(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
2020 年 1 月 2 日 (02.01.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/000817 A1

(51) 国际专利分类号:
G06F 3/06 (2006.01)

(21) 国际申请号: PCT/CN2018/112052

(22) 国际申请日: 2018 年 10 月 26 日 (26.10.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810689442.X 2018年6月28日 (28.06.2018) CN

(71) 申请人: 郑州云海信息技术有限公司 (**ZHENGZHOU YUNHAI INFORMATION TECHNOLOGY CO., LTD.**) [CN/CN]; 中国河南省郑州市郑东新区心怡路278号基运投资大厦16层, Henan 450018 (CN)。

(72) 发明人: 甄天桥 (**ZHEN, Tianqiao**); 中国河南省郑州市郑东新区心怡路278号基运投资大厦16层, Henan 450018 (CN)。

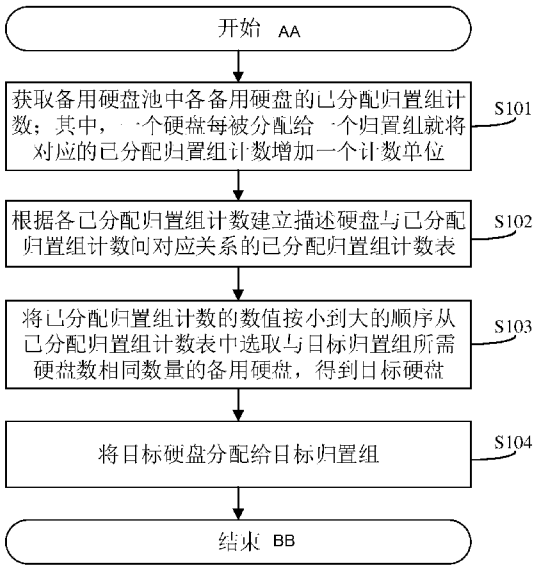
(74) 代理人: 北京集佳知识产权代理有限公司 (**UNITALEN ATTORNEYS AT LAW**); 中国北京市朝阳区建国门外大街22号赛特广场7层, Beijing 100004 (CN)。

(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,

WO 2020/000817 A1

(54) **Title:** METHOD, SYSTEM, AND APPARATUS FOR ALLOCATING HARD DISKS BELONGING TO PLACEMENT GROUP, AND STORAGE MEDIUM

(54) **发明名称:** 一种归置组所属硬盘分配方法、系统、装置及存储介质



S101 ACQUIRE THE COUNT OF STANDBY HARD DISKS ALREADY ALLOCATED TO A PLACEMENT GROUP IN A STANDBY HARD DISK POOL, THE CORRESPONDING ALLOCATED PLACEMENT GROUP COUNT BEING INCREASED BY ONE COUNT UNIT EACH TIME ONE HARD DISK IS ALLOCATED TO A PLACEMENT GROUP

S102 ON THE BASIS OF THE ALLOCATED PLACEMENT GROUP COUNT, ESTABLISH AN ALLOCATED PLACEMENT GROUP COUNT TABLE DESCRIBING THE CORRESPONDING RELATIONSHIP BETWEEN THE HARD DISKS AND THE ALLOCATED PLACEMENT GROUP COUNT

S103 SELECT THE SAME NUMBER OF STANDBY HARD DISKS AS THE NUMBER OF HARD DISKS REQUIRED BY A TARGET PLACEMENT GROUP FROM THE ALLOCATED PLACEMENT GROUP COUNT TABLE IN ASCENDING ORDER OF THE NUMERICAL VALUE OF THE ALLOCATED PLACEMENT GROUP COUNT, IN ORDER TO OBTAIN TARGET HARD DISKS

S104 ALLOCATE THE TARGET HARD DISKS TO THE TARGET PLACEMENT GROUP

AA START

BB FINISH

图 1

(57) **Abstract:** Disclosed in the present application is a method for allocating the hard disks belonging to a placement group, newly introducing the concept of an allocated placement group count, i.e. each time one hard disk is allocated to a placement group, the corresponding allocated placement group count is increased by one count unit; on this basis, establishing an allocated placement group count table reflecting the allocated placement group count of each of all of the standby hard disks, in order to select a certain number of target hard disks in ascending order of the numerical value of the count, the target hard disks having a lower allocated placement group count than other standby hard disks; as the number of placement groups supported by said hard disks is lower, said disks are more suitable to be component hard disks constituting a new placement group, making the distribution of the placement groups on the hard disks more uniform, and correspondingly also making the uniformity of data distribution higher. Also disclosed in the present

[见续页]

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, 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))。

application are a system and an apparatus for allocating the hard disks belonging to a placement group, and a computer readable storage medium, having the aforementioned beneficial effects.

(57) 摘要：本申请公开了一种归置组所属硬盘分配方法，新引入了已分配归置组计数这一概念，即每当一个硬盘被分配给一个归置组时就将其对应的已分配归置组计数增加一个计数单位，并据此建立起反映所有备选硬盘各自的已分配归置组计数的已分配归置组计数表，以将计数数值按从小到大的顺序选取出一定数量的目标硬盘，这些目标硬盘相对其他备用硬盘具有较小的已分配归置组计数，即说明这些硬盘上承载的归置组的数量较少，因为更合适作为组成新归置组的组成硬盘，由于归置组在硬盘上的分布更加均匀，相应的也使得数据分布的均匀程度更高。本申请还同时公开了一种归置组所属硬盘分配系统、装置及计算机可读存储介质，具有上述有益效果。

一种归置组所属硬盘分配方法、系统、装置及存储介质

本申请要求于2018年06月28日提交中国专利局、申请号201810689442.X、发明名称为“一种归属组所属硬盘分配方法、系统、装置及存储介质”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

本申请涉及数据均衡分布技术领域，特别涉及一种归置组所属硬盘分配方法、系统、装置及计算机可读存储介质。

背景技术

分布式存储系统由多个存储池组成，每个存储池都有其对应的数据分布规则，比如副本规则或者纠删规则。其中，副本规则是形成与原数据文件完全相同的副本数据，即以完整的数据文件进行备份存储；以4+2的纠删规则为例，表示将一份完整的数据分成互不相同的4份，再根据这4份数据，按固定的纠删算法计算出2份冗余数据出来，然后保存这6份数据，之后在任何时候只要能从6份数据中读取出任意4份数据来，就能恢复出原始的数据。

一般来说元数据由于其数据量较小、重要度高，通常保存在基于副本规则形成的副本存储池中，而完整的数据文件则保存在纠删池中。为了方便管理数据分布，一般存储池还会划分成多个归置组（Placement Group，PG），每个归置组按副本规则或者纠删规则包含若干块硬盘。比如4+2的纠删规则要保存6份数据，通常就需要6块硬盘来分别保存这6份数据，即每个归置组就会包含6块硬盘。归置组在硬盘上的分布情况直接决定了数据在集群中到底是不是均匀分布的。

当前为每个归置组分配所属硬盘的策略一般都是按某种随机算法来选择所需的硬盘的，但随机算法只能在一定程度上保证归置组在硬盘上分布的随机性，随着要求的提高，基于随机算法的硬盘分配方式对应的数据分布均匀度已经无法满足当前的要求，因此需要进一步提升数据分布的均匀程度。

因此，如何克服现有归置组所属硬盘分配方法存在的各项技术缺陷，提供一种所属硬盘分配更加科学、合理，使得数据分布均匀程度更高的归置组所属硬盘分配机制是本领域技术人员亟待解决的问题。

发明内容

本申请的目的是提供一种归置组所属硬盘分配方法，新引入了已分配归置组计数这一概念，即每当一个硬盘被分配给一个归置组时就将其对应的已分配归置组计数增加一个计数单位，并据此建立起反映所有备选硬盘各自的已分配归置组计数的已分配归置组计数表，以将计数数值按从小到大的顺序选取出一定数量的目标硬盘，这些目标硬盘相对其他备用硬盘具有较小的已分配归置组计数，即说明这些硬盘上承载的归置组的数量较少，因为更合适作为组成新归置组的组成硬盘，由于归置组在硬盘上的分布更加均匀，相应的也使得数据分布的均匀程度更高。

本申请的另一目的在于提供了一种归置组所属硬盘分配系统、装置及计算机可读存储介质。

为实现上述目的，本申请提供一种归置组所属硬盘分配方法，包括:

获取备用硬盘池中各备用硬盘的已分配归置组计数；其中，一个硬盘每被分配给一个归置组就将对应的已分配归置组计数增加一个计数单位;

根据各所述已分配归置组计数建立描述硬盘与已分配归置组计数间对应关系的已分配归置组计数表;

将所述已分配归置组计数的数值按小到大的顺序从所述已分配归置组计数表中选取出与目标归置组所需硬盘数相同数量的备用硬盘，得到目标硬盘;

将所述目标硬盘分配给所述目标归置组。

可选的，将所述已分配归置组计数的数值按小到大的顺序从所述已分配归置组计数表中选取出与目标归置组所需硬盘数相同数量的备用硬盘，得到目标硬盘，包括:

按照从小到大的顺序每次从所述已分配归置组计数表中选取出连续的两块备用硬盘，直至所述目标硬盘的数量与所述所需硬盘数相同;

判断连续的两块备用硬盘是否处于相同的存储设备上;

若否，则将处于不同存储设备上两块备用硬盘均作为所述目标硬盘;

若是，则选择处于相同存储设备上的两块备用硬盘中的任一块作为所述目标硬盘。

可选的，该归置组所属硬盘分配方法还包括:

判断每个硬盘是否满足预设的备用硬盘池加入规则；其中，所述备用硬盘池加入规则至少包括：上一次增加所述已分配归置组计数距当前时间的时长大于预设时长、所述已分配归置组计数小于所述计数上限、各硬盘的当前存储空间占用率中小于占用率阈值的至少一项；

若是，则将对应的硬盘作为所述备用硬盘加入所述备用硬盘池。

可选的，该归置组所属硬盘分配方法还包括：

当有新硬盘添加进存储系统时，确定每个所述归置组中拥有最大已分配归置组计数的硬盘，得到每个所述归置组中的待替换硬盘；

利用所述新硬盘替换每个所述归置组中的待替换硬盘。

为实现上述目的，本申请还提供了一种归置组所属硬盘分配系统，包括：

备用硬盘已分配归置组计数获取单元，用于获取备用硬盘池中各备用硬盘的已分配归置组计数；其中，一个硬盘每被分配给一个归置组就将对应的已分配归置组计数增加一个计数单位；

已分配归置组计数表建立单元，用于根据各所述已分配归置组计数建立描述硬盘与已分配归置组计数间对应关系的已分配归置组计数表；

较小归置组计数硬盘选取单元，用于将所述已分配归置组计数的数值按小到大的顺序从所述已分配归置组计数表中选取出与目标归置组所需硬盘数相同数量的备用硬盘，得到目标硬盘；

目标硬盘分配单元，用于将所述目标硬盘分配给所述目标归置组。

可选的，所述较小归置组计数硬盘选取单元包括：

连续两块备用硬盘选取子单元，用于按照从小到大的顺序每次从所述已分配归置组计数表中选取出连续的两块备用硬盘，直至所述目标硬盘的数量与所述所需硬盘数相同；

相同存储设备判断子单元，用于判断连续的两块备用硬盘是否处于相同的存储设备上；

非相同存储设备处理子单元，用于当连续的两块备用硬盘分别处于不同的存储设备上时，将处于不同存储设备上两块备用硬盘均作为所述目标硬盘；

相同存储设备处理子单元，用于当连续的两块备用硬盘处于相同的存储设备上时，选择处于相同存储设备上的两块备用硬盘中的任一块作为所述目标硬盘。

可选的，该归置组所属硬盘分配系统还包括:

备用硬盘判断单元，用于判断每个硬盘是否满足预设的备用硬盘池加入规则；其中，所述备用硬盘池加入规则至少包括：上一次增加所述已分配归置组计数距当前时间的时长大于预设时长、所述已分配归置组计数小于所述计数上限、各硬盘的当前存储空间占用率中小于占用率阈值的至少一项；

备用硬盘确定单元，用于当硬盘满足所述备用硬盘池加入规则时，将对应的硬盘作为所述备用硬盘加入所述备用硬盘池。

可选的，该归置组所属硬盘分配系统还包括:

待替换硬盘确定单元，用于当有新硬盘添加进存储系统时，确定每个所述归置组中拥有最大已分配归置组计数的硬盘，得到每个所述归置组中的待替换硬盘；

替换单元，用于利用所述新硬盘替换每个所述归置组中的待替换硬盘。

为实现上述目的，本申请还提供了一种归置组所属硬盘分配装置，包括:

存储器，用于存储计算机程序；

处理器，用于执行所述计算机程序时实现如上述内容所描述的归置组所属硬盘分配方法的步骤。

为实现上述目的，本申请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上存储有计算机程序，所述计算机程序被处理器执行时实现如上述内容所描述的归置组所属硬盘分配方法的步骤。

本申请所提供的归置组所属硬盘分配方法：获取备用硬盘池中各备用硬盘的已分配归置组计数；其中，一个硬盘每被分配给一个归置组就将对应的已分配归置组计数增加一个计数单位；根据各所述已分配归置组计数建立描述硬盘与已分配归置组计数间对应关系的已分配归置组计数表；将所述已分配归置组计数的数值按小到大的顺序从所述已分配归置组计数表中选取出与目标归置组所需硬盘数相同数量的备用硬盘，得到目标硬盘；将所述目标硬盘分配给所述目标归置组。

显然，该方法新引入了已分配归置组计数这一概念，即每当一个硬盘被分配给一个归置组时就将其对应的已分配归置组计数增加一个计数单位，并据此建立起反映所有备选硬盘各自的已分配归置组计数的已分配归置组计数表，以将计数数值按从小到大的顺序选取出一定数量的目标硬盘，这些目标硬盘相对其他备用硬盘具有较小的已分配归置组计数，即说明这些硬盘上承载的归置组的数量较少，因为更合适作为组成新归置组的组成硬盘，由于归置组在硬盘上的分布更加均匀，相应的也使得数据分布的均匀程度更高。本申请同时还提供了一种归置组所属硬盘分配系统、装置及计算机可读存储介质，具有上述有益效果，在此不再赘述。

附图说明

图1为本申请实施例提供的一种归置组所属硬盘分配方法的流程图；

图2为本申请实施例提供的归置组所属硬盘分配方法中一种较小已分配归置组计数的备用硬盘选取方法的流程图；

图3为本申请实施例提供的归置组所属硬盘分配方法中一种判断硬盘是否满足加入备用硬盘池作为备用硬盘的方法的流程图；

图4为本申请实施例所提供的一种归置组所属硬盘分配系统的结构框图。

具体实施方式

本申请的核心是提供一种归置组所属硬盘分配方法，新引入了已分配归置组计数这一概念，即每当一个硬盘被分配给一个归置组时就将其对应的已分配归置组计数增加一个计数单位，并据此建立起反映所有备选硬盘各自的已分配归置组计数的已分配归置组计数表，以将计数数值按从小到大的顺序选取出一定数量的目标硬盘，这些目标硬盘相对其他备用硬盘具有较小的已分配归置组计数，即说明这些硬盘上承载的归置组的数量较少，因为更合适作为组成新归置组的组成硬盘，由于归置组在硬盘上的分布更加均匀，相应的也使得数据分布的均匀程度更高。本申请还同时提供了一种归置组所属硬盘分配系统、装置及计算机可读存储介质，具有上述有益效果。

为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请实施例中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申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例，都属于本申请保护的范围。

实施例一

以下结合图1，图1为本申请实施例提供的一种归置组所属硬盘分配方法的流程图，其具体包括以下步骤:

S101：获取备用硬盘池中各备用硬盘的已分配归置组计数；其中，一个硬盘每被分配给一个归置组就将对应的已分配归置组计数增加一个计数单位；

本申请新引入了“已分配归置组计数”的概念，顾名思义，这个概念指的是一个硬盘被分配给不同归置组的个数，即一个硬盘每被分配给一个归置组就将该硬盘的已分配归置组计数增加一个计数单位。

本步骤旨在获取处于备用硬盘池中每个备用硬盘各自的已分配归置组计数，其中，该备用硬盘池是所有备用硬盘的集合，备用硬盘也可以理解为处于待分配状态的硬盘，可以被分配给新创立的归置组做数据存储用。

进一步的，可以设定一个备用硬盘判别机制，对每个硬盘是否可以作为备用硬盘加入备用硬盘池已进行后续的待分配处理操作，具体方式有很多，可以基于不同的原则来实现，例如为了防止短时间内频繁将一个硬盘分配给多个归置组，可以设定将其分配给不同的归置组的时间间隔，即必须从上一次将其分配给一个归置组经过该时间间隔后才能将其再分配给另一个归置组；也可以设定一定时间段内最大分配给不同归置组的个数；还可以根据一个硬盘的当前状态是否处于异常状态等等，此处并不做具体限定，设定备用硬盘判别机制的主要目的还是要使数据分布均匀程度更高，在此目的的指导下可以拥有不同的解决方案，可根据实际情况灵活选择，在此不再一一赘述。

S102：根据各已分配归置组计数建立描述硬盘与已分配归置组计数间对应关系的已分配归置组计数表；

在S101的基础上，本步骤旨在建立描述各硬盘与各已分配归置组计数间对应关系的已分配归置组计数表，即每条对应关系的一端是硬盘的识别标识，另一端是该识别标识对应硬盘的已分配归置组计数，具体的，已分配归置组

计数的表现形式多种多样，可以直接采用数字编号的形式出现，也可以按照二进制代码的方式存在等等，目的在于可以根据该已分配归置组计数判别出相应硬盘已经被分配给了多少个归置组。

进一步的，在本步骤提供的对应关系的基础上，还可以再链接每个硬盘的属性信息，可以包括具体所属哪台存储设备、生产日期、加入集群的时间、日志等等信息，以便能够根据额外链接的信息判断出一个硬盘的其它状态信息，并据此做出其它判断。

S103：将已分配归置组计数的数值按小到大的顺序从已分配归置组计数表中选取出与目标归置组所需硬盘数相同数量的备用硬盘，得到目标硬盘；

在S102的基础上，本步骤旨在从该已分配归置组计数表中选取出预设数量的备用硬盘，并将其作为待分配给目标归置组的目标硬盘使用。其中，选取的标准遵循为选取拥有较小的已分配归置组计数的备用硬盘，因为要为目标归置组分配其所属的硬盘，且尽可能的使数据分布均匀程度更高，因此有必要选取一些拥有较小已分配归置组计数的备用硬盘，而不是随机选取或选取一些拥有较大已分配归置组计数的备用硬盘作为目标硬盘。

此处所说的拥有较小已分配归置组计数的备用硬盘可通过下面这个例子来理解：

假定该已分配归置组计数表中包含20个备用硬盘的已分配归置组计数的数值，而目标归置组所需硬盘数为5，也就是说需要分配给目标归置组5个备选硬盘构成这个归置组，此时则按照各已分配归置组计数的数值从小到大的顺序选取5个出来，之所以不说是选取5个最小的已分配归置组计数对应的备用硬盘，是因为通常“最”一词仅表示其中一个，即在选出的5个中也仅存在一个最小的，而按照从小到大选取出的5个则是表中20个中较小，选不出另外5个比他们拥有更小已分配归置组计数的备用硬盘了。

当然，具体如何选取出与目标归置组所需硬盘数相同数量的备用硬盘的方式多种多样，上面给出的例子仅作为其中一次性选出所需硬盘的方式存在，还可以每次选择一个拥有当前最小已分配归置组计数的备用硬盘，在下次选择时出于选择不同硬盘的基础要求，就会再次找到一个拥有最小已分配归置组计数的备用硬盘，依次类推直至选取得到与所需硬盘数相同数量的备用硬

盘作为目标硬盘。进一步的，在每次为目标归置组选取一块合适的目标硬盘时，可以按照一定的选取算法随机选取一定数量的备用硬盘，而非每次都在包含全部备用硬盘的已分配归置组计数信息的表中选取最小的那个，比如可以每次从表中挑出3个备用硬盘，并从这3个中选取最小的，这样虽然会牺牲一定的数据分布均匀程度，但会节省选取时间，是否按照此种方式进行选择可根据实际情况对数据分布均匀程度的具体要求和对耗时的具体要求灵活选择，此处并不做具体限定。

更进一步的，实际上上线的存储设备一般都不是以硬盘为单位，而是包含多个硬盘的存储设备，例如磁盘阵列、存储服务器等等，对于一个磁盘整列或存储服务器都会包含数量众多的硬盘，在组成一个归置组的所有硬盘都是同一存储设备上不同硬盘的基础上，可能会因该存储设备出现故障时，因单节点故障造成该归置组整体不可用的现象，因此在还可以根据S102中提及的硬盘属性信息判断选取分属不同存储设备上的备用硬盘，但总体原则还是要选取拥有较小已分配归置组计数的硬盘。在此指导思想下本领域技术人员可得到多种实现方式，在此不再赘述。

S104：将目标硬盘分配给目标归置组。

在S103的基础上，本步骤旨在确定好的目标硬盘分配给目标归置组，根据S103中阐述的选取目标硬盘的不同方式，可具体包括一次性将所有目标硬盘都分配目标归置组的方式，和每次将一块目标硬盘分配给目标归置组，直至数量满足该目标归置组所需。当然，在采用后一种分配方式时，目标归置组在未得到与所需硬盘数相同数量的目标硬盘之前，还无法正式承接数据的存储任务。

进一步的，存储系统每添加进一块新硬盘（已分配归置组计数为0）时，为了使集群内数据分布的均匀程度更高，还可以使用该新硬盘替换每个归置组中拥有最大已分配归置组计数的硬盘，直至已将该新硬盘分配了与之前已有硬盘数量一致的归置组。

举例如下：当前存储系统内有两个归置组，第一归置组由6个硬盘组成，其中拥有最大已分配归置组计数的硬盘为003号硬盘，第二归置组由5个硬盘组成，其中拥有最大已分配归置组计数的硬盘为006号硬盘，因此一种采用新

硬盘替换每个归置组中拥有最大已分配归置组计数的硬盘的方式为：使用新硬盘（020号）替换003号硬盘被分配给第一归置组，003号硬盘被解除与第一归置组的分配关系；使用新硬盘（020号）替换006号硬盘被分配给第二归置组，006号硬盘被解除与第二归置组的分配关系。

当然，当归置组数量众多时，还会出现替换下来的原有硬盘会解除很多与不同归置组的分配关系，致使在一增一减的情况下出现替换下来的原硬盘的已分配归置组计数小于新硬盘的已分配归置组计数，出于使数据分布尽可能均匀的原则，可以再出现此种情况下严格按照已分配归置组计数的数值大小来判断要使用哪块硬盘替换哪块硬盘，当然，由于新硬盘的数据稳定存储稳定性更高，还可以出于此种考虑适当使新硬盘被分配给更多的归置组。具体实际情况下还有很多其它需要考虑的实际因素，可根据实际情景灵活选择合适的方式，此处并不做具体限定。

基于上述技术方案，本申请实施例提供的一种归置组所属硬盘分配方法，新引入了已分配归置组计数这一概念，即每当一个硬盘被分配给一个归置组时就将其对应的已分配归置组计数增加一个计数单位，并据此建立起反映所有备选硬盘各自的已分配归置组计数的已分配归置组计数表，以将计数数值按从小到大的顺序选取出一定数量的目标硬盘，这些目标硬盘相对其他备用硬盘具有较小的已分配归置组计数，即说明这些硬盘上承载的归置组的数量较少，因为更合适作为组成新归置组的组成硬盘，由于归置组在硬盘上的分布更加均匀，相应的也使得数据分布的均匀程度更高。

实施例二

以下结合图2，图2为本申请实施例提供的归置组所属硬盘分配方法中一种较小已分配归置组计数的备用硬盘选取方法的流程图，本实施例旨在考虑实际情况下可能存在的各目标硬盘都处于相同存储设备上致使单节点故障出现时导致严重故障的问题：

S201：按照从小到大的顺序每次从已分配归置组计数表中选取出连续的两块备用硬盘，直至目标硬盘的数量与所需硬盘数相同；

S202：判断连续的两块备用硬盘是否处于相同的存储设备上；

S203：将处于不同存储设备上两块备用硬盘均作为目标硬盘；

S204：选择处于相同存储设备上的两块备用硬盘中的任一块作为目标硬盘。

本实施例每次按照从小到大的顺序每次从已分配归置组计数表中选取出连续的两块备用硬盘，即通过两两比对的方式判断这两块备用硬盘是否处于相同存储上，若分别处于不同的存储设备上，则这两块备用硬盘都可以被作为目标硬盘使用，即目标硬盘的计数加2；若两块备用硬盘为同一存储设备上的两块不同硬盘，则从中任选其一作为目标硬盘使用，即目标硬盘的计数加1，直至目标硬盘的计数与所需硬盘数相等。

当然，本实施例仅以连续的两块硬盘做判断，还会出现第一次选取的两块硬盘与第二次选取的两块硬盘依然处于同一存储设备的情况，即本实施例只能够在一定程度上防止单节点故障的出现，但基于此指导思想只需要再进行一次判断即可，本领域技术人员可据此轻易得到一个效果更佳的实现方式，以尽可能在提高数据分布均匀程度的基础上避免单节点故障导致的严重问题出现，在此不一一赘述。

实施例三

以下结合图3，图3为本申请实施例提供的归置组所属硬盘分配方法中一种判断硬盘是否满足加入备用硬盘池作为备用硬盘的方法的流程图。

S301：判断每个硬盘是否满足预设的备用硬盘池加入规则；

其中，该备用硬盘池加入规则至少包括：上一次增加已分配归置组计数距当前时间的时长大于预设时长、已分配归置组计数小于计数上限、各硬盘的当前存储空间占用率中小于占用率阈值中的至少一项。

当然也不仅仅包括这些参数，可根据实际情况自行加入其它种类的判别规则，且各规则可单独使用得出判断结论，也可以灵活的进行多项组合得到更加负责、更符合实际情况的判断方式，目的都在于尽可能的提高数据分布的均匀程度。

S302：不能够作为备用硬盘加入备用硬盘池；

本步骤建立在S301的判断结果为一个硬盘不满足备用硬盘池加入规则的

基础上，以上一次增加已分配归置组计数距当前时间的时长大于预设时长为例，则经过判断发现将该硬盘上一次分配给一个归置组的时间距今的时长不大于该预设时长，因此判定该硬盘不适合作为备用硬盘加入备用硬盘池。

S303：将对应的硬盘作为备用硬盘加入备用硬盘池；

本步骤建立在S301的判断结果为一个硬盘满足备用硬盘池加入规则的基础上，依然以上一次增加已分配归置组计数距当前时间的时长大于预设时长为例，则经过判断发现将该硬盘上一次分配给一个归置组的时间距今的时长大于该预设时长，因此判定该硬盘适合作为备用硬盘加入备用硬盘池。

S304：获取备用硬盘池中各备用硬盘的已分配归置组计数；其中，一个硬盘每被分配给一个归置组就将对应的已分配归置组计数增加一个计数单位。

基于上述任一实施例，本实施例通过为备用硬盘设置加入规则，能够先判断一次每个硬盘是否适合被分配给一个归置组，能够更好地提高数据分布的均匀程度和数据的稳定性。

实施例四

本实施就一个具体应用场景提供一种适用于该场景下的实现方案：

PG 分布均衡性的优化方法的具体实施过程如下：

（1）为 PG 分配每块硬盘时，都按随机性算法一次选择出若干块盘（比如选出 3 块盘，个数适当即可，不需太多）；

（2）从选出的若干块盘中选择分布 PG 个数最少的一个硬盘作为最终选择结果，并将该盘的 PG 分布个数加 1；

（3）循环上述选择，直到为该 PG 分配完所需个数的硬盘，且保证每次选择出的目标硬盘与之前选择出的所有目标硬盘均属于不同的存储设备；

（4）当有硬盘故障退出集群时，以上述方法为该盘上分布的 PG 重新分配一块硬盘即可；

（5）当有新盘加入时（比如扩容，或者某个节点上插入新盘），则循环遍历每个 PG，每次选择该 PG 的所有硬盘中 PG 分布最多的一块盘来替换成新加入的硬盘中的某一块，这样每个 PG 每次替换一块硬盘并逐步替换，直

到新加入的硬盘上分布的PG和已有硬盘上分布的PG个数相当为止。

因为情况复杂，无法一一列举进行阐述，本领域技术人员应能意识到根据本申请提供的基本方法原理结合实际情况可以存在很多的例子，在不付出足够的创造性劳动下，应均在本申请的保护范围内。

下面请参见图4，图4为本申请实施例所提供的一种归置组所属硬盘分配系统的结构框图，该归置组所属硬盘分配系统可以包括:

备用硬盘已分配归置组计数获取单元 100，用于获取备用硬盘池中各备用硬盘的已分配归置组计数；其中，一个硬盘每被分配给一个归置组就将对应的已分配归置组计数增加一个计数单位;

已分配归置组计数表建立单元 200，用于根据各已分配归置组计数建立描述硬盘与已分配归置组计数间对应关系的已分配归置组计数表;

较小归置组计数硬盘选取单元 300，用于按照从小到大的顺序从已分配归置组计数表中选取与目标归置组所需硬盘数相同数量的备用硬盘，得到目标硬盘;

目标硬盘分配单元400，用于将目标硬盘分配给目标归置组。

其中，较小归置组计数硬盘选取单元300可以包括:

连续两块备用硬盘选取子单元，用于按照从小到大的顺序每次从已分配归置组计数表中选取出连续的两块备用硬盘，直至目标硬盘的数量与所需硬盘数相同;

相同存储设备判断子单元，用于判断连续的两块备用硬盘是否处于相同的存储设备上;

非相同存储设备处理子单元，用于当连续的两块备用硬盘分别处于不同的存储设备上时，将处于不同存储设备上两块备用硬盘均作为目标硬盘;

相同存储设备处理子单元，用于当连续的两块备用硬盘处于相同的存储设备上时，选择处于相同存储设备上的两块备用硬盘中的任一块作为目标硬盘。

进一步的，该归置组所属硬盘分配系统还可以包括:

备用硬盘判断单元，用于判断每个硬盘是否满足预设的备用硬盘池加入规则；其中，备用硬盘池加入规则至少包括：上一次增加已分配归置组计数

距当前时间的时长大于预设时长、已分配归置组计数小于计数上限、各硬盘的当前存储空间占用率中小于占用率阈值的至少一项;

备用硬盘确定单元，用于当硬盘满足备用硬盘池加入规则时，将对应的硬盘作为备用硬盘加入备用硬盘池。

进一步的，该归置组所属硬盘分配系统还可以包括:

待替换硬盘确定单元，用于当有新硬盘添加进存储系统时，确定每个归置组中拥有最大已分配归置组计数的硬盘，得到每个归置组中的待替换硬盘;

替换单元，用于利用新硬盘替换每个归置组中的待替换硬盘。

基于上述实施例，本申请还提供了一种归置组所属硬盘分配装置，该装置可以包括存储器和处理器，其中，该存储器中存有计算机程序，该处理器调用该存储器中的计算机程序时，可以实现上述实施例所提供的步骤。当然，该装置还可以包括各种必要的网络接口、电源以及其它零部件等。

本申请还提供了一种计算机可读存储介质，其上存有计算机程序，该计算机程序被执行终端或处理器执行时可以实现上述实施例所提供的步骤。该存储介质可以包括: U盘、移动硬盘、只读存储器(Read-Only Memory ，ROM)、随机存取存储器（Random Access Memory ， RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。对于实施例公开的装置而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。

专业人员还可以进一步意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想。对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以对本申请进行若干改进和修饰，这些改进和修饰也落入本申请权利要求的保护范围内。

还需要说明的是，在本说明书中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其它变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其它要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括要素的过程、方法、物品或者设备中还存在另外的相同要素。

权 利 要 求

1、一种归置组所属硬盘分配方法，包括:

获取备用硬盘池中各备用硬盘的已分配归置组计数；其中，一个硬盘每被分配给一个归置组就将对应的已分配归置组计数增加一个计数单位;

根据各所述已分配归置组计数建立描述硬盘与已分配归置组计数间对应关系的已分配归置组计数表;

将所述已分配归置组计数的数值按小到大的顺序从所述已分配归置组计数表中选取出与目标归置组所需硬盘数相同数量的备用硬盘，得到目标硬盘;

将所述目标硬盘分配给所述目标归置组。

2、根据权利要求 1 所述方法，将所述已分配归置组计数的数值按小到大的顺序从所述已分配归置组计数表中选取出与目标归置组所需硬盘数相同数量的备用硬盘，得到目标硬盘，包括:

按照从小到大的顺序每次从所述已分配归置组计数表中选取出连续的两块备用硬盘，直至所述目标硬盘的数量与所述所需硬盘数相同;

判断连续的两块备用硬盘是否处于相同的存储设备上;

若否，则将处于不同存储设备上两块备用硬盘均作为所述目标硬盘;

若是，则选择处于相同存储设备上的两块备用硬盘中的任一块作为所述目标硬盘。

3、根据权利要求 1 或 2 所述方法，所述方法还包括:

判断每个硬盘是否满足预设的备用硬盘池加入规则；其中，所述备用硬盘池加入规则至少包括：上一次增加所述已分配归置组计数距当前时间的时长大于预设时长、所述已分配归置组计数小于所述计数上限、各硬盘的当前存储空间占用率中小于占用率阈值的至少一项;

若是，则将对应的硬盘作为所述备用硬盘加入所述备用硬盘池。

4、根据权利要求 3 所述方法，所述方法还包括:

当有新硬盘添加进存储系统时，确定每个所述归置组中拥有最大已分配归置组计数的硬盘，得到每个所述归置组中的待替换硬盘;

利用所述新硬盘替换每个所述归置组中的待替换硬盘。

5、一种归置组所属硬盘分配系统，包括:

备用硬盘已分配归置组计数获取单元，用于获取备用硬盘池中各备用硬盘的已分配归置组计数；其中，一个硬盘每被分配给一个归置组就将对应的已分配归置组计数增加一个计数单位；

已分配归置组计数表建立单元，用于根据各所述已分配归置组计数建立描述硬盘与已分配归置组计数间对应关系的已分配归置组计数表；

较小归置组计数硬盘选取单元，用于将所述已分配归置组计数的数值按小到大的顺序从所述已分配归置组计数表中选取出与目标归置组所需硬盘数相同数量的备用硬盘，得到目标硬盘；

目标硬盘分配单元，用于将所述目标硬盘分配给所述目标归置组。

6、根据权利要求5所述系统，所述较小归置组计数硬盘选取单元包括：

连续两块备用硬盘选取子单元，用于按照从小到大的顺序每次从所述已分配归置组计数表中选取出连续的两块备用硬盘，直至所述目标硬盘的数量与所述所需硬盘数相同；

相同存储设备判断子单元，用于判断连续的两块备用硬盘是否处于相同的存储设备上；

非相同存储设备处理子单元，用于当连续的两块备用硬盘分别处于不同的存储设备上时，将处于不同存储设备上两块备用硬盘均作为所述目标硬盘；

相同存储设备处理子单元，用于当连续的两块备用硬盘处于相同的存储设备上时，选择处于相同存储设备上的两块备用硬盘中的任一块作为所述目标硬盘。

7、根据权利要求5或6所述系统，所述系统还包括：

备用硬盘判断单元，用于判断每个硬盘是否满足预设的备用硬盘池加入规则；其中，所述备用硬盘池加入规则至少包括：上一次增加所述已分配归置组计数距当前时间的时长大于预设时长、所述已分配归置组计数小于所述计数上限、各硬盘的当前存储空间占用率中小于占用率阈值的至少一项；

备用硬盘确定单元，用于当硬盘满足所述备用硬盘池加入规则时，将对应的硬盘作为所述备用硬盘加入所述备用硬盘池。

8、根据权利要求7所述系统，所述系统还包括：

待替换硬盘确定单元，用于当有新硬盘添加进存储系统时，确定每个所

述归置组中拥有最大已分配归置组计数的硬盘，得到每个所述归置组中的待替换硬盘；

替换单元，用于利用所述新硬盘替换每个所述归置组中的待替换硬盘。

9、一种归置组所属硬盘分配装置，包括：

存储器，用于存储计算机程序；

处理器，用于执行所述计算机程序时实现如权利要求 1 至 4 任一项所述的归置组所属硬盘分配方法的步骤。

10、一种计算机可读存储介质，所述计算机可读存储介质上存储有计算机程序，所述计算机程序被处理器执行时实现如权利要求 1 至 4 任一项所述的归置组所属硬盘分配方法的步骤。

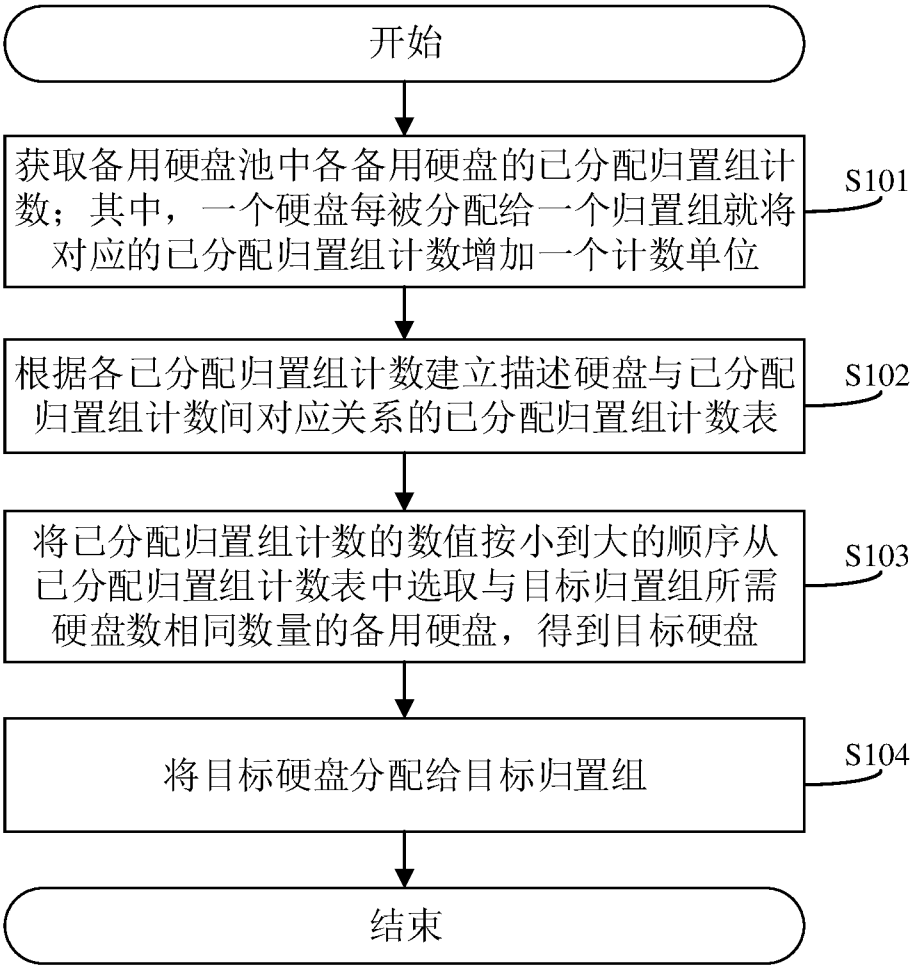

开始
获取备用硬盘池中各备用硬盘的已分配归置组计数；其中，一个硬盘每被分配给一个归置组就将对应的已分配归置组计数增加一个计数单位
S101
根据各已分配归置组计数建立描述硬盘与已分配归置组计数间对应关系的已分配归置组计数表
S102
将已分配归置组计数的数值按小到大的顺序从已分配归置组计数表中选取与目标归置组所需硬盘数相同数量的备用硬盘，得到目标硬盘
S103
将目标硬盘分配给目标归置组
S104
结束

图 1

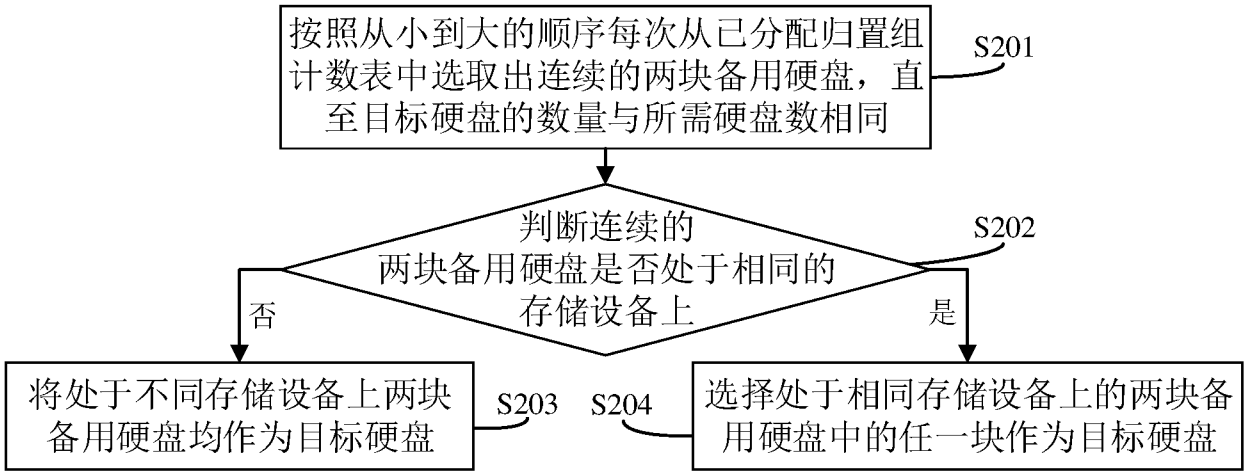

按照从小到大的顺序每次从已分配归置组计数表中选取出连续的两块备用硬盘，直至目标硬盘的数量与所需硬盘数相同
S201
判断连续的两块备用硬盘是否处于相同的存储设备上
S202
否
是
将处于不同存储设备上两块备用硬盘均作为目标硬盘
S203
S204
选择处于相同存储设备上的两块备用硬盘中的任一块作为目标硬盘

图 2

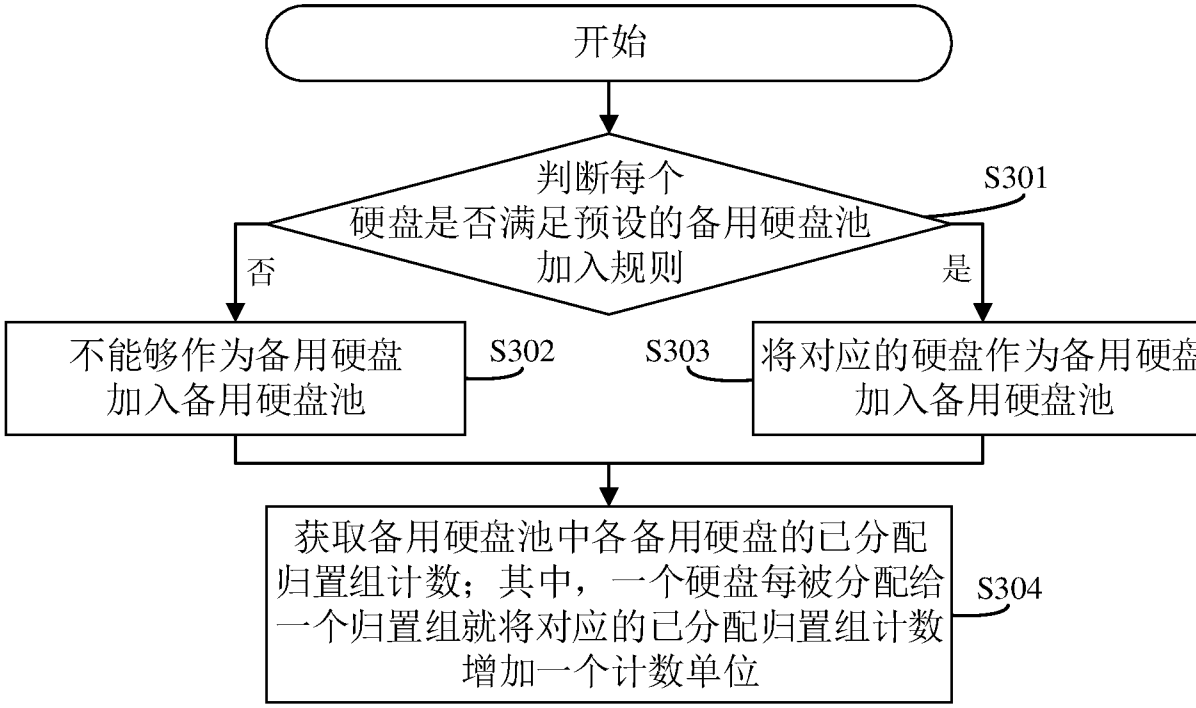

开始
判断每个硬盘是否满足预设的备用硬盘池加入规则
S301
否
是
不能够作为备用硬盘加入备用硬盘池
S302
S303
将对应的硬盘作为备用硬盘加入备用硬盘池
获取备用硬盘池中各备用硬盘的已分配归置组计数；其中，一个硬盘每被分配给一个归置组就将对应的已分配归置组计数增加一个计数单位
S304

图 3

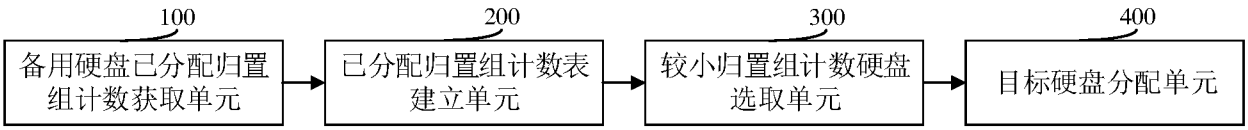

100
备用硬盘已分配归置组计数获取单元
200
已分配归置组计数表建立单元
300
较小归置组计数硬盘选取单元
400
目标硬盘分配单元

图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/112052

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, CNKI, DWPI, SIPOABS: 归置组, 放置组, 硬盘, 存储器, 分配, 分派, 计数, 表, 顺序, 次序, 排序, 映射, 均衡, 均匀, 数据, 选择, 选取, placement group, PG, hard disk, memory, distribute, allocate, allot, assign, count, table, list, sequence, queue, order, mapping, balance, data, select, choose

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 106055277 A (CHONGQING UNIVERSITY) 26 October 2016 (2016-10-26) entire document	1-10
A	CN 107977319 A (SK HYNIX INC.) 01 May 2018 (2018-05-01) entire document	1-10
A	US 2011153917 A1 (HITACHI, LTD.) 23 June 2011 (2011-06-23) entire document	1-10
A	US 2014297982 A1 (ARRIS GROUP, INC.) 02 October 2014 (2014-10-02) entire document	1-10

☐ 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
06 March 2019	**19 March 2019**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/112052

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	106055277	A	26 October 2016	CN	109196459	A	11 January 2019
				WO	2017206649	A1	07 December 2017
CN	107977319	A	01 May 2018	US	2018113620	A1	26 April 2018
				KR	2018045091	A	04 May 2018
US	2011153917	A1	23 June 2011	JP	2011128895	A	30 June 2011
				JP	4912456	B2	11 April 2012
US	2014297982	A1	02 October 2014	US	9483191	B2	01 November 2016

国际检索报告

国际申请号
PCT/CN2018/112052

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,CNKI,DWPI,SIPOABS:归置组，放置组，硬盘，存储器，分配，分派，计数，表，顺序，次序，排序，映射，均衡，均匀，数据，选择，选取，placement group, PG, hard disk, memory, distribute, allocate, allot, assign, count, table, list, sequence, queue, order, mapping, balance, data, select, choose

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 106055277 A (重庆大学) 2016年 10月 26日 (2016-10-26) 全文	1-10
A	CN 107977319 A (爱思开海力士有限公司) 2018年 5月 1日 (2018-05-01) 全文	1-10
A	US 2011153917 A1 (HITACHI, LTD.) 2011年 6月 23日 (2011-06-23) 全文	1-10
A	US 2014297982 A1 (ARRIS GROUP, INC.) 2014年 10月 2日 (2014-10-02) 全文	1-10

☐ 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
2019年 3月 6日	2019年 3月 19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 邓隽 电话号码 86-(10)-62411644

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/112052

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	106055277	A	2016年 10月 26日	CN	109196459	A	2019年 1月 11日
				WO	2017206649	A1	2017年 12月 7日
CN	107977319	A	2018年 5月 1日	US	2018113620	A1	2018年 4月 26日
				KR	2018045091	A	2018年 5月 4日
US	2011153917	A1	2011年 6月 23日	JP	2011128895	A	2011年 6月 30日
				JP	4912456	B2	2012年 4月 11日
US	2014297982	A1	2014年 10月 2日	US	9483191	B2	2016年 11月 1日