2017
US	2016301602	A1	13 October 2016	US	2015236980	A1	20 August 2015

国际检索报告

国际申请号
PCT/CN2021/106804

A. 主题的分类

H04L 1/00(2006.01)i; H04L 12/24(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)

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）)

CNKI,CNPAT,WPI,EPODOC:数据，组，全连通，全互连，聚合，蜻蜓网络，data，group，all?to?all，fuse，dragonfly network

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 104009907 A (国际商业机器公司) 2014年 8月 27日 (2014-08-27) 说明书第[0012]-[0072]段、图2-9	1-26
Y	CN 101035040 A (南京邮电大学) 2007年 9月 12日 (2007-09-12) 说明书第10-11页	1-26
A	CN 111245747 A (华为技术有限公司) 2020年 6月 5日 (2020-06-05) 全文	1-26
A	US 2016301602 A1 (AMAZON TECHNOLOGIES, INC.) 2016年 10月 13日 (2016-10-13) 全文	1-26

☐ 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
2021年 9月 29日	2021年 10月 14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 黄菲 电话号码 86-(10)-53961743

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/106804

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	104009907	A	2014年 8月 27日	GB	201303181	D0	2013年 4月 10日
				JP	2014164756	A	2014年 9月 8日
				US	2014245324	A1	2014年 8月 28日
				US	2016239356	A1	2016年 8月 18日
				US	2017220398	A1	2017年 8月 3日
				GB	2511089	A	2014年 8月 27日
CN	101035040	A	2007年 9月 12日		无		
CN	111245747	A	2020年 6月 5日	EP	3573313	A1	2019年 11月 27日
				US	2019230050	A1	2019年 7月 25日
				JP	2018501703	A	2018年 1月 18日
				WO	2017101114	A1	2017年 6月 22日
				US	2018262446	A1	2018年 9月 13日
				CN	107211036	A	2017年 9月 26日
				EP	3211858	A1	2017年 8月 30日
US	2016301602	A1	2016年 10月 13日	US	2015236980	A1	2015年 8月 20日