(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 3 月 30 日 (30.03.2023)

WIPO | PCT

(10) 国际公布号
WO 2023/045385 A1

(51) 国际专利分类号:
G06F 3/06 (2006.01)

(21) 国际申请号: PCT/CN2022/095948

(22) 国际申请日: 2022 年 5 月 30 日 (30.05.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202111116330.3 2021年9月23日 (23.09.2021) 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: **涂剑洪(TU, Jianhong)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **董如良(DONG, Ruliang)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **张进毅(ZHANG, Jinyi)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **吴沛(WU, Pei)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 深圳市深佳知识产权代理事务所 (普通合伙) (**SHENPAT INTELLECTUAL PROPERTY AGENCY**); 中国广东省深圳市罗湖区南湖街道春风路庐山大厦B座18C2、18D、18E、18E2, Guangdong 518001 (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, 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。

(54) **Title:** DATA PROCESSING METHOD AND RELATED DEVICE

(54) **发明名称:** 数据处理方法以及相关设备

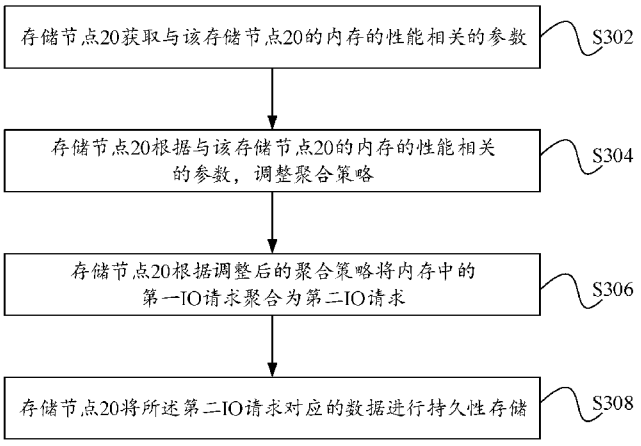

图 3

S302 A storage node 20 acquires a parameter related to the performance of a memory of the storage node 20

S304 The storage node 20 adjusts an aggregation policy according to the parameter related to the performance of the memory of the storage node 20

S306 The storage node 20 aggregates first IO requests in the memory into a second IO request according to the adjusted aggregation policy

S308 The storage node 20 performs persistent storage on data corresponding to the second IO request

(57) **Abstract:** Provided in the present application is a data processing method, comprising: acquiring a parameter related to the performance of a memory of a storage node; adjusting an aggregation policy according to the parameter; aggregating first input/output (IO) requests in the memory into a second IO request according to the adjusted aggregation policy; and performing persistent storage on data corresponding to the second IO request. A storage node aggregates small IO requests into a large IO request, thereby avoiding the effect of the small IO requests on the performance of the entire distributed storage system; moreover, an aggregation policy is adaptively adjusted, so as to avoid the overall performance of the distributed storage system from being affected by a memory being quickly filled due to the aggregation that is performed according to an aggregation policy having a fixed parameter.

WO 2023/045385 A1

[见续页]

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要：本申请提供了一种数据处理方法，包括：获取与存储节点的内存的性能相关的参数，根据该参数调整聚合策略，根据调整后的聚合策略将内存中的第一输入输出IO请求聚合为第二IO请求，并将第二IO请求对应的数据进行持久性存储。一方面，存储节点通过将小IO请求聚合为大IO请求，避免了小IO请求对整个分布式存储系统的性能的影响，另一方面，通过自适应调整聚合策略，可以避免按照参数固定的聚合策略进行聚合导致内存很快被写满，进而影响分布式存储系统整体的性能。

数据处理方法以及相关设备

本申请要求于2021年09月23日提交中国国家知识产权局、申请号为202111116330.3、发明名称为“数据处理方法以及相关设备”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。

技术领域

本申请涉及存储技术领域，尤其涉及一种数据处理方法、装置以及存储节点、计算机可读存储介质、计算机程序产品。

背景技术

随着互联网的发展，尤其是移动互联网的发展，产生了海量的数据。如何提高数据的存储效率，降低数据的存储成本成为业界重点关注的问题。在分布式存储系统中，为了避免单个存储节点故障导致数据丢失，通常会针对每份数据存储多个副本，以保障数据安全。然而，多副本存储需要耗费较多的存储空间，导致成本急剧上升。

基于此，业界还提出了纠删码（erasure code，EC）算法。其中，EC 算法是指对于 N 份原始数据，增加 M 份校验数据，并能通过 N+M 份数据中的任意 N 份数据，还原为 N 份原始数据。当 N+M 份数据分布式地存储在不同的存储节点或硬盘，则可以容忍 M 个存储节点或硬盘的故障。在保障数据安全的基础上，EC 算法针对 N 份数据仅需增加 M 份校验数据，而不是存储多个备份，如此大幅度提升了单位存储空间下的用户可用容量（也称作得盘率）。其中，EC 配比（原始数据的占比，即 N/（N+M））越大，得盘率越高。

EC 算法虽然具有较高可靠性和得盘率，但是在处理小随机读写（例如是小块随机读写或小对象随机读写）时的性能较差。EC 配比越大，得盘率越高，会有越多的大随机读写被拆分为多个小随机读写，对整个存储系统的性能影响也越大。为此，可以先将输入输出（input output，IO）请求对应的数据进行缓存，然后对 IO 请求进行聚合，将聚合后的 IO 请求对应的数据进行持久性存储。其中，数据在缓存后，可以返回写成功通知，无需等到持久性存储完成后再返回写成功通知，如此减少了延时。

相关技术通常采用参数固定的聚合策略对 IO 请求进行聚合。例如，采用固定的聚合等待时间（具体是内存中的 IO 请求等待聚合的最长时间，超过该时间则放弃聚合）对多个小 IO 请求进行聚合。然而，分布式存储系统中存储节点的内存的性能可以受到至少一种因素影响。例如，业务由低谷时段变为高峰时段时，内存的输入带宽（将接收到 IO 请求写入内存的带宽）大于内存的输出带宽（将内存中的 IO 请求进行持久性存储时的带宽），采用参数固定的聚合策略，如低谷时段的聚合策略，对内存中的 IO 请求进行聚合，可以导致内存很快被写满。又例如，分布式存储系统中用于形成内存池的多个存储节点中至少有一个存储节点发生故障，导致内存池的容量减少，如果采用参数固定的聚合策略，如故障发生前的聚合策略，对多个小 IO 请求进行聚合，可以导致内存池容易被写满。由此影响了分布式存储系统整体的性能。

发明内容

本申请提供了一种数据处理方法，该方法通过自适应调整聚合策略，并按照调整后的聚合策略将小IO请求聚合为大IO请求，避免了小IO请求对整个分布式存储系统的性能的影响，并且避免按照参数固定的聚合策略进行聚合导致内存很快被写满，进而影响分布式存储系统整体的性能。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方法，该方法可以由存储节点执行。具体地，存储节点获取与该存储节点的内存的性能相关的参数，根据与该存储节点的内存的性能相关的参数自适应调整聚合策略，并根据调整后的策略，将内存中的多个小 IO 请求（可以称为第一 IO 请求）聚合为大 IO 请求（可以称为第二 IO 请求），将大 IO 请求对应的数据进行持久性存储。

一方面，存储节点通过将小 IO 请求聚合为大 IO 请求，避免了小 IO 请求对整个分布式存储系统的性能的影响，另一方面，通过自适应调整聚合策略，例如调整聚合等待时间、聚合阈值（用于和 IO 请求对应的数据的数据量比较，以确定 IO 请求是否参与聚合）、聚合分条大小（聚合后的 IO 请求对应的数据的最大数据量）和聚合并发数量（同时启动持久性存储的 IO 请求的最大数量）等聚合参数中的一种或多种，如此可以避免按照参数固定的聚合策略进行聚合导致内存很快被写满，进而影响分布式存储系统整体的性能。

在一些可能的实现方式中，与存储节点的内存的性能相关的参数包括与该存储节点属于一个集群的其他存储节点的故障信息。与存储节点的内存的性能相关的参数也可以包括内存的输入带宽及内存的输出带宽，其中，输入带宽为存储节点将接收到 IO 请求写入内存的带宽，输出带宽为存储节点将内存中的 IO 请求进行持久性存储时的带宽。在一些实施例中，与存储节点的内存的性能相关的参数可以既包括上述故障信息，也包括内存的输入带宽和内存的输出带宽。

存储节点通过感知其他存储节点的故障信息，和/或内存的输入带宽、输出带宽，可以实现根据上述信息自适应调整聚合策略，以保障存储节点的内存的性能，进而保障分布式存储系统整体的性能。

在一些可能的实现方式中，聚合策略包括聚合阈值、聚合等待时间、聚合分条大小和聚合并发数量中的一个或多个。根据获取的与存储节点的性能相关的参数不同，存储节点可以执行不同的调整方式。

当输入带宽大于输出带宽时，存储节点可以降低聚合阈值，如此可以使得更多的 IO 请求能够透写，以避免内存被快速写满。其中，透写是指不参与聚合，直接写入存储池。为了平衡输入带宽和输出带宽，降低内存的压力，存储节点也可以增大聚合并发数量，以加快聚合进度，从而使得内存中的数据可以被快速聚合，并存储至存储池。

当输入带宽小于输出带宽时，存储节点可以延长所述聚合等待时间，增大所述聚合分条大小，和/或减小所述聚合并发数量。如此，可以合并更多重复的数据，提高聚合效率。其中，重复的数据是指相同地址或相同范围的数据。

当获取与所述存储节点属于一个集群的其他存储节点的故障信息时，存储节点可以增

大所述聚合并发数量，缩短所述聚合等待时间，减小所述聚合分条大小，和/或增大所述聚合阈值。其中，增大聚合并发数量，可以提高由内存写入存储池的能力；缩短聚合等待时间，减小聚合分条大小，可以将数据及时写入存储池；降低聚合阈值，可以增加写入内存的 IO 请求不经过聚合，直接写入存储池的概率。

在一些可能的实现方式中，存储节点还可以获取所接收的 IO 请求的访问特性，根据访问特性对所述 IO 请求进行分类。相应地，存储节点将各个类别的第一 IO 请求分别聚合为第二 IO 请求。

该方法通过在聚合机制上引入分类机制，对 IO 请求按照访问特性进行分类，然后对不同类别的 IO 请求分别进行聚合，可以显著降低因为聚合导致的垃圾回收对系统性能的影响。

在一些可能的实现方式中，访问特性包括请求访问的数据在统计周期内的访问次数，和/或所述数据的最近访问时间。其中，访问次数和/或最近访问时间可以用于表征数据的热度。存储节点可以根据所述数据在统计周期内的访问次数和/或所述数据的最近访问时间，将所述 IO 请求分为多个类别。

其中，所述多个类别中第一类别的 IO 请求对应的数据的访问次数高于所述多个类别中第二类别的 IO 请求对应的数据的访问次数，或者所述多个类别中第一类别的 IO 请求对应的数据的最近访问时间晚于所述多个类别中第二类别的 IO 请求对应的数据的最近访问时间。

如此，可以实现按照热度对 IO 请求进行分类。将具有相同生命周期（冷热程度）的数据聚合，可以显著降低因为聚合导致的垃圾回收对系统性能波动的影响。

在一些可能的实现方式中，热数据被访问的频率高于冷数据，相应地，热数据被修改的概率高于冷数据。基于此，存储节点还可以结合所接收 IO 请求的访问特性，调整聚合策略。

具体地，存储节点可以将聚合策略调整为：所述第一类别的 IO 请求的聚合等待时间大于所述第二类别的 IO 请求的聚合等待时间，和/或，所述第一类别的 IO 请求的聚合分条大小大于所述第二类别的 IO 请求的聚合分条大小。

基于调整后的聚合策略，热 IO 请求对应的数据能够多缓存一段时间，如此可以有效合并冗余的数据，提高聚合效率。

在一些可能的实现方式中，存储节点可以采用 EC 算法，将所述第二 IO 请求对应的数据进行持久性存储。其中，存储节点可以将所述第二 IO 请求对应的数据拆分，获得 N 份源数据，所述 N 为大于 1 的正整数，然后存储节点根据所述 N 份源数据，确定 M 份校验数据，所述 M 为大于 1 的正整数，接着存储节点将所述 N 份源数据和所述 M 份校验数据中的一份数据在本地进行持久性存储，并向与所述存储节点属于一个集群的 N+M-1 个目标存储节点分别转发剩余的 N+M-1 份数据进行持久性存储。其中，存储节点向每个目标存储节点发送一份数据。

在该方法中，存储节点与所述 N+M-1 个目标存储节点之间的转发路径满足预设条件。例如，存储节点与 N+M-1 个目标存储节点放置在同一机架时，或连接在同一路由器时，存储节点与 N+M-1 个目标存储节点之间的转发路径大幅缩短。当转发路径小于预设长度时，则可以确定转发路径满足预设条件。存储节点按照该转发路径进行转发，可以大幅减少网

络开销。

第二方面，本申请实施例提供了一种数据处理装置。所述装置包括：

参数采集模块，用于获取与存储节点的内存的性能相关的参数；

策略管理模块，用于根据所述参数调整聚合策略，所述聚合策略用于对所述内存中的输入输出 IO 请求进行聚合；

聚合模块，用于根据调整后的聚合策略将内存中的第一 IO 请求聚合为第二 IO 请求；

存储模块，用于将所述第二 IO 请求对应的数据进行持久性存储。

在一些可能的实现方式中，所述与所述存储节点的内存的性能相关的参数包括与所述存储节点属于一个集群的其他存储节点的故障信息及/或所述内存的输入带宽及所述内存的输出带宽，所述输入带宽为所述存储节点将接收到 IO 请求写入所述内存的带宽，所述输出带宽为所述存储节点将所述内存中的 IO 请求进行持久性存储时的带宽。

在一些可能的实现方式中，所述聚合策略包括聚合阈值、聚合等待时间、聚合分条大小和聚合并发数量中的一个或多个;

当所述输入带宽大于所述输出带宽时，所述策略管理模块具体用于：

降低所述聚合阈值，和/或增大所述聚合并发数量;

当所述输入带宽小于所述输出带宽时，所述策略管理模块具体用于：

延长所述聚合等待时间，增大所述聚合分条大小，和/或减小所述聚合并发数量；

当获取与所述存储节点属于一个集群的其他存储节点的故障信息时，所述策略管理模块具体用于：

增大所述聚合并发数量，缩短所述聚合等待时间，减小所述聚合分条大小，和/或增大所述聚合阈值。

在一些可能的实现方式中，所述参数采集模块还用于：

获取所接收的 IO 请求的访问特性；

所述装置还包括：

分类模块，用于根据所述访问特性对所述 IO 请求进行分类；

所述聚合模块具体用于：

将各个类别的第一 IO 请求分别聚合为第二 IO 请求。

在一些可能的实现方式中，所述访问特性包括请求访问的数据在统计周期内的访问次数，和/或所述数据的最近访问时间；

所述分类模块具体用于：

根据所述数据在统计周期内的访问次数和/或所述数据的最近访问时间，将所述 IO 请求分为多个类别，所述多个类别中第一类别的 IO 请求对应的访问次数高于所述多个类别中第二类别的 IO 请求对应的访问次数，或者所述多个类别中第一类别的 IO 请求对应的最近访问时间晚于所述多个类别中第二类别的 IO 请求对应的最近访问时间。

在一些可能的实现方式中，所述第一类别的 IO 请求的聚合等待时间大于所述第二类别的 IO 请求的聚合等待时间，和/或，所述第一类别的 IO 请求的聚合分条大小大于所述第二类别的 IO 请求的聚合分条大小。

在一些可能的实现方式中，所述存储模块具体用于：

将所述第二 IO 请求对应的数据拆分，获得 N 份源数据，所述 N 为大于 1 的正整数；

根据所述 N 份源数据，确定 M 份校验数据，所述 M 为大于 1 的正整数；

将所述 N 份源数据和所述 M 份校验数据中的一份数据在本地进行持久性存储，并向与所述存储节点属于一个集群的 N+M-1 个目标存储节点分别转发剩余的 N+M-1 份数据进行持久性存储，每个节点发送一份数据，所述存储节点与所述 N+M-1 个目标存储节点之间的转发路径满足预设条件。

第三方面，本申请提供一种存储节点，所述存储节点包括处理器和存储器。所述处理器、所述存储器进行相互的通信。所述处理器用于执行所述存储器中存储的计算机可读指令，以使得存储节点执行如第一方面或第一方面的任一种实现方式中的数据处理方法。

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，所述指令指示存储节点执行上述第一方面或第一方面的任一种实现方式所述的数据处理方法。

第五方面，本申请提供了一种包含指令的计算机程序产品，当其在存储节点上运行时，使得存储节点执行上述第一方面或第一方面的任一种实现方式所述的数据处理方法。

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分布式存储系统的系统架构图；

图 2 为本申请实施例提供的一种分布式存储系统的系统架构图；

图 3 为本申请实施例提供的一种数据处理方法的流程图；

图 4 为本申请实施例提供的一种数据处理方法的流程图；

图 5 为本申请实施例提供的一种数据处理装置的结构示意图。

具体实施方式

本申请实施例中的术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。

本申请可以应用于分布式存储系统的应用场景中。分布式存储系统是指将数据分散存储在多台独立的存储节点上的系统。传统的网络存储系统采用集中式的存储阵列存放所有数据，存储阵列的性能既是系统性能的瓶颈，也是可靠性和安全性的焦点，不能满足大规模存储应用的需要。分布式存储系统采用可扩展的系统结构，利用多台存储节点分担存储负荷，它不但提高了系统的可靠性、可用性和存取效率，还易于扩展。

参见图 1 所示的分布式存储系统的架构图，该架构具体是结构化存储（memory fabric，MF）架构。如图 1 所示，分布式存储系统包括计算节点集群和存储节点集群。其中，计算节点集群包括一个或多个计算节点 100（图 1 中示出了两个计算节点 100，但不限于两个计

算节点 100)。计算节点 100 是用户侧的一种计算设备，如服务器、台式计算机等。在硬件层面，计算节点 100 中设置有处理器和内存（图 1 中未示出)。在软件层面，计算节点 100 上运行有应用程序（application）101（简称应用）和客户端程序 102（简称客户端)。应用 101 是对用户呈现的各种应用程序的统称。客户端 102 用于接收由应用 101 触发的数据访问请求，并且与存储节点 20 交互，向存储节点 20 发送所述数据访问请求，也称作 IO 请求。客户端 102 还用于接收来自存储节点的数据，并向应用 101 转发所述数据。可以理解的是，当客户端 102 是软件程序时，客户端 102 的功能由计算节点 100 所包含的处理器运行内存中的程序来实现。客户端 102 也可以由位于计算节点 100 内部的硬件组件来实现。计算节点集群中的任意一个客户端 102 可以访问存储节点集群中的任意一个存储节点 20。

存储节点集群包括一个或多个存储节点 20（图 1 中示出了三个存储节点 20，但不限于三个存储节点 20)，各个存储节点 20 之间可以互联。存储节点如服务器、台式计算机或者存储阵列的控制器、硬盘框等。在功能上，存储节点 20 主要用于对数据进行计算或处理等。另外，所述存储节点集群还包括管理节点(图 1 未示出)。管理节点用于创建并管理内存池。各个存储节点 20 从存储节点中选举出一个节点让它承担管理节点的职能。管理节点可以与任意一个存储节点 20 通信。

在硬件上，如图 1 所示，存储节点 20 至少包括处理器、存储器和控制单元。其中，处理器是中央处理器（central processing unit，CPU)，用于处理来自存储节点 20 外部的数据，或者存储节点 20 内部生成的数据。存储器，是指用于存储数据的装置，它可以是内存，也可以是硬盘。控制单元用于根据读/写数据请求，从存储器（例如是硬盘）中读取数据，或向存储器中写入数据。在读写数据的过程中，控制单元需要将读/写数据请求中携带的地址转换为存储器能够识别的地址。

内存是指与处理器直接交换数据的内部存储器，它可以随时读写数据，而且速度很快，作为操作系统或其他正在运行中的程序的临时数据存储器。内存包括至少两种存储器，例如内存既可以是随机存取存储器（Random Access Memory，RAM)，也可以是只读存储器（Read Only Memory, ROM)。举例来说，随机存取存储器可以是动态随机存取存储器（Dynamic Random Access Memory, DRAM)、存储级存储器（Storage Class Memory, SCM）或者是静态随机存取存储器（Static Random Access Memory，SRAM）等；只读存储器可以是可编程只读存储器（Programmable Read Only Memory, PROM)、可抹除可编程只读存储器（Erasable Programmable Read Only Memory, EPROM）等。内存还可以是双列直插式存储器模块或双线存储器模块（Dual In-line Memory Module，简称 DIMM)，即由动态随机存取存储器（DRAM）组成的模块。后面的描述中，均以 DRAM 和 SCM 为例进行说明，但不代表存储节点 20 不包含其他类型的存储器。

与内存不同的是，硬盘读写数据的速度比内存慢，通常用于持久性地存储数据。以存储节点 20a 为例，其内部可以设置一个或多个硬盘；或者，在存储节点 20a 的外部还可以挂载一个硬盘框，在硬盘框中设置多个硬盘。无论哪一种部署方式，这些硬盘都可以视作存储节点 20a 所包含的硬盘。硬盘类型为固态硬盘、机械硬盘，或者其他类型的硬盘。类似的，存储节点集群中的其他存储节点，如存储节点 20b、存储节点 20c 也可以包含各种类型的硬盘。一个存储节点 20 中可以包含一个或多个同一种类型的存储器。

需要说明的是，存储节点 20 和计算节点 100 也可以集成在同一个物理设备中。参见图 2 所示的分布式存储系统的另一种架构图，本实施例将上述集成的设备统一称为存储节点。应用部署在存储节点 20 内部，所以应用可直接通过存储节点 20 中的客户端触发写数据请求或读数据请求，由该存储节点 20 处理，或者发送给其他存储节点 20 处理。此时，客户端向本地存储节点 20 发送的读写数据请求具体是指客户端向处理器发送数据访问请求。除此之外，存储节点 20 所包含的组件（例如 IO 控制器）及其功能与图 1 中的存储节点 20 包含的组件（例如控制单元）类似，这里不再赘述。

图 1 或图 2 所示的存储系统中，存储节点 20 采用了盘控一体的结构。控制单元 201 或 IO 控制器接收到 IO 请求，可以将 IO 请求的数据暂时保存在内存中，当内存中的数据总量达到一定阈值时，控制单元 201 或 IO 控制器将内存中存储的数据发送给硬盘进行持久化存储。在另一些可能的实现方式中，存储节点 20 也可以采用盘控分离的结构。具体地，硬盘通常放置在硬盘框中，控制单元 201 或 IO 控制器可以通过网络与硬盘框通信，从而将数据存储至硬盘中。

各个存储节点 20 中包括不同类型的存储介质，例如同时包括 DRAM、SCM 以及硬盘等存储介质，这些不同类型的存储介质均提供内存接口可直接被处理器访问。并且这些存储节点所包含的各种类型内存可以组成一个内存池。数据在内存池可根据访问频率在不同类型的存储介质之间换入换出。

在一些实施例中，内存池可以只包含部分类型的存储器，例如只包含较高性能的存储器，例如 DRAM 和 SCM，而排除硬盘等性能相对较低的存储器。当内存池仅包含存储集群中的较高性能的存储器（例如 DRAM 和 SCM）时，管理节点还可以将存储集群中的较低性能的存储器（例如硬盘）构建成一个存储池。与内存池类似，存储池也跨越了至少两个存储节点，其存储空间由所述至少两个存储节点中的一种或多种类型的硬盘构成。

当存储集群中既包含内存池又包含存储池时，存储池用于持久化地存储数据，特别是访问频率较低的数据，而内存池则用于临时存储数据，特别是访问频率较高的数据。内存池用于缓存数据时，也可以称作缓存池。具体的，当内存池中存储的数据的数据量到达设定的阈值时，所述内存池中的部分数据将会被写入存储池中存储。

在以上介绍的应用场景中，随着数据的爆炸式增长，数据存储需求与日俱增。考虑到多副本（例如是三副本）的存储方式需要占用较多的存储空间，业界提出了 EC 算法替代三副本存储模式进行数据存储，以提高得盘率。EC 算法具有较高可靠性、较高得盘率，然而在处理小随机读写（例如小块随机读写或小对象随机读写）时的性能较差。并且，EC 配比越大，会有越多的大随机读写被拆分为多个小随机读写，对整个存储系统的性能影响也越大。为此，可以将 IO 请求对应的数据进行缓存，然后对 IO 请求进行聚合，将聚合后的 IO 请求对应的数据进行持久性存储。一方面可以通过聚合 IO 请求降低对存储系统的性能的影响，另一方面数据在缓存后即可返回写成功通知，从而减少延时。

目前，在对 IO 请求进行聚合时，通常是采用参数固定的聚合策略对 IO 请求进行聚合。例如，采用固定的聚合等待时间对多个小 IO 请求进行聚合。然而，分布式存储系统中存储节点的内存的性能可以受到至少一种因素影响。例如，业务由低谷时段变为高峰时段时，

内存的输入带宽（将接收到IO请求写入内存的带宽）大于内存的输出带宽（将内存中的IO请求进行持久性存储时的带宽），采用参数固定的聚合策略，如低谷时段的聚合策略，对内存中的IO请求进行聚合，可以导致内存很快被写满。又例如，分布式存储系统中用于形成内存池的多个存储节点中至少有一个存储节点发生故障，导致内存池的容量减少，如果采用参数固定的聚合策略，如故障发生前的聚合策略，对多个小IO请求进行聚合，可以导致内存池容易被写满。由此影响了分布式存储系统整体的性能。

有鉴于此，本申请实施例提供了一种数据处理方法，该方法可以由存储节点（例如是图1或图2中的存储节点20）执行。具体地，存储节点获取与该存储节点的内存的性能相关的参数，根据与该存储节点的内存的性能相关的参数自适应调整聚合策略，并根据调整后的策略，将内存中的多个小IO请求（可以称为第一IO请求）聚合为大IO请求（可以称为第二IO请求），将大IO请求对应的数据进行持久性存储。

一方面，存储节点通过将小IO请求聚合为大IO请求，避免了小IO请求对整个分布式存储系统的性能的影响，另一方面，通过自适应调整聚合策略，例如调整聚合等待时间、聚合阈值（用于和IO请求对应的数据的数据量比较，以确定IO请求是否参与聚合）、聚合分条大小（聚合后的IO请求对应的数据的最大数据量）和聚合并发数量（同时启动持久性存储的IO请求的最大数量）等聚合参数中的一种或多种，如此可以避免按照参数固定的聚合策略进行聚合导致内存很快被写满，进而影响分布式存储系统整体的性能。

接下来，从存储节点20的角度对本申请实施例的数据处理方法进行详细说明。

参见图3所示的数据处理方法的流程图，该方法包括：

S302：存储节点20获取与该存储节点20的内存的性能相关的参数。

内存的性能与内存的容量、内存的水位（已写入内存的数据的占比）等指标具有相关性。其中，内存的输入带宽和内存的输出带宽可以影响内存的水位，基于此，与该存储节点20的内存的性能相关的参数可以包括内存的输入带宽和内存的输出带宽。

在分布式存储系统中，存储节点20的内存可以与属于一个集群的其他存储节点的内存形成内存池。当属于一个集群的其他存储节点故障时，内存池的容量也会相应地受到影响。基于此，与该存储节点20的内存的性能相关的参数可以包括与所述存储节点20属于一个集群的其他存储节点的故障信息。该故障信息例如可以为故障节点的节点标识和内存容量。

存储节点20可以获取与该存储节点20属于一个集群的其他存储节点的故障信息，和/或获取内存的输入带宽、内存的输出带宽。下面对存储节点20获取其他存储节点的故障信息以及获取内存的输入带宽、内存的输出带宽的实现方式进行说明。

存储节点20可以从集群的管理节点获取其他存储节点的故障信息。具体地，集群的管理节点可以对属于该集群的存储节点的健康状态进行管理。例如，属于该集群的存储节点可以周期性地向管理节点发送心跳消息，以告知管理节点该存储节点的健康状态，当管理节点连续N个周期未收到某个或某些存储节点的心跳消息时，管理节点可以确定该存储节点发生故障，并记录故障信息，该故障信息例如包括发生故障的存储节点的节点标识和内容容量。相应地，存储节点20可以从管理节点获取上述故障信息。需要说明的是，存储节点20可以按照预设时间间隔从管理节点获取故障信息。其中，预设时间间隔可以等于上述

心跳消息的上报周期，也可以是上述上报周期的整数倍。

存储节点20可以周期性地统计写入内存的IO请求（也可以称作前台IO）的信息和由内存写入存储的IO请求（即从内存进行持久化存储的IO请求，也可以称作后台IO）的信息。根据上述IO请求对应的数据的存储格式不同，存储节点20统计的信息可以是不同的。

以存储格式为对象存储为例，写入内存的IO请求中可以包括由对象名和对象数据形成的键值对。相应地，存储节点20可以根据IO请求采集对象名、访问的对象范围、读写时延等信息。其中，访问的对象范围可以通过在对象内的地址表征，例如对象的大小为10兆字节（megabytes，MB），访问的对象范围可以是第6至8MB。基于此，存储节点20可以根据写入内存（例如是内存池）的IO请求的信息，如访问对象的范围，获得统计周期内写入内存的数据量，根据该数据量和统计周期的比值，确定内存的输入带宽。类似地，存储节点20可以根据写入存储（例如是存储池）的IO请求的信息，获得统计周期内写入存储的数据量，根据该数据量和统计周期的比值，确定内存的输出带宽。统计周期可以根据经验值设定，例如可以设置为10秒（second，s）。

可以理解，本申请的数据处理方法也适用于块存储的数据。当数据采用块存储时，存储节点20采集的IO请求的信息可以包括访问地址、访问长度、读写时间和访问时延中的一种或多种。其中，存储节点20可以根据统计周期内写入内存的IO请求的访问长度，获得该统计周期内写入内存的数据量，根据该数据量和统计周期的比值，确定内存的输入带宽。类似地，存储节点20可以采用与确定内存的输入带宽相似的方式，确定内存的输出带宽，在此不再赘述。

S304：存储节点20根据与该存储节点20的内存的性能相关的参数，调整聚合策略。

聚合策略用于对所述内存中的IO请求进行聚合。其中，聚合策略可以包括聚合等待时间、聚合阈值、聚合分条大小和聚合并发数量等聚合参数中的一种或多种。下面对不同参数情况下，聚合策略的调整方式进行详细说明。

在第一种可能的实现方式中，输入带宽大于输出带宽，存储节点20可以降低聚合阈值，如此可以使得更多的IO请求能够透写，以避免内存被快速写满。其中，透写是指不参与聚合，直接写入存储池。为了平衡输入带宽和输出带宽，降低内存的压力，存储节点20也可以增大聚合并发数量，以加快聚合进度，从而使得内存中的数据可以被快速聚合，并存储至存储池。需要说明的是，存储节点20执行上述调整操作可以是在内存的水位达到预设水位时开始执行。

在第二种可能的实现方式中，输入带宽小于输出带宽，存储节点20可以延长所述聚合等待时间，增大所述聚合分条大小，和/或减小所述聚合并发数量。如此，可以合并更多重复的数据，提高聚合效率。其中，重复的数据是指相同地址或相同范围的数据。为了便于理解，本申请实施例结合一具体示例进行说明。假设调整前的聚合等待时间为1s，调整后的聚合等待时间为2s，当第1s内包括对一个大小为10MB的对象A的第6-8MB范围的IO请求，第2s内也包括对上述对象的第6至8MB范围的IO请求，这两个IO请求对应的数据为重复的数据，通过合并重复的数据，可以提高聚合效率。

在第三种可能的实现方式中，与存储节点20属于一个集群的其他节点存在故障，存储节点20增大所述聚合并发数量，缩短所述聚合等待时间，减小所述聚合分条大小，和/或

增大所述聚合阈值。其中，增大聚合并发数量，可以提高由内存写入存储池的能力；缩短聚合等待时间，减小聚合分条大小，可以将数据及时写入存储池；降低聚合阈值，可以增加写入内存的IO请求不经过聚合，直接写入存储池的概率。

需要说明的是，在调整上述聚合参数时，存储节点20可以针对每种参数设置基础调整幅度，然后基于内存的输入带宽和内存的输出带宽的差值以及发生故障的存储节点的内存容量中的至少一种，确定调整系数。存储节点20基于上述调整系数和基础调整幅度，确定对各聚合参数的调整幅度。其中，带宽的差值和调整系数可以呈正相关，带宽的差值越大，调整系数越大，对于聚合参数的调整幅度也越大。类似地，发生故障的存储节点的内存容量和调整系数也可以呈正相关，发生故障的存储节点的内存容量越大，对于聚合参数的调整幅度也越大。

还需要说明的是，内存的输入带宽和内存的输出带宽相等时，例如调整聚合策略后，内存的输入带宽和内存的输出带宽由不相等变为相等时，存储节点20可以保持当前聚合策略。

S306:存储节点20根据调整后的聚合策略将内存中的第一IO请求聚合为第二IO请求。

具体地,存储节点20可以将内存中IO请求对应的数据的数据量与聚合阈值进行比较，当数据量大于聚合阈值（例如为调整后的聚合阈值）时，则可以透写该IO请求中的数据，当数据量小于或等于聚合阈值时，则可以将该IO请求加入聚合队列。

存储节点20可以按照聚合并发数量，从聚合队列中获取多个聚合请求，当多个聚合请求中到达内存时间最早的聚合请求的等待时间达到聚合等待时间（例如是调整后的聚合等待时间），或者多个聚合请求对应的数据的数据量之和达到聚合分条大小（例如是调整后的聚合分条大小），则可以将当前获取的多个聚合请求进行聚合，由此实现将第一IO请求聚合为第二IO请求。其中，第一IO请求可以是聚合队列中的IO请求，通常是小IO请求，第二IO请求可以是聚合后的IO请求，通常是大IO请求。

其中，存储节点20还可以记录第二IO请求对应的数据的元数据。例如，数据采用对象存储时，存储节点20可以记录第二IO请求对应的数据的对象名、访问的对象范围，以便于后续可以根据该元数据快速访问上述数据。在另一些实施例中，数据采用块存储时，存储节点20可以记录第二IO请求对应的数据的块标识和块地址。

S308：存储节点20将所述第二IO请求对应的数据进行持久性存储。

存储节点20可以采用EC算法,将第二IO请求对应的数据进行持久性存储。具体地，存储节点20可以将所述第二IO请求对应的数据拆分，获得N份源数据，其中， N为大于1的正整数。然后，存储节点20可以根据所述N份源数据，确定M份校验数据，其中，M为大于1的正整数。接着，存储节点20可以将所述N份源数据和所述M份校验数据中的一份数据在本地进行持久性存储，并向与所述存储节点属于一个集群的N+M-1个目标存储节点分别转发剩余的N+M-1份数据进行持久性存储。其中，存储节点20向N+M-1个目标存储节点中的每个目标存储节点发送一份数据，并且，向每个目标存储节点发送的数据不同，如此可以实现数据分布式存储。

当分布式存储系统采用高配比的EC算法进行数据存储时，分布式存储系统会有更多的网络转发，进而导致网络开销增加。为此，存储节点20可以将剩余的N+M-1份数据分

别写入与存储节点 20 之间的转发路径满足预设条件的 N+M-1 个目标存储节点进行持久性存储。每个目标存储节点存储剩余的 N+M-1 份数据中的一份数据。

其中，与存储节点 20 之间的转发路径可以用于表征和存储节点 20 的亲和性，当转发路径越短，则亲和性越高。基于此，存储节点 20 也可以是向与存储节点 20 之间的亲和性满足要求的 N+M-1 个目标存储节点分别转发剩余的 N+M-1 份数据进行持久性存储。转发路径满足预设条件或者亲和性满足要求的 N+M-1 个目标存储节点可以是与存储节点 20 位于相同机架的 N+M-1 个目标存储节点，或者是与存储节点 20 连接至同一路由器的 N+M-1 个目标存储节点。如此可以减少转发路径的长度，进而减少网络开销。

下面以第二 IO 请求对应的数据为对象进行示例说明。在具体实现时，存储节点 20 可以是主对象存储设备 OSD（Object-based Storage，device），主 OSD 可以将聚合后的大对象切分为 M 个源对象，并根据 M 个源对象生成 N 个校验对象。主 OSD 可以通过对象 ID 生成器为上述大对象生成对象 ID，然后可以将上述对象 ID 作为输入，结合路由算法和分布式存储系统的拓扑（也称作视图）确定源对象和校验对象需要转发的节点，例如是 OSD1，OSD2，…，OSD N+M-1，主 OSD 可以计算该主 OSD 和需要转发的 OSD1 至 OSD N+M-1 的亲和性。

当亲和性不满足要求，例如低于预设值时，主 OSD 可以重新生成对象 ID，并通过路由算法和分布式存储系统的拓扑重新确定源对象和校验对象需要转发的节点，计算主 OSD 与重新确定的 OSD 之间的亲和性。当亲和性满足要求时，主 OSD 可以将 M 个源对象和 N 个校验对象转发至上述亲和性满足要求的节点，以减少网络开销。

在一些实施例中，主 OSD 也可以设置最大重试次数，当主 OSD 重试次数达到上述最大重试次数，亲和性仍无法满足要求时，还可以从主 OSD 确定的多组节点中选择亲和性最佳的节点进行转发，如此可以减少网络开销。

考虑到对象存储的低成本、大容量等特性，在将第二 IO 请求对应的数据进行持久性存储时，可以采用对象存储格式进行存储。其中，第二 IO 请求对应的数据采用块存储时，还可以对该数据进行格式转换，例如基于对象存储的存储协议进行格式转换，获得对象。如此，存储节点 20 可以基于 EC 算法，对上述对象进行持久性存储。

在一些可能的实现方式中，不同 IO 请求对应的数据的热度可以是不同的，基于此，存储节点 20 还可以按照请求访问的数据的热度对 IO 请求进行分类，并针对不同类别的 IO 请求分别进行聚合。由于同一类别的 IO 请求访问的数据的热度相当，有较大概率在同一时间段进行垃圾回收（garbage collection，GC）。其中，垃圾回收是指写入存储池的聚合数据中的部分数据被修改或删除，导致出现无效存储空间时，将剩余的有效数据重新聚合，以回收原聚合数据的存储空间。如此可以显著提高垃圾回收的效率，进而降低因为聚合导致的垃圾回收对系统性能的影响。

接下来，结合附图对本申请实施例提供的数据处理方法的另一实现方式进行详细说明。

参见图 4 所示的数据处理方法的流程图，该方法包括：

S402：存储节点 20 获取与所述存储节点的内存的性能相关的参数。

S404：存储节点 20 根据与所述存储节点的内存的性能相关的参数，调整聚合策略。

其中，S402 至 S404 的相关实现可以参见图 3 所示实施例中相关内容描述，在此不再赘述。

S406：存储节点 20 获取所接收的 IO 请求的访问特性。

IO 请求的访问特性可以包括请求访问的数据在统计周期内的访问次数，和/或所述数据的最近访问时间。具体地，存储节点 20 可以获取统计周期内接收的 IO 请求，根据 IO 请求携带的信息，确定请求访问的数据在统计周期内的访问次数和/或所述数据的最近访问时间。

以请求访问的数据为对象中的数据为例，存储节点 20 接收的 IO 请求可以携带对象名和访问对象的范围，存储节点 20 可以根据对象名，统计同一对象的不同范围的访问次数，或者统计同一对象的不同范围的至少一个访问时间，并从同一对象的每个范围的至少一个访问时间中确定每个范围的最近访问时间。

S408：存储节点 20 根据访问特性对所述 IO 请求进行分类。

访问特性如访问次数和/或最近访问时间可以用于表征数据的热度。数据的热度可以通过定量的热度值表征，也可以通过定性的热度级别表征。在一些实施例中，热度级别可以包括热和冷两个级别，也可以包括热、温、冷三个级别。热度级别还可以进一步细分，例如热度级别可以包括极热、热、温、冷、极冷五个级别。

以基于统计周期内的访问次数，获得数据的热度为例。存储节点 20 可以获取各数据在统计周期内的访问次数，将该次数输入热度函数，获得函数值，该函数值即为上述数据的热度值。存储节点 20 还可以基于上述热度值确定热度级别。例如，数据的热度值小于预设阈值，则确定该数据为冷数据，数据的热度值大于预设阈值，则确定该数据为热数据。

基于此，存储节点 20 可以按照请求访问的数据的热度，对 IO 请求进行分类。例如，存储节点 20 可以按照热度将数据分为冷和热两个类别；又例如，存储节点 20 可以按照热度将数据分为冷、温、热三个类别。本实施例对类别的数量不作限制。

S410：存储节点 20 根据调整后的聚合策略，将各个类别的第一 IO 请求分别聚合为第二 IO 请求。

存储节点 20 可以针对每个类别的 IO 请求分别维护对应的聚合队列。具体地，存储节点可以将各个类别的 IO 请求对应的数据的数据量与聚合阈值进行比较，当数据量大于聚合阈值（例如为调整后的聚合阈值）时，则可以透写该 IO 请求中的数据，当数据量小于或等于聚合阈值时，则可以将该 IO 请求加入该类别对应的聚合队列。

然后，存储节点 20 可以从各类别对应的聚合队列中取出 IO 请求进行聚合，以将各个类别的第一 IO 请求分别聚合为第二 IO 请求。聚合过程可以参见图 3 所示实施例相关内容描述，在此不再赘述。

进一步地，热数据被访问的频率高于冷数据，相应地，热数据被修改的概率高于冷数据。基于此，在 S404 中，存储节点 20 还可以结合所接收 IO 请求的访问特性，调整聚合策略。

具体地，多个类别中第一类别的 IO 请求对应的数据的访问次数高于所述多个类别中第二类别的 IO 请求对应的数据的访问次数，或者所述多个类别中第一类别的 IO 请求对应的数据的最近访问时间晚于所述多个类别中第二类别的 IO 请求对应的数据的最近访问时间，即第一类别的 IO 请求对应的数据的热度高于第二类别的 IO 请求的热度，存储节点 20 可以

设置第一类别的IO请求的聚合等待时间大于所述第二类别的IO请求的聚合等待时间，或者设置第一类别的IO请求的聚合分条大小大于所述第二类别的IO请求的聚合分条大小。在一些实施例中，存储节点20也可以设置第一类别的IO请求的聚合等待时间大于第二类别的IO请求的聚合等待时间，且第一类别的IO请求的聚合分条大小大于第二类别的IO请求的聚合分条大小。

以多个类别包括冷和热两个类别为例，存储节点20可以针对热IO请求，延长聚合等待时间，增大聚合分条大小，使得热IO请求对应的数据能够多缓存一段时间，如此可以有效合并冗余的数据。对于极热的数据，存储节点20甚至可以长期缓存，不写入存储池，以减少开销。

上文结合图1至图4对本申请实施例提供的数据处理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5所示的数据处理装置的结构示意图，该装置500包括：参数采集模块502、策略管理模块504、聚合模块506和存储模块508。

其中，参数采集模块502用于执行图3所示实施例中S302对应的方法步骤，策略管理模块504用于执行图3所示实施例中S304对应的方法步骤，聚合模块506用于执行图3所示实施例中S306对应的方法步骤，存储模块508用于执行图3所示实施例中S308对应的方法步骤。

在一些可能的实现方式中，与所述存储节点的内存的性能相关的参数可以参见S302中相关内容描述，在此不再赘述。

在一些可能的实现方式中，所述聚合策略包括聚合阈值、聚合等待时间、聚合分条大小和聚合并发数量中的一个或多个，策略管理模块504对上述聚合策略的调整过程可以参见S304相关内容描述。

在一些可能的实现方式中，该装置500还可以包括分类模块。其中，参数采集模块502还用于执行图4所示实施例中S406对应的方法步骤，分类模块用于执行图4所示实施例中S408对应的方法步骤，相应地，聚合模块506具体用于执行图4所示实施例中S410对应的方法步骤。

在一些可能的实现方式中，所述访问特性包括请求访问的数据在统计周期内的访问次数，和/或所述数据的最近访问时间，分类模块用于执行图4所示实施例中S408对应的方法步骤，以对IO请求进行分类。

在一些可能的实现方式中，该装置500还可以结合访问特性对聚合策略进行调整，例如可以调整为：第一类别的IO请求的聚合等待时间大于所述第二类别的IO请求的聚合等待时间，和/或，所述第一类别的IO请求的聚合分条大小大于所述第二类别的IO请求的聚合分条大小。

在一些可能的实现方式中，存储模块508用于执行图3所示实施例中S308对应的方法步骤，以减少采用高配比的EC算法进行持久性存储产生的网络开销。

根据本申请实施例的数据处理装置500可对应于执行本申请实施例中描述的方法，并且数据处理装置500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3、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存储节点 20。该存储节点 20 可以是笔记本电脑、台式机等终端设备，也可以是云环境或边缘环境中的服务器。该存储节点20 具体用于实现如图 5 所示实施例中数据处理装置 500 的功能。

本申请实施例提供的存储节点20 的硬件结构可以参见图 1 或图 2，如图 1 或图 2 所示，存储节点 20 可以包括处理器和存储器，存储器中存储有计算机可读指令，所述处理器执行所述计算机可读指令，使得所述存储节点 20 执行前述实施例所述的数据处理方法（或实现前述数据处理装置 500 的功能）。

具体地，在实现图 5 所示装置的实施例的情况下，且图 5 中所描述的数据处理装置 500 的各模块如参数采集模块 502、策略管理模块 504、聚合模块 506、存储模块 508 的功能为通过软件实现的情况下，执行图 5 中各模块的功能所需的软件或程序代码可以存储在存储器中。处理器执行存储器中存储的程序代码，以使得存储节点 20 执行前述数据处理方法。

本申请实施例还提供了一种计算机可读存储介质。所述计算机可读存储介质可以是计算设备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的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）等。该计算机可读存储介质包括指令，所述指令指示计算设备或计算设备集群执行上述数据处理方法。

本申请实施例还提供了一种计算机程序产品。所述计算机程序产品包括一个或多个计算机指令。在计算设备上加载和执行所述计算机指令时，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、计算设备或数据中心进行传输。所述计算机程序产品可以为一个软件安装包，在需要使用前述数据处理方法的任一方法的情况下，可以下载该计算机程序产品并在计算设备或计算设备集群上执行该计算机程序产品。

上述各个附图对应的流程或结构的描述各有侧重，某个流程或结构中没有详述的部分，可以参见其他流程或结构的相关描述。

权 利 要 求

1. 一种数据处理方法，应用于存储节点，其特征在于，所述方法包括：

获取与所述存储节点的内存的性能相关的参数；

根据所述参数调整聚合策略，所述聚合策略用于对所述内存中的输入输出 IO 请求进行聚合；

根据调整后的聚合策略将所述内存中的第一 IO 请求聚合为第二 IO 请求；

将所述第二 IO 请求对应的数据进行持久性存储。

2. 如权利要求 1 所述的方法，其特征在于，所述与所述存储节点的内存的性能相关的参数包括与所述存储节点属于一个集群的其他存储节点的故障信息及/或所述内存的输入带宽及所述内存的输出带宽，所述输入带宽为所述存储节点将接收到 IO 请求写入所述内存的带宽，所述输出带宽为所述存储节点将所述内存中的 IO 请求进行持久性存储时的带宽。

3. 如权利要求 2 所述的方法，其特征在于，所述聚合策略包括聚合阈值、聚合等待时间、聚合分条大小和聚合并发数量中的一个或多个;

当所述输入带宽大于所述输出带宽时，所述根据所述参数调整聚合策略包括：

降低所述聚合阈值，和/或增大所述聚合并发数量;

当所述输入带宽小于所述输出带宽时，所述根据所述参数调整聚合策略包括：

延长所述聚合等待时间，增大所述聚合分条大小，和/或减小所述聚合并发数量；

当获取与所述存储节点属于一个集群的其他存储节点的故障信息时，则所述根据所述参数调整聚合策略包括：

增大所述聚合并发数量，缩短所述聚合等待时间，减小所述聚合分条大小，和/或降低所述聚合阈值。

4. 如权利要求 1 至 3 任意一项所述的方法，其特征在于，所述方法包括：

获取所接收的 IO 请求的访问特性；

根据访问特性对所述 IO 请求进行分类；

根据调整后的聚合策略将所述第一 IO 请求聚合为所述第二 IO 请求包括：

将各个类别的第一 IO 请求分别聚合为第二 IO 请求。

5. 如权利要求 1 至 4 任一项所述的方法，其特征在于，所述将所述第二 IO 请求对应的数据进行持久性存储，包括：

将所述第二 IO 请求对应的数据拆分，获得 N 份源数据，所述 N 为大于 1 的正整数；

根据所述 N 份源数据，确定 M 份校验数据，所述 M 为大于 1 的正整数；

将所述 N 份源数据和所述 M 份校验数据中的一份数据在本地进行持久性存储，并向与所述存储节点属于一个集群的 N+M-1 个目标存储节点分别转发剩余的 N+M-1 份数据进行持久性存储，每个节点发送一份数据，所述存储节点与所述 N+M-1 个目标存储节点之间的转发路径满足预设条件。

6. 一种数据处理装置，其特征在于，所述装置包括：

参数采集模块，用于获取与所述存储节点的内存的性能相关的参数；

策略管理模块，用于根据所述参数调整聚合策略，所述聚合策略用于对所述内存中的输入输出 IO 请求进行聚合；

聚合模块，用于根据调整后的聚合策略将所述内存中的第一 IO 请求聚合为第二 IO 请求；

存储模块，用于将所述第二 IO 请求对应的数据进行持久性存储。

7. 如权利要求 6 所述的装置，其特征在于，所述与所述存储节点的内存的性能相关的参数包括与所述存储节点属于一个集群的其他存储节点的故障信息及/或所述内存的输入带宽及所述内存的输出带宽，所述输入带宽为所述存储节点将接收到 IO 请求写入所述内存的带宽，所述输出带宽为所述存储节点将所述内存中的 IO 请求进行持久性存储时的带宽。

8. 如权利要求 7 所述的装置，其特征在于，所述聚合策略包括聚合阈值、聚合等待时间、聚合分条大小和聚合并发数量中的一个或多个;

当所述输入带宽大于所述输出带宽时，所述策略管理模块具体用于：

降低所述聚合阈值，和/或增大所述聚合并发数量;

当所述输入带宽小于所述输出带宽时，所述策略管理模块具体用于：

延长所述聚合等待时间，增大所述聚合分条大小，和/或减小所述聚合并发数量；

当获取与所述存储节点属于一个集群的其他存储节点的故障信息时，所述策略管理模块具体用于：

增大所述聚合并发数量，缩短所述聚合等待时间，减小所述聚合分条大小，和/或增大所述聚合阈值。

9. 如权利要求 6 至 8 任意一项所述的装置，其特征在于，所述参数采集模块还用于：

获取所接收的 IO 请求的访问特性；

所述装置还包括：

分类模块，用于根据所述访问特性对所述 IO 请求进行分类；

所述聚合模块具体用于：

将各个类别的第一 IO 请求分别聚合为第二 IO 请求。

10. 如权利要求 6 至 9 任一项所述的装置，其特征在于，所述存储模块具体用于：

将所述第二 IO 请求对应的数据拆分，获得 N 份源数据，所述 N 为大于 1 的正整数；

根据所述 N 份源数据，确定 M 份校验数据，所述 M 为大于 1 的正整数；

将所述 N 份源数据和所述 M 份校验数据中的一份数据在本地进行持久性存储，并向与所述存储节点属于一个集群的 N+M-1 个目标存储节点分别转发剩余的 N+M-1 份数据进行持久性存储，每个节点发送一份数据，所述存储节点与所述 N+M-1 个目标存储节点之间的转发路径满足预设条件。

11. 一种存储节点，其特征在于，所述存储节点包括处理器和存储器，所述存储器中存储有计算机可读指令，所述处理器执行所述计算机可读指令，使得所述存储节点执行如权利要求 1 至 5 任一项所述的方法。

12. 一种计算机可读存储介质，其特征在于，包括计算机可读指令，当所述计算机可读指令在存储节点上运行时，使得所述存储节点执行如权利要求 1 至 5 任一项所述的方法。

13. 一种计算机程序产品，其特征在于，包括计算机可读指令，当所述计算机可读指令在存储节点上运行时，使得所述存储节点执行如权利要求 1 至 5 任一项所述的方法。

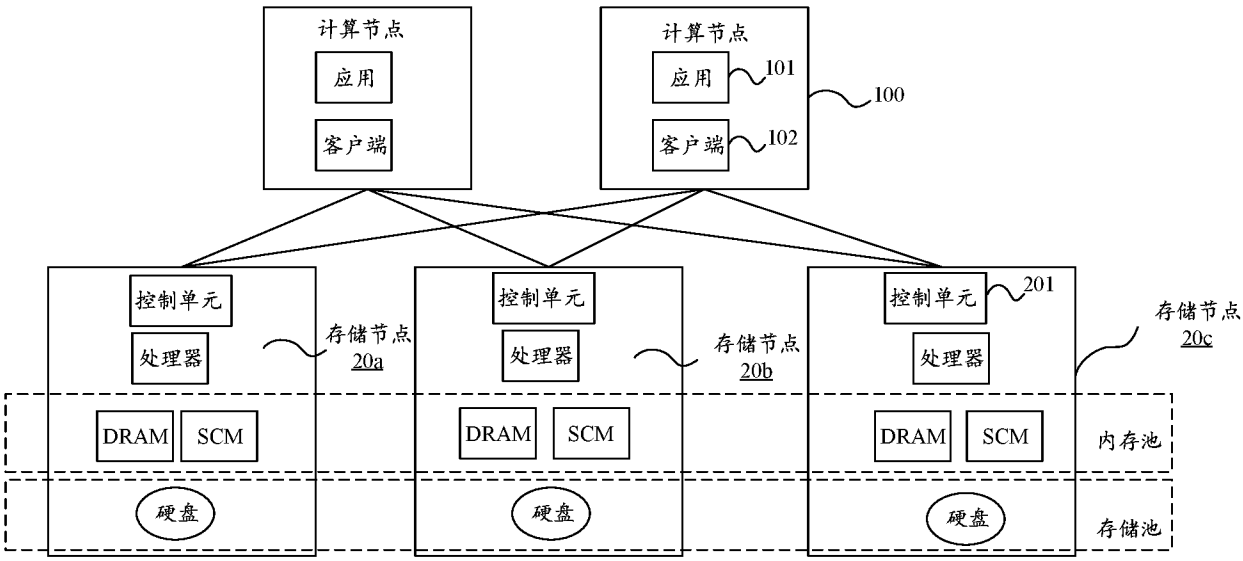
计算节点
应用
客户端
计算节点
应用
101
客户端
102
100
控制单元
处理器
存储节点
20a
DRAM
SCM
硬盘
控制单元
处理器
存储节点
20b
DRAM
SCM
硬盘
控制单元
201
处理器
存储节点
20c
DRAM
SCM
硬盘
内存池
存储池

图 1

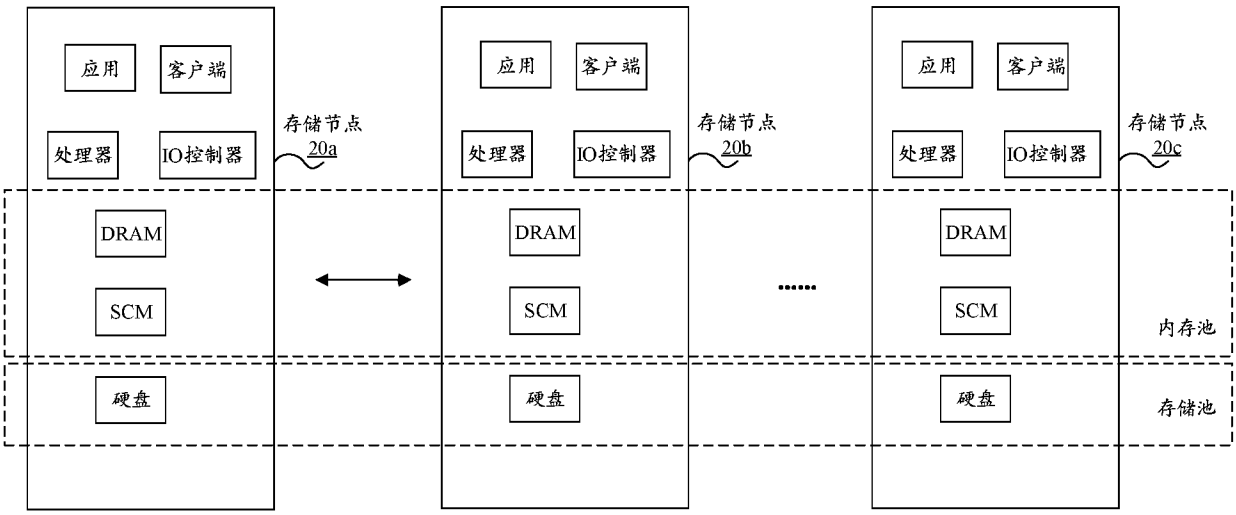
应用
客户端
处理器
IO控制器
存储节点
20a
DRAM
SCM
硬盘
应用
客户端
处理器
IO控制器
存储节点
20b
DRAM
SCM
硬盘
应用
客户端
处理器
IO控制器
存储节点
20c
DRAM
SCM
硬盘
内存池
存储池

图 2

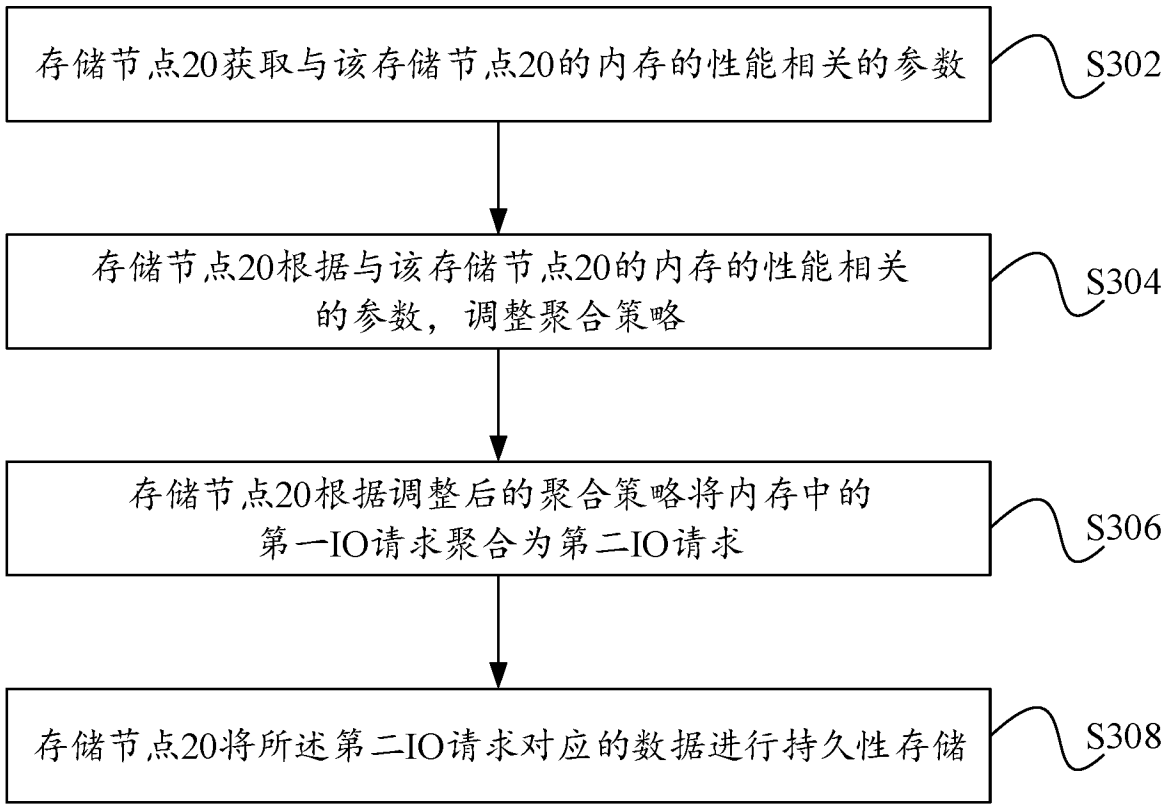
存储节点20获取与该存储节点20的内存的性能相关的参数
S302
存储节点20根据与该存储节点20的内存的性能相关的参数，调整聚合策略
S304
存储节点20根据调整后的聚合策略将内存中的第一IO请求聚合为第二IO请求
S306
存储节点20将所述第二IO请求对应的数据进行持久性存储
S308

图 3

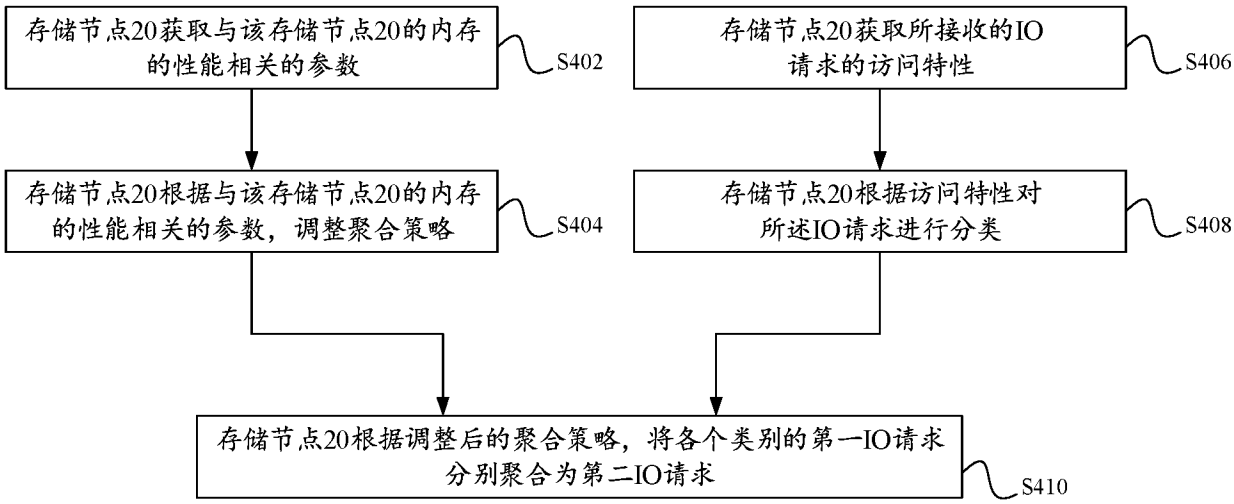
存储节点20获取与该存储节点20的内存的性能相关的参数
S402
存储节点20根据与该存储节点20的内存的性能相关的参数，调整聚合策略
S404
存储节点20获取所接收的IO请求的访问特性
S406
存储节点20根据访问特性对所述IO请求进行分类
S408
存储节点20根据调整后的聚合策略，将各个类别的第一IO请求分别聚合为第二IO请求
S410

图 4

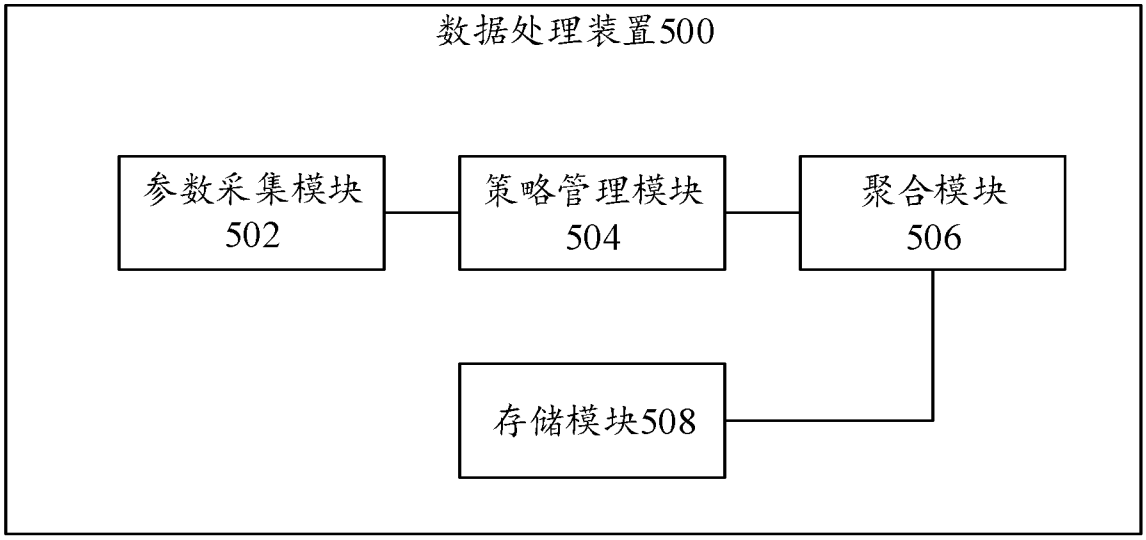

图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/095948

A. CLASSIFICATION OF SUBJECT MATTER

G06F 3/06(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)

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)

CNABS; CNTXT; VEN; USTXT; ENTXT; WOTXT; CNKI: 输入输出, 访问, 读, 写, 请求, 聚合, 合并, 自适应, 动态, 灵活, 调整, 修改, 内存, 带宽, 分类, IO, access, read, write, request, aggregate, incorporate, selfadaption, dynamic, flexible, adjust, amend, memory, bandwidth, classify

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104571931 A (INSTITUTE OF ACOUSTICS, CHINESE ACADEMY OF SCIENCES et al.) 29 April 2015 (2015-04-29) description, paragraphs 6-30 and 38-61, and figures 1-2	1-3, 6-8, 11-13
Y	CN 104571931 A (INSTITUTE OF ACOUSTICS, CHINESE ACADEMY OF SCIENCES et al.) 29 April 2015 (2015-04-29) description, paragraphs 6-30 and 38-61, and figures 1-2	4-5, 9-10, 11-13
X	CN 111371848 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 03 July 2020 (2020-07-03) description, paragraphs 43-95, and figures 1-3	1-3, 6-8, 11-13
Y	CN 111371848 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 03 July 2020 (2020-07-03) description, paragraphs 43-95, and figures 1-3	4-5, 9-10, 11-13
Y	CN 104636201 A (CHINA TELECOM CORP., LTD.) 20 May 2015 (2015-05-20) description, paragraphs 33-57, and figures 2-3	4, 9, 11-13

☑ 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
23 July 2022	**08 August 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.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	高原 等 (GAO, Yuan et al.). "一种用于透明计算的多媒体I/O访问控制方法 (A Multimedia I/O Access Control Policy for Transparent Computing)" *湖南大学学报 (Journal of Hunan University)*, Vol. 40, No. 3, 25 March 2013 (2013-03-25), pages 93-98	4, 9, 11-13
Y	CN 111381770 A (ZHEJIANG UNIVIEW TECHNOLOGIES CO., LTD.) 07 July 2020 (2020-07-07) description, paragraphs 34-35	5, 10, 11-13
A	CN 106569893 A (ALIBABA GROUP HOLDING LIMITED) 19 April 2017 (2017-04-19) entire document	1-13
A	CN 111737212 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 02 October 2020 (2020-10-02) entire document	1-13
A	US 5408644 A (COMPAQ COMPUTER CORP.) 18 April 1995 (1995-04-18) entire document	1-13

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/095948

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	104571931	A	29 April 2015	CN	104571931	B	03 October 2017
CN	111371848	A	03 July 2020	WO	2021164163	A1	26 August 2021
CN	104636201	A	20 May 2015	CN	104636201	B	02 February 2018
CN	111381770	A	07 July 2020	WO	2020140523	A1	09 July 2020
				CN	111381770	B	06 July 2021
CN	106569893	A	19 April 2017	CN	106569893	B	05 February 2021
CN	111737212	A	02 October 2020		None		
US	5408644	A	18 April 1995	EP	0573308	A2	08 December 1993
				CA	2097782	A1	06 December 1993

国际检索报告

国际申请号
PCT/CN2022/095948

A. 主题的分类

G06F 3/06(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)

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）)

CNABS;CNTXT;VEN;USTXT;ENTXT;WOTXT;CNKI: 输入输出，访问，读，写，请求，聚合，合并，自适应，动态，灵活，调整，修改，内存，带宽，分类，IO，access，read，write，request，aggregate，incorporate，selfadaption，dynamic，flexible，adjust，amend，memory，bandwidth，classify

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	CN 104571931 A (中国科学院声学研究所 等) 2015年4月29日 (2015 - 04 - 29) 说明书第6-30、38-61段、图1-2	1-3、6-8、11-13
Y	CN 104571931 A (中国科学院声学研究所 等) 2015年4月29日 (2015 - 04 - 29) 说明书第6-30、38-61段、图1-2	4-5、9-10、11-13
X	CN 111371848 A (苏州浪潮智能科技有限公司) 2020年7月3日 (2020 - 07 - 03) 说明书第43-95段、图1-3	1-3、6-8、11-13
Y	CN 111371848 A (苏州浪潮智能科技有限公司) 2020年7月3日 (2020 - 07 - 03) 说明书第43-95段、图1-3	4-5、9-10、11-13
Y	CN 104636201 A (中国电信股份有限公司) 2015年5月20日 (2015 - 05 - 20) 说明书第33-57段、图2-3	4、9、11-13
Y	高原 等. "一种用于透明计算的多媒体I/O访问控制方法" 湖南大学学报, 第40卷, 第3期, 2013年3月25日 (2013 - 03 - 25), 第93-98页	4、9、11-13
Y	CN 111381770 A (浙江宇视科技有限公司) 2020年7月7日 (2020 - 07 - 07) 说明书第34-35段	5、10、11-13

☑ 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7月23日	2022年8月8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 赖女女 电话号码 86-(20)-28958938

国际检索报告

国际申请号
PCT/CN2022/095948

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
A	CN 106569893 A (阿里巴巴集团控股有限公司) 2017年4月19日 (2017 - 04 - 19) 全文	1-13
A	CN 111737212 A (苏州浪潮智能科技有限公司) 2020年10月2日 (2020 - 10 - 02) 全文	1-13
A	US 5408644 A (COMPAQ COMPUTER CORP) 1995年4月18日 (1995 - 04 - 18) 全文	1-13

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2022/095948

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	104571931	A	2015年4月29日	CN	104571931	B	2017年10月3日
CN	111371848	A	2020年7月3日	WO	2021164163	A1	2021年8月26日
CN	104636201	A	2015年5月20日	CN	104636201	B	2018年2月2日
CN	111381770	A	2020年7月7日	WO	2020140523	A1	2020年7月9日
				CN	111381770	B	2021年7月6日
CN	106569893	A	2017年4月19日	CN	106569893	B	2021年2月5日
CN	111737212	A	2020年10月2日		无		
US	5408644	A	1995年4月18日	EP	0573308	A2	1993年12月8日
				CA	2097782	A1	1993年12月6日