

(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
2023 年 1 月 12 日 (12.01.2023)



(10) 国际公布号
WO 2023/280207 A1

(51) 国际专利分类号:
G06F 16/242 (2019.01) **G06F 16/2455** (2019.01)

(21) 国际申请号: PCT/CN2022/104127

(22) 国际申请日: 2022 年 7 月 6 日 (06.07.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110766780.0 2021 年 7 月 7 日 (07.07.2021) CN

(71) 申请人: 清华大学 (TSINGHUA UNIVERSITY)
[CN/CN]; 中国北京市海淀区清华园,
Beijing 100084 (CN)。

(72) 发明人: 俞博文 (YU, Bowen); 中国北京市海淀区清华园, Beijing 100084 (CN)。 陈文光 (CHEN, Wenguang); 中国北京市海淀区清华园, Beijing 100084 (CN)。

(74) 代理人: 北京市柳沈律师事务所 (LIU, SHEN & ASSOCIATES); 中国北京市海淀区彩和坊路 10 号 1 号楼 10 层, Beijing 100080 (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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, 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, 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,

(54) Title: DATA PROCESSING METHOD, EXECUTION WORKSTATION, DISTRIBUTED COMPUTING SYSTEM, AND STORAGE MEDIUM

(54) 发明名称: 数据处理方法、执行工作站、分布式计算系统和存储介质

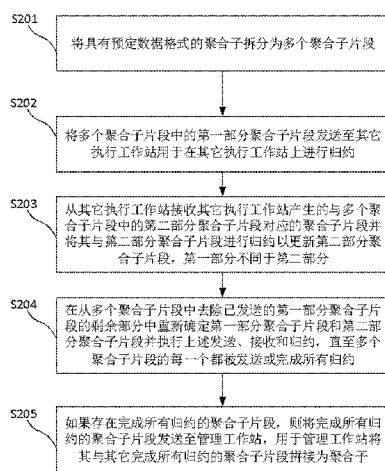


图 2A

(57) Abstract: A data processing method, an execution workstation, a distributed computing system, and a computer-readable storage medium. The data processing method comprises: splitting an aggregate of a pre-determined data format into a plurality of aggregate fragments (S201); sending a first portion of aggregate fragments among the plurality of aggregate fragments to another execution workstation to perform reduction on the other execution workstation (S202); receiving from another execution workstation aggregate fragments produced by the other execution workstation and corresponding to a second portion of aggregate fragments among the plurality of aggregate fragments, and performing reduction on said received aggregate fragments and the second portion of aggregate fragments, so as to update the second portion of aggregate fragments (S203); newly determining a first portion of aggregate fragments and a second portion of aggregate fragments, and executing the aforementioned sending, receiving, and reducing until every fragment among the plurality of aggregate fragments is sent or completes all reductions (S204); and sending aggregate fragments that have completed all reductions to a management workstation (S205).

- S201 Split aggregate of pre-determined data format into plurality of aggregate fragments
S202 Send first portion of aggregate fragments among plurality of aggregate fragments to other execution workstation to perform reduction on other execution workstation
S203 Receive from other execution workstation aggregate fragments produced by other execution workstation and corresponding to second portion of aggregate fragments among plurality of aggregate fragments, and perform reduction on said received aggregate fragments and second portion of aggregate fragments, so as to update second portion of aggregate fragments, first portion being different from second portion
S204 Newly determine first portion of aggregate fragments and second portion of aggregate fragments from remaining portion of plurality of aggregate fragments after removing sent first portion of aggregate fragments, and execute aforementioned sending, receiving, and reducing until every fragment among plurality of aggregate fragments is sent or completes all reductions
S205 If there are aggregate fragments that have completed all reductions, send aggregate fragments that have completed all reductions to management workstation, to allow management workstation to combine said fragments and other aggregate fragments that have completed all reductions into aggregate

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))。

(57) 摘要：一种数据处理方法、执行工作站、分布式计算系统和计算机可读存储介质。该数据处理方法包括：将具有预定数据格式的聚合子拆分为多个聚合子片段(S201)；将多个聚合子片段中的第一部分聚合子片段发送至其它执行工作站用于在其它执行工作站上进行归约(S202)；从其它执行工作站接收其它执行工作站产生的与多个聚合子片段中的第二部分聚合子片段对应的聚合子片段并将其与第二部分聚合子片段进行归约以更新第二部分聚合子片段(S203)；重新确定第一部分聚合子片段和第二部分聚合子片段并执行上述发送、接收和归约，直至多个聚合子片段的每一个都被发送或完成所有归约(S204)；以及将完成所有归约的聚合子片段发送至管理工作站(S205)。

数据处理方法、执行工作站、分布式计算系统和存储介质

本申请要求于 2021 年 7 月 7 日递交的中国专利申请第 202110766780.0 号的
5 的优先权，在此全文引用上述中国专利申请公开的内容以作为本申请的一部分。

技术领域

本公开的实施例涉及一种数据处理方法、执行工作站、分布式计算系统和
10 计算机可读存储介质。

背景技术

现如今是数据时代，数据信息化与我们的生活和工作息息相关，由于数据
量的急剧增加，单机难以处理，因此设计出了分布式计算用于大数据处理框架。
15 Spark 是当今最流行的大数据处理框架之一。Spark 经由其弹性分布式数据集编
程模型连接各种数据分析域，Spark 的机器学习框架 MLlib 被广泛地使用，这
是由于 MLlib 可以被很容易地集成到 Spark 的数据分析工作流中。树聚合是
MLlib 的一个重要的研究热点，Spark 中的树聚合可以被大致分为两个连续的
步骤：计算与归约。在计算步骤中，当前分区的值在本地计算。在归约步骤中，
20 计算出的值（称为聚合子）被进一步全局归约为单个聚合子。

发明内容

本公开至少一实施例提供一种由执行工作站执行的用于分布式计算的数据
处理方法，包括：将具有预定数据格式的聚合子拆分为多个聚合子片段；将
25 多个聚合子片段中的第一部分聚合子片段发送至其他执行工作站用于在其他
执行工作站上进行归约；从其他执行工作站接收其他执行工作站产生的与多个
聚合子片段中的第二部分聚合子片段对应的聚合子片段并将其与第二部分聚
合子片段进行归约以更新第二部分聚合子片段，第一部分不同于第二部分；在
从多个聚合子片段中去除已发送的第一部分聚合子片段的剩余部分中重新确
30 定第一部分聚合子片段和第二部分聚合子片段并执行上述发送、接收和归约，
直至多个聚合子片段的每一个都被发送或完成所有归约；以及如果存在完成所

有归约的聚合子片段，则将完成所有归约的聚合子片段发送至管理工作站，用于管理工作站将其与其他完成所有归约的聚合子片段拼接为聚合子。

例如，在本公开至少一实施例提供的数据处理方法中，聚合子被拆分的聚合子片段的数目等于参与聚合的执行工作站的数目；并且第一部分聚合子片段和第二部分聚合子片段的数目分别为一个。

例如，在本公开至少一实施例提供的数据处理方法中，聚合子被拆分的聚合子片段的数目等于参与聚合的执行工作站的数目乘以每个执行工作站的并行通信线路数；并且第一部分聚合子片段和第二部分聚合子片段的数目分别为每个执行工作站的并行通信线路数。

例如，在本公开至少一实施例提供的数据处理方法中，在从多个聚合子片段中去除已发送的第一部分聚合子片段的剩余部分中重新确定的第一部分聚合子片段是在前一轮的接收和归约中被更新的第二部分聚合子片段。

本公开至少一实施例提供一种用于分布式计算的执行工作站，包括：处理器和存储器，存储器存储一个或多个计算机程序指令，一个或多个计算机程序指令被存储在存储器中，并由处理器执行时实现本公开至少一实施例提供的数据处理方法的步骤。

本公开至少一实施例提供一种由分布式计算系统执行的数据处理方法，分布式计算系统包括多个执行工作站和管理工作站，数据处理方法包括：多个执行工作站中的每个执行以下步骤：将具有预定数据格式的聚合子拆分为多个聚合子片段；将多个聚合子片段中的第一部分聚合子片段发送至其他执行工作站用于在其他执行工作站上进行归约；从其他执行工作站接收其他执行工作站产生的与多个聚合子片段中的第二部分聚合子片段对应的聚合子片段并将其与第二部分聚合子片段进行归约以更新第二部分聚合子片段，第一部分不同于第二部分；在从多个聚合子片段中去除已发送的第一部分聚合子片段的剩余部分中重新确定第一部分聚合子片段和第二部分聚合子片段并执行上述发送、接收和归约，直至多个聚合子片段的每一个都被发送或完成所有归约；并且如果存在完成所有归约的聚合子片段，则将完成所有归约的聚合子片段发送至管理工作站；以及管理工作站接收多个执行工作站发送的完成所有归约的聚合子片段，并将所接收的聚合子片段拼接为聚合子。

例如，在本公开至少一实施例提供的数据处理方法中，聚合子被拆分的聚合子片段的数目等于多个执行工作站的数目；并且第一部分聚合子片段和第二

部分聚合子片段的数目分别为一个。

例如，在本公开至少一实施例提供的数据处理方法中，多个执行工作站的数量为 N ，并被顺序编号为 $0, 1, \dots, i, \dots, N-1$ ；多个聚合子片段的数量为 N ，并被顺序编号为 $0, 1, \dots, i, \dots, N-1$ ；多个执行工作站如下执行上述发送、接收和归约的步骤：第一步：设 $j=N-1$ ，第二步：对于 $i=0$ 至 $N-1$ ，第 i 个执行工作站将其第 $i+j+1$ 个聚合子片段发送至第 $i+1$ 个执行工作站，以及从第 $i-1$ 个执行工作站接收第 $i+j$ 个聚合子片段并进行归约，其中，若 $i-1 < 0$ ，则用 $i+N-1$ 代替 $i-1$ ，若 $i+j > N-1$ ，则用 $i+j-N$ 代替 $i+j$ ，若 $i+j+1 > N-1$ ，则用 $i+j+1-N$ 代替 $i+j+1$ ；若 $j > 1$ ，则将 j 值减 1 后重复第二步；第三步：对于 $i=0$ 至 $N-1$ ，第 i 个执行工作站将第 $i+1$ 个聚合子片段发送至管理工作站，并且若 $i+1=N$ ，则用 0 替代 $i+1$ 。

例如，在本公开至少一实施例提供的数据处理方法中，聚合子被拆分的聚合子片段的数目等于多个执行工作站的数目乘以每个执行工作站的并行通信线路数；并且第一部分聚合子片段和第二部分聚合子片段的数目分别为每个执行工作站的并行通信线路数。

本公开至少一实施例提供一种分布式计算系统，分布式计算系统包括多个执行工作站和管理工作站，多个执行工作站中的每个用于：将具有预定数据格式的聚合子拆分为多个聚合子片段；将多个聚合子片段中的第一部分聚合子片段发送至其他执行工作站用于在其他执行工作站上进行归约；从其他执行工作站接收其他执行工作站产生的与多个聚合子片段中的第二部分聚合子片段对应的聚合子片段并将其与第二部分聚合子片段进行归约以更新第二部分聚合子片段，第一部分不同于第二部分；在从多个聚合子片段中去除已发送的第一部分聚合子片段的剩余部分中重新确定第一部分聚合子片段和第二部分聚合子片段并执行上述发送、接收和归约，直至多个聚合子片段的每一个都被发送或完成所有归约；并且如果存在完成所有归约的聚合子片段，则将完成所有归约的聚合子片段发送至管理工作站；以及管理工作站用于接收多个执行工作站发送的完成所有归约的聚合子片段，并将所接收的聚合子片段拼接为聚合子。

本公开至少一实施例提供一种计算机可读存储介质，非暂时性存储有计算机可读指令，当计算机可读指令由处理器执行时实现本公开至少一实施例提供的数据处理方法。

根据本公开实施例的数据处理方法、执行工作站、分布式计算系统和计算机可读存储介质能够将聚合子进行拆分后归约，提高数据聚合的并行性，从而

提高处理性能。

附图说明

5 为了更清楚地说明本公开实施例的技术方案，下面将对实施例的附图作简单地介绍，显而易见地，下面描述中的附图仅仅涉及本公开的一些实施例，而非对本公开的限制。

图 1 示出了本公开至少一实施例提供的拆分归约和非拆分归约之间的对比示意图；

10 图 2A 示出了本公开至少一实施例提供的一种由执行工作站执行的用于分布式计算的数据处理方法的流程图；

图 2B 示出了本公开至少一实施例提供的由分布式计算系统执行的数据处理方法的流程图；

图 3 示出了本公开至少一实施例提供的执行工作站的通信基础结构的拓扑示意图；

15 图 4 示出了本公开至少一实施例提供的多个执行工作站执行发送、接收和归约过程的流程图；

图 5 示出了图 4 所提供的算法在执行工作站的数量为 4 的情况下的示意图；

图 6 示出了本公开至少一实施例提供的一种分布式计算系统的示意框图；

20 图 7 示出了本公开至少一实施例提供的一种用于分布式计算的执行工作站的示意框图；以及

图 8 示出了本公开至少一实施例提供的一种计算机可读存储介质的示意图。

具体实施方式

25 为使本公开实施例的目的、技术方案和优点更加清楚，下面将结合本公开实施例的附图，对本公开实施例的技术方案进行清楚、完整地描述。显然，所描述的实施例是本公开的一部分实施例，而不是全部的实施例。基于所描述的本公开的实施例，本领域普通技术人员在无需创造性劳动的前提下所获得的所有其它实施例，都属于本公开保护的范围。

30 除非另外定义，本公开使用的技术术语或者科学术语应当为本公开所属领域内具有一般技能的人士所理解的通常意义。本公开中使用的“第一”、“第二”

以及类似的词语并不表示任何顺序、数量或者重要性，而只是用来区分不同的组成部分。同样，“一个”、“一”或者“该”等类似词语也不表示数量限制，而是表示存在至少一个。“包括”或者“包含”等类似的词语意指出现该词前面的元件或者物件涵盖出现在该词后面列举的元件或者物件及其等同，而不排除其它元件或者物件。“连接”或者“相连”等类似的词语并非限定于物理的或者机械的连接，而是可以包括电性的连接，不管是直接的还是间接的。“上”、“下”、“左”、“右”等仅用于表示相对位置关系，当被描述对象的绝对位置改变后，则该相对位置关系也可能相应地改变。

MLlib 的可扩展性较差，其瓶颈在于树聚合中的归约步骤。可扩展的归约算法（例如，Rabenseifner 归约算法）的性能比树聚合的性能好，然而，Spark 中的不可拆分对象接口由于缺少对象拆分功能，从而使得 MLlib 不支持可扩展的归约算法。因此，发明人设计了一种新的技术方案，其能够将对象进行拆分以获得更多的并行性，从而能够适用于可扩展的归约算法。

本公开至少一个实施例提供一种数据处理方法、执行工作站、分布式计算系统和计算机可读存储介质。例如，该数据处理方法用于分布式计算并由执行工作站执行，其可以包括：将具有预定数据格式的聚合子拆分为多个聚合子片段；将多个聚合子片段中的第一部分聚合子片段发送至其他执行工作站用于在其他执行工作站上进行归约；从其他执行工作站接收其他执行工作站产生的与多个聚合子片段中的第二部分聚合子片段对应的聚合子片段并将其与第二部分聚合子片段进行归约以更新第二部分聚合子片段，第一部分不同于第二部分；在从多个聚合子片段中去除已发送的第一部分聚合子片段的剩余部分中重新确定第一部分聚合子片段和第二部分聚合子片段并执行上述发送、接收和归约，直至多个聚合子片段的每一个都被发送或完成所有归约；以及如果存在完成所有归约的聚合子片段，则将完成所有归约的聚合子片段发送至管理工作站，用于管理工作站将其与其他完成所有归约的聚合子片段拼接为聚合子。

该实施例的数据处理方法可以将要被归约的对象拆分成小块（片段）分别归约以获得更高的并行性，从而提高处理性能，并因此能够适用于可扩展的归约算法。

需要说明的是，大数据处理框架包括但不限于 Spark，本公开至少一实施例提供的数据处理方法还可以适用于其它大数据处理框架。

本公开至少一实施例提供的分布式计算系统可以包括管理工作站和执行

工作站和内部存储器。执行工作站是分布式计算系统中进行任务处理的计算设备，例如，其可以运行 Spark 架构中的执行器（executor）；管理工作站是分布式计算系统中负责管理各个执行工作站的计算设备，例如，其可以运行 Spark 架构中的驱动器（driver），管理工作站可以协调、调度、并监控每个执行工作站的5 任务。执行工作站的任務結果最終可以汇总到管理工作站中。管理工作站通过通信网络与多个执行工作站交互，并且执行工作站通过通信网络与其它执行工作站交互，以接收或发送消息。通信网络用以在管理工作站和多个执行工作站之间以及多个执行工作站之间提供通信链路的介质。通信网络可以包括各种连接类型，例如有线或无线通信链路，具体地例如 WIFI、3G、4G、5G 10 和光纤电缆等。

在本公开的实施例中，归约表示将两个或多个数据进行合并，合并后的总数据量小于合并前的总数据量，例如可以将两个或多个数据进行求和等运算。聚合子表示具有预定数据格式的数据单元，其可以用于聚合（合并）或者其它计算。预定数据结构可以根据具体的应用而预先确定，例如，可以是数组，也15 可以是用户定义的其它数据结构。

图 1 示出了本公开至少一实施例提供的拆分归约和非拆分归约之间的对比示意图。

如图 1 所示，在示意图左右两侧示出的归约过程都展示了将多个聚合子归约为一个聚合子。该示意图左侧示出了非拆分归约，而示意图右侧示出了拆分20 归约。

图中四个聚合子 V_1 、 V_2 、 V_3 、 V_4 被归约为一个聚合子。非拆分归约将聚合子视为不可拆分的对象，直接将 V_1 、 V_2 、 V_3 、 V_4 归约为 V^* 。而对于拆分归约，每个聚合子 V_i ($i=1, 2, 3, 4$) 都是可拆分的并且每个聚合子 V_i ($i=1, 2, 3, 4$) 可以被拆分为例如三个片段 $V_{i,1}$ 、 $V_{i,2}$ 、 $V_{i,3}$ 。聚合子 V_i ($i=1, 2, 3, 4$) 25 的第一个片段 $V_{1,1}$ 、 $V_{2,1}$ 、 $V_{3,1}$ 、 $V_{4,1}$ 为一组，类似地，第二个片段 $V_{1,2}$ 、 $V_{2,2}$ 、 $V_{3,2}$ 、 $V_{4,2}$ 为一组，第三个片段 $V_{1,3}$ 、 $V_{2,3}$ 、 $V_{3,3}$ 、 $V_{4,3}$ 为一组。每组片段可以被独立地归约以分别形成结果 $V_{*,1}$ 、 $V_{*,2}$ 、 $V_{*,3}$ 。与非拆分归约相比，拆分归约获得了 3 倍的并行性，因此提高了处理性能。

图 2A 示出了本公开至少一实施例提供的一种由执行工作站执行的用于分布式计算的数据处理方法的流程图。30

如图 2A 所示，该数据处理方法包括步骤 S201~步骤 S205。

步骤 S201-S205 由执行工作站执行。

步骤 S201: 将具有预定数据格式的聚合子拆分为多个聚合子片段。

步骤 S202: 将多个聚合子片段中的第一部分聚合子片段发送至其他执行工作站用于在其他执行工作站上进行归约。

5 步骤 S203: 从其他执行工作站接收其他执行工作站产生的与多个聚合子片段中的第二部分聚合子片段对应的聚合子片段并将其与第二部分聚合子片段进行归约以更新第二部分聚合子片段, 第一部分不同于第二部分。

10 步骤 S204: 在从多个聚合子片段中去除已发送的第一部分聚合子片段的剩余部分中重新确定第一部分聚合子片段和第二部分聚合子片段并执行上述发送、接收和归约 (即步骤 S202 和 S203), 直至多个聚合子片段的每一个都被发送或完成所有归约。

步骤 S205: 如果存在完成所有归约的聚合子片段, 则将完成所有归约的聚合子片段发送至管理工作站, 用于管理工作站将其与其他完成所有归约的聚合子片段拼接为聚合子。

15 对于步骤 S201, 聚合子通常是作为数据聚合或归约的数据单位, 其数据格式可以根据具体的应用而预先确定, 例如, 可以是数组, 也可以是用户定义的其他数据结构。在本公开的实施例, 聚合子被拆分为多个聚合子片段, 每个聚合子片段可以独立地进行归约, 从而提高了并行性。聚合子片段的划分可以根据应用的需要而确定, 例如, 可以根据希望并行处理的数量确定聚合子片段的数目, 再根据聚合子片段的数目对聚合子进行划分, 例如均匀划分或不均匀划分。例如, 聚合子片段的数目可以等于参与聚合的执行工作站的数目, 在此情况下, 每个执行工作站可以同时各自进行一个聚合子片段的归约, 从而可以充分利用多个执行工作站的并行性。

25 对于步骤 S202, 第一部分聚合子片段是指发送至其它执行工作站并在其它执行工作站上进行归约的聚合子片段。其它执行工作站是指分布式计算系统中不同于本执行工作站且参与聚合的一个或多个执行工作站。第一部分聚合子片段可以是一个或多个片段, 这些片段被发送至其它执行工作站上与其它工作站上的对应聚合子片段进行归约。“对应聚合子片段”表示该聚合子片段在其聚合子中的位置与被发送的聚合子片段在其聚合子中的位置相同。在归约时, 两个对应的聚合子片段被合并计算为一个片段。在第一部分具有多个片段时, 该第一部分可以发送给同一个其它执行工作站, 也可以分别发送给不同的其它执

行工作站。

对于步骤 S203, 第二部分聚合子片段是指与接收自其它执行工作站的聚合子片段进行归约的聚合子片段, 即, 第二部分聚合子片段是在本执行工作站进行归约的聚合子片段。需要说明的是, 步骤 S203 中的其它执行工作站与步骤 5 202 中的其它执行工作站可以相同也可以不同。此外, 在聚合子被拆分的聚合子片段的数目等于参与聚合的执行工作站的数目的情况下, 第一部分聚合子片段和第二部分聚合子片段的数目可以分别为一个, 从而分布式系统中的所有执行工作站可以每次分别归约一个聚合子片段, 以便充分利用多个执行工作站的并行性。

10 对于步骤 S204, 在去除已经发送的聚合子片段后重新确定第一部分聚合子片段和第二部分聚合子片段, 并迭代执行上述步骤 S202 和 S203, 直至多个聚合子片段的每一个都被发送或完成所有归约。这里所述的“所有归约”表示该聚合子片段已经聚合了分布式系统需要归约的所有聚合子的对应片段, 例如, 分布式系统共有 M 个执行工作站, 每个执行工作站有一个聚合子需要归约, 则 15 如果某个聚合子片段已经聚合了 M 个执行工作站的 M 个聚合子中的对应片段, 则表示该聚合子片段已经完成了所有归约。对于第一部分和第二部分的重新确定可以根据应用的需要设定, 例如, 在从多个聚合子片段中去除已发送的第一部分聚合子片段的剩余部分中重新确定的第一部分聚合子片段可以是在前一轮的接收和归约中被更新的第二部分聚合子片段, 即优先将已经归约过的聚合子片段进行发送。对于步骤 S205, 执行工作站将步骤 S204 中完成归约的聚合子片段发送给管理工作站, 以便管理工作站将完成所有归约的聚合子片段拼接为结果聚合子。 20

图 2B 示出了本公开至少一实施例提供的由分布式计算系统执行的数据处理方法的流程图。分布式计算系统包括管理工作站和多个执行工作站

25 如图 2B 所示, 该数据处理方法包括步骤 S301~步骤 S306。

步骤 S301-S305 由每个执行工作站执行, 步骤 S306 由管理工作站执行。

步骤 S301: 将具有预定数据格式的聚合子拆分为多个聚合子片段。

步骤 S302: 将多个聚合子片段中的第一部分聚合子片段发送至其他执行工作站用于在其他执行工作站上进行归约。

30 步骤 S303: 从其他执行工作站接收其他执行工作站产生的与多个聚合子片段中的第二部分聚合子片段对应的聚合子片段并将其与第二部分聚合子片段

进行归约以更新第二部分聚合子片段，第一部分不同于第二部分。

步骤 S304: 在从多个聚合子片段中去除已发送的第一部分聚合子片段的剩余部分中重新确定第一部分聚合子片段和第二部分聚合子片段并执行上述发送、接收和归约，直至多个聚合子片段的每一个都被发送或完成所有归约。

5 步骤 S305: 如果存在完成所有归约的聚合子片段，则将完成所有归约的聚合子片段发送至管理工作站。

步骤 S306: 管理工作站接收多个执行工作站发送的完成所有归约的聚合子片段，并将所接收的聚合子片段拼接为聚合子。

10 步骤 S301-S305 和图 2A 所示的步骤 S201-S205 相同，已经在上面进行了详细说明，在此不再赘述。

对于步骤 S306，管理工作站接收完成所有归约的聚合子片段，然后将这些聚合子片段拼接成一个结果聚合子。

15 图 3 示出了本公开至少一实施例提供的执行工作站的通信基础结构的拓扑示意图。需要说明的是，图 3 中每个执行工作站之间的连接关系仅表示它们之间的逻辑通信关系，并不表示实际网络连接关系，实际网络连接可以采用任意合适的方式，只要能够实现图 3 中所示的逻辑通信关系即可。

20 如图 3 所示，例如，分布式计算系统存在四个执行工作站，分别为执行工作站 0、执行工作站 1、执行工作站 2、执行工作站 3。四个执行工作站按并行的定向环的形式通信。每个执行工作站可以向其下一个执行工作站发送消息，并且从其前一个工作站接受消息，例如，执行工作站 3 向执行工作站 0 发送消息并从执行工作站 2 接受消息。图 3 中的箭头表示消息信道，相邻两个执行工作站之间可以有一个或多个并行消息信道，图 3 中示出了 2 个并行消息信道（并行通信线路），分别标示为“0”和“1”。

25 在本公开的一些实施例中，聚合子被拆分的聚合子片段的数目可以等于参与聚合的执行工作站的数目乘以每个执行工作站的并行通信线路数，并且第一部分聚合子片段和第二部分聚合子片段的数目分别为每个执行工作站的并行通信线路数。在此情况下，可以充分利用通信资源提高并行性。

30 例如，在图 3 所示的情况下，聚合子被拆分的聚合子片段的数目等于参与聚合的执行工作站的数目 4 乘以每个执行工作站的并行通信线路数 2，即被拆分成 8 个聚合子片段。第一部分聚合子片段和第二部分聚合子片段的数目分别为 2。从而，在每一轮处理时，一个执行工作站可以通过 2 个通信线路同时发

送 2 个聚合子片段并同时接收 2 个聚合子片段。

图 4 示出了本公开至少一个实施例提供的多个执行工作站执行发送、接收和归约过程的流程图。多个执行工作站的数量为 N ，并被顺序编号为 $0, 1, \dots, i, \dots, N-1$ ；多个聚合子片段的数量也为 N ，并被顺序编号为 $0, 1, \dots, i, \dots, N-1$ 。

5 如图 4 所示，该方法可以包括步骤 S401~步骤 S402。

步骤 S401：设 $j=N-1$ 。

步骤 S402：对于 $i=0$ 至 $N-1$ ，执行工作站 i 将其聚合子片段 $i+j+1$ 发送至执行工作站 $i+1$ ，以及从执行工作站 $i-1$ 接收聚合子片段 $i+j$ 并进行归约，其中，若 $i-1 < 0$ ，则用 $i+N-1$ 代替 $i-1$ ，若 $i+j > N-1$ ，则用 $i+j-N$ 代替 $i+j$ ，若 $i+j+1 > N-1$ ，
10 则用 $i+j+1-N$ 代替 $i+j+1$ ；

若 $j > 1$ ，则将 j 值减 1 后重复步骤 S402。

步骤 S403：对于 $i=0$ 至 $N-1$ ，执行工作站 i 将聚合子片段 $i+1$ 发送至管理工作站，并且若 $i+1=N$ ，则用 0 替代 $i+1$ 。

例如， $N=4$ ，即执行工作站的数量为 4，多个聚合子片段的数量为 4。设
15 $j=N-1=3$ 。

在第一轮迭代中，对于执行工作站 0，由于 $i-1 < 0$ ，则设 $i-1=i+N-1=3$ ，并且由于 $i+j+1 > N-1$ ，则设 $i+j+1=i+j+1-N=0$ ，执行工作站 0 将其聚合子片段 0 发送至执行工作站 1，以及从执行工作站 3 接收聚合子片段 3 并进行归约。对于执行工作站 1，由于 $i+j=1+3 > 3$ ，则设 $i+j=i+j-N=0$ ，并且由于 $i+j+1 > N-1$ ，则设
20 $i+j+1=i+j+1-N=1$ ，执行工作站 1 将其聚合子片段 1 发送至执行工作站 2，以及从执行工作站 0 接收聚合子片段 0 并进行归约。对于执行工作站 2，由于 $i+j=2+3 > 3$ ，则设 $i+j=i+j-N=1$ ，并且由于 $i+j+1 > N-1$ ，则设 $i+j+1=i+j+1-N=2$ ，执行工作站 2 将其聚合子片段 2 发送给执行工作站 3，以及从执行工作站 1 接受聚合子片段 1 并进行归约。对于执行工作站 3，由于 $i+j=3+3 > 3$ ，则设
25 $i+j=i+j-N=2$ ，并且由于 $i+j+1 > N-1$ ，则设 $i+j+1=i+j+1-N=3$ ，执行工作站 3 将其聚合子片段 3 发送至执行工作站 0，以及从执行工作站 2 接收聚合子片段 2 并进行归约。由于 $j=3 > 1$ ，则将 j 值减 1，此时 $j=2$ ，并进行第二轮迭代。

在第二轮迭代中，对于执行工作站 0，由于 $i-1 < 0$ ，则设 $i-1=i+N-1=3$ ，执行工作站 0 将其聚合子片段 3 发送给执行工作站 1，以及从执行工作站 3 接收
30 聚合子片段 2 并进行归约。对于执行工作站 1，执行工作站 1 将其聚合子片段 0 发送给执行工作站 2，以及从执行工作站 0 接收聚合子片段 3 并进行归约。

对于执行工作站 2，执行工作站 2 将其聚合子片段 1 发送给执行工作站 3，以及从执行工作站 1 接收聚合子片段 0 并进行归约。对于执行工作站 3，执行工作站 3 将其聚合子片段 2 发送给执行工作站 0，以及从执行工作站 2 接收聚合子片段 1 并进行归约。由于 $j=2>1$ ，则将 j 值减 1，此时 $j=1$ ，并进行第三轮迭代。

在第三轮迭代中，对于执行工作站 0，由于 $i-1<0$ ，则设 $i-1=i+N-1=3$ ，执行工作站 0 将其聚合子片段 2 发送给执行工作站 1，以及从执行工作站 3 接收聚合子片段 1 并进行归约。对于执行工作站 1，执行工作站 1 将其聚合子片段 3 发送给执行工作站 2，以及从执行工作站 0 接收聚合子片段 2 并进行归约。

10 对于执行工作站 2，由于 $i+j+1>N-1$ ，则设 $i+j+1=i+j+1-N=0$ ，执行工作站 2 将其聚合子片段 0 发送给执行工作站 3，以及从执行工作站 1 接收聚合子片段 3 并进行归约。对于执行工作站 3，由于 $i+j=3+1>3$ ，则设 $i+j=i+j-N=0$ ，并且由于 $i+j+1>N-1$ ，则设 $i+j+1=i+j+1-N=1$ ，执行工作站 3 将其聚合子片段 1 发送给执行工作站 0，以及从执行工作站 2 接收聚合子片段 0 并进行归约。由于 $j=1$ ，
15 循环结束，并进入到步骤 S403。

对于步骤 S403，执行工作站 0 将聚合子片段 1 发送至管理工作站，执行工作站 1 将聚合子片段 2 发送至管理工作站，执行工作站 2 将聚合子片段 3 发送至管理工作站。对于执行工作站 3，由于 $i+1=N$ ，则 $i+1=0$ ，执行工作站 3 将聚合子片段 0 发送至管理工作站。

20 图 5 示出了图 4 所提供的算法在执行工作站的数量为 4 的情况下的示意图。

如图 5 所示，4 个圆角矩形表示 4 个执行工作站，圆角矩形左侧的数字 0、1、2、3 表示执行工作站的编号，每个执行工作站将聚合子分成 4 个片段，用片段 0、片段 1、片段 2、片段 3 表示。标有圆形符号的矩形表示要被发送的当前值，标有正方形符号的矩形表示接收并合并来自其它执行工作站的聚合子片段的下一个值，标有三角形符号的矩形表示接收并合并来自其它执行工作站的聚合子片段的最终值。在第一轮迭代（迭代 0）中，执行工作站 0 将其聚合子片段 0 发送给执行工作站 1，以及从执行工作站 3 接收聚合子片段 3，执行工作站 1 将其聚合子片段 1 发送给执行工作站 2，以及从执行工作站 0 接收聚合子片段 0，执行工作站 2 将其聚合子片段 2 发送给执行工作站 3，以及从执行
25 工作站 1 接收聚合子片段 1，执行工作站 3 将其聚合子片段 3 发送给执行工作站 0，以及从执行工作站 2 接收聚合子片段 2。在第二轮迭代（迭代 1）中，执
30

行工作站 0 将其聚合子片段 3 发送给执行工作站 1，以及从执行工作站 3 接收聚合子片段 2，执行工作站 1 将其聚合子片段 0 发送给执行工作站 2，以及从执行工作站 0 接收聚合子片段 3，执行工作站 2 将其聚合子片段 1 发送给执行工作站 3，以及从执行工作站 1 接收聚合子片段 0，执行工作站 3 将其聚合子片段 2 发送给执行工作站 0，以及从执行工作站 2 接收聚合子片段 1。在第三轮迭代（迭代 2）中，执行工作站 0 将其聚合子片段 2 发送给执行工作站 1 并进行合并形成最终值，以及从执行工作站 3 接收聚合子片段 1 并与当前值进行合并形成最终值，执行工作站 1 将其聚合子片段 3 发送给执行工作站 2 并进行合并形成最终值，以及从执行工作站 0 接收聚合子片段 2 并与当前值合并形成最终值，执行工作站 2 将其聚合子片段 0 发送给执行工作站 3 并进行合并形成最终值，以及从执行工作站 1 接收聚合子片段 3 并与当前值合并形成最终值，执行工作站 3 将其聚合子片段 1 发送给执行工作站 0 并进行合并形成最终值，以及从执行工作站 2 接收聚合子片段 0 并与当前值合并形成最终值。

图 6 示出了本公开至少一实施例提供的一种分布式计算系统的示意框图。
该分布式计算系统 600 包括多个执行工作站 610 和管理工作站 620。

多个执行工作站 610 中的每个配置为：将具有预定数据格式的聚合子拆分为多个聚合子片段；将多个聚合子片段中的第一部分聚合子片段发送至其他执行工作站用于在其他执行工作站上进行归约；从其他执行工作站接收其他执行工作站产生的与多个聚合子片段中的第二部分聚合子片段对应的聚合子片段并将其与第二部分聚合子片段进行归约以更新第二部分聚合子片段，第一部分不同于第二部分；在从多个聚合子片段中去除已发送的第一部分聚合子片段的剩余部分中重新确定第一部分聚合子片段和第二部分聚合子片段并执行上述发送、接收和归约，直至多个聚合子片段的每一个都被发送或完成所有归约；并且如果存在完成所有归约的聚合子片段，则将完成所有归约的聚合子片段发送至管理工作站。

执行工作站 610 例如可以执行图 2B 描述的步骤 S301-S305。

管理工作站 620 配置为接收多个执行工作站发送的完成所有归约的聚合子片段，并将所接收的聚合子片段拼接为聚合子。

管理工作站 620 例如可以执行图 2B 描述的步骤 S306。

例如，执行工作站 610 和管理工作站 620 可以实现为硬件、软件、固件以及它们的任意可行的组合。例如，执行工作站 610 和管理工作站 620 可以为专

用或通用的电路、芯片或装置等，也可以为处理器和存储器的结合。关于上述各个模块的具体实现形式，本公开的实施例对此不作限制。

需要说明的是，本公开的实施例中，用于分布式计算系统 600 的各个模块与前述的数据处理方法的各个步骤对应，关于用于分布式计算系统 600 的具体功能可以参考关于数据处理方法的相关描述，此处不再赘述。图 6 所示的用于分布式计算系统 600 的组件和结构只是示例性的，而非限制性的，根据需要，该用于分布式计算系统 600 还可以包括其它组件和结构。

本公开的至少一个实施例还提供了一种用于分布式计算的执行工作站，该用于分布式计算的执行工作站包括处理器和存储器，存储器包括一个或多个计算机程序模块。一个或多个计算机程序模块被存储在存储器中并被配置为由处理器执行，一个或多个计算机程序模块包括用于实现上述的数据处理方法的指令。该用于分布式计算的执行工作站可以将要被归约的对象拆分成小块以获得更高的并行性，从而提高处理性能，并因此能够适用于可扩展的归约算法。

图 7 示出了本公开至少一实施例提供的一种用于分布式计算的执行工作站 700 的示意框图。如图 7 所示，该用于分布式计算的执行工作站 700 包括处理器 710 和存储器 720。存储器 720 用于存储非暂时性计算机可读指令（例如一个或多个计算机程序模块）。处理器 710 用于运行非暂时性计算机可读指令，非暂时性计算机可读指令被处理器 710 运行时可以执行上文所述的数据处理方法中的一个或多个步骤。存储器 720 和处理器 710 可以通过总线系统和/或其它形式的连接机构（未示出）互连。

例如，处理器 710 可以是中央处理单元（CPU）、图形处理单元（GPU）或者具有数据处理能力和/或程序执行能力的其它形式的处理单元。例如，中央处理单元（CPU）可以为 X86 或 ARM 架构等。处理器 710 可以为通用处理器或专用处理器，可以控制用于分布式计算的执行工作站 700 中的其它组件以执行期望的功能。

例如，存储器 720 可以包括一个或多个计算机程序产品的任意组合，计算机程序产品可以包括各种形式的计算机可读存储介质，例如易失性存储器和/或非易失性存储器。易失性存储器例如可以包括随机存取存储器（RAM）和/或高速缓冲存储器（cache）等。非易失性存储器例如可以包括只读存储器（ROM）、硬盘、可擦除可编程只读存储器（EPROM）、便携式紧致盘只读存储器（CD-ROM）、USB 存储器、闪存等。在计算机可读存储介质上可以存

储一个或多个计算机程序模块，处理器 710 可以运行一个或多个计算机程序模块，以实现用于分布式计算的执行工作站 700 的各种功能。在计算机可读存储介质中还可以存储各种应用程序和各种数据以及应用程序使用和/或产生的各种数据等。

5 需要说明的是，本公开的实施例中，用于分布式计算的执行工作站 700 的具体功能和技术效果可以参考上文中关于数据处理方法的描述，此处不再赘述。

本公开的至少一个实施例还提供了一种计算机可读存储介质，该计算机可读存储介质用于存储非暂时性计算机可读指令，当非暂时性计算机可读指令由
10 计算机执行时可以实现上述的数据处理方法。利用该计算机可读存储介质，可以将要被归约的对象拆分成小块以获得更多的并行性，从而提高了归约的可扩展性。

图 8 为本公开一些实施例提供的一种存储介质的示意图。如图 8 所示，存储介质 800 用于存储非暂时性计算机可读指令 810。例如，当非暂时性计算机
15 可读指令 810 由计算机执行时可以执行根据上文所述的数据处理方法中的一个或多个步骤。

例如，该存储介质 800 可以应用于上述用于分布式计算的执行工作站 700 中。例如，存储介质 800 可以为图 7 所示的用于分布式计算的执行工作站 700 中的存储器 720。例如，关于存储介质 800 的相关说明可以参考图 7 所示的
20 用于分布式计算的执行工作站 700 中的存储器 720 的相应描述，此处不再赘述。需要说明的是，本公开实施例附图只涉及到本公开实施例涉及到的结构，其它结构可参考通常设计。在不冲突的情况下，本公开的实施例及实施例中的特征可以相互组合以得到新的实施例。

以上所述，仅为本公开的具体实施方式，但本公开的保护范围并不局限于
25 此，本公开的保护范围应以所述权利要求要求的保护范围为准。

权利要求书

1、一种由执行工作站执行的用于分布式计算的数据处理方法，包括：

将具有预定数据格式的聚合子拆分为多个聚合子片段；

5 将所述多个聚合子片段中的第一部分聚合子片段发送至其它执行工作站用于在其它执行工作站上进行归约；

从其它执行工作站接收其它执行工作站产生的与所述多个聚合子片段中的第二部分聚合子片段对应的聚合子片段并将其与所述第二部分聚合子片段进行归约以更新所述第二部分聚合子片段，所述第一部分不同于所述第二部分；

10 在从所述多个聚合子片段中去除已发送的第一部分聚合子片段的剩余部分中重新确定第一部分聚合子片段和第二部分聚合子片段并执行上述发送、接收和归约，直至所述多个聚合子片段的每一个都被发送或完成所有归约；以及

15 如果存在完成所有归约的聚合子片段，则将所述完成所有归约的聚合子片段发送至管理工作站，用于所述管理工作站将其与其它完成所有归约的聚合子片段拼接为聚合子。

2、根据权利要求 1 所述的数据处理方法，其中

所述聚合子被拆分的聚合子片段的数目等于参与聚合的执行工作站的数目；并且

20 所述第一部分聚合子片段和所述第二部分聚合子片段的数目分别为一个。

3、根据权利要求 1 所述的数据处理方法，其中

所述聚合子被拆分的聚合子片段的数目等于参与聚合的执行工作站的数目乘以每个执行工作站的并行通信线路数；并且

25 所述第一部分聚合子片段和所述第二部分聚合子片段的数目分别为每个执行工作站的并行通信线路数。

4、根据权利要求 1-3 中的任一项所述的数据处理方法，其中

在从所述多个聚合子片段中去除已发送的第一部分聚合子片段的剩余部分中重新确定的第一部分聚合子片段是在前一轮的接收和归约中被更新的第二部分聚合子片段。

30 5、一种用于分布式计算的执行工作站，包括处理器和存储器，所述存储器存储一个或多个计算机程序指令，其中，所述一个或多个计算机程序指令被

存储在所述存储器中，并由所述处理器执行时实现权利要求 1-4 任一项所述的数据处理方法的步骤。

6、一种由分布式计算系统执行的数据处理方法，所述分布式计算系统包括多个执行工作站和管理工作站，其中所述数据处理方法包括：

5 所述多个执行工作站中的每个执行以下步骤：

将具有预定数据格式的聚合子拆分为多个聚合子片段；

将所述多个聚合子片段中的第一部分聚合子片段发送至其它执行工作站用于在其它执行工作站上进行归约；

10 从其它执行工作站接收其它执行工作站产生的与所述多个聚合子片段中的第二部分聚合子片段对应的聚合子片段并将其与所述第二部分聚合子片段进行归约以更新所述第二部分聚合子片段，所述第一部分不同于所述第二部分；

15 在从所述多个聚合子片段中去除已发送的第一部分聚合子片段的剩余部分中重新确定第一部分聚合子片段和第二部分聚合子片段并执行上述发送、接收和归约，直至所述多个聚合子片段的每一个都被发送或完成所有归约；并且

如果存在完成所有归约的聚合子片段，则将所述完成所有归约的聚合子片段发送至管理工作站；以及

20 所述管理工作站接收所述多个执行工作站发送的完成所有归约的聚合子片段，并将所接收的聚合子片段拼接为聚合子。

7、根据权利要求 6 所述的数据处理方法，其中

所述聚合子被拆分的聚合子片段的数目等于所述多个执行工作站的数目；并且

所述第一部分聚合子片段和所述第二部分聚合子片段的数目分别为一个。

25 8、根据权利要求 6 或 7 所述的数据处理方法，其中

所述多个执行工作站的数量为 N ，并被顺序编号为 $0, 1, \dots, i, \dots, N-1$ ；

所述多个聚合子片段的数量为 N ，并被顺序编号为 $0, 1, \dots, i, \dots, N-1$ ；

所述多个执行工作站如下执行上述发送、接收和归约的步骤：

第一步：设 $j=N-1$ ，

30 第二步：对于 $i=0$ 至 $N-1$ ，第 i 个执行工作站将其第 $i+j+1$ 个聚合子片段发送至第 $i+1$ 个执行工作站，以及从第 $i-1$ 个执行工作站接收第 $i+j$ 个聚

合子片段并进行归约，其中，若 $i-1 < 0$ ，则用 $i+N-1$ 代替 $i-1$ ，若 $i+j > N-1$ ，则用 $i+j-N$ 代替 $i+j$ ，若 $i+j+1 > N-1$ ，则用 $i+j+1-N$ 代替 $i+j+1$ ；

若 $j > 1$ ，则将 j 值减 1 后重复第二步；

5 第三步：对于 $i=0$ 至 $N-1$ ，第 i 个执行工作站将第 $i+1$ 个聚合子片段发送至所述管理工作站，其中若 $i+1=N$ ，则用 0 替代 $i+1$ 。

9、根据权利要求 6-8 中的任一项所述的数据处理方法，其中

所述聚合子被拆分的聚合子片段的数目等于所述多个执行工作站的数目乘以每个执行工作站的并行通信线路数；并且

10 所述第一部分聚合子片段和所述第二部分聚合子片段的数目分别为每个执行工作站的并行通信线路数。

10、一种分布式计算系统，所述分布式计算系统包括多个执行工作站和管理工作站，其中

所述多个执行工作站中的每个用于：

将具有预定数据格式的聚合子拆分为多个聚合子片段；

15 将所述多个聚合子片段中的第一部分聚合子片段发送至其它执行工作站用于在其它执行工作站上进行归约；

20 从其它执行工作站接收其它执行工作站产生的与所述多个聚合子片段中的第二部分聚合子片段对应的聚合子片段并将其与所述第二部分聚合子片段进行归约以更新所述第二部分聚合子片段，所述第一部分不同于所述第二部分；

在从所述多个聚合子片段中去除已发送的第一部分聚合子片段的剩余部分中重新确定第一部分聚合子片段和第二部分聚合子片段并执行上述发送、接收和归约，直至所述多个聚合子片段的每一个都被发送或完成所有归约；并且

25 如果存在完成所有归约的聚合子片段，则将所述完成所有归约的聚合子片段发送至管理工作站；以及

所述管理工作站用于接收所述多个执行工作站发送的完成所有归约的聚合子片段，并将所接收的聚合子片段拼接为聚合子。

30 11、一种计算机可读存储介质，非暂时性存储有计算机可读指令，其中，当所述计算机可读指令由处理器执行时实现权利要求 1-4 任一项所述的数据处理方法。

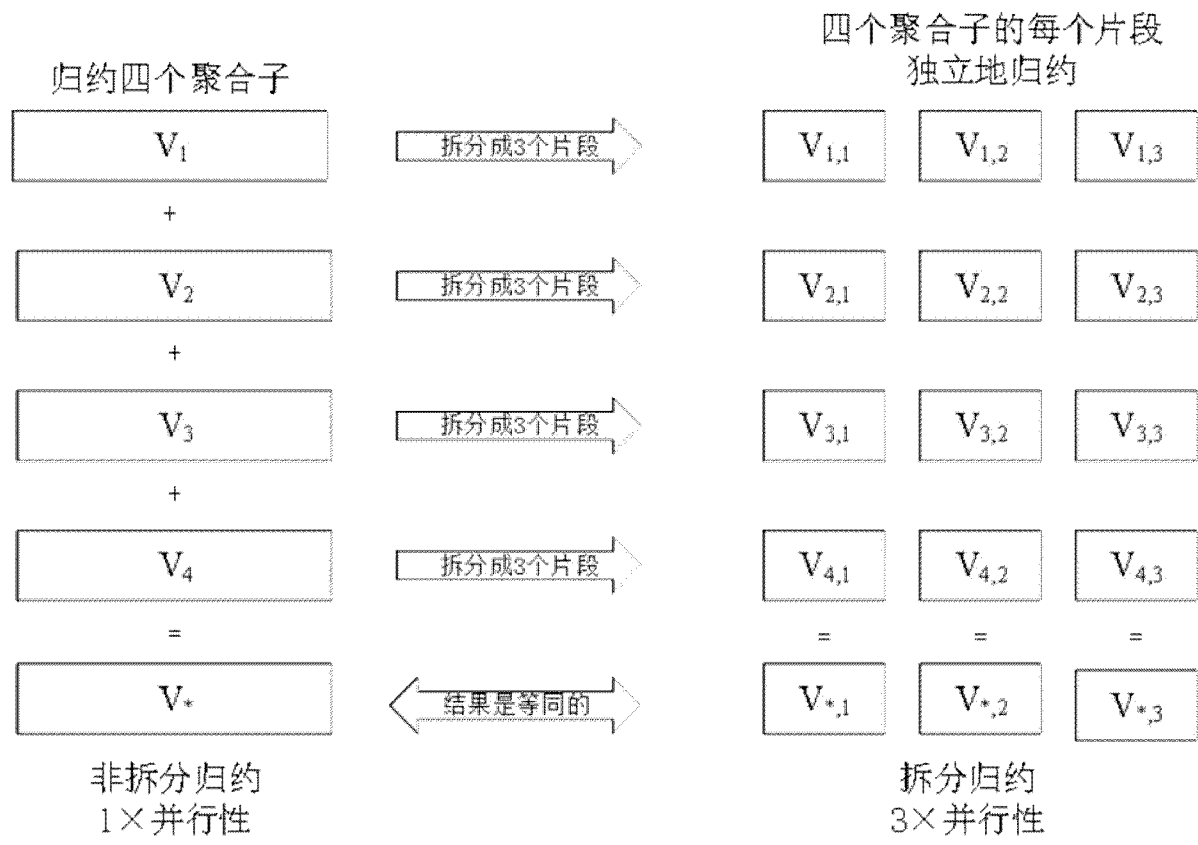


图 1

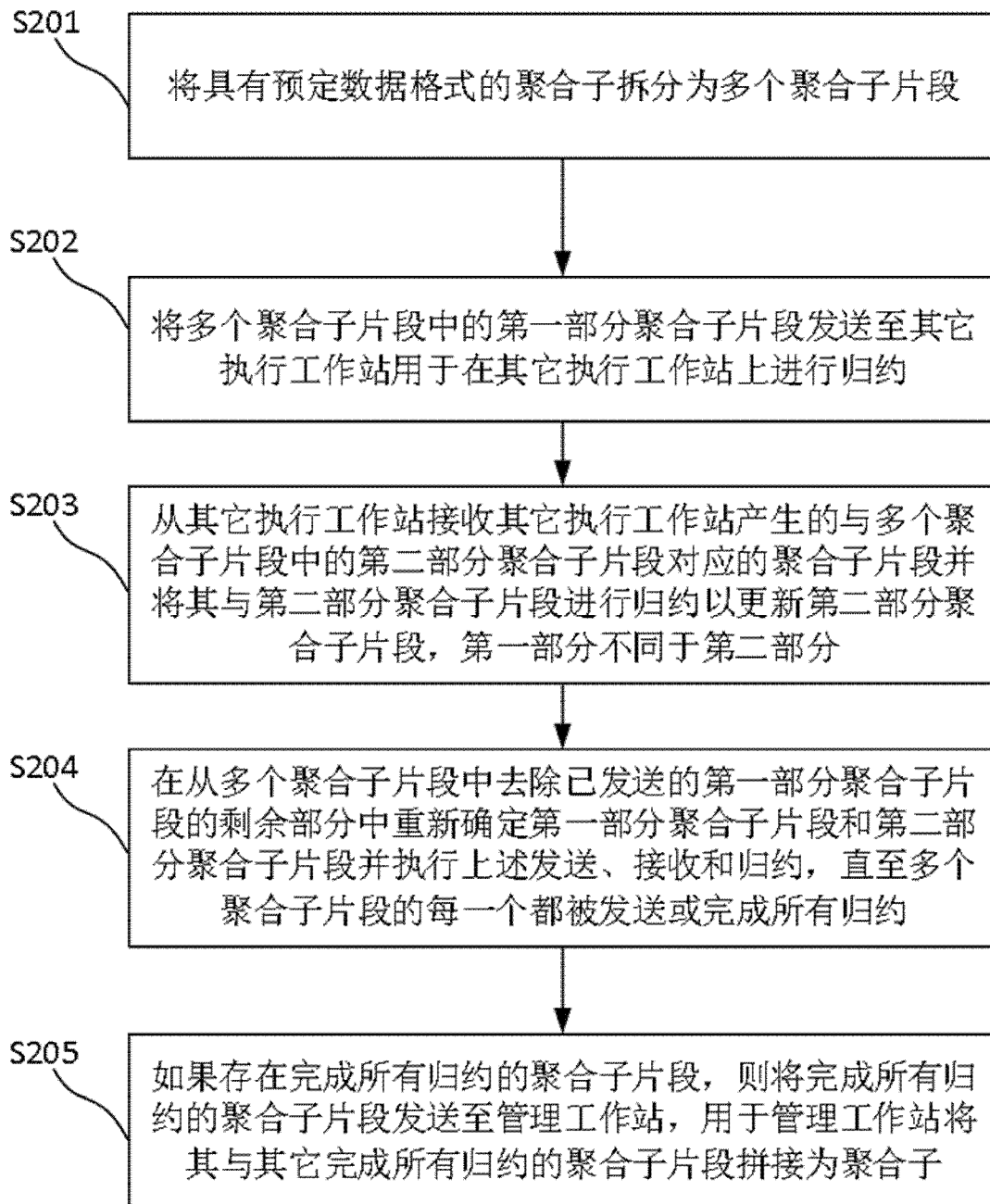


图 2A

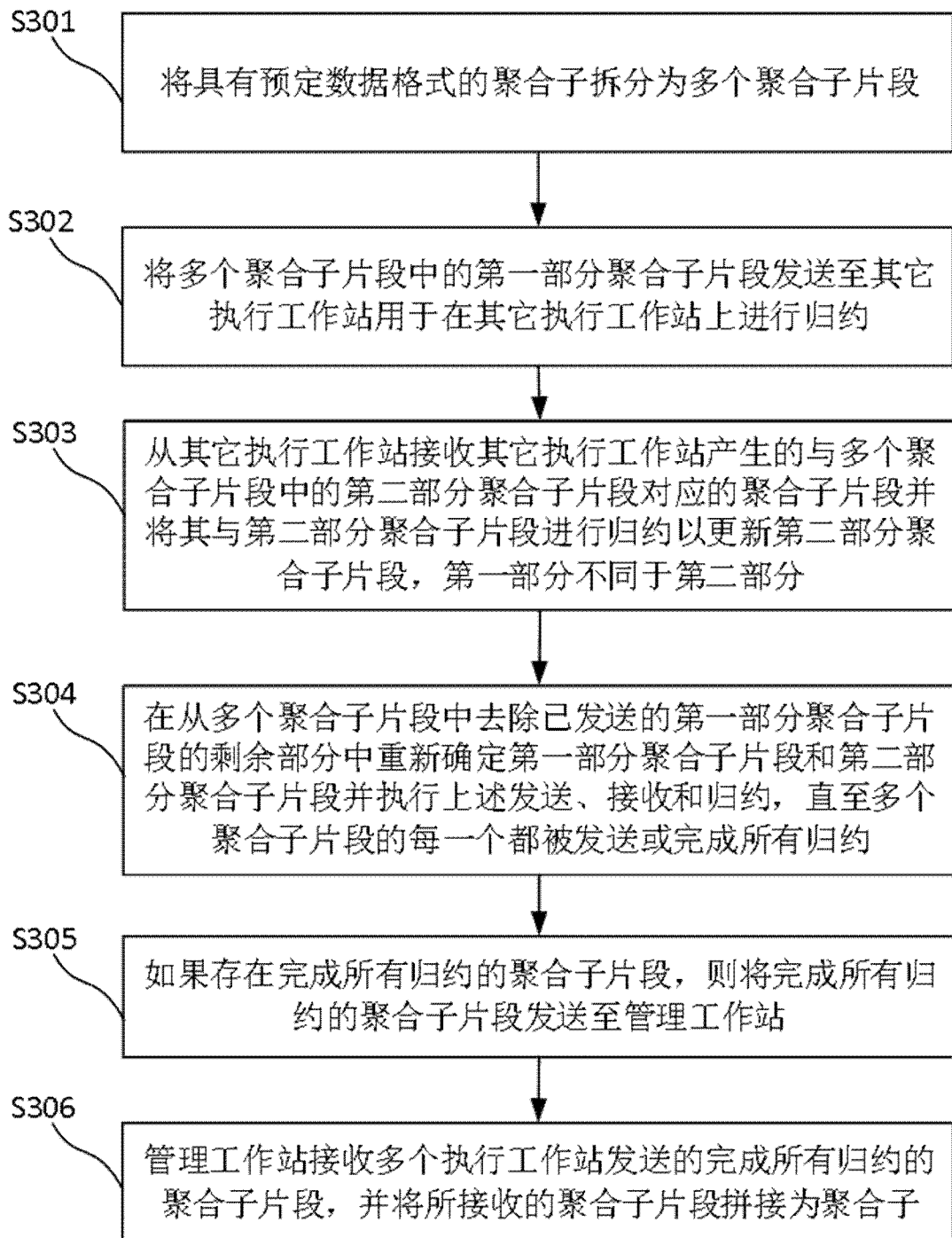


图 2B

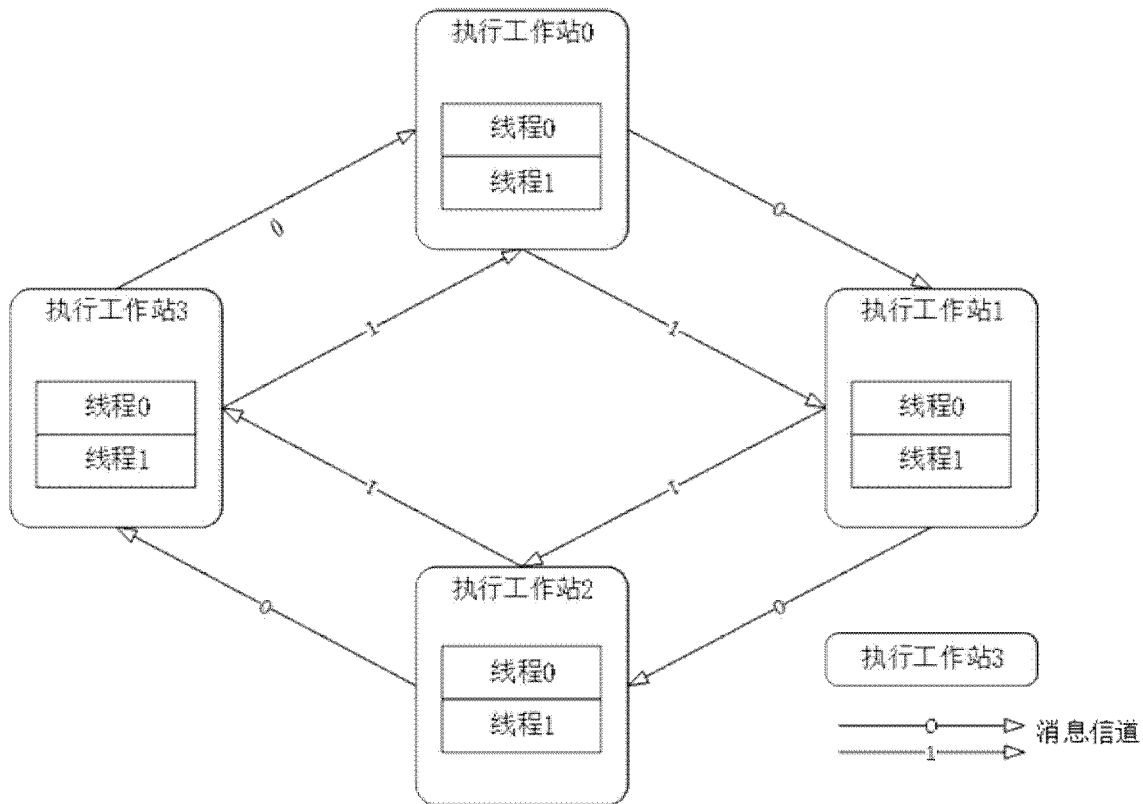


图 3

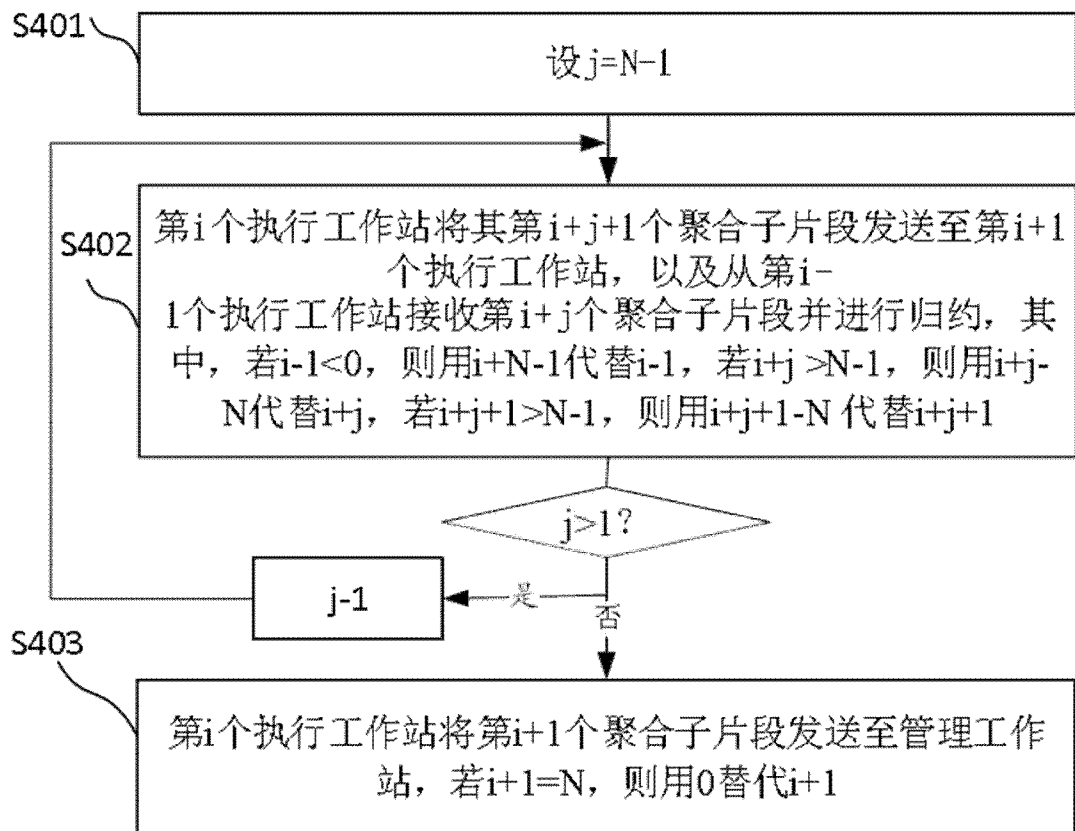


图 4

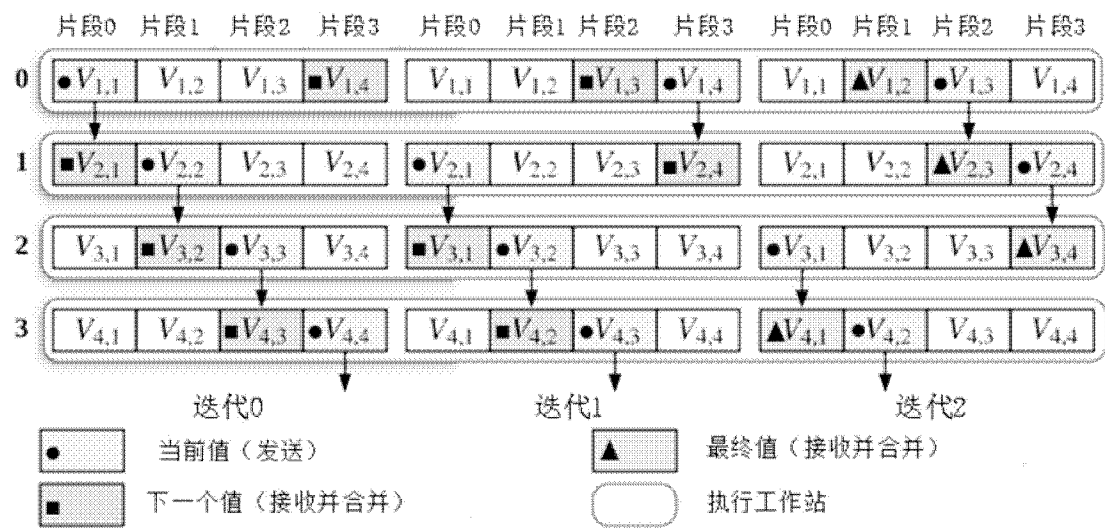


图 5

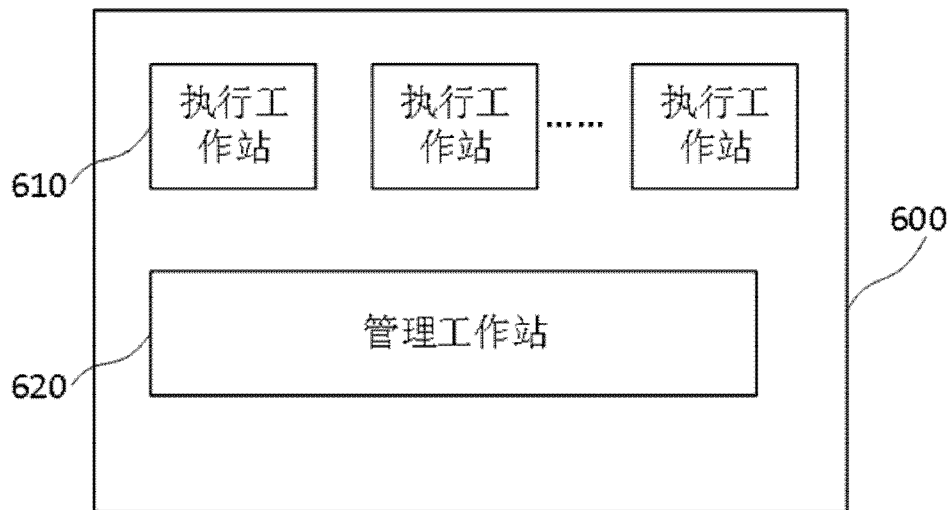


图 6

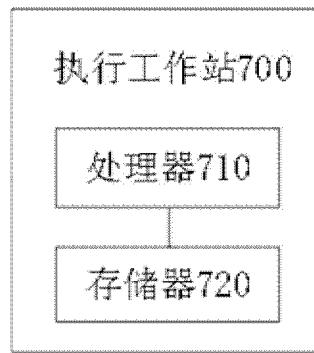


图 7

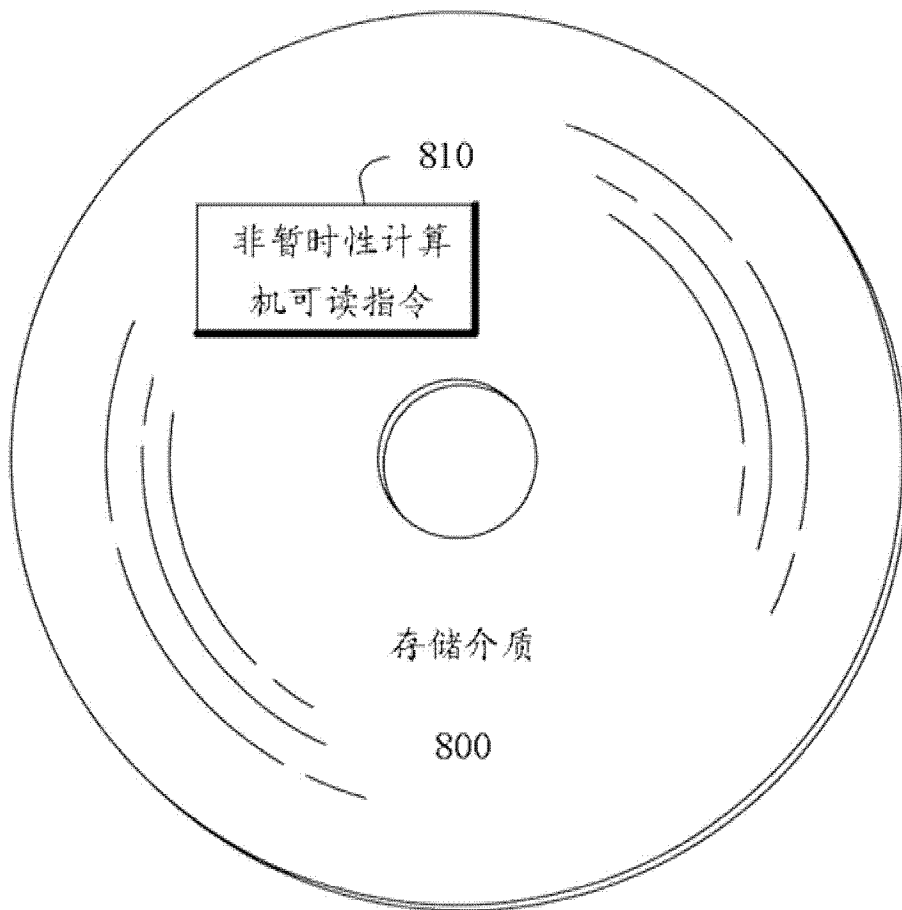


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/104127

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/242(2019.01)i; G06F 16/2455(2019.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

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)

CNPAT, CNKI, WPI, EPODOC, IEEE: 清华, 俞博文, 片段, 单元, 序列, 数据块, 聚合, 拆, 分, 归约, 拼接, 合并, 并行, split +, distribut+, MLib, aggregat+, reduc+, parallel+, spark

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	YU, Bowen et al. "Sparker: Efficient Reduction for More Scalable Machine Learning with Spark" <i>50th International Conference on Parallel Processing</i> , 05 October 2021 (2021-10-05), pages 4-7	1-11
A	CN 111949681 A (YUSUR TECHNOLOGY CO., LTD.) 17 November 2020 (2020-11-17) description, paragraphs 0159-0163, and figure 6	1-11
A	CN 112506950 A (SHENZHEN QIANHAI WEBANK CO., LTD.) 16 March 2021 (2021-03-16) entire document	1-11
A	CN 111694876 A (BEIJING QIHOO TECHNOLOGY CO., LTD.) 22 September 2020 (2020-09-22) entire document	1-11
A	US 2017262517 A1 (HOROWITZ, E. et al.) 14 September 2017 (2017-09-14) entire document	1-11



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

20 September 2022

Date of mailing of the international search report

29 September 2022

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/CN2022/104127

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	111949681	A	17 November 2020	WO	2021258512	A1	30 December 2021
CN	112506950	A	16 March 2021	None			
CN	111694876	A	22 September 2020	None			
US	2017262517	A1	14 September 2017	None			

国际检索报告

国际申请号

PCT/CN2022/104127

A. 主题的分类 G06F 16/242(2019.01)i; G06F 16/2455(2019.01)i 按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类		
B. 检索领域 检索的最低限度文献(标明分类系统和分类号) G06F 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)) CNPAT, CNKI, WPI, EPDOC, IEEE:清华, 俞博文, 片段, 单元, 序列, 数据块, 聚合, 拆, 分, 归约, 拼接, 合并, 并行, split+, distribut+, MLlib, aggregat+, reduc+, parallel+, spark		
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
PX	YU, Bowen et al. "Sparker: Efficient Reduction for More Scalable Machine Learning with Spark" 50th International Conference on Parallel Processing, 2021年10月5日 (2021 - 10 - 05), 第4-7页	1-11
A	CN 111949681 A (中科驭数北京科技有限公司) 2020年11月17日 (2020 - 11 - 17) 说明书第0159-0163段及附图6	1-11
A	CN 112506950 A (深圳前海微众银行股份有限公司) 2021年3月16日 (2021 - 03 - 16) 全文	1-11
A	CN 111694876 A (北京奇虎科技有限公司) 2020年9月22日 (2020 - 09 - 22) 全文	1-11
A	US 2017262517 A1 (HOROWITZ, Eliot等) 2017年9月14日 (2017 - 09 - 14) 全文	1-11
☐ 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
2022年9月20日		2022年9月29日
ISA/CN的名称和邮寄地址		授权官员
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 传真号 (86-10)62019451		王欣玥 电话号码 86-(10)-53961379

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/104127

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	111949681	A	2020年11月17日	WO	2021258512	A1	2021年12月30日
CN	112506950	A	2021年3月16日	无			
CN	111694876	A	2020年9月22日	无			
US	2017262517	A1	2017年9月14日	无			