

(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
2017 年 8 月 17 日 (17.08.2017)



(10) 国际公布号
WO 2017/136999 A1

- (51) 国际专利分类号:
H04L 29/08 (2006.01)
- (21) 国际申请号: PCT/CN2017/072818
- (22) 国际申请日: 2017 年 2 月 3 日 (03.02.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201610084419.9 2016 年 2 月 14 日 (14.02.2016) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 吕志强 (LV, Zhiqiang); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 陶阳宇 (TAO, Yangyu); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 陆一峰 (LU, Yifeng); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 李超 (LI, Chao); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 吴永军 (WU, Yongjun); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang

311121 (CN)。 李治 (LI, Zhi); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 刘耀莉 (LIU, Yaoli); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (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, 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, 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,

[见续页]

(54) Title: INTERMEDIATE DATA TRANSMISSION METHOD, SYSTEM, AND DISTRIBUTED SYSTEM

(54) 发明名称: 中间数据传输方法及系统、分布式系统

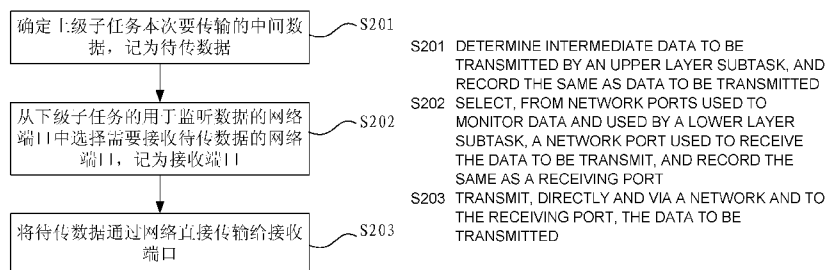


图 2

(57) Abstract: The invention relates to an intermediate data transmission method, system, and distributed system. The intermediate data transmission method comprises: determining intermediate data to be transmitted by an upper layer subtask, and recording the same as data to be transmitted; selecting, from network ports used to monitor data and used by a lower layer subtask, a network port used to receive the data to be transmitted, and recording the same as a receiving port; and transmitting, directly and via a network and to the receiving port, the data to be transmitted. In the embodiment of the present invention, the transmission of the intermediate data does not go through a distributed system, preventing the transmission speed of the intermediate data from being affected by hard disk IO speed, and increasing the transmission speed of the intermediate data.

(57) 摘要: 本申请涉及一种中间数据传输方法及系统、分布式系统。其中, 中间数据传输方法包括: 确定所述上级任务本次要传输的中间数据, 记为待传数据; 从所述下级任务的用于监听数据的网络端口中选择需要接收所述待传数据的网络端口, 记为接收端口; 将所述待传数据通过网络直接传输给所述接收端口。本发明实施例中, 中间数据的传输不需要经过分布式存储系统, 避免了磁盘 IO 速率对中间数据传输速率的影响, 提高了中间数据传输速率。



WO 2017/136999 A1



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))。

中间数据传输方法及系统、分布式系统

本申请要求 2016 年 02 月 14 日递交的申请号为 201610084419.9、发明名称为“中间数据传输方法及系统、分布式系统”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5

技术领域

本发明涉及通信领域，尤其涉及一种中间数据传输方法及系统、分布式系统。

背景技术

10 在分布式系统中，用户任务通常可以被分解为几级不同的子任务。这些子任务之间存在依赖关系，一些子任务的输出结果会作为另一些子任务的输入数据。这就涉及到了子任务之间的数据传递，这些在子任务之间传递的数据称为中间数据。

目前，在相关技术中，中间数据的传输是通过分布式存储系统来完成的，其过程是：产生中间数据的上级子任务通过分布式存储系统的接口将中间数据以文件的形式写到分布式存储系统的磁盘上；需要将该中间数据作为输入数据的下级子任务从分布式存储系统的磁盘中读取该中间数据，以做进一步处理。

以最常见的 MapReduce 任务为例。MapReduce 任务可以被分解为 MapTask 和 ReduceTask 这两级子任务。当通过 MapReduce 任务完成数据排序时，MapTask 可以在不同的机器上同时启动若干个进程，每个进程分别读取一部分输入数据，并对这部分数据进行排序，然后将排序的结果输出。这些输出数据会作为 ReduceTask 的输入，做进一步排序，以达到全局有序。这一过程中，在 MapTask 和 ReduceTask 这两级子任务之间传递的数据是排序过程中产生的中间计算结果，是不需要被呈现给用户的。

图 1 为相关技术中 MapTask 和 ReduceTask 之间的中间数据传输过程示意图。如图 1 所示，相关技术中，MapTask 和 ReduceTask 之间的中间数据传输过程如下：

25 MapTask 输出中间数据，通过分布式存储系统的用户接口将中间数据以文件形式写入分布式存储系统，分布式存储系统将中间数据存储到一个或多个存储节点中，即将这些中间数据持久化到一台或多台机器的磁盘上；

ReduceTask 通过分布式存储系统的用户接口，从存储中间数据的存储节点中读取中间数据。

相关技术中，分布式系统的中间数据传输存在如下问题：

1、中间数据传输需要经过磁盘的读写来实现，而传统机械硬盘的平均读写速率只能达到 100MB/s 左右。因此，这种中间数据传输方式严重受到磁盘 IO（输入输出）速率的影响，传输速率低，导致用户任务执行效率低下。

5 2、分布式存储系统中可能会出现单个存储节点不可用的情况，为了保证数据不会丢失，分布式存储系统通常会为一个文件生成多份拷贝，并存放到不同的存储节点上，这一过程造成了同样的数据会在网络中进行多次传输，占用网络带宽。

发明内容

10 本发明的目的在于提供一种中间数据传输方法及系统、分布式系统，提高分布式系统的中间数据传输速率。

为实现上述目的，本发明提出了一种中间数据传输方法，用于分布式系统，所述分布式系统的用户任务包括多级子任务，所述子任务中产生中间数据的子任务称为上级子任务，所述子任务中依赖所述中间数据进行处理的子任务称为下级子任务，所述方法包
15 括：

确定所述上级子任务本次要传输的中间数据，记为待传数据；

从所述下级子任务的用于监听数据的网络端口中选择需要接收所述待传数据的网络端口，记为接收端口；

将所述待传数据通过网络直接传输给所述接收端口。

20 本发明实施例中的中间数据传输方法，中间数据的传输不需要经过分布式存储系统，避免了磁盘 IO 速率对中间数据传输速率的影响，提高了中间数据传输速率。

为实现上述目的，本发明还提出了一种中间数据传输系统，用于分布式系统，所述分布式系统的用户任务包括多级子任务，所述子任务中产生中间数据的子任务称为上级子任务，所述子任务中依赖所述中间数据进行处理的子任务称为下级子任务，所述中间
25 数据传输系统包括：

确定模块，用于确定所述上级子任务本次要传输的中间数据，记为待传数据；

选择模块，用于从所述下级子任务的用于监听数据的网络端口中选择需要接收所述待传数据的网络端口，记为接收端口；

30 传输模块，用于将所述确定模块确定的待传数据通过网络直接传输给所述选择模块选择的接收端口。

本发明实施例中的中间数据传输系统，中间数据的传输不需要经过分布式存储系统，避免了磁盘 IO 速率对中间数据传输速率的影响，提高了中间数据传输速率。

为实现上述目的，本发明还提出了一种分布式系统，包括前述任一项所述的中间数据传输系统。

- 5 本发明实施例中的分布式系统，中间数据的传输不需要经过分布式存储系统，避免了磁盘 IO 速率对中间数据传输速率的影响，提高了中间数据传输速率。

附图说明

图 1 为相关技术中 MapTask 和 ReduceTask 之间的中间数据传输过程示意图；

- 10 图 2 为本发明实施例一中中间数据传输方法的流程图；

图 3 为根据图 2 所示的中间数据传输方法进行传输时 MapTask 和 ReduceTask 之间的中间数据传输过程示意图之一；

图 4 为根据图 2 所示的中间数据传输方法进行传输时 MapTask 和 ReduceTask 之间的中间数据传输过程示意图之二；

- 15 图 5 为本发明实施例二中中间数据传输方法的流程图；

图 6 为本发明实施例三中中间数据传输方法的流程图；

图 7 为本发明实施例四中中间数据传输方法的流程图；

图 8 为本发明实施例五中中间数据传输系统的结构框图；

图 9 为本发明实施例六中中间数据传输系统的结构框图；

- 20 图 10 为本发明实施例七中中间数据传输系统的传输模块的结构框图；

图 11 为本发明实施例八中中间数据传输系统的传输模块的结构框图；

图 12 为本发明实施例九中分布式系统的结构框图。

具体实施方式

- 25 以下结合附图对本发明的原理和特征进行描述，所举实施例只用于解释本发明，并非用于限定本发明的范围。对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，根据本发明精神所获得的所有实施例，都属于本发明的保护范围。

- 需要说明的是，本发明各实施例中的中间数据传输方法和中间数据传输系统均用于分布式系统，该分布式系统的用户任务包括多级子任务，其中，将这些子任务中产生中间数据的子任务称为上级子任务，将这些子任务中依赖中间数据进行处理的子任务称为
- 30

下级子任务。上级子任务和下级子任务是相对而言的。例如，在具有多级子任务的用户任务中，一个子任务对于子任务 A 来说是下级子任务，但对于子任务 B 来说是却可能是下级子任务。

其中，用户任务可以是 MapReduce 任务、DAG(Directed Acyclic Graph, 有向无环图)任务等。

图 2 为本发明实施例一中中间数据传输方法的流程图。如图 2 所示，本实施例中，中间数据传输方法可以包括如下步骤：

步骤 S201，确定上级子任务本次要传输的中间数据，记为待传数据；

在上级子任务有多个进程的情况下，上级子任务本次要传输的中间数据（也即待传数据）可以是上级子任务的一个进程产生的中间数据，也可以是上级子任务的多个或全部进程产生的中间数据。

步骤 S202 从下级子任务的用于监听数据的网络端口中选择需要接收待传数据的网络端口，记为接收端口；

下级子任务用于监听数据的网络端口用来接收上级子任务发送的待传数据。

下级子任务可以根据需要设置网络端口。比如，下级子任务可以为每个进程分别设置一个网络端口，也可以为所有进程设置一个或多个共同网络端口。

步骤 S203，将待传数据通过网络直接传输给接收端口。

也就是说，将待传数据由上级子任务所在的网络通过网络直接传输到下任务所在的网络，中间不再像背景技术中提到的相关技术那样经过分布式存储系统的写入和读取过程。

以 MapReduce 任务为例。MapReduce 任务的两级子任务 MapTask 和 ReduceTask 之间的中间数据传输过程如图 3 所示。

在此基础上，用户可以根据需要制定具体的传输策略。

例如，传输策略之一可以是：为上级子任务的每个进程分别设置一个第一网络端口，为下级子任务的每个进程分别设置一个第二网络端口，将待传数据由产生待传数据的一个或多个进程对应的第一网络端口分别传输到所有的第二网络端口。也即，将中间数据由产生中间数据的进程直接发送给下级子任务的所有进程。仍以 MapReduce 任务为例。按照此传输策略，MapReduce 任务的两级子任务 MapTask 和 ReduceTask 之间的中间数据传递过程如图 4 所示。

传输策略之二可以是：为上级子任务的所有进程设置一个或多个第一共同网络端口，

为下级子任务的所有进程设置一个或多个第二共同网络端口，将上级子任务的各个进程产生的中间数据由第一共同网络端口通过网络直接传输到第二共同网络端口，第二共同网络端口再将中间数据分发给下级子任务的各个进程。

5 本发明实施例中的中间数据传输方法，将分布式系统中的中间数据由产生该中间数据的上级子任务通过网络直接传输给下级子任务，中间不需要经过分布式存储系统，避免了磁盘 IO 速率对中间数据传输速率的影响，提高了中间数据传输速率。

图 5 为本发明实施例二中中间数据传输方法的流程图。如图 5 所示，本实施例中，中间数据传输方法可以包括如下步骤：

10 步骤 S501，启动下级子任务，使下级子任务的用于监听数据的网络端口处于监听状态；

由于本发明实施例中，中间数据是由上级子任务通过网络直接传输给下级子任务的，因此在启动上级子任务产生中间数据前，下级子任务应处于能够接收中间数据的状态。也就是说，下级子任务应当先于上级子任务启动，这样才能保证下级子任务能够接收到上级子任务产生的中间数据。

15 步骤 S502，将下级子任务的用于监听数据的网络端口的信息通知给上级子任务；

网络端口的信息一般可以包括 IP 地址等。上级子任务可以根据下级子任务的用于监听数据的网络端口的信息确定中间数据的目的地址。

在具体应用中，下级子任务可以将网络端口的信息先上报给调度器，调度器再将下级子任务的用于监听数据的网络端口信息发送给上级子任务。

20 步骤 S503，在下级子任务启动之后启动上级子任务，产生中间数据；

在下级子任务启动后，下级子任务的用于监听数据的网络端口已经处于监听状态的情况下，再启动上级子任务。即先启动下级子任务启动，后启动上级子任务。

步骤 S504，确定上级子任务本次要传输的中间数据，记为待传数据；

25 步骤 S505 从下级子任务的用于监听数据的网络端口中选择需要接收待传数据的网络端口，记为接收端口；

步骤 S506，将待传数据通过网络直接传输给接收端口。

30 本发明实施例中的中间数据传输方法，将分布式系统中的中间数据由产生该中间数据的上级子任务通过网络直接传输给下级子任务，中间不需要经过分布式存储系统，避免了磁盘 IO 速率对中间数据传输速率的影响，提高了中间数据传输速率。再者，本发明实施例中的中间数据传输方法，下级子任务先于上级子任务启动，以保证下级子任务能

够接收到上级子任务产生的中间数据。

图 6 为本发明实施例三中中间数据传输方法的流程图。如图 6 所示，本实施例中，中间数据传输方法可以包括如下步骤：

步骤 S601，确定上级子任务的单个进程本次要传输的中间数据，记为待传数据，其中，上级子任务的每个进程对应一个第一网络端口；

本实施例中，以进程为单位进行中间数据的发送。上级子任务的每个进程各自确定自己要传输的中间数据，然后将各自将自己要传输的中间数据通过网络传输给下级子任务的所有进程。

步骤 S602，下级子任务的每个进程对应一个第二网络端口，将所有第二网络端口均选为待传数据的接收端口；

在本实施例中，上、下级子任务的每个进程都分别对应一个网络端口。上级子任务的进程对应的第一网络端口为中间数据的发送端口，下级子任务的进程对应的第二网络端口为中间数据的接收端口。相应地，将上级子任务的进程称为发送进程，将下级子任务的进程称为接收进程。中间数据由发送端口通过网络直接传输到接收端口。该传输方式例如前述图 4 所示的中间数据传输方式。

步骤 S603，将待传数据由发送进程对应的第一网络端口通过网络直接传输给所有第二网络端口，其中，发送进程为产生待传数据的进程。

由上述步骤可见，本实施例中，中间数据由发送进程通过网络直接传输到接收进程，不必经由上级子任务收集各个发送进程产生的中间数据统一发送，也不必经由下级子任务统一接收中间数据后再分别发给各个接收进程，因此可以进一步提高中间数据传输速率。不仅如此，接收进程可以在接收到中间数据后立即进行处理，而不必等到所有发送进程产生完中间数据并传输到接收进程后才能进行处理，因此，中间数据由发送进程通过网络直接传输到接收进程，还可以让接收进程在发送进程仍在产生中间数据的时候就能够对已经传输到接收进程的一部分中间数据先进行处理，也就是说，接收进程和发送进程可以并行运行，这显著提高了上下级子任务之间的并行执行能力，缩短了用户任务整体的运行时间。

由于通过网络传输数据本身会有一定的延时，如果发送进程每次只发送少量数据，那么发送数据的次数就会变多，这样由于网络传输产生的 overhead（间接成本）就会很大。另一方面，如果发送进程每次发送的数据量都较大，当需要接收进程比较多时，就会导致发送进程使用的内存很高。为了解决发送进程何时发送中间数据的问题，在本发

明其他实施例中，在步骤 S603 之前，也即在将待传数据由发送进程对应的第一网络端口通过网络直接传输给所有第二网络端口之前，还可以包括如下步骤：

判断待传数据是否大于或等于 msgSize，msgSize 表示发送进程每次传输中的发送数据长度；

- 5 在待传数据大于或等于 msgSize 的情况下，启动待传数据的传输，并确定待传数据的发送数据长度为 msgSize；

其中， $\text{msgSize} = \max(\min(\text{BuffSize}, \text{MaxMsgSize} * n), \text{MinMsgSize} * n) / n$ ，

- 10 BuffSize 表示预设的发送进程的可用内存上限，MaxMsgSize 表示预设的发送进程在一次发送中传输的最大数据长度，MinMsgSize 表示预设的发送进程在一次发送中传输的最小数据长度，n 表示第二网络端口的数量，“*”表示乘以运算，“/”表示除以运算。n 的含义是接收进程的数量，在本实施例中，接收进程的数量等于接收端口的数量，也等于第二网络端口的数量，因此，接收进程的数量就是第二网络端口的数量。

- 15 上述步骤根据用户任务的规模（指接收进程的数量）及每个发送进程可用的内存大小计算发送进程每次传输中的发送数据长度，实现动态调节，这样就可以避免由于每次发送数据过少导致的间接成本过高的问题以及每次发送数据过大导致的内存需求过高的问题。

- 20 待传数据通常是以消息的形式发送的。在本发明实施例中，承载待传数据的消息中可以携带第一进程标识和消息编号，其中，第一进程标识用于表明该消息的发送进程，消息编号用于表明该消息是该发送进程发送的第几条消息。也就是说，第一进程标识用于说明消息是由“谁”发送的问题，消息编号是单调递增的，它用于说明该消息在发送进程发送的所有消息中所处的位置问题或者说序号问题。

接收进程可以根据承载待传数据的消息中携带的第一进程标识和消息编号判断是否已经接收到该消息，以便根据具体情况处理该消息。

- 25 第二网络端口对应的进程可以记录当前已经从发送进程收到的最大消息编号，如果接收到的消息携带的消息编号大于该最大消息编号，说明接收进程第一次接收该消息，如果接收到的消息携带的消息编号小于或等于该最大消息编号，说明接收进程已经收到过该消息。

据此，在本发明实施例中，在将待传数据由发送进程对应的第一网络端口通过网络直接传输给所有第二网络端口之后，还可以包括：

- 30 第二网络端口接收到承载待传数据的消息后，在该消息携带的消息编号大于第二网

络端口对应的进程记录的当前已经从该发送进程收到的最大消息编号的情况下，第二网络端口对应的进程保存该待传数据，并将记录的最大消息编号更新为该消息携带的消息编号。

或者，在本发明实施例中，在将待传数据由发送进程对应的第一网络端口通过网络直接传输给所有第二网络端口之后，还可以包括：第二网络端口接收到承载待传数据的消息后，在该消息携带的消息编号小于或等于第二网络端口对应的进程记录的当前已经从该发送进程收到的最大消息编号的情况下，第二网络端口对应的进程丢弃该消息，并维持记录的最大消息编号不变。

本发明实施例中的中间数据传输方法，将分布式系统中的中间数据由产生该中间数据的上级子任务通过网络直接传输给下级子任务，中间不需要经过分布式存储系统，避免了磁盘 IO 速率对中间数据传输速率的影响，提高了中间数据传输速率。而且，本发明实施例的中间数据传输系统，还显著提高了上下级子任务之间的并行执行能力，缩短了用户任务整体的运行时间。

在中间数据采用图 6 所示实施例中的由发送进程直接传输到接收进程的传输方式的情况下，可以采用超时重发机制应对可能出现的消息丢失（即发送进程发送的中间数据无法到达接收进程）现象。该超时重发机制的具体内容可以是，发送进程向接收进程发送中间数据后，如果接收进程接收到该中间数据，就向发送进程返回应答响应，告知发送进程该接收进程已经接收到其发送的中间数据；发送进程接收到应答响应后，就可以向返回应答响应的接收进程发送下一次的中间数据了。如果发送进程向接收进程发送中间数据后，接收进程没有接收到该中间数据，就不会向发送进程返回应答响应。发送进程在设定的时限内没有接收到应答响应，就认为没有返回应答响应的接收进程没有接收到该中间数据，因此会向没有返回应答响应的接收进程重发本次传输的中间数据，而不会向没有返回应答响应的接收进程发送下一次的中间数据，直到发送进程接收到该接收进程返回的应答响应为止。这样，既可以保证接收进程能够接收到中间数据，又可以使同样的中间数据不会在网络中进行多次传输，因此能够节约网络带宽。

据此，本发明的中间数据传输方法可以采用图 7 所示的流程。

图 7 为本发明实施例四中中间数据传输方法的流程图。如图 7 所示，本实施例中，中间数据传输方法可以包括如下步骤：

步骤 S701，确定上级子任务的单个进程本次要传输的中间数据，记为待传数据，其中，上级子任务的每个进程对应一个第一网络端口；

步骤 S702, 下级子任务的每个进程对应一个第二网络端口, 将所有第二网络端口均选为待传数据的接收端口;

步骤 S703, 将待传数据由发送进程对应的第一网络端口通过网络直接传输给所有第二网络端口, 其中, 发送进程为产生待传数据的进程;

5 步骤 S704, 在传输之后, 判断在设定时间内发送进程是否接收到第二网络端口对应的进程返回的应答响应, 如果在设定时间内发送进程接收到第二网络端口对应的进程返回的应答响应, 则执行步骤 S705, 否则执行步骤 S706;

10 为了区分应答响应来自于哪一个接收进程, 应答响应中可以携带第二进程标识, 该第二进程标识用于表明发出该应答响应的接收进程, 发送进程根据第二进程标识判断发出该应答响应的接收进程。

步骤 S705, 允许发送进程通过网络向返回应答响应的进程传输下一次的数据;

步骤 S706, 发送进程再次将待传数据通过网络直接传输给未返回应答响应的一个或多个第二网络端口, 返回步骤 S704。

15 本发明实施例中的中间数据传输方法, 将分布式系统中的中间数据由产生该中间数据的上级子任务通过网络直接传输给下级子任务, 中间不需要经过分布式存储系统, 避免了磁盘 IO 速率对中间数据传输速率的影响, 提高了中间数据传输速率。而且, 本发明实施例的中间数据传输系统, 还显著提高了上下级子任务之间的并行执行能力, 缩短了用户任务整体的运行时间。同时, 在本发明实施例中, 同样的中间数据不会在网络中进行多次传输, 因此本发明实施例的中间数据传输系统, 还能够节约网络带宽。

20 本发明还提出了中间数据传输系统, 用以实施上述各实施例的中间数据传输方法。上述对中间数据传输方法的说明均适用于中间数据传输系统的相应部分。下面本发明各实施例中的中间数据传输系统都用于分布式系统, 该分布式系统的用户任务包括多级子任务, 这些子任务中产生中间数据的子任务称为上级子任务, 这些子任务中依赖中间数据进行处理的子任务称为下级子任务。

25 图 8 为本发明实施例五中中间数据传输系统的结构框图。如图 8 所示, 本实施例中, 中间数据传输系统 800 可以包括确定模块 810、选择模块 820 和传输模块 830。其中, 确定模块 810 用于确定上级子任务本次要传输的中间数据, 记为待传数据。选择模块 820 用于从下级子任务的用于监听数据的网络端口中选择需要接收待传数据的网络端口, 记为接收端口。传输模块 830 用于将确定模块 810 确定的待传数据通过网络直接传输给选
30 择模块 820 选择的接收端口。

其中，在上级子任务有多个进程的情况下，确定模块 810 所确定的待传数据可以是上级子任务的一个进程产生的中间数据，也可以是上级子任务的多个或全部进程产生的中间数据。

其中，选择模块 820 选择的接收端口可以是下级子任务的单个进程对应的网络端口，
5 也可以为下级子任务的所有进程对应的一个或多个共同网络端口。

其中，传输模块 830 可以使用用户根据需要制定的具体传输策略将待传数据通过网络直接传输给接收端口。例如前述本发明实施例一中列举的传输策略。

本发明实施例中的中间数据传输系统，将分布式系统中的中间数据由产生该中间数据的上级子任务通过网络直接传输给下级子任务，中间不需要经过分布式存储系统，避免
10 了磁盘 IO 速率对中间数据传输速率的影响，提高了中间数据传输速率。

图 9 为本发明实施例六中中间数据传输系统的结构框图。如图 9 所示，本实施例中，中间数据传输系统可以包括第一启动模块 910、通知模块 920、第二启动模块 930 确定模块 940、选择模块 950 和传输模块 960。其中，第一启动模块 910 用于启动下级子任务，使下级子任务的用于监听数据的网络端口处于监听状态。通知模块 920 与第一启动模块
15 910 相连，用于将下级子任务的用于监听数据的网络端口的信息通知给上级子任务。第二启动模块 930 分别与通知模块 920 和确定模块 940 相连，用于在下级子任务启动之后启动上级子任务，产生中间数据。确定模块 940 用于确定上级子任务本次要传输的中间数据，记为待传数据。选择模块 950 用于从下级子任务的用于监听数据的网络端口中选择需要接收待传数据的网络端口，记为接收端口。传输模块 960 用于将确定模块 940 确
20 定的待传数据通过网络直接传输给选择模块 950 选择的接收端口。

本发明实施例中的中间数据传输系统，将分布式系统中的中间数据由产生该中间数据的上级子任务通过网络直接传输给下级子任务，中间不需要经过分布式存储系统，避免了磁盘 IO 速率对中间数据传输速率的影响，提高了中间数据传输速率。再者，本发明
25 实施例中的中间数据传输系统，下级子任务先于上级子任务启动，以保证下级子任务能够接收到上级子任务产生的中间数据。

图 10 为本发明实施例七中中间数据传输系统的传输模块的结构框图。如图 10 所示，本实施例中，中间数据传输系统的传输模块 1000 可以包括进程间传输单元 1010。进程间传输单元 1010 用于在待传数据由上级子任务的单个进程产生，上级子任务的每个进程对应一个第一网络端口，下级子任务的每个进程对应一个第二网络端口，所有第二网络
30 端口均为接收端口的情况下，将待传数据由发送进程对应的第一网络端口通过网络直接

传输给所有第二网络端口，其中，发送进程指产生待传数据的进程。

本实施例中，中间数据由发送进程通过网络直接传输到接收进程，不必经由上级子任务收集各个发送进程产生的中间数据统一发送，也不必经由下级子任务统一接收中间数据后再分别发给各个接收进程，因此可以进一步提高中间数据传输速率。不仅如此，接收进程可以在接收到中间数据后立即进行处理，而不必等到所有发送进程产生完中间数据并传输到接收进程后才能进行处理，因此，中间数据由发送进程通过网络直接传输到接收进程，还可以让接收进程在发送进程仍在产生中间数据的时候就能够对已经传输到接收进程的一部分中间数据先进行处理，也就是说，接收进程和发送进程可以并行运行，这显著提高了上下级子任务之间的并行执行能力，缩短了用户任务整体的运行时间。

10 在图 10 所示实施例的基础上，传输模块 1000 还可以进一步包括判断单元和启动单元。其中，判断单元用于判断待传数据是否大于或等于 msgSize ， msgSize 表示发送进程每次传输中的发送数据长度。启动单元用于在判断单元的判断结果为待传数据大于或等于 msgSize 的情况下，启动进程间传输单元 1010 进行待传数据的传输，并确定待传数据的发送数据长度为 msgSize 。其中， $\text{msgSize} = \max(\min(\text{BuffSize}, \text{MaxMsgSize} * n), \text{MinMsgSize} * n) / n$ ， BuffSize 表示预设的发送进程的可用内存上限， MaxMsgSize 表示预设的发送进程在一次发送中传输的最大数据长度， MinMsgSize 表示预设的发送进程在一次发送中传输的最小数据长度， n 表示第二网络端口的数量，“*”表示乘以运算，“/”表示除以运算。

20 上述判断单元和启动单元根据用户任务的规模（指接收进程的数量）及每个发送进程可用的内存大小计算发送进程每次传输中的发送数据长度，实现动态调节，这样就可以避免由于每次发送数据过少导致的间接成本过高的问题以及每次发送数据过大导致的内存需求过高的问题。

25 在本发明实施例中，承载待传数据的消息中可以携带第一进程标识和消息编号，其中，第一进程标识用于表明该消息的发送进程，消息编号用于表明该消息是该发送进程发送的第几条消息。

接收进程可以根据承载待传数据的消息中携带的第一进程标识和消息编号判断是否已经接收到该消息，以便根据具体情况处理该消息。

30 第二网络端口对应的进程可以记录当前已经从发送进程收到的最大消息编号，如果接收到的消息携带的消息编号大于该最大消息编号，说明接收进程第一次接收该消息，如果接收到的消息携带的消息编号小于或等于该最大消息编号，说明接收进程已经收到

过该消息。

据此，在本发明实施例中，传输模块 1000 可以包括保存单元。保存单元用于在第二网络端口接收到承载待传数据的消息后，在该消息携带的消息编号大于第二网络端口对应的进程记录的当前已经从该发送进程（指发出待传数据的进程）收到的最大消息编号的情况下，令第二网络端口对应的进程保存待传数据，并将记录的最大消息编号更新为该消息携带的消息编号。

在本发明实施例中，传输模块 1000 还可以包括丢弃单元。丢弃单元用于在第二网络端口接收到承载待传数据的消息后，在消息携带的消息编号小于或等于第二网络端口对应的进程记录的当前已经从发送进程（指发出待传数据的进程）收到的最大消息编号的情况下，令第二网络端口对应的进程丢弃该消息，并维持记录的最大消息编号不变。

本发明实施例中的中间数据传输系统，将分布式系统中的中间数据由产生该中间数据的上级子任务通过网络直接传输给下级子任务，中间不需要经过分布式存储系统，避免了磁盘 IO 速率对中间数据传输速率的影响，提高了中间数据传输速率。而且，本发明实施例的中间数据传输系统，还显著提高了上下级子任务之间的并行执行能力，缩短了用户任务整体的运行时间。

在中间数据采用图 10 所示实施例中的进程间传输单元进行传输的情况下，传输模块可以采用超时重发机制应对可能出现的消息丢失（即发送进程发送的中间数据无法到达接收进程）现象。

据此，中间数据传输系统的传输模块可以采用如图 11 所示的结构。图 11 为本发明实施例八中中间数据传输系统的传输模块的结构框图。如图 11 所示，本实施例中，传输模块 1100 可以包括进程间传输单元 1110、响应接收单元 1120。其中，进程间传输单元 1110 同前述的进程间传输单元 1000 的功能相同，此处不再赘述。响应接收单元 1120 用于在进程间传输单元 1110 进行传输之后，令发送进程接收第二网络端口对应的进程返回的应答响应，该应答响应用于表示第二网络端口对应的进程已经接收到待传数据。

参见图 11，传输模块 1110 还可以包括许可单元 1130。许可单元 1130 用于在响应接收单元 1120 接收到应答响应后，允许发送进程通过网络向返回应答响应的进程传输下一数据的数据。

参见图 11，传输模块 1110 还可以包括重传单元 1140。重传单元 1140 用于在进程间传输单元 1110 进行传输之后，在设定时间内响应接收单元 1120 未接收到一个或多个第二网络端口对应的进程返回的应答响应的情况下，令发送进程再次将待传数据通过网络

直接传输给未返回应答响应的一个或多个第二网络端口。

为了区分应答响应来自于哪一个接收进程，应答响应中可以携带第二进程标识，该第二进程标识用于表明发出该应答响应的接收进程，发送进程根据第二进程标识判断发出该应答响应的接收进程。

- 5 本发明实施例中的中间数据传输系统，将分布式系统中的中间数据由产生该中间数据的上级子任务通过网络直接传输给下级子任务，中间不需要经过分布式存储系统，避免了磁盘 IO 速率对中间数据传输速率的影响，提高了中间数据传输速率。而且，本发明实施例的中间数据传输系统，还显著提高了上下级子任务之间的并行执行能力，缩短了用户任务整体的运行时间。同时，在本发明实施例中，同样的中间数据不会在网络中进行多次传输，因此本发明实施例的中间数据传输系统还能够节约网络带宽。

图 12 为本发明实施例九中分布式系统的结构框图。如图 12 所示，本实施例中，分布式系统 1200 可以包括中间数据传输系统。该中间数据传输系统可以是本发明前述实施例中的任一种中间数据传输系统。

- 15 本发明实施例中的分布式系统中包括中间数据传输系统，将分布式系统中的中间数据由产生该中间数据的上级子任务通过网络直接传输给下级子任务，中间不需要经过分布式存储系统，避免了磁盘 IO 速率对中间数据传输速率的影响，提高了中间数据传输速率。而且，本发明实施例的分布式系统，还能够显著提高上下级子任务之间的并行执行能力，缩短用户任务整体的运行时间。同时，在本发明实施例中，同样的中间数据不会在网络中进行多次传输，因此本发明实施例的分布式系统还能够节约网络带宽。

- 20 以上所述仅为本发明的较佳实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

权利要求书

1、一种中间数据传输方法，用于分布式系统，所述分布式系统的用户任务包括多级子任务，所述子任务中产生中间数据的子任务称为上级子任务，所述子任务中依赖所述中间数据进行处理的子任务称为下级子任务，其特征在于，所述方法包括：

5 确定所述上级子任务本次要传输的中间数据，记为待传数据；

 从所述下级子任务的用于监听数据的网络端口中选择需要接收所述待传数据的网络端口，记为接收端口；

 将所述待传数据通过网络直接传输给所述接收端口。

10 2、根据权利要求 1 所述的中间数据传输方法，其特征在于，在所述确定所述上级子任务本次要传输的中间数据之前，还包括：

 启动所述下级子任务，使所述下级子任务的用于监听数据的网络端口处于监听状态；

 将所述下级子任务的用于监听数据的网络端口的信息通知给所述上级子任务；

 在所述下级子任务启动之后启动所述上级子任务，产生中间数据。

15 3、根据权利要求 1 所述的中间数据传输方法，其特征在于，所述将所述待传数据通过网络直接传输给所述接收端口包括：

 在所述待传数据由所述上级子任务的单个进程产生，所述上级子任务的每个进程对应一个第一网络端口，所述下级子任务的每个进程对应一个第二网络端口，所有第二网络端口均为接收端口的情况下，将所述待传数据由发送进程对应的第一网络端口通过网络
20 直接传输给所有第二网络端口，所述发送进程为产生所述待传数据的进程。

 4、根据权利要求 3 所述的中间数据传输方法，其特征在于，在将所述待传数据由发送进程对应的第一网络端口通过网络直接传输给所有第二网络端口之前，还包括：

 判断所述待传数据是否大于或等于 `msgSize`，`msgSize` 表示产生所述发送进程每次传输中的发送数据长度；

25 在所述待传数据大于或等于 `msgSize` 的情况下，启动所述待传数据的传输，并确定所述待传数据的发送数据长度为 `msgSize`；

 其中， $\text{msgSize} = \max(\min(\text{BuffSize}, \text{MaxMsgSize} * n), \text{MinMsgSize} * n) / n$ ，

`BuffSize` 表示预设的所述发送进程的可用内存上限，`MaxMsgSize` 表示预设的所述发送进程在一次发送中传输的最大数据长度，`MinMsgSize` 表示预设的所述发送进程在
30 一次发送中传输的最小数据长度，`n` 表示所述第二网络端口的数量，“*”表示乘以运算，

“/” 表示除以运算。

5、根据权利要求 3 所述的中间数据传输方法，其特征在于，承载所述待传数据的消息中携带第一进程标识和消息编号，所述第一进程标识用于表明所述消息的发送进程，所述消息编号用于表明所述消息是所述发送进程发送的第几条消息。

5 6、根据权利要求 5 所述的中间数据传输方法，其特征在于，在将所述待传数据由发送进程对应的第一网络端口通过网络直接传输给所有第二网络端口之后，还包括：

所述第二网络端口接收到承载所述待传数据的消息后，在所述消息携带的消息编号大于所述第二网络端口对应的进程记录的当前已经从所述发送进程收到的最大消息编号的情况下，所述第二网络端口对应的进程保存所述待传数据，并将记录的所述最大消息编号更新为所述消息携带的消息编号。

10

7、根据权利要求 5 所述的中间数据传输方法，其特征在于，在将所述待传数据由发送进程对应的第一网络端口通过网络直接传输给所有第二网络端口之后，还包括：

所述第二网络端口接收到承载所述待传数据的消息后，在所述消息携带的消息编号小于或等于所述第二网络端口对应的进程记录的当前已经从所述发送进程收到的最大消息编号的情况下，所述第二网络端口对应的进程丢弃所述消息，并维持记录的所述最大消息编号不变。

15

8、根据权利要求 3 所述的中间数据传输方法，其特征在于，在将所述待传数据由发送进程对应的第一网络端口通过网络直接传输给所有第二网络端口之后，还包括：

在所述传输之后，所述发送进程接收所述第二网络端口对应的进程返回的应答响应，所述应答响应用于表示所述第二网络端口对应的进程已经接收到所述待传数据。

20

9、根据权利要求 8 所述的中间数据传输方法，其特征在于，在将所述待传数据由发送进程对应的第一网络端口通过网络直接传输给所有第二网络端口之后，还包括：

在所述发送进程接收所述第二网络端口对应的进程返回的应答响应后，允许所述发送进程通过网络向返回所述应答响应的进程传输下一一次的数据。

25 10、根据权利要求 8 所述的中间数据传输方法，其特征在于，在将所述待传数据由发送进程对应的第一网络端口通过网络直接传输给所有第二网络端口之后，还包括：

在所述传输之后，在设定时间内所述发送进程未接收到一个或多个第二网络端口对应的进程返回的应答响应的情况下，所述发送进程再次将所述待传数据通过网络直接传输给所述一个或多个第二网络端口。

30 11、根据权利要求 8 所述的中间数据传输方法，其特征在于，所述应答响应中携带

第二进程标识, 所述第二进程标识用于表明发出所述应答响应的进程, 所述发送进程根据所述第二进程标识判断发出所述应答响应的进程。

12、一种中间数据传输系统, 用于分布式系统, 所述分布式系统的用户任务包括多级子任务, 所述子任务中产生中间数据的子任务称为上级子任务, 所述子任务中依赖所述中间数据进行处理的子任务称为下级子任务, 其特征在于, 所述中间数据传输系统包括:

确定模块, 用于确定所述上级子任务本次要传输的中间数据, 记为待传数据;

选择模块, 用于从所述下级子任务的用于监听数据的网络端口中选择需要接收所述待传数据的网络端口, 记为接收端口;

10 传输模块, 用于将所述确定模块确定的待传数据通过网络直接传输给所述选择模块选择的接收端口。

13、根据权利要求 12 所述的中间数据传输系统, 其特征在于, 还包括:

第一启动模块, 用于启动所述下级子任务, 使所述下级子任务的用于监听数据的网络端口处于监听状态;

15 通知模块, 与所述第一启动模块相连, 用于将所述下级子任务的用于监听数据的网络端口的信息通知给所述上级子任务;

第二启动模块, 分别与所述通知模块和所述确定模块相连, 用于在所述下级子任务启动之后启动所述上级子任务, 产生中间数据。

14、根据权利要求 12 所述的中间数据传输系统, 其特征在于, 所述传输模块包括:

20 进程间传输单元, 用于在所述待传数据由所述上级子任务的单个进程产生, 所述上级子任务的每个进程对应一个第一网络端口, 所述下级子任务的每个进程对应一个第二网络端口, 所有第二网络端口均为接收端口的情况下, 将所述待传数据由发送进程对应的第一网络端口通过网络直接传输给所有第二网络端口, 所述发送进程指产生所述待传数据的进程。

25 15、根据权利要求 14 所述的中间数据传输系统, 其特征在于, 所述传输模块还包括:

判断单元, 用于判断所述待传数据是否大于或等于 msgSize, msgSize 表示所述发送进程每次传输中的发送数据长度;

30 启动单元, 用于在所述判断单元的判断结果为所述待传数据大于或等于 msgSize 的情况下, 启动所述进程间传输单元进行所述待传数据的传输, 并确定所述待传数据的发

送数据长度为 msgSize;

其中, $\text{msgSize} = \max(\min(\text{BuffSize}, \text{MaxMsgSize} * n), \text{MinMsgSize} * n) / n$,

BuffSize 表示预设的所述发送进程的可用内存上限, MaxMsgSize 表示预设的所述发送进程在一次发送中传输的最大数据长度, MinMsgSize 表示预设的所述发送进程在
5 一次发送中传输的最小数据长度, n 表示所述第二网络端口的数量, “*”表示乘以运算, “/”表示除以运算。

16、根据权利要求 14 所述的中间数据传输系统, 其特征在于, 承载所述待传数据的消息中携带第一进程标识和消息编号, 所述第一进程标识用于表明所述消息的发送进程, 所述消息编号用于表明所述消息是所述发送进程发送的第几条消息。

10 17、根据权利要求 16 所述的中间数据传输系统, 其特征在于, 所述传输模块还包括:

保存单元, 用于在所述第二网络端口接收到承载所述待传数据的消息后, 在所述消息携带的消息编号大于所述第二网络端口对应的进程记录的当前已经从所述发送进程收到的最大消息编号的情况下, 令所述第二网络端口对应的进程保存所述待传数据, 并
15 将记录的所述最大消息编号更新为所述消息携带的消息编号。

18、根据权利要求 16 所述的中间数据传输系统, 其特征在于, 所述传输模块还包括:

丢弃单元, 用于在所述第二网络端口接收到承载所述待传数据的消息后, 在所述消息携带的消息编号小于或等于所述第二网络端口对应的进程记录的当前已经从所述发
20 送进程收到的最大消息编号的情况下, 令所述第二网络端口对应的进程丢弃所述消息, 并维持记录的所述最大消息编号不变。

19、根据权利要求 14 所述的中间数据传输系统, 其特征在于, 所述传输模块还包括:

响应接收单元, 用于在所述进程间传输单元进行传输之后, 令所述发送进程接收所
25 述第二网络端口对应的进程返回的应答响应, 所述应答响应用于表示所述第二网络端口对应的进程已经接收到所述待传数据。

20、根据权利要求 19 所述的中间数据传输系统, 其特征在于, 所述传输模块还包括:

许可单元, 用于在所述响应接收单元接收到所述应答响应后, 允许所述发送进程通
30 过网络向返回所述应答响应的进程传输下一一次的数据。

21、根据权利要求 19 所述的中间数据传输系统，其特征在于，所述传输模块还包括：

重传单元，用于在所述进程间传输单元进行传输之后，在设定时间内所述响应接收单元未接收到一个或多个第二网络端口对应的进程返回的应答响应的情况下，令所述发送进程再次将所述待传数据通过网络直接传输给所述一个或多个第二网络端口。

22、根据权利要求 19 所述的中间数据传输系统，其特征在于，所述应答响应中携带第二进程标识，所述第二进程标识用于表明发出所述应答响应的进程，所述发送进程根据所述第二进程标识判断发出所述应答响应的进程。

23、一种分布式系统，其特征在于，包括权利要求 12 至 22 任一项所述的中间数据传输系统。

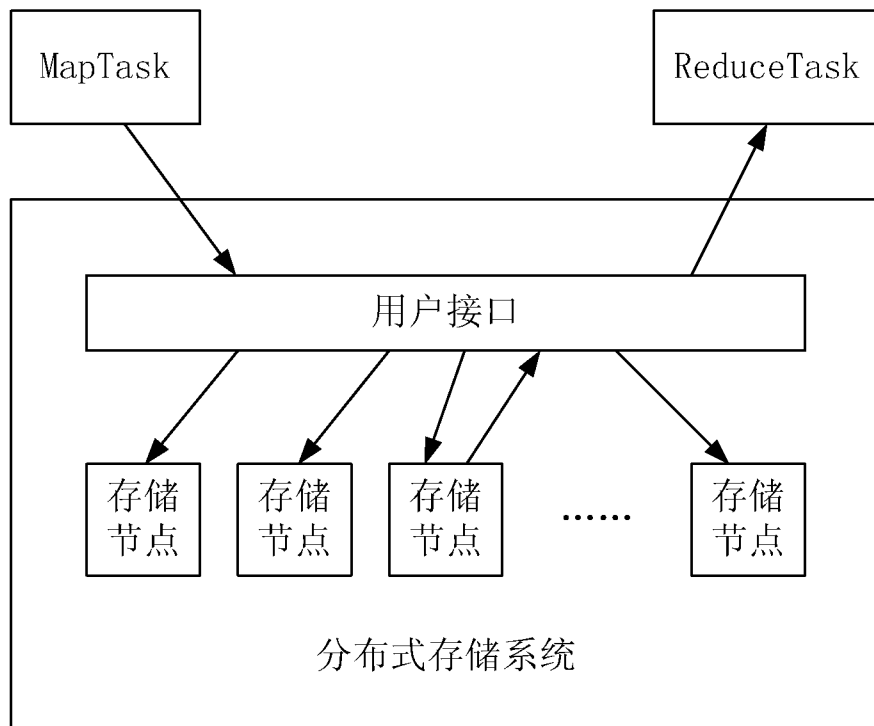


图 1

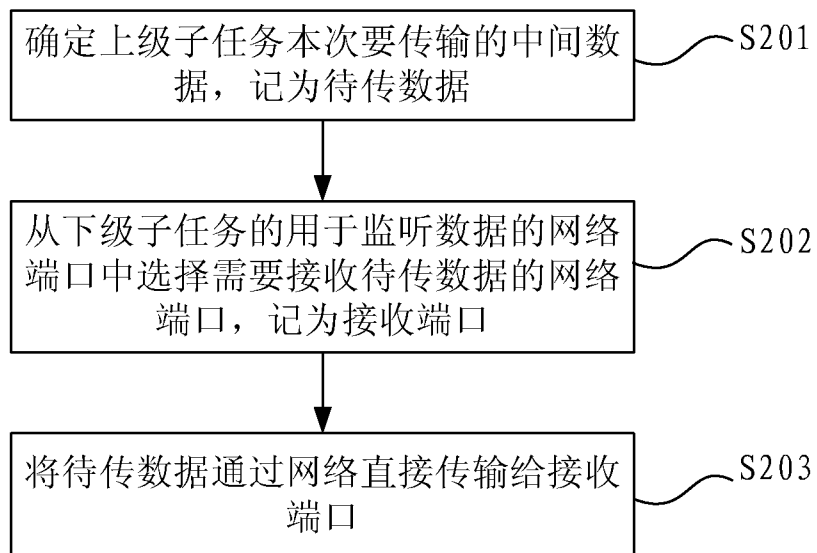


图 2

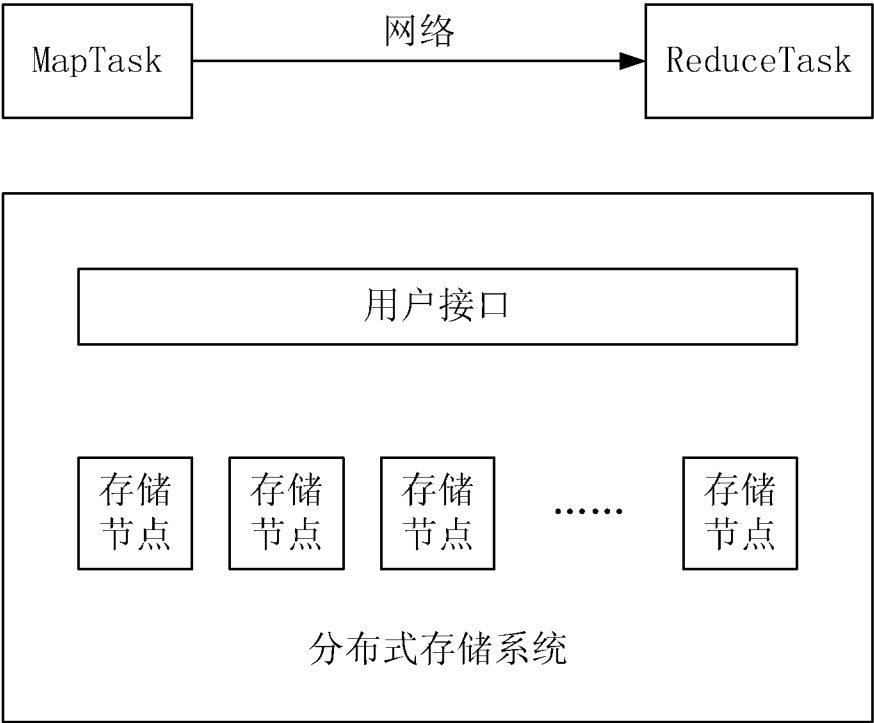


图 3

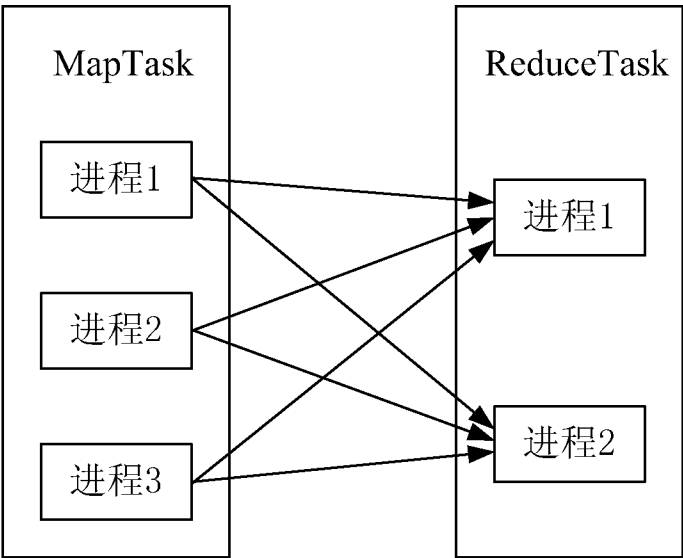


图 4

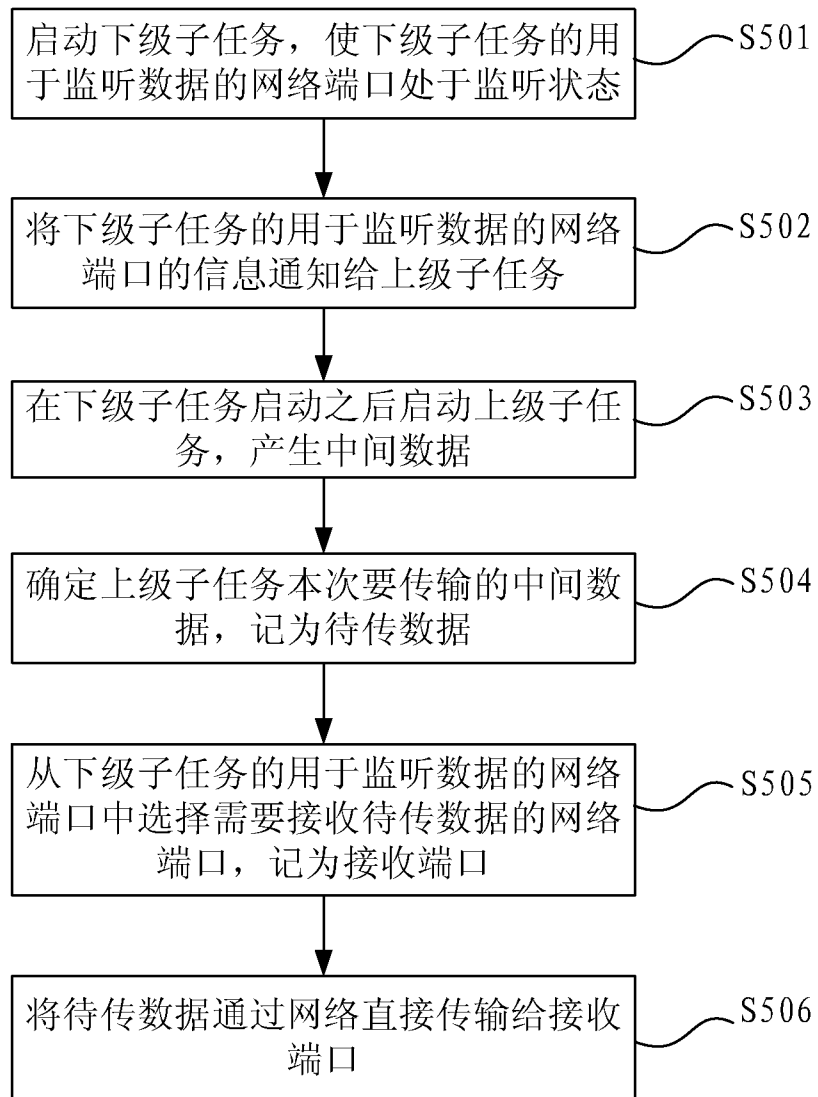


图 5

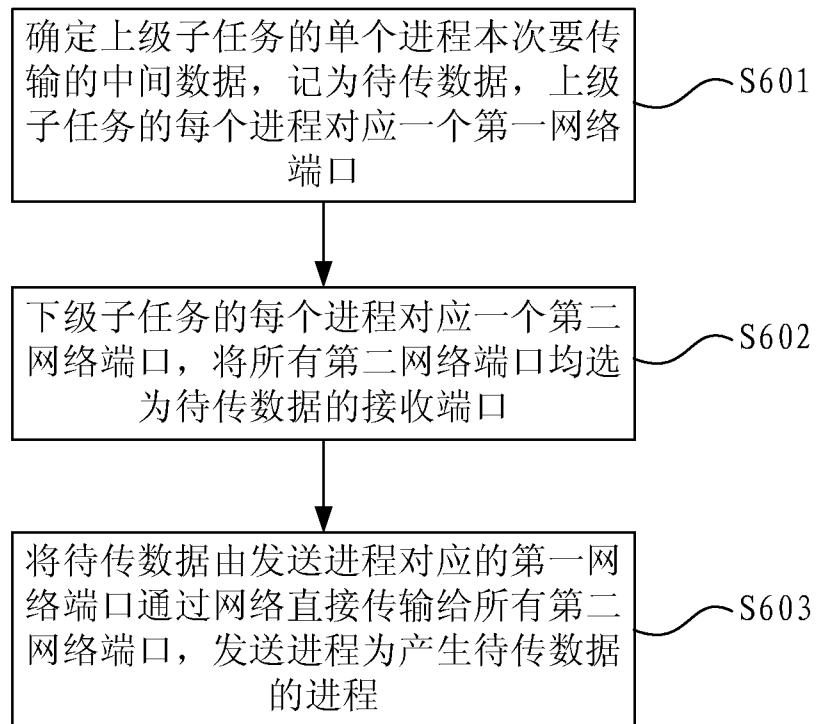


图 6

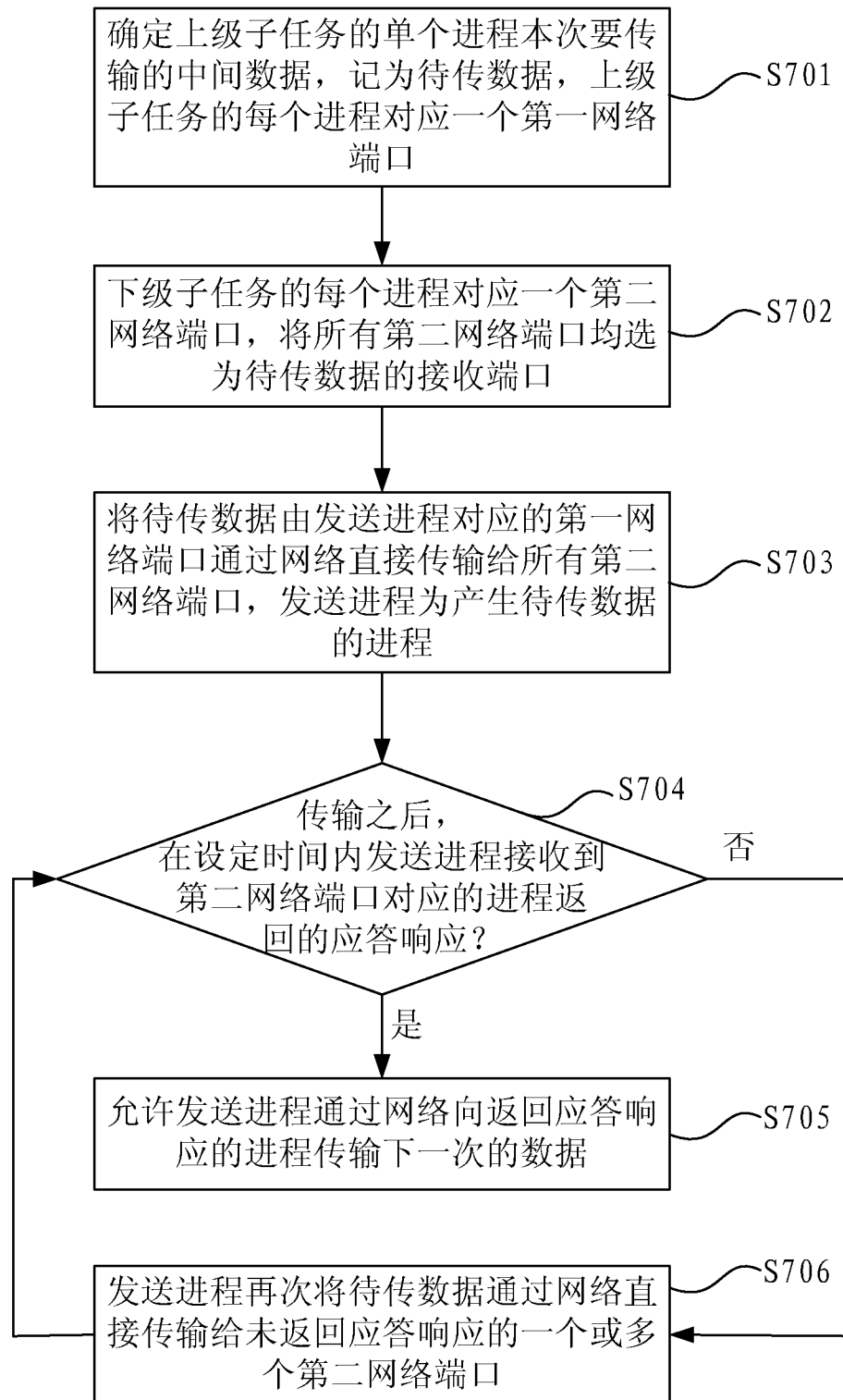


图 7

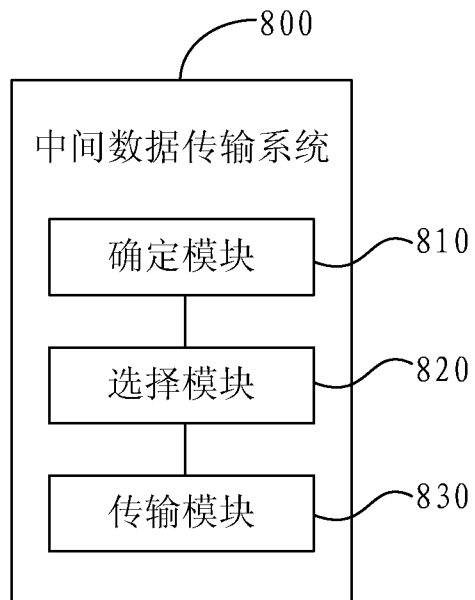


图 8

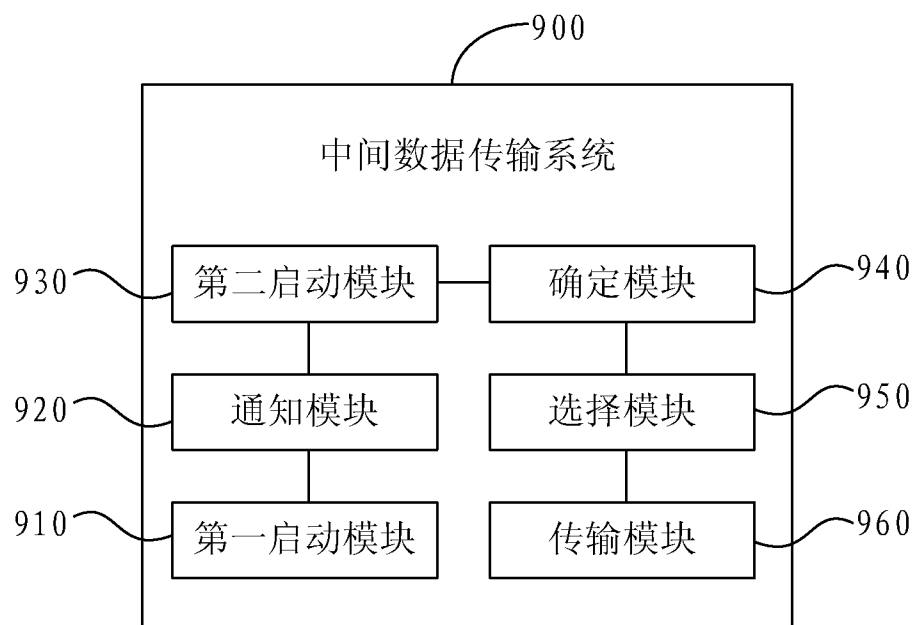


图 9

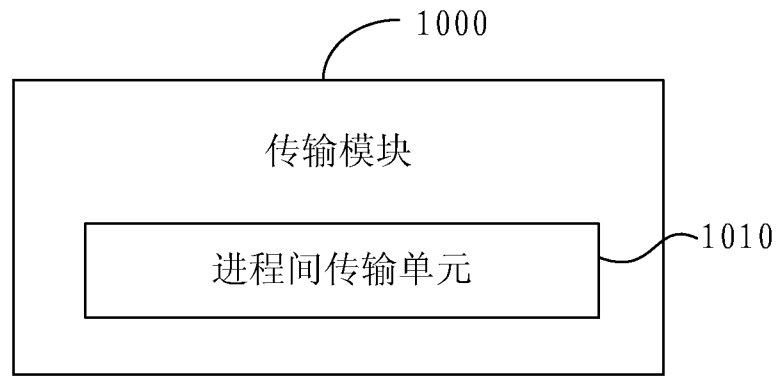


图 10

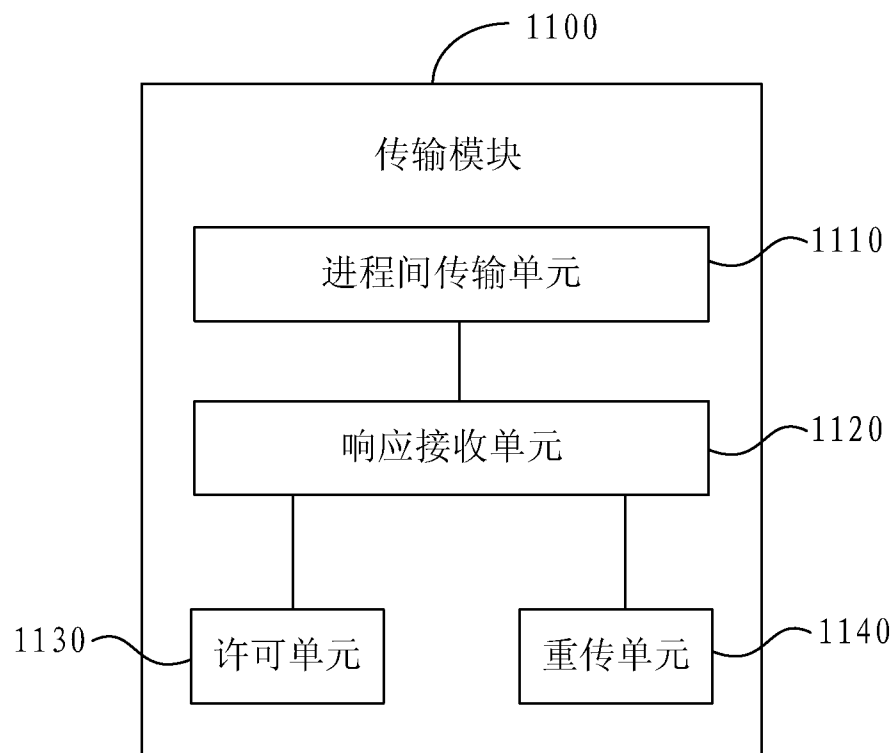


图 11

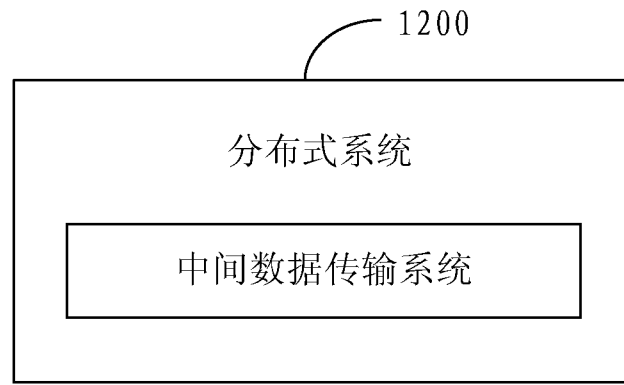


图 12

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/072818

A. CLASSIFICATION OF SUBJECT MATTER

H04L 29/08 (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

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, VEN, CNTXT, CNKI: intermediate data, intermediate result, distributed, task, subtask, process, thread, port, interface, direct, transmit, transport, transfer, not, store

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 103279390 A (CHINESE ACADEMY SCIENCE INFORMATION ENGINEERING INSTITUTE) 04 September 2013 (04.09.2013) see description, paragraphs [0015]-[0021], [0039]-[0056] and [0089]-[0103], and figures 4 and 5	1-23
A	CN 104063486 A (SICHUAN FEDSTORE TECHNOLOGY CO., LTD.) 24 September 2014 (24.09.2014) the whole document	1-23
A	CN 105138679 A (GUILIN ELECTRONIC TECHNOLOGY UNIVERSITY) 09 December 2015 (09.12.2015) the whole document	1-23
A	US 2015095308 A1 (CLOUDERA INC.) 02 April 2015 (02.04.2015) the whole document	1-23

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 18 April 2017	Date of mailing of the international search report 02 May 2017
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer WU, Xu Telephone No. (86-10) 62411267

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/072818

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103279390 A	04 September 2013	CN 103279390 B	28 September 2016
CN 104063486 A	24 September 2014	None	
CN 105138679 A	09 December 2015	None	
US 2015095308 A1	02 April 2015	US 9477731 B2	25 October 2016
		AU 2014240211 A1	16 April 2015
		GB 2520411 A	20 May 2015
		EP 2857995 A2	08 April 2015
		KR 20150039118 A	09 April 2015
		EP 2857995 A3	12 August 2015
		CA 2865451 A1	01 April 2015
		AU 2014240211 B2	02 February 2017
		JP 2015072688 A	16 April 2015

A. 主题的分类 H04L 29/08 (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 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, VEN, CNTXT, CNKI: 中间数据, 中间结果, 分布式, 任务, 子任务, 进程, 线程, 端口, 接口, 直接, 传输, 传送, 传递, 不, 存储, 储存, intermediate data, intermediate result, distributed, task, subtask, process, thread, port, interface, direct, transmit, transport, transfer, not, store																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 103279390 A (中国科学院信息工程研究所) 2013年 9月 4日 (2013 - 09 - 04) 参见说明书第0015-0021、0039-0056和0089-0103段, 图4-5</td> <td>1-23</td> </tr> <tr> <td>A</td> <td>CN 104063486 A (四川中亚联邦科技有限公司) 2014年 9月 24日 (2014 - 09 - 24) 全文</td> <td>1-23</td> </tr> <tr> <td>A</td> <td>CN 105138679 A (桂林电子科技大学) 2015年 12月 9日 (2015 - 12 - 09) 全文</td> <td>1-23</td> </tr> <tr> <td>A</td> <td>US 2015095308 A1 (CLOUDERA INC) 2015年 4月 2日 (2015 - 04 - 02) 全文</td> <td>1-23</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 103279390 A (中国科学院信息工程研究所) 2013年 9月 4日 (2013 - 09 - 04) 参见说明书第0015-0021、0039-0056和0089-0103段, 图4-5	1-23	A	CN 104063486 A (四川中亚联邦科技有限公司) 2014年 9月 24日 (2014 - 09 - 24) 全文	1-23	A	CN 105138679 A (桂林电子科技大学) 2015年 12月 9日 (2015 - 12 - 09) 全文	1-23	A	US 2015095308 A1 (CLOUDERA INC) 2015年 4月 2日 (2015 - 04 - 02) 全文	1-23
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
X	CN 103279390 A (中国科学院信息工程研究所) 2013年 9月 4日 (2013 - 09 - 04) 参见说明书第0015-0021、0039-0056和0089-0103段, 图4-5	1-23															
A	CN 104063486 A (四川中亚联邦科技有限公司) 2014年 9月 24日 (2014 - 09 - 24) 全文	1-23															
A	CN 105138679 A (桂林电子科技大学) 2015年 12月 9日 (2015 - 12 - 09) 全文	1-23															
A	US 2015095308 A1 (CLOUDERA INC) 2015年 4月 2日 (2015 - 04 - 02) 全文	1-23															
☐ 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 2017年 4月 18日		国际检索报告邮寄日期 2017年 5月 2日															
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (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 吴旭 电话号码 (86-10) 62411267															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/072818

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	103279390	A	2013年 9月 4日	CN	103279390	B	2016年 9月 28日
CN	104063486	A	2014年 9月 24日	无			
CN	105138679	A	2015年 12月 9日	无			
US	2015095308	A1	2015年 4月 2日	US	9477731	B2	2016年 10月 25日
				AU	2014240211	A1	2015年 4月 16日
				GB	2520411	A	2015年 5月 20日
				EP	2857995	A2	2015年 4月 8日
				KR	20150039118	A	2015年 4月 9日
				EP	2857995	A3	2015年 8月 12日
				CA	2865451	A1	2015年 4月 1日
				AU	2014240211	B2	2017年 2月 2日
				JP	2015072688	A	2015年 4月 16日