

(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
2018 年 8 月 30 日 (30.08.2018)



(10) 国际公布号
WO 2018/153251 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2018/075458
- (22) 国际申请日: 2018 年 2 月 6 日 (06.02.2018)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201710091836.0 2017年2月21日 (21.02.2017) CN
- (71) 申请人: 中兴通讯股份有限公司 (ZTE CORPORATION) [CN/CN]; 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。
- (72) 发明人: 张广艳 (ZHANG, Guangyan); 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦中兴通讯股份有限公司转交, Guangdong 518057 (CN)。 罗圣美 (LUO, Shengmei); 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦中兴通讯股份有限公司转交, Guangdong 518057 (CN)。 都涛 (DU, Tao); 中国广东省深圳市南山区高新技术产业园科技

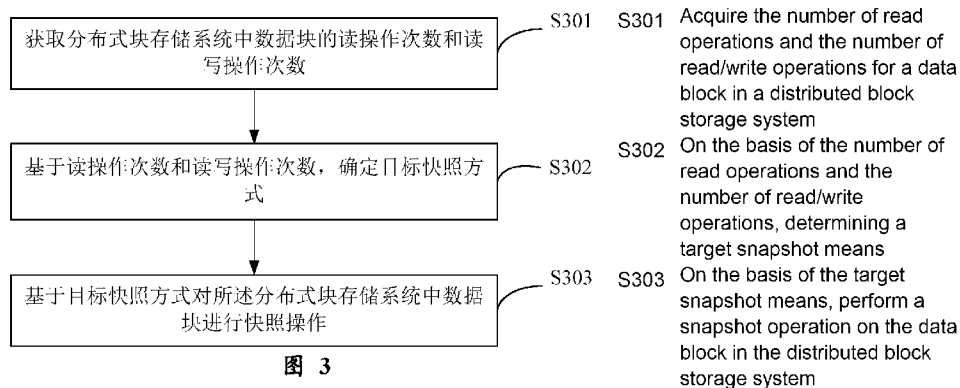
南路中兴通讯大厦中兴通讯股份有限公司转交, Guangdong 518057 (CN)。 杨洪章 (YANG, Hongzhang); 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦中兴通讯股份有限公司转交, Guangdong 518057 (CN)。

(74) 代理人: 北京安信方达知识产权代理有限公司 (AFD CHINA INTELLECTUAL PROPERTY LAW OFFICE); 中国北京市海淀区学清路 8 号 B 座 1601A, Beijing 100192 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(54) Title: METHOD FOR PROCESSING SNAPSHOTS AND DISTRIBUTED BLOCK STORAGE SYSTEM

(54) 发明名称: 一种快照的处理方法及分布式块存储系统



(57) Abstract: A method for processing snapshots and a distributed block storage system, the method comprising: acquiring the number of read operations and the number of read/write operations for a data block in a distributed block storage system (S301); on the basis of the number of read operations and the number of read/write operations, determining a target snapshot means, wherein the target snapshot means comprises copying while writing or re-orienting while writing (S302); on the basis of the target snapshot means, performing a snapshot operation on the data block in the distributed block storage system (S303).

(57) 摘要: 一种快照的处理方法及分布式块存储系统, 该方法可包括: 获取所述分布式块存储系统中数据块的读操作次数和读写操作次数 (S301); 基于所述读操作次数和读写操作次数, 确定目标快照方式, 其中, 所述目标快照方式包括写时复制或写时重定向 (S302); 基于所述目标快照方式对所述分布式块存储系统中数据块进行快照操作 (S303)。

(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)。

根据细则4.17的声明:

- 关于申请人有权申请并被授予专利 (细则4.17(ii))
- 发明人资格 (细则4.17(iv))

本国际公布:

- 包括国际检索报告 (条约第21条(3))。

一种快照的处理方法及分布式块存储系统

技术领域

- 5 本文涉及但不限于通信技术领域，特别涉及一种快照的处理方法及分布式块存储系统。

背景技术

- 10 分布式块存储系统具有可靠、高效、可扩展、接口灵活等特性，被广泛用于云存储、高性能计算、设备虚拟化等场景。典型的分布式存储系统通常由客户端、元数据服务器和数据服务器三个部分构成。快照技术为存储系统提供了在线备份、数据恢复、数据重用等能力，是分布式块存储系统不可缺少的一部分。

15 发明内容

以下是对本文详细描述的主题的概述。本概述并非是为了限制权利要求的保护范围。

本发明实施例提供一种快照的处理方法及分布式块存储系统。

本发明实施例提供一种快照的处理方法，包括：

- 20 获取所述分布式块存储系统中数据块的读操作次数和读写操作次数；

基于所述读操作次数和读写操作次数，确定目标快照方式，其中，所述目标快照方式包括写时复制或写时重定向；

基于所述目标快照方式对所述分布式块存储系统中数据块进行快照操作。

- 25 本发明实施例还提供一种分布式块存储系统，包括：

数据获取模块，设置为获取所述分布式块存储系统中数据块的读操作次数和读写操作次数；

快照方式确定模块，设置为基于所述读操作次数和读写操作次数，确定

目标快照方式，其中，所述目标快照方式包括写时复制或写时重定向；

处理模块，设置为基于所述目标快照方式对所述分布式块存储系统中数据块进行快照操作。

5 本发明实施例还提供一种计算机存储介质，所述计算机存储介质中存储有计算机可执行的一个或多个程序，所述一个或多个程序被所述计算机执行时使所述计算机执行如上述提供的一种快照的处理方法。

本发明实施例具有如下优点或有益效果：

10 获取所述分布式块存储系统中数据块的读操作次数和读写操作次数；基于所述读操作次数和读写操作次数，确定目标快照方式，其中，所述目标快照方式包括写时复制或写时重定向；基于所述目标快照方式对所述分布式块存储系统中数据块进行快照操作。这样在分布式存储系统的快照中，当数据块以读为主时，确定以写时复制对数据块进行快照操作；当数据块以写为主时，确定以写时重定向对数据块进行快照操作，从而可以同时满足写入性能、回滚性能以及读性能都较佳。

15 在阅读并理解了附图和详细描述后，可以明白其他方面。

附图概述

图 1 为写时复制快照示意图；

图 2 为写时重定快照示意图；

20 图 3 为本发明实施例提供的一种快照的处理方法的流程示意图；

图 4 为本发明实施例提供的另一种快照的处理方法的流程示意图；

图 5 为本发明实施例提供的分布式块存储系统的部分架构示意图；

图 6 为本发明实施例提供的一种分布式块存储系统的结构示意图；

25 图 7 为本发明实施例提供的一种分布式块存储系统中快照方式确定模块的结构示意图；

图 8 为本发明实施例提供的另一种分布式块存储系统中快照方式确定模块的结构示意图；

图 9 为本发明实施例提供的另一种分布式块存储系统的结构示意图。

本发明的实施方式

下面将结合附图及具体实施例进行详细描述。

如图 1 和图 2 所示，分布式块存储系统通常使用写时复制（Copy on Write, COW）和写时重定向（Redirect on Write, ROW）两种快照方式。其中，COW 具有更好的读性能，而 ROW 具有更好的写性能和回滚性能。但是，由于 COW 进行写入和回滚操作时，需要先进行数据复制，因此写性能和回滚性能不佳；ROW 在读操作时，由于需要对数据进行重定位导致读性能不佳。

如图 3 所示，本发明实施例提供一种快照的处理方法的流程示意图，该方法应用于分布式块存储系统，包括以下步骤：

步骤 S301、获取分布式块存储系统中数据块的读操作次数和读写操作次数。

本发明实施例中，分布式存储系统可以通过磁盘操作监控工具获取每个数据块的读操作次数和读写操作次数，例如：在分布式存储系统输入 iostat 命令，实时统计存储该数据块的数据服务器的每秒进行读写（I/O）操作的次数（Input/ Output Operations Per Second, IOPS）等负载数据，从而获取该数据块的读操作次数和读写操作次数。

步骤 S302、基于读操作次数和读写操作次数，确定目标快照方式，其中，目标快照方式包括写时复制或写时重定向。

本发明实施例中，由于 COW 具有更好的读性能，而 ROW 具有更好的写性能和回滚性能，可以基于数据块的读操作次数和读写操作次数确定数据块的读写情况。若确定该数据块以读为主，可以将 COW 作为对数据块进行快照操作的快照方式；若确定该数据块以写为主，可以将 ROW 作为对数据块进行快照操作的快照方式。

其中，上述基于读操作次数和读写操作次数，确定目标快照方式，可以通过数据块的读操作次数与读写操作次数的比值或差值大小确定目标快照方式，比值或差值越大，表明该数据块以读为主；比值或差值越小，表明该数据块以写为主。例如：若读操作次数与读写操作次数的比值为 10，该数

据块以读为主，可以确定目标快照方式为 COW；若读操作次数与读写操作次数的比值为 0.1，该数据块以写为主，可以确定目标操作方式为 ROW。

步骤 S303、基于目标快照方式对分布式块存储系统中数据块进行快照操作。

- 5 本发明实施例中，上述步骤 S302 确定目标快照方式时，分布式存储系统可以采用该目标快照方式对数据块进行快照操作。其中，上述快照操作可以包括创建快照操作、读操作、写操作、回滚操作以及删除快照操作等中的任意一种。

- 10 本发明实施例中，获取分布式块存储系统中数据块的读操作次数和读写操作次数；基于读操作次数和读写操作次数，确定目标快照方式，其中，目标快照方式包括写时复制或写时重定向；基于目标快照方式对分布式块存储系统中数据块进行快照操作。这样在分布式存储系统的快照中，当数据块以读为主时，确定以写时复制对数据块进行快照操作；当数据块以写为主时，确定以写时重定向对数据块进行快照操作，从而可以同时满足写入性能、回滚性能以及读性能都较佳。
- 15

如图 4 所示，本发明实施例提供另一种快照的处理方法的流程示意图，包括以下步骤：

步骤 S401、获取分布式块存储系统中数据块的读操作次数和读写操作次数。

- 20 本发明实施例中，当客户端发起快照操作请求时，分布式块存储系统会响应该快照操作请求，通过元数据服务器中存储的快照元数据，查找该快照操作请求对应的数据块在数据服务器中的位置，并通过磁盘操作监控工具获取每个数据块的读操作次数和读写操作次数，例如：通过 iostat 命令统计存储该数据块的数据服务器的 IOPS 等负载数据，获取该数据块的读操作次数
- 25 和读写操作次数。

步骤 S402、判断读操作次数与读写操作次数的预设参数值是否大于或等于第一预设阈值，其中，预设参数值为读操作次数与读写操作次数的比值或差值。

本发明实施例中，将数据块的读操作次数与读写操作次数的比值或差值

作为预设参数值，可以直观反映数据块的读写情况，读操作次数与读写操作次数的比值或差值越大时，表明该数据块读操作较多；读操作次数与读写操作次数的比值或差值越大时，表明该数据块写操作较多。

其中，若读操作次数与读写操作次数的比值或差值大于或等于第一预设
5 阈值，则可以确定该数据块以读为主。例如：当预设参数值为读操作次数与读写操作次数的比值时，可以设置第一预设阈值为 1.5，若预设参数值大于或等于 1.5，则可以确定该数据块以读为主；当预设参数值为读操作次数与读写操作次数的差值时，可以设置第一预设阈值为 50，若预设参数值大于或等于 50，则可以确定该数据块以读为主。

10 步骤 S403、若读操作次数与读写操作次数的预设参数值大于或等于第一预设阈值，确定目标快照方式为写时复制。

本发明实施例中，当上述步骤 S403 确定读操作次数与读写操作次数的
15 预设参数值大于或等于第一预设阈值，可以确定该数据块以读为主，从而可以确定目标快照方式为具有更好的读性能的写时复制，从而使分布式存储系统具有较佳的读性能。

步骤 S404、若读操作次数与读写操作次数的预设参数值小于或等于第一预设阈值，确定目标快照方式为写时重定向。

本发明实施例中，当上述步骤 S403 确定读操作次数与读写操作次数的
20 预设参数值小于第一预设阈值，可以确定该数据块不以读为主，该数据块可能读写接近或者以写为主，可以确定目标快照方式为具有更好的写性能和回滚性能的写时重定向，从而使分布式存储系统具有较佳的写性能和回滚性能。

可选的，上述确定目标快照方式为写时重定向之前还可以包括：

判断读操作次数与读写操作次数的预设参数值是否大于第二预设阈值，
25 其中，第二预设阈值小于第一预设阈值；

若读操作次数与读写操作次数的预设参数值大于第二预设阈值，判断前端负载值是否大于预设负载阈值，其中，前端负载值为存储数据块的数据服务器的负载值；

若前端负载值小于或等于预设负载阈值，确定目标快照方式为写时复

制；

若前端负载值大于预设负载阈值或者读操作次数与读写操作次数的预设参数值小于或等于第二预设阈值，执行确定目标快照方式为写时重定向的步骤。

5 本发明实施例中，当数据块的读操作次数与读写操作次数的预设参数值小于第二预设阈值时，可以确定数据块以写为主，从而可以确定目标快照方式为具有较佳的写性能和回滚性能的写时重定向；而当数据块的读操作次数与读写操作次数的预设参数值小于第一预设阈值且大于第二预设阈值时，此时数据块的读写接近。

10 其中，为了保证分布式块存储系统在数据块的读写接近情况下仍同时满足写入性能、回滚性能以及读性能都较佳，可以通过获取存储该数据块的数据服务器的前端负载值，并将该前端负载值与预设负载阈值进行比较，若前端负载值大于预设负载阈值，表明该数据服务器负载高，可以确定目标快照方式为 ROW，从而降低数据块的写请求的开销；若前端负载值小于或等于
15 预设负载阈值，表明数据服务器负载低，可以确定目标快照方式为 COW，从而提升数据块的读请求的性能。

例如：当预设参数值为读操作次数与读写操作次数的比值时，可以设定第一预设阈值为 1.5，第二预设阈值为 0.7，而预设负载阈值为 10000，若读操作次数与读写操作次数的比值小于或等于 0.7，表明数据块以写为主，确
20 定目标快照方式为 ROW；若读操作次数与读写操作次数的比值大于 0.7 且小于 1.5，表明数据块的读写接近，此时若对应数据服务器的前端负载值为 15000，则确定目标快照方式为 ROW。

需要说明的是，上述第一预设阈值、第二预设阈值以及预设负载阈值可以根据实际需要进行设定，在此并不进行限定。

25 步骤 S405、基于目标快照方式对分布式块存储系统中数据块进行快照操作。

本发明实施例中，当确定目标快照方式时，分布式存储系统可以采用该目标快照方式对数据块进行快照操作。

可选的，上述步骤 S405 之前，还可以包括：判断当前快照方式是否与

目标快照方式匹配；

若当前快照方式与目标快照方式不匹配，切换当前快照方式为目标快照方式。

本发明实施例中，当上述目标快照方式确定时，分布式存储系统可以将目标快照方式与数据块的当前快照方式进行匹配，若目标快照方式与当前快照方式不匹配，可以将当前快照方式切换为目标快照方式。例如：若当前的快照方式为 COW，而目标快照方式为 ROW，可以将该数据块的当前快照方式由 COW 切换为 ROW；若当前的快照方式为 ROW，而目标快照方式为 COW，可以将该数据块的当前快照方式由 ROW 切换为 COW。

可选的，上述快照操作包括创建快照操作、读操作、写操作、回滚操作以及删除快照操作中的任意一种，从而使分布式存储系统在任意一种快照操作下都同时满足写入性能、回滚性能以及读性能都较佳。

本发明实施例中，以上述快照的处理方法在读操作、写操作、回滚操作以及删除快照操作中的实际应用为例进行说明。其中，如图 5 所示，分布式存储系统包括多个数据元数据服务器以及多个数据服务器，每一数据服务器包括多个数据块，数据块为分布式存储系统中的最小存储单元。分布式存储系统的快照元数据存储在元数据服务器上，分布式存储系统所包含的数据块的快照元数据信息存储于该数据块所在的数据服务器上。

其中，数据块的当前数据和快照数据以文件的形式存放于数据服务器的同一目录下，快照文件、快照方式为 COW 时的数据文件以及快照方式为 ROW 时的数据文件均具有不同的命名格式，例如：快照文件的命名格式为数据块 id+ “_” +snap_id；快照方式为 COW 时的数据文件的命名格式为数据块 id+ “_” + “head”；快照方式为 ROW 时的数据文件的命名格式为 id+“_”+“head_id”。并且快照方式为 COW 时操作数据文件，可以直接生成数据文件的文件名，无需经过查找或计算数据块的位置；快照方式为 ROW 时操作数据文件，从快照元数据信息中读取 head_id，从而生成数据文件的文件名，进行 IO 操作。

在快照操作为创建快照操作时，分布式块存储系统将创建快照请求发送到对应的数据块所在的数据服务器上，根据数据块的读操作次数、读写操作

次数以及前端负载值，确定该数据块的目标快照方式，设置当前快照方式为目标快照方式，以当前快照方式创建快照，并更新数据块的快照元数据信息。

5 在快照操作为读取数据的读操作时，分布式块存储系统将读请求发送到对应的数据块所在的数据服务器上，根据数据块的读操作次数、读写操作次数以及前端负载值，确定该数据块的目标快照方式，设置当前快照方式为目标快照方式，生成目标快照方式对应的数据文件的文件名，利用快照元数据信息的快照序号到快照文件的对应关系的属性，从该数据块的数据目录下获取数据文件。

10 在快照操作为写入数据的写操作时，分布式块存储系统将写请求发送到对应的数据块所在的数据服务器上，根据数据块的读操作次数、读写操作次数以及前端负载值，确定该数据块的目标快照方式，设置当前快照方式为目标快照方式，以当前快照方式进行复制写或者重定向写，并更新数据块的快照元数据信息。

15 在快照操作为回滚操作时，分布式块存储系统将回滚请求发送到对应的数据块所在的数据服务器上，根据数据块的读操作次数、读写操作次数以及前端负载值，确定该数据块的目标快照方式，设置当前快照方式为目标快照方式，以当前快照方式进行复制回滚或者直接修改指针，并更新数据块的快照元数据信息。

20 在快照操作为删除快照操作时，分布式块存储系统将删除快照请求发送到对应的数据块所在的数据服务器上，根据数据块的读操作次数、读写操作次数以及前端负载值，确定该数据块的目标快照方式，设置当前快照方式为目标快照方式，判断是否需要删除实际对应的快照文件，若为是，删除快照文件，并更新数据块的快照元数据信息。

25 本发明实施例中，通过判断数据块的读操作次数和读写操作次数的比值或差值与第一预设阈值和第二预设阈值的关系，若比值或差值大于或等于第一预设阈值，将 COW 作为对数据块进行快照操作的快照方式；若比值或差值小于或等于第二预设阈值，将 ROW 作为对数据块进行快照操作的快照方式；若比值或差值大于第二预设阈值且小于第一预设阈值，判断存储数据块的数据服务器的前端负载值是否大于预设负载阈值，若前端负载值大于预设

负载阈值，将 ROW 作为对数据块进行快照操作的快照方式；若前端负载值小于或等于预设负载阈值，将 COW 作为对数据块进行快照操作的快照方式。这样，在分布式存储系统的快照中，可以同时满足写入性能、回滚性能以及读性能都较佳。

5 如图 6 所示，本发明实施例提供一种分布式块存储系统的结构示意图，所述分布式块存储系统 60 包括：

数据获取模块 61，设置为获取所述分布式块存储系统中数据块的读操作次数和读写操作次数；

快照方式确定模块 62，设置为基于所述读操作次数和读写操作次数，
10 确定目标快照方式，其中，所述目标快照方式包括写时复制或写时重定向；

处理模块 63，设置为基于所述目标快照方式对所述分布式块存储系统中数据块进行快照操作。

可选的，如图 7 所示，所述快照方式确定模块 62 可以包括：

第一判断单元 621，设置为判断所述读操作次数与所述读写操作次数的
15 预设参数值是否大于第一预设阈值，其中，所述预设参数值为所述读操作次数与
所述读写操作次数的比值或差值；

第一确定单元 622，设置为若所述读操作次数与所述读写操作次数的预设参数值大于第一预设阈值，确定所述目标快照方式为写时复制；

第二确定单元 623，设置为若所述读操作次数与所述读写操作次数的预
20 设参数值小于或等于第一预设阈值，确定所述目标快照方式为写时重定向。

可选的，如图 8 所示，所述快照方式确定模块 62 还可以包括：

第二判断单元 624，设置为判断所述读操作次数与所述读写操作次数的预设参数值是否大于第二预设阈值，其中，所述第二预设阈值小于所述第一预设阈值；

25 第三判断单元 625，设置为若所述读操作次数与所述读写操作次数的预设参数值大于第二预设阈值，判断前端负载值是否大于预设负载阈值，其中，所述前端负载值为存储所述数据块的数据服务器的负载值；

第三确定单元 626，设置为若所述前端负载值小于或等于预设负载阈值，确定所述目标快照方式为写时复制；

执行单元 627, 设置为若所述前端负载值大于预设负载阈值或者所述读操作次数与所述读写操作次数的预设参数值小于或等于第二预设阈值, 执行所述确定所述目标快照方式为写时重定向的步骤。

可选的, 如图 9 所示, 所述分布式块存储系统 60 还可以包括:

- 5 判断模块 64, 设置为判断当前快照方式是否与所述目标快照方式匹配;
 切换模块 65, 设置为若所述当前快照方式与所述目标快照方式不匹配, 切换所述当前快照方式为所述目标快照方式。

可选的, 所述快照操作可以包括创建快照操作、读操作、写操作、回滚操作以及删除快照操作中的任意一种。

- 10 需要说明的是, 本实施例的分布式块存储系统可以是图 1 至图 5 所示的实施例中的分布式块存储系统, 图 1 至图 5 所示实施例介绍的分布式块存储系统的任意实施方式都可以被上述分布式块存储系统实现, 此处不再赘述。

本领域普通技术人员可以理解实现上述实施例方法的全部或者部分步骤是可以通过程序指令相关的硬件来完成, 所述的程序可以存储于一计算机

- 15 可读介质中, 该程序在执行时, 包括以下步骤:

 获取所述分布式块存储系统中数据块的读操作次数和读写操作次数;

 基于所述读操作次数和读写操作次数, 确定目标快照方式, 其中, 所述目标快照方式包括写时复制或写时重定向;

- 20 基于所述目标快照方式对所述分布式块存储系统中数据块进行快照操作。

可选的, 所述基于所述读操作次数和读写操作次数, 确定目标快照方式的步骤, 包括:

- 判断所述读操作次数与所述读写操作次数的预设参数值是否大于或等于第一预设阈值, 其中, 所述预设参数值为所述读操作次数与所述读写操作
25 次数的比值或差值;

 若所述读操作次数与所述读写操作次数的预设参数值大于或等于第一预设阈值, 确定所述目标快照方式为写时复制;

 若所述读操作次数与所述读写操作次数的预设参数值小于第一预设阈值, 确定所述目标快照方式为写时重定向。

可选的,所述确定所述目标快照方式为写时重定向的步骤之前,还包括:

判断所述读操作次数与所述读写操作次数的预设参数值是否大于第二预设阈值,其中,所述第二预设阈值小于所述第一预设阈值;

5 若所述读操作次数与所述读写操作次数的预设参数值大于第二预设阈值,判断前端负载值是否大于预设负载阈值,其中,所述前端负载值为存储所述数据块的数据服务器的负载值;

若所述前端负载值小于或等于预设负载阈值,确定所述目标快照方式为写时复制;

10 若所述前端负载值大于预设负载阈值或者所述读操作次数与所述读写操作次数的预设参数值小于或等于第二预设阈值,执行所述确定所述目标快照方式为写时重定向的步骤。

可选的,所述基于所述目标快照方式对所述分布式块存储系统中数据块进行快照操作步骤之前,还包括:

判断当前快照方式是否与所述目标快照方式匹配;

15 若所述当前快照方式与所述目标快照方式不匹配,切换所述当前快照方式为所述目标快照方式。

可选的,所述快照操作包括创建快照操作、读操作、写操作、回滚操作以及删除快照操作中的任意一种。

20

本领域普通技术人员可以理解,上文中所公开方法中的全部或某些步骤、系统、装置中的功能模块/单元可以被实施为软件、固件、硬件及其适当的组合。在硬件实施方式中,在以上描述中提及的功能模块/单元之间的划分不一定对应于物理单元的划分;例如,一个物理组件可以具有多个功能,25 或者一个功能或步骤可以由若干物理组件合作执行。某些组件或所有组件可以被实施为由处理器,如数字信号处理器或微处理器执行的软件,或者被实施为硬件,或者被实施为集成电路,如专用集成电路。这样的软件可以分布在计算机可读介质上,计算机可读介质可以包括计算机存储介质(或非暂时性介质)和通信介质(或暂时性介质)。如本领域普通技术人员公知的,术

语计算机存储介质包括用于存储信息（诸如计算机可读指令、数据结构、程序模块或其他数据）的任何方法或技术中实施的易失性和非易失性、可移除和不可移除介质。计算机存储介质包括但不限于 RAM、ROM、EEPROM、闪存或其他存储器技术、CD-ROM、数字多功能盘（DVD）或其他光盘存储、磁盒、磁带、磁盘存储或其他磁存储装置、或者可以用于存储期望的信息并且可以被计算机访问的任何其他的介质。此外，本领域技术人员公知的是，通信介质通常包含计算机可读指令、数据结构、程序模块或者诸如载波或其他传输机制之类的调制数据信号中的其他数据，并且可包括任何信息递送介质。

- 5
- 10 以上所述是本发明的可选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明所述原理的前提下，还可以作出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

工业实用性

- 15 本发明实施例可实现在分布式存储系统的快照中，当数据块以读为主时，确定以写时复制对数据块进行快照操作；当数据块以写为主时，确定以写时重定向对数据块进行快照操作，从而可以同时满足写入性能、回滚性能以及读性能都较佳。

权 利 要 求 书

1、一种快照的处理方法，应用于分布式块存储系统，包括：

获取所述分布式块存储系统中数据块的读操作次数和读写操作次数（S301）；

5 基于所述读操作次数和读写操作次数，确定目标快照方式，其中，所述目标快照方式包括写时复制或写时重定向（S302）；

 基于所述目标快照方式对所述分布式块存储系统中数据块进行快照操作（S303）。

10 2、如权利要求1所述的方法，其中，所述基于所述读操作次数和读写操作次数，确定目标快照方式（S302）的步骤，包括：

 判断所述读操作次数与所述读写操作次数的预设参数值是否大于或等于第一预设阈值，其中，所述预设参数值为所述读操作次数与所述读写操作次数的比值或差值（S401、S402）；

15 若所述读操作次数与所述读写操作次数的预设参数值大于或等于第一预设阈值，确定所述目标快照方式为写时复制（S403）；

 若所述读操作次数与所述读写操作次数的预设参数值小于第一预设阈值，确定所述目标快照方式为写时重定向（S404）。

3、如权利要求2所述的方法，还包括：

20 所述确定所述目标快照方式为写时重定向（S404）的步骤之前，判断所述读操作次数与所述读写操作次数的预设参数值是否大于第二预设阈值，其中，所述第二预设阈值小于所述第一预设阈值；

 若所述读操作次数与所述读写操作次数的预设参数值大于第二预设阈值，判断前端负载值是否大于预设负载阈值，其中，所述前端负载值为存储所述数据块的数据服务器的负载值；

25 若所述前端负载值小于或等于预设负载阈值，确定所述目标快照方式为写时复制；

 若所述前端负载值大于预设负载阈值或者所述读操作次数与所述读写操作次数的预设参数值小于或等于第二预设阈值，执行所述确定所述目标快

照方式为写时重定向的步骤。

4、如权利要求 1 至 3 中任意一项所述的方法，还包括：

所述基于所述目标快照方式对所述分布式块存储系统中数据块进行快照操作(S303)步骤之前，判断当前快照方式是否与所述目标快照方式匹配；

5 若所述当前快照方式与所述目标快照方式不匹配，切换所述当前快照方式为所述目标快照方式。

5、如权利要求 1 至 3 中任意一项所述的方法，其中，所述快照操作包括创建快照操作、读操作、写操作、回滚操作以及删除快照操作中的任意一种。

10 6、一种分布式块存储系统(60)，包括：

数据获取模块(61)，设置为获取所述分布式块存储系统中数据块的读操作次数和读写操作次数；

快照方式确定模块(62)，设置为基于所述读操作次数和读写操作次数，确定目标快照方式，其中，所述目标快照方式包括写时复制或写时重定向；

15 处理模块(63)，设置为基于所述目标快照方式对所述分布式块存储系统中数据块进行快照操作。

7、如权利要求 6 所述的分布式块存储系统(60)，其中，所述快照方式确定模块(62)包括：

20 第一判断单元(621)，设置为判断所述读操作次数与所述读写操作次数的预设参数值是否大于第一预设阈值，其中，所述预设参数值为所述读操作次数与所述读写操作次数的比值或差值；

第一确定单元(622)，设置为若所述读操作次数与所述读写操作次数的预设参数值大于第一预设阈值，确定所述目标快照方式为写时复制；

25 第二确定单元(623)，设置为若所述读操作次数与所述读写操作次数的预设参数值小于或等于第一预设阈值，确定所述目标快照方式为写时重定向。

8、如权利要求 7 所述的分布式块存储系统(60)，所述快照方式确定模块(62)还包括：

第二判断单元(624)，设置为判断所述读操作次数与所述读写操作次

数的预设参数值是否大于第二预设阈值，其中，所述第二预设阈值小于所述第一预设阈值；

第三判断单元（625），设置为若所述读操作次数与所述读写操作次数的预设参数值大于第二预设阈值，判断前端负载值是否大于预设负载阈值，其中，所述前端负载值为存储所述数据块的数据服务器的负载值；

第三确定单元（626），设置为若所述前端负载值小于或等于预设负载阈值，确定所述目标快照方式为写时复制；

执行单元（627），设置为若所述前端负载值大于预设负载阈值或者所述读操作次数与所述读写操作次数的预设参数值小于或等于第二预设阈值，执行所述确定所述目标快照方式为写时重定向的步骤。

9、如权利要求 6 至 8 中任意一项所述的分布式块存储系统（60），所述分布式块存储系统（60）还包括：

判断模块（64），设置为判断当前快照方式是否与所述目标快照方式匹配；

切换模块（65），设置为若所述当前快照方式与所述目标快照方式不匹配，切换所述当前快照方式为所述目标快照方式。

10、如权利要求 6 至 8 中任意一项所述的分布式块存储系统（60），其中，所述快照操作包括创建快照操作、读操作、写操作、回滚操作以及删除快照操作中的任意一种。

11、一种计算机可读存储介质，存储有计算机可执行指令，所述计算机可执行指令被处理器执行时实现权利要求 1 至 5 中任一项所述的方法。

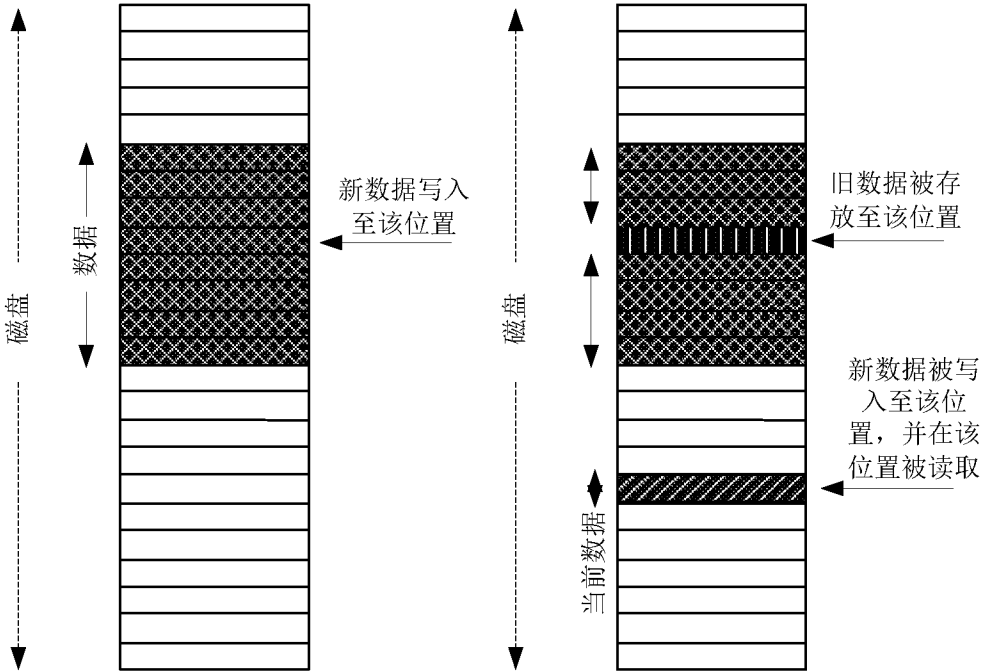


图 1

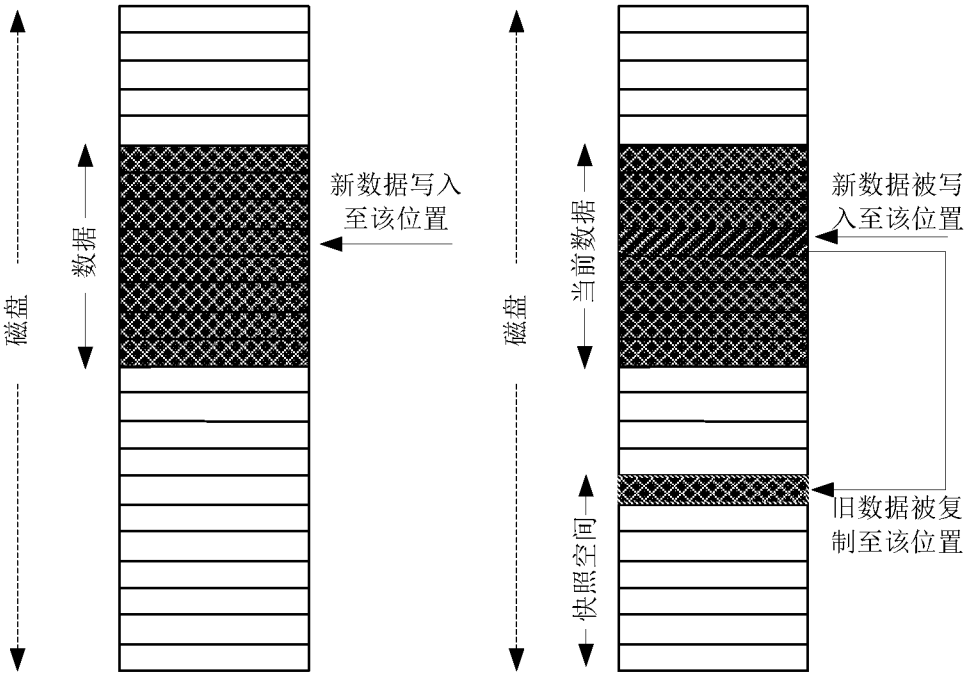


图 2

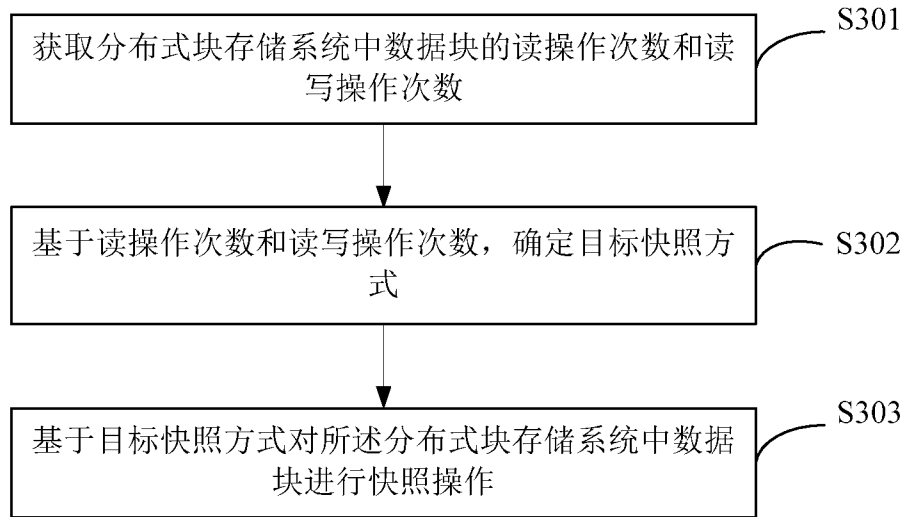


图 3

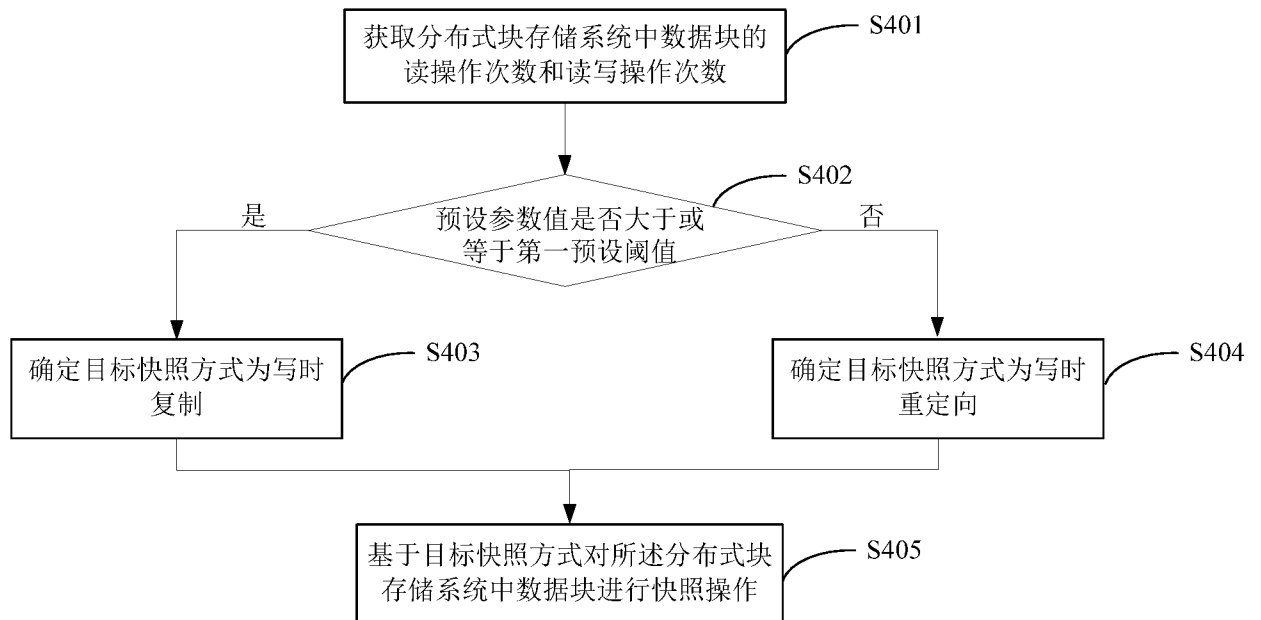


图 4

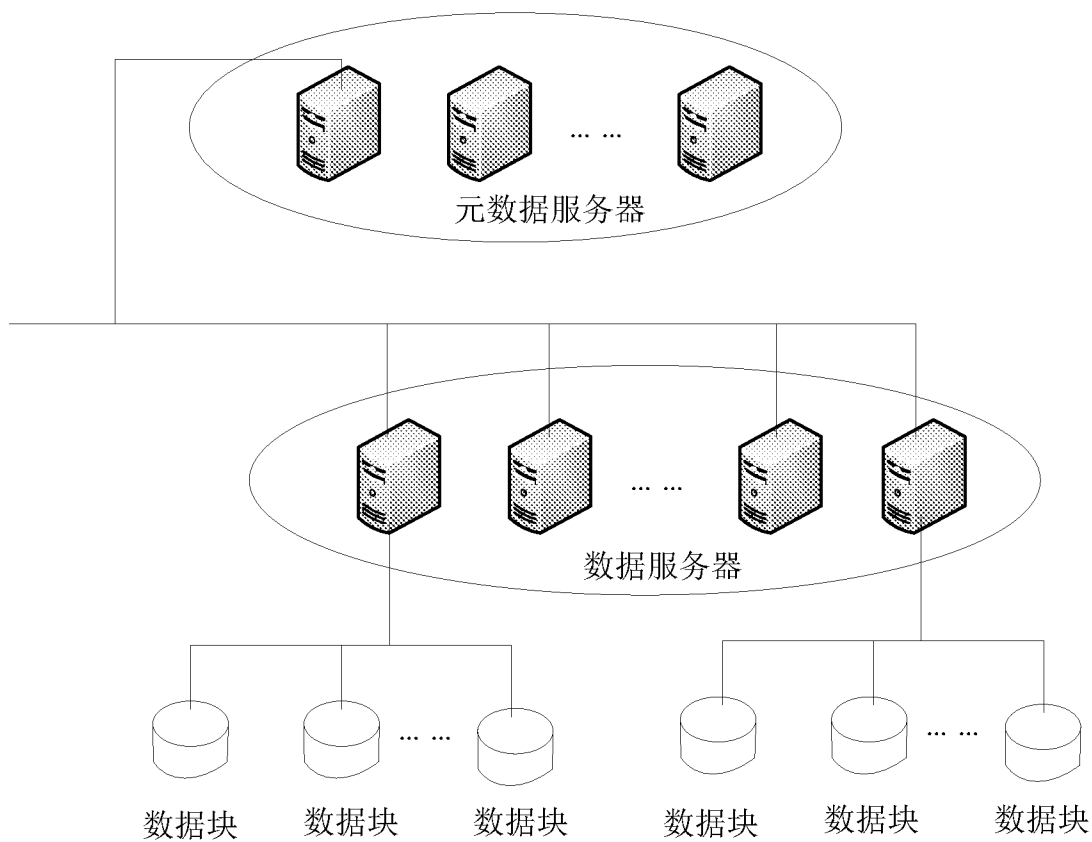


图 5

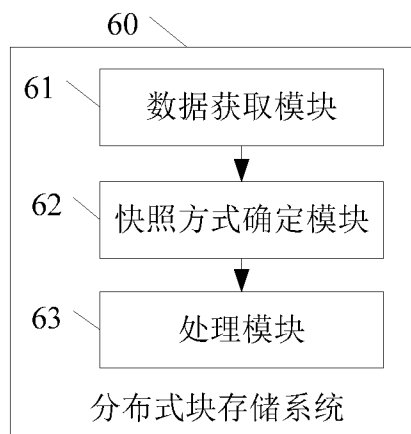


图 6

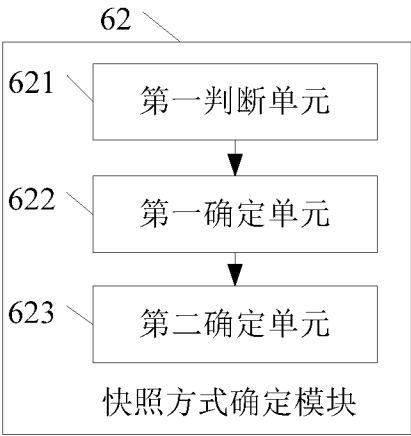


图 7

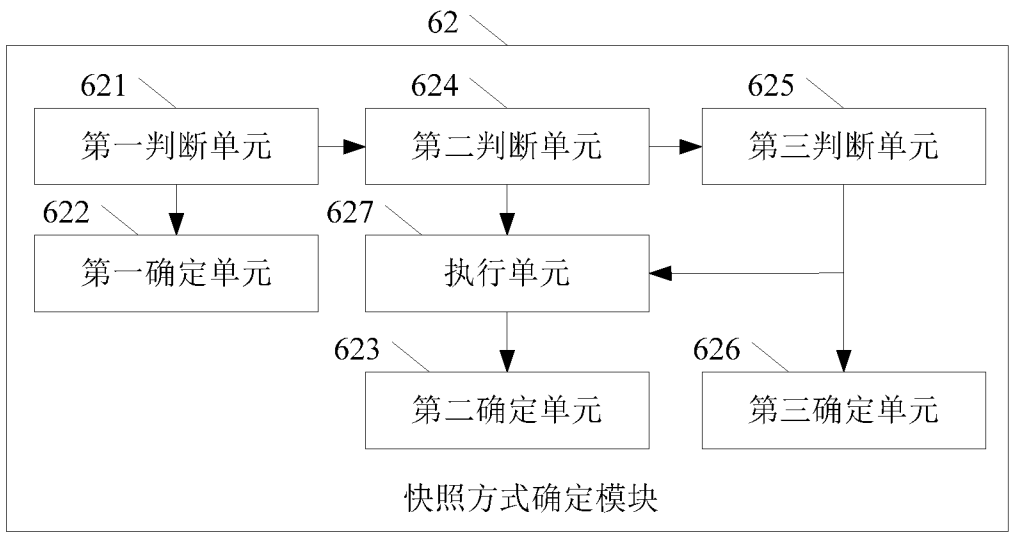


图 8

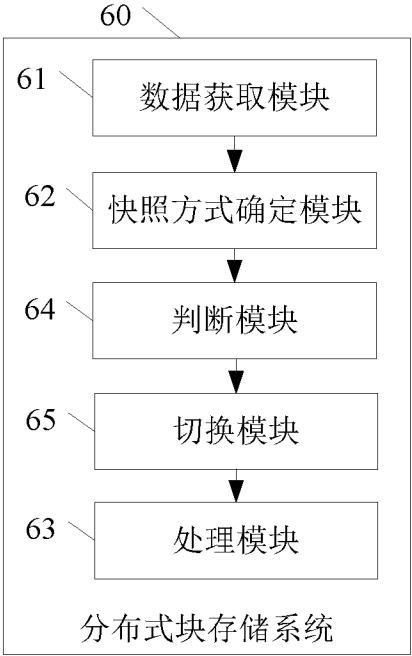


图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/075458

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (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)

WPI, EPODOC, CNPAT, CNKI, IEEE: 快照, 写时拷贝, 写时复制, 写前拷贝, 写时重定向, 读, 写, 比值, 阈值, 次数, snapshot,
Copy w On w Write, COW, Redirect w On w Write, ROW, read+, writ+, ratio, threshold, number?, count+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 101183383 A (INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 21 May 2008 (21.05.2008), description, page 1, paragraphs 2 and 3, page 7, paragraph 3 and page 10, paragraphs 2 and 3	1-11
A	CN 101482838 A (HANGZHOU H3C TECHNOLOGIES CO., LTD.) 15 July 2009 (15.07.2009), entire document	1-11
A	CN 102971698 A (HUAWEI TECHNOLOGIES CO., LTD.) 13 March 2013 (13.03.2013), entire document	1-11
A	US 2006206677 A1 (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 14 September 2006 (14.09.2006), entire document	1-11

☐ 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 26 April 2018	Date of mailing of the international search report 07 May 2018
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 ZHAO, Ting Telephone No. (86-10) 53961350

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/CN2018/075458

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 101183383 A	21 May 2008	None	
CN 101482838 A	15 July 2009	None	
CN 102971698 A	13 March 2013	WO 2014000249 A1	03 January 2014
US 2006206677 A1	14 September 2006	None	

国际检索报告

国际申请号

PCT/CN2018/075458

A. 主题的分类 G06F 17/30 (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)) WPI, EPODOC, CNPAT, CNKI, IEEE: 快照, 写时拷贝, 写时复制, 写前拷贝, 写时重定向, 读, 写, 比值, 阈值, 次数, snapshot, Copy w On w Write, COW, Redirect w On w Write, ROW, read+, wirt+, ratio, threshold, number?, count+																	
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>A</td> <td>CN 101183383 A (中国科学院计算技术研究所) 2008年 5月 21日 (2008 - 05 - 21) 说明书第1页第2, 3段, 第7页第3段, 第10页第2, 3段</td> <td>1-11</td> </tr> <tr> <td>A</td> <td>CN 101482838 A (杭州华三通信技术有限公司) 2009年 7月 15日 (2009 - 07 - 15) 全文</td> <td>1-11</td> </tr> <tr> <td>A</td> <td>CN 102971698 A (华为技术有限公司) 2013年 3月 13日 (2013 - 03 - 13) 全文</td> <td>1-11</td> </tr> <tr> <td>A</td> <td>US 2006206677 A1 (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 2006年 9月 14日 (2006 - 09 - 14) 全文</td> <td>1-11</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	A	CN 101183383 A (中国科学院计算技术研究所) 2008年 5月 21日 (2008 - 05 - 21) 说明书第1页第2, 3段, 第7页第3段, 第10页第2, 3段	1-11	A	CN 101482838 A (杭州华三通信技术有限公司) 2009年 7月 15日 (2009 - 07 - 15) 全文	1-11	A	CN 102971698 A (华为技术有限公司) 2013年 3月 13日 (2013 - 03 - 13) 全文	1-11	A	US 2006206677 A1 (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 2006年 9月 14日 (2006 - 09 - 14) 全文	1-11
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
A	CN 101183383 A (中国科学院计算技术研究所) 2008年 5月 21日 (2008 - 05 - 21) 说明书第1页第2, 3段, 第7页第3段, 第10页第2, 3段	1-11															
A	CN 101482838 A (杭州华三通信技术有限公司) 2009年 7月 15日 (2009 - 07 - 15) 全文	1-11															
A	CN 102971698 A (华为技术有限公司) 2013年 3月 13日 (2013 - 03 - 13) 全文	1-11															
A	US 2006206677 A1 (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 2006年 9月 14日 (2006 - 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 2018年 4月 26日		国际检索报告邮寄日期 2018年 5月 7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		授权官员 赵婷 电话号码 86-(10)-53961350															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/075458

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	101183383	A	2008年 5月 21日	无	
CN	101482838	A	2009年 7月 15日	无	
CN	102971698	A	2013年 3月 13日	WO 2014000249 A1	2014年 1月 3日
US	2006206677	A1	2006年 9月 14日	无	