

(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
2012 年 9 月 20 日 (20.09.2012)



(10) 国际公布号
WO 2012/122915 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2012/072107
- (22) 国际申请日: 2012 年 3 月 9 日 (09.03.2012)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201110060440.2 2011 年 3 月 14 日 (14.03.2011) CN
- (71) 申请人 (对除美国外的所有指定国): **成都市华为赛门铁克科技有限公司 (CHENGDU HUAWEI SYMANTEC TECHNOLOGIES CO.,LTD.)** [CN/CN]; 中国四川省成都市高新西区清水河片区电子科大高新科技园天辰路 88 号, Sichuan 611731 (CN)。
- (72) 发明人; 及
- (75) 发明人/申请人 (仅对美国): **魏明昌 (WEI, Mingchang)** [CN/CN]; 中国四川省成都市高新西区清水河片区电子科大高新科技园天辰路 88 号, Sichuan 611731 (CN)。
张巍 (ZHANG, Wei) [CN/CN]; 中国四川省成都市高新西区清水河片区电子科大高新科技园天辰路 88 号, Sichuan 611731 (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, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。
- (84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, 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, 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))。

(54) Title: METHOD AND APPARATUS FOR READING AND WRITING DATA IN FILE SYSTEM

(54) 发明名称: 一种文件系统内读写数据的方法及装置

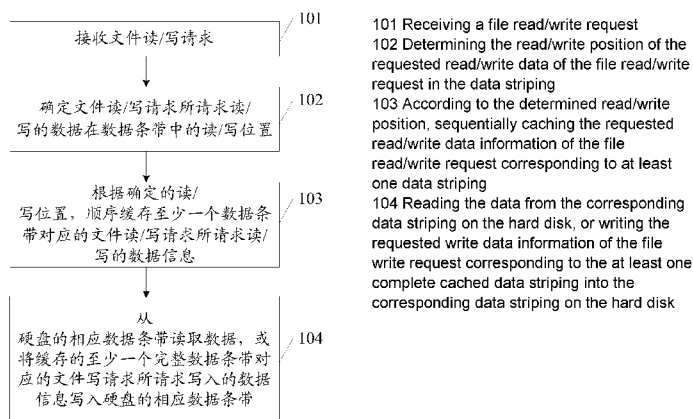


图 1 / Fig. 1

(57) Abstract: The present invention is applicable to the technical field of communication. Disclosed in an embodiment of the present invention are a method and apparatus for reading and writing data in a file system. In the embodiment of the present invention: the file system supports file-level independent disk redundancy algorithm operation, determines the read/write position of the data requesting to be read/written in a data striping after receiving a file read/write request, and sequentially caches data information requesting to be read/written corresponding to at least one data striping, and correspondingly processes the data information of the at least one cached data striping. Thus, after caching the data information requesting to be read/written corresponding to a complete data striping and after reading/writing the data using data striping as unit, there is no need to first write the file read/write request into a LUN virtual block device after every instance of receiving the file read/write request, and then accessing the hard disk by a RAID module, thus increasing data read/write efficiency.

[见续页]

WO 2012/122915 A1



(57) 摘要:

本发明实施例公开了文件系统内读写数据的方法及装置,应用于通信技术领域。本发明实施例中:文件系统支持文件级独立磁盘冗余算法操作,在接收到文件读/写请求后,确定请求读/写的的数据在数据条带中的读/写位置,并顺序缓存至少一个数据条带对应的请求读/写的的数据信息,并对缓存的至少一个数据条带的的数据信息进行相对应处理。这样在缓存完整一个数据条带对应的请求读/写的的数据信息后,并以数据条带为单位进行数据读/写,可以不需要每次在接收到文件读/写请求后,都先将文件读/写请求写入到 LUN 虚拟块设备中,再由 RAID 模块访问硬盘,从而提高了数据读/写效率。

一种文件系统内读写数据的方法及装置

本申请要求于 2011 年 3 月 14 日提交中国专利局、申请号为 201110060440.2、发明名称为“一种文件系统内读写数据的方法及装置”的中国
5 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

技术领域

本发明涉及通信技术领域，特别涉及一种文件系统内读写数据的方法及装置。

10 背景技术

文件系统是用来管理和存储文件信息的文件管理系统，文件系统可以通过逻辑单元（LUN）的虚拟块设备，向独立磁盘冗余阵列（Redundant Array of Independent Disk, RAID）模块读写数据。其中RAID模块中可以包括一个或多个数据存储块（记为RAIDn），每个数据存储块包括多个硬盘，数据存储在这些硬盘中，并带有附加的冗余信息。
15 些硬盘中，并带有附加的冗余信息。

在进行数据读写时，当文件系统接收到数据读写请求即输入输出（IO）请求，对该数据读写请求进行解析后储存到LUN虚拟块设备中相应位置；而RAID模块就会从LUN虚拟块设备中读取了数据读写请求，并对数据读写请求解析后，采用RAID的数据保护方法对请求的数据进行处理，比如将数据存储到硬
20 盘中，或从硬盘中读取数据，其中RAID的数据保护方法即为RAID模块进行数据读写时所采用的方法，包括多种类型，即RAIDn（n=1,2,...）的数据保护方法。现有RAID模块会以数据条带为单位来存储数据，每个数据条带中存储有多个分片数据。

发明人在研究中发现，现有对文件数据处理时，是文件系统将读写请求写
25 到LUN虚拟块设备中，再由RAID模块向硬盘读写数据的，数据读写的效率较低。

发明内容

本发明实施例提供文件系统内读写数据的方法及装置，提高数据读/写效率。

30 本发明实施例提供一种文件系统内读写数据的方法，所述文件系统支持文

件级独立磁盘冗余算法操作, 包括:

接收文件读/写请求, 确定所述文件读/写请求所请求读/写的的数据在数据条带中的读/写位置; 所述数据条带是独立磁盘冗余阵列RAID的数据读写条带;

5 根据所述确定的读/写位置, 顺序缓存至少一个数据条带对应的文件读/写请求所请求读/写的的数据信息;

10 根据所述缓存的至少一个完整数据条带对应的文件读请求所请求读出的数据信息, 从硬盘的相应数据条带读取数据, 或根据所述缓存的至少一个完整数据条带对应的文件写请求所请求写入的数据信息, 将所述缓存的至少一个完整数据条带对应的文件写请求所请求写入的数据信息写入所述硬盘的相应数据条带中。

本发明实施例提供一种文件系统内读写数据的装置, 所述文件系统支持文件级独立磁盘冗余算法操作, 所述装置包括: 请求接收模块, 位置确定模块, 条带缓存模块和操作模块;

所述请求接收模块, 用于接收文件读/写请求;

15 所述位置确定模块, 用于确定所述文件读/写请求所请求读/写的的数据在数据条带中的读/写位置; 所述数据条带是独立磁盘冗余阵列RAID的数据读写条带;

所述条带缓存模块, 用于根据所述位置确定模块确定的读/写位置, 顺序缓存至少一个数据条带对应的文件读/写请求所请求读/写的的数据信息;

20 所述操作模块, 用于根据缓存的至少一个完整数据条带对应的所述文件读请求所请求读出的数据信息, 从硬盘的相应数据条带读取数据, 或根据所述缓存的至少一个完整数据条带对应的文件写请求所请求写入的数据信息, 将所述缓存的至少一个完整数据条带对应的文件写请求所请求写入的数据信息写入所述硬盘的相应数据条带中。

25 本发明实施例中: 文件系统支持文件级独立磁盘冗余算法操作, 在接收到文件读/写请求后, 确定请求读/写的的数据在数据条带中的读/写位置, 并顺序缓存至少一个数据条带对应的请求读/写的的数据信息, 并对缓存的至少一个完整数据条带的的数据信息进行相对应处理。这样在缓存完整一个数据条带对应的请求读/写的的数据信息后, 并以数据条带为单位进行数据读/写, 可以不需要每次
30 在接收到文件读/写请求后, 都先将文件读/写请求写入到LUN虚拟块设备中,

再由RAID模块访问硬盘，从而提高了数据读/写效率；

且本发明实施例中的方法不需要经过文件系统和RAID模块分别对文件操作请求进行解析后，再由RAID模块向硬盘进行文件操作，而是由文件系统采用RAID的数据保护算法将数据信息缓存成数据条带，并直接访问硬盘，可以
5 提高数据存储的可靠性，并简化了文件处理的过程。

附图说明

为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。
10 出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

图1是本发明实施例提供的一种文件系统内读写数据的方法的流程图；

图2是本发明实施例提供的一种文件系统内读写数据方法的结构示意图；

图3是本发明实施例提供的一种文件系统内读写数据装置的结构示意图；

图4是本发明实施例中文件系统与上层应用系统之间的关系结构示意图；

15 图5是本发明实施例提供的另一种文件系统内读写数据装置的结构示意图。

具体实施方式

下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是
20 全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

本发明实施例提供一种文件系统内读写数据的方法，本发明实施例的文件操作方法是文件系统所执行的方法，流程图如图1所示，包括：

101，接收文件读/写请求，这里文件读/写请求即为IO请求。

25 102，确定文件读/写请求所请求读/写的的数据在数据条带中的读/写位置，这里数据条带是RAID的数据读写条带。

本发明实施例中，文件系统支持文件级独立磁盘冗余算法操作，比如文件系统可以把单个文件按照RAID条带分布进行储存，即文件系统中的数据以数据条带形式存储在多个硬盘中，文件系统直接对这些硬盘进行文件操作，比如
30 可以直接向硬盘写数据或读数据等。其中这里的硬盘可以是任何非易失性存储

介质，比如传统硬盘，固态硬盘，磁带，或可擦写光盘等。

本实施例中文件系统对硬盘进行的文件操作方法可以是传统的任一类型的RAID数据保护方法即RAIDn ($n=1,2,\dots$) 的数据保护方法，其中RAIDn是一个数据存储块，每个数据存储块的数据保护方法的类型都不同：例如RAID1
5 是数据多份镜像冗余的保护方法；而RAID5将数据存储和数据条带中，且每个数据条带中的任一分片数据都可以通过该数据条带中的检验块和其他分片数据计算出，其中一个数据条带中的分片数据储存在不同的硬盘上。

这里数据条带是采用某些类型的RAID数据保护算法进行存储数据的单位，一个数据条带中可以包含多个分片数据。在文件读/写请求中，都会包含
10 有文件数据相对于初始存储数据的偏移位置，文件系统可以根据这个偏移位置来确定请求读/写的的数据在数据条带中的读/写位置，这里读/写位置可以是在数据条带中的第几个分片，或第几个硬盘等。

可以理解，在每个硬盘上都可以有个单盘文件系统，在单盘文件系统上建立有至少两个分片文件，文件数据可以分布到这些分片文件中，而一个文件读
15 /写请求对应的文件数据就可以分布在至少两个硬盘上单盘文件系统的分片文件中，且组成RAID的数据保护关系。

103，根据确定的读/写位置，顺序缓存至少一个数据条带对应的文件读/写请求所请求读/写的的数据信息。

如果接收到文件写请求，则这里缓存至少一个数据条带对应的文件写请求
20 所请求写入的数据本身，如果接收到的是文件读请求，则这里缓存的是至少一个数据条带对应的文件读请求所请求读出数据的信息。

可以理解，如果一个文件读/写请求中请求读/写的的数据没有占满一个数据条带，则需要等待接收其它的文件读/写请求，直到所请求读/写数据能占满一个数据条带，才执行步骤104。例如：一个数据条带允许存储7个分片数据，而
25 接收到的一个文件读/写请求所请求读/写的的数据形成4个分片数据，则文件系统需要等待接收其它文件读/写请求，直到其它文件读/写请求所请求读/写的的数据可以形成另外3个分片数据，才执行步骤104。

如果一个文件读/写请求中请求读/写的的数据超出一个数据条带的存储空间，则只要读/写数据能占满一个数据条带，就执行步骤104，并继续缓存该文件
30 读/写请求中请求读/写的，且在该数据条带之外存储的数据信息。例如：一

个数据条带允许存储4个分片数据，而接收到的一个文件读/写请求所请求读/写的数据形成7个分片数据，则文件系统在顺序缓存了4个分片数据，就执行步骤104，并将继续缓存该文件读/写请求中请求读/写的其它3个分片数据信息。

本实施例中可以同时多个数据条带对应的所请求读/写的数据信息进行缓存。

104，根据缓存的至少一个完整数据条带对应的文件读请求所请求读出的数据信息，从硬盘的相应数据条带读取数据，或根据缓存的至少一个完整数据条带对应的文件写请求所请求写入的数据信息，将缓存的至少一个完整数据条带对应的文件写请求所请求写入的数据信息写入硬盘的相应数据条带。

10 需要说明的是，文件系统在执行步骤103到104时，对于同一个数据条带来说是顺序执行；而对于不同的数据条带来说是并行执行，即在缓存一个数据条带对应的数据信息的同时，文件系统可以对缓存好的其他数据条带对应的数据信息进行相应读/写操作。

15 本实施例中，文件系统是缓存一个完整的数据条带之后，才访问硬盘进行数据的读写，可见，文件系统是以一个数据条带为单位来访问硬盘的，而不是以一个文件读写请求来访问硬盘的，而数据条带是数据在硬盘中的存储位置，这样会提高文件系统进行数据的读写速率。

20 本发明实施例中：文件系统支持文件级独立磁盘冗余算法操作，在接收到文件读/写请求后，确定请求读/写的数据在数据条带中的读/写位置，并顺序缓存至少一个数据条带对应的请求读/写的数据信息，并对缓存的至少一个完整数据条带的的数据信息进行相对应处理。这样在缓存完整一个数据条带对应的请求读/写的数据信息后，并以数据条带为单位进行数据读/写，可以不需要每次在接收到文件读/写请求后，都先将文件读/写请求写入到LUN虚拟块设备中，再由RAID模块访问硬盘，从而提高了数据读/写效率；

25 且本发明实施例中的方法不需要经过文件系统和RAID模块分别对文件操作请求进行解析后，再由RAID模块向硬盘进行文件操作，而是由文件系统采用RAID的数据保护算法将数据信息缓存成数据条带，并直接访问硬盘，可以提高数据存储的可靠性，并简化了文件处理的过程。

在一个具体的实施例中，文件系统在步骤101中接收到文件读/写请求后，

还可以先确定文件读/写请求所请求读/写的文件类型，如果文件类型是文件数据，即文件的内容数据，则执行步骤102。

如果文件类型是文件元数据，即文件系统中的属性数据，比如：文件格式、文件大小等数据，则文件系统会向储存有文件元数据的任一硬盘中读取文件元数据，或按照多份镜像冗余算法，将文件元数据备份成多份并存储到至少两个硬盘中。

本实施例中文件系统对不同文件类型采用不同的方法。比如对于一些文件类型的数据比较重要（如文件元数据），可以选用RAID1的多份镜像冗余算法；而一些类型的数据（如文件数据）可以根据应用场景需要，选择RAID5，RAID6等保护方法，并以数据条带的形式对硬盘中的数据进行操作，具体选择哪种RAID算法，可以根据预置策略来确定。

这样使得在进行文件读/写操作时，就可以对比较重要的文件类型的数据重点保护，比如可以将重要的数据写到系统中的至少两个硬盘上等，从而提高了数据存储的可靠性。

且为了进一步确保数据的可靠性，当文件系统在接收到同步（sync）操作，清除（flush）操作或关闭（close）操作等操作指令时，如果当前正在缓存的数据信息没有占满一个数据条带，也要执行步骤104中的数据写入操作，即将步骤103当前缓存的数据条带对应的数据信息写入到相应的数据条带。这样保证了在执行同步、清除或关闭文件等操作之前，将数据信息都备份到硬盘中。

在本发明实施例中，文件系统是先将写入硬盘的数据以数据条带为单位进行缓存，如果在将缓存的数据写入硬盘之前，需要读取这些缓存的数据条带的信息，则为了使得数据读取的及时准确，在另一个具体的实施例中：如果步骤101中接收的是文件读请求，文件系统在执行上述步骤104的读取数据时，需要先确定缓存的至少一个完整数据条带对应的文件读请求所请求读出的数据信息，是否储存在文件系统当前缓存的其它数据条带中，如果是则从文件系统当前缓存的数据条带中读取相应的数据；如果否，才从硬盘的相应数据条带读取数据。

需要说明的是，在其它的具体实施例中，如果系统接收到文件写请求，则可以在开始缓存一个数据条带对应的数据信息开始计时，当计时时间到达预置的一个时长还未能存满一个完整的数据条带对应的数据信息时，则可以将当前

未存满的一个数据条带对应的数据信息写入相应的数据条带中。这样可以避免缓存整个数据条带对应的数据信息的缓存时间过长,而不能将文件写入到硬盘中。

以下以一个具体的实施例为例来说明,参考图2所示,数据储存到10个硬盘中,每个硬盘上有多个分片数据,图中只画出硬盘上的11个分片数据,其中斜线填充的部分存储有校验块P,每个文件数据条带中包含有横向分布的9个分片数据和1个校验块P。

如果文件系统接收到文件读请求,确定该文件读请求所请求读取的数据的位置是在第2个数据条带中的硬盘1、2、4和5,则顺序缓存该数据条带对应的文件读请求所请求读出的数据信息。因为所请求读取的数据没有占满该数据条带中9个分片数据的空间,则文件系统会等待接收另一个文件读请求,直到其它文件读请求所请求读取的数据为另外5个分片数据,则根据缓存的该数据条带对应的数据信息,向硬盘的第2数据条带读取数据。

如果文件系统接收到文件写请求,确定该文件写请求所请求写入的数据的位置是在第3个数据条带和第4个数据条带中的硬盘1到4中,则顺序缓存该数据条带对应的文件写请求所请求写入的数据。因为所请求读取的数据超过该数据条带中9个分片数据的空间,则文件系统会先将缓存的一个数据条带对应的数据写入到硬盘的第3个数据条带,并等待接收另一个文件写请求,直到其它文件写请求所请求写入的数据可以形成另外5个分片数据,再对第4个数据条带对应的数据进行写入。

本发明实施例提供一种文件系统内读写数据的装置,该文件系统支持文件级独立磁盘冗余算法操作,本实施例中的装置的结构示意图如图3所示,包括:请求接收模块10,位置确定模块20,条带缓存模块30和操作模块40,其中:

请求接收模块10,用于接收文件读/写请求;

位置确定模块20,用于确定请求接收模块10接收的文件读/写请求所请求读/写的数据在数据条带中的读/写位置;数据条带是独立磁盘冗余阵列RAID的数据读写条带;

条带缓存模块30,用于根据位置确定模块20确定的读/写位置,顺序缓存至少一个数据条带对应的文件读/写请求所请求读/写的数据信息;

操作模块40,用于根据条带缓存模块30缓存的至少一个完整数据条带对应的文件读请求所请求读出的数据信息,从硬盘的相应数据条带读取数据,或根据缓存的至少一个完整数据条带对应的文件写请求所请求写入的数据信息,将缓存的至少一个完整数据条带对应的文件写请求所请求写入的数据信息写入
5 硬盘的相应数据条带中。

参考图4所示,本发明实施例中的装置可以是设备在执行应用程序过程中对文件进行的操作装置,该装置中可以包括用户文件系统,该装置可以向硬盘进行数据操作,每个硬盘中有一个单盘文件系统。上层应用系统是面对用户的接口,当上层应用系统接收到文件读写请求后会发送给用户文件系统,而用户
10 文件系统会采用RAID的数据保护方法直接访问硬盘,使得文件分布在多个硬盘上单盘文件系统中的分片文件中,并组成RAID保护关系。

本发明实施例的装置中:请求接收模块10在接收到文件读/写请求后,位置确定模块20确定请求读/写的数据在数据条带中的读/写位置,并由条带缓存模块30顺序缓存至少一个数据条带对应的请求读/写的数据信息,由操作模块
15 40对缓存的至少一个完整数据条带的数据信息进行相对应处理。这样在缓存完整一个数据条带对应的请求读/写的数据信息后,并以数据条带为单位进行数据读/写,可以不需要每次在接收到文件读/写请求后,都先将文件读/写请求写入到LUN虚拟块设备中,再由RAID模块访问硬盘,从而提高了数据读/写效率;

且本发明实施例中的方法不需要经过文件系统和RAID模块分别对文件操作请求进行解析后,再由RAID模块向硬盘进行文件操作,而是由文件系统采用RAID数据保护算法将数据信息缓存成数据条带,并直接访问硬盘,可以提高数据存储的可靠性,并简化了文件处理的过程。
20

参考图5所示,在一个具体的实施例中,文件系统内读写数据的装置还可以包括文件元数据操作模块60和操作接收模块70,其中:

25 文件元数据操作模块60,用于若请求接收模块10接收的文件读/写请求所请求读/写的文件类型是文件元数据,当接收的是文件读请求时,则向储存有文件元数据的任一硬盘中读取文件元数据,或当接收的是文件写请求时,按照多份镜像冗余算法,将文件元数据备份成多份并存储到至少两个硬盘中;

操作接收模块70,用于接收同步操作、清除操作或关闭操作指令。

30 在本实施例的装置中,当请求接收模块10接收到文件读/写请求后,如果

该请求所请求读/写的文件类型是文件数据，则由位置确定模块20确定在数据条带中的读/写位置；如果文件类型中是文件元数据，则文件元数据操作模块60向存储有文件元数据的任一硬盘中读取文件元数据，或按照多份镜像冗余算法，将文件元数据备份成多份并存储到至少两个硬盘中。这样使得在进行文件读/写操作时，就可以对比较重要的文件类型的数据重点保护，比如可以将重要的数据写到至少两个硬盘上等，从而提高了数据存储的可靠性。

且为了进一步确保数据的可靠性，当操作接收模块70接收到同步操作、清除操作或关闭操作等操作指令时，操作模块40会将条带缓存模块30当前缓存的数据条带对应的数据信息写入到硬盘中相应的数据条带。这样保证了在执行同步、清除或关闭文件等操作指令之前，将数据信息都备份到硬盘中。

且参考图5所示，在另一个具体的实施例，文件系统内读写数据的装置还可以包括：缓存读取模块80和计时模块90，其中：

缓存读取模块80，用于确定条带缓存模块30缓存的至少一个完整数据条带对应的文件读请求所请求读出的数据信息，是否储存在当前缓存的其它数据条带中，如果是，则从当前缓存的数据条带中读取相应的数据；如果否，通知操作模块40从硬盘的相应数据条带读取数据；

计时模块90，用于当缓存一个完整数据条带对应的所请写入的数据信息所用的时间超过预置的时长，发出写操作指令以通知操作模块40将当前缓存的未存满的数据条带对应的数据信息写入到硬盘的数据条带中。

这样当请求接收模块10接收的是文件读取请求时，在操作操作模块40进行数据的读写前，通过缓存读取模块80确定需要读取的数据信息是否储存在条带缓存模块30缓存的其它数据条带中，如果是，则从其它数据条带读取数据；如果否，则通知操作操作40从硬盘读取数据。

如果请求接收模块10接收到的是文件写请求时，则计时模块90可以在开始缓存一个数据条带对应的数据信息开始计时，当计时时间到达预置的一个时长还未能存满一个完整的数据条带对应的数据信息时，可以通知操作模块40将当前未存满的一个数据条带对应的数据信息写入相应的数据条带中。这样避免缓存整个数据条带对应的数据信息的缓存时间过长，而不能将文件写入到硬盘中。

本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步

骤是可以通程序来指令相关的硬件来完成,该程序可以存储于一计算机可读存储介质中,存储介质可以包括:只读存储器(ROM)、随机存取存储器(RAM)、磁盘或光盘等。

- 5 以上对本发明实施例所提供的文件系统内读写数据的方法及装置,进行了详细介绍,本文中应用了具体个例对本发明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发明的方法及其核心思想;同时,对于本领域的一般技术人员,依据本发明的思想,在具体实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本发明的限制。

权 利 要 求

1、一种文件系统内读写数据的方法，其特征在于，所述文件系统支持文件级独立磁盘冗余算法操作，包括：

接收文件读/写请求，确定所述文件读/写请求所请求读/写的的数据在数据条带中的读/写位置；所述数据条带是独立磁盘冗余阵列RAID的数据读写条带；

根据所述确定的读/写位置，顺序缓存至少一个数据条带对应的文件读/写请求所请求读/写的的数据信息；

根据缓存的至少一个完整数据条带对应的文件读请求所请求读出的数据信息，从硬盘的相应数据条带读取数据，或根据所述缓存的至少一个完整数据条带对应的文件写请求所请求写入的数据信息，将所述缓存的至少一个完整数据条带对应的文件写请求所请求写入的数据信息写入所述硬盘的相应数据条带中。

2、如权利要求1所述的方法，其特征在于，

当接收文件读/写请求所请求读/写的文件类型为文件元数据时，所述确定所述文件读/写请求所请求读/写的的数据在数据条带中的读/写位置，包括：

当接收的是文件写请求时，按照RAID1或多份镜像冗余算法的方式，确定所述文件元数据写入的数据条带的位置以使所述文件元数据备份到至少两个硬盘中；

当接收的是文件写请求时，确定向存储有所述文件元数据的任一硬盘中读取所述文件元数据。

3、如权利要求1所述的方法，其特征在于，若所述文件读/写请求所请求读/写的文件类型是文件元数据，还包括：

当接收的是文件读请求时，向储存有文件元数据的任一硬盘中读取所述文件元数据，或当接收的是文件写请求时，按照多份镜像冗余算法，将所述文件元数据备份成多份并存储到至少两个硬盘中。

4、如权利要求1所述的方法，其特征在于，所述硬盘的单盘文件系统上建立有至少两个分片文件，所述文件读/写请求对应的文件数据分布在至少两个硬盘上单盘文件系统的分片文件中，且组成RAID的数据保护关系。

5、如权利要求1至4任一项所述的方法，其特征在于，所述方法还包括：

接收到同步操作、清除操作、或关闭操作指令时，将当前缓存的数据条带

对应的数据信息写入到所述硬盘的相应数据条带。

6、如权利要求1至4任一项所述的方法，其特征在于，还包括：

如果缓存一个完整数据条带对应的请写入的数据信息所用的时间超过预置的时长，将当前缓存的所述未存满的数据条带对应的数据信息写入到所述硬盘的数据条带中。

7、如权利要求1至4任一项所述的方法，其特征在于，所述从硬盘的相应数据条带读取数据之前还包括：

确定所述缓存的至少一个完整数据条带对应的文件读请求所请求读出的数据信息，是否储存在当前缓存的其它数据条带中，如果是，则从当前缓存的数据条带中读取相应的数据；如果否，则从硬盘的相应数据条带读取数据。

8、根据权利要求1所述的方法，其特征在于，当文件读/写请求所请求读/写的文件类型是文件数据时，所述确定所述文件读/写请求所请求读/写的的数据在数据条带中的读/写位置，包括：

按照预置策略选择的RAID算法，确定所述文件读/写请求所请求读/写的的数据在数据条带中的位置。9、一种文件系统内读写数据的装置，其特征在于，所述文件系统支持文件级独立磁盘冗余算法操作，所述装置包括：请求接收模块，位置确定模块，条带缓存模块和操作模块；

所述请求接收模块，用于接收文件读/写请求；

所述位置确定模块，用于确定所述文件读/写请求所请求读/写的的数据在数据条带中的读/写位置；所述数据条带是独立磁盘冗余阵列RAID的数据读写条带；

所述条带缓存模块，用于根据所述位置确定模块确定的读/写位置，顺序缓存至少一个数据条带对应的文件读/写请求所请求读/写的的数据信息；

所述操作模块，用于根据缓存的至少一个完整数据条带对应的所述文件读请求所请求读出的数据信息，从所述硬盘的相应数据条带读取数据，或根据所述缓存的至少一个完整数据条带对应的文件写请求所请求写入的数据信息，将所述缓存的至少一个完整数据条带对应的文件写请求所请求写入的数据信息写入所述硬盘的相应数据条带中。

10、根据权利要求9所述的装置，其特征在于，所述装置还包括：

文件元数据操作模块，用于若所述文件读/写请求所请求读/写的文件类型

是文件元数据,则当接收的是文件读请求时,向所述储存有文件元数据的任一硬盘中读取所述文件元数据,或当接收的是文件写请求时,按照多份镜像冗余算法,将所述文件元数据备份成多份并存储到至少两个硬盘中。

11、如权利要求9所述的装置,其特征在于:

- 5 所述位置确定模块,具体用于当接收文件读/写请求所请求读/写的文件类型为文件元数据,并且为文件写请求时,按照RAID1或多份镜像冗余算法的方式,确定所述文件元数据写入的数据条带的位置以使所述文件元数据备份到至少两个硬盘中;当接收文件读/写请求所请求读/写的文件类型为文件元数据,并且为文件读时,确定向存储有所述文件元数据的任一硬盘中读取所述文件元数据;
- 10 数据;

当文件读/写请求所请求读/写的文件类型是文件数据时,按照预置策略选择的RAID算法,确定所述文件读/写请求所请求读/写的数据在数据条带中的位置。

- 12、如权利要求9至11任一项所述的装置,其特征在于,所述装置还包括:
- 15 操作接收模块,用于接收同步操作、清除操作或关闭操作指令;

所述操作模块,还用于当所述操作接收模块接收到操作指令时,将所述条带缓存模块当前缓存的数据条带对应的数据信息写入到所述硬盘中相应的数据条带。

- 13、如权利要求9至11任一项所述的装置,其特征在于,所述装置还包括:
- 20 缓存读取模块,用于确定所述条带缓存模块缓存的至少一个完整数据条带对应的文件读请求所请求读出的数据信息,是否储存在当前缓存的其它数据条带中,如果是,则从当前缓存的数据条带中读取相应的数据;如果否,通知所述操作模块从硬盘的相应数据条带读取数据。

- 14、如权利要求9至11任一项所述的装置,其特征在于,所述装置还包括:
- 25 计时模块,用于当缓存一个完整数据条带对应的所请写入的数据信息所用的时间超过预置的时长,发出写操作指令以通知所述操作模块将当前缓存的所述未存满的数据条带对应的数据信息写入到所述硬盘的数据条带中。

— 1/3 —

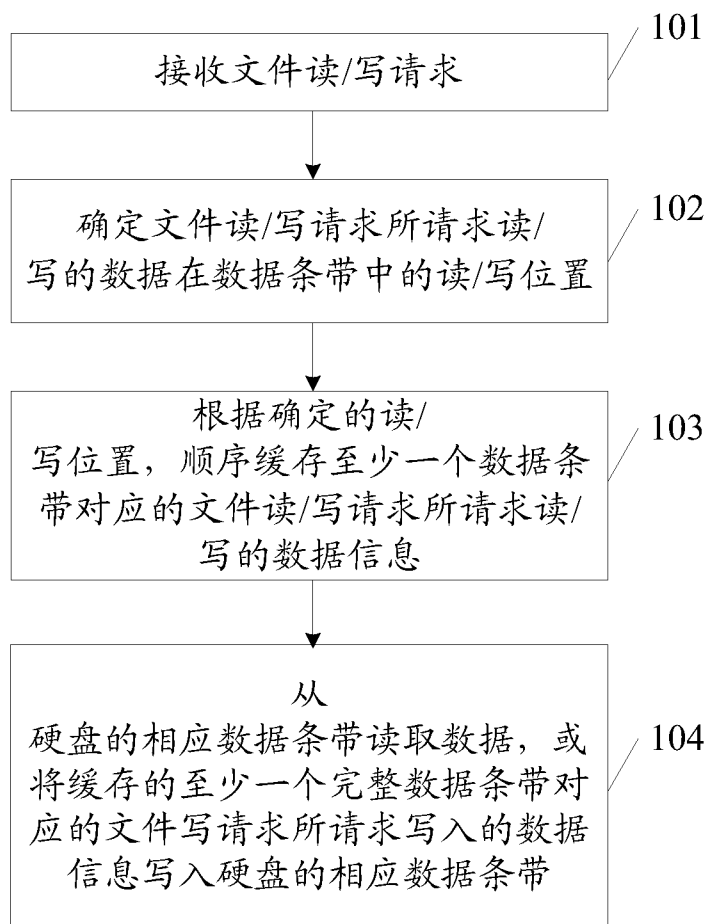


图 1

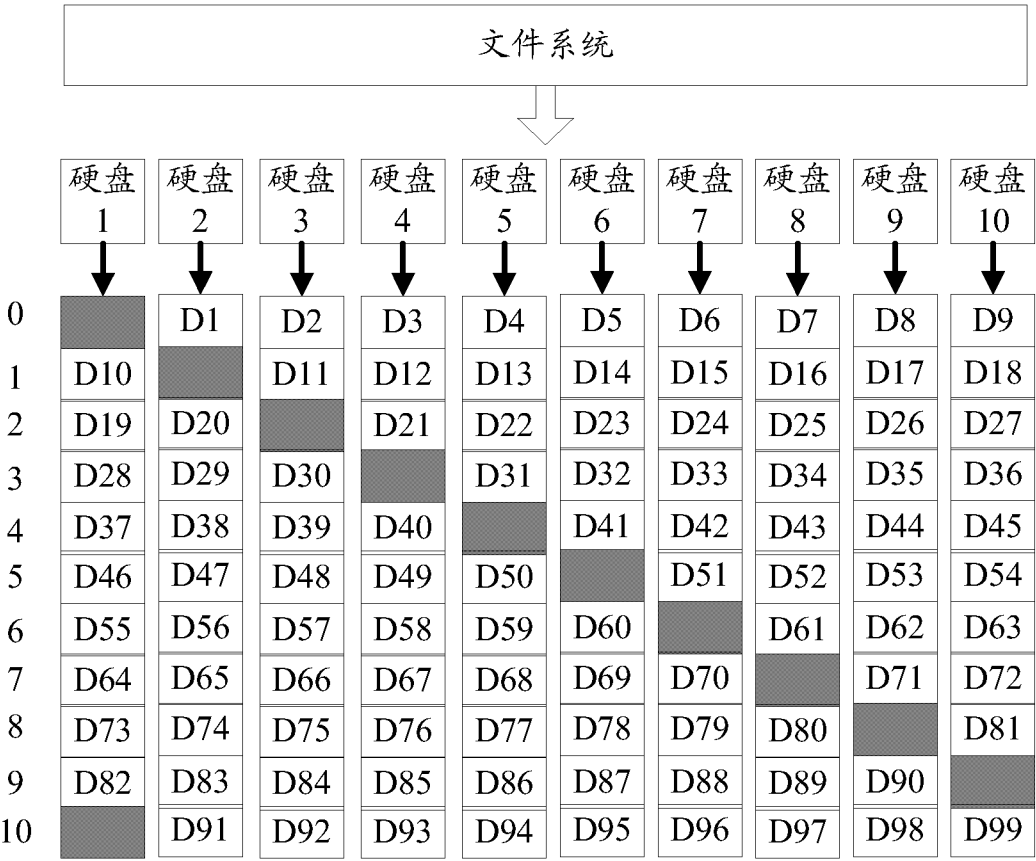


图 2

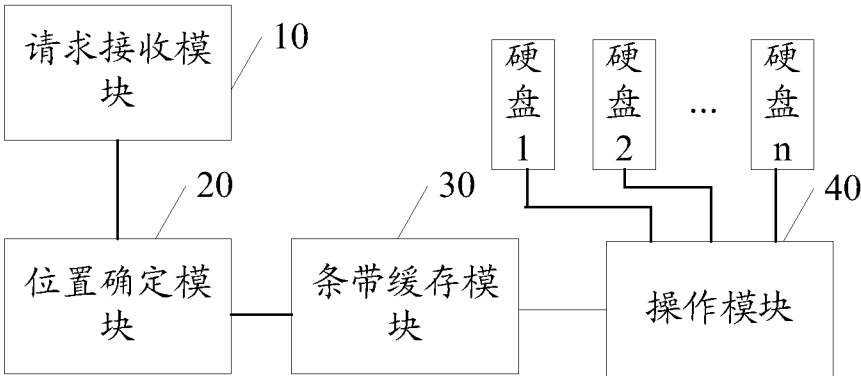


图 3

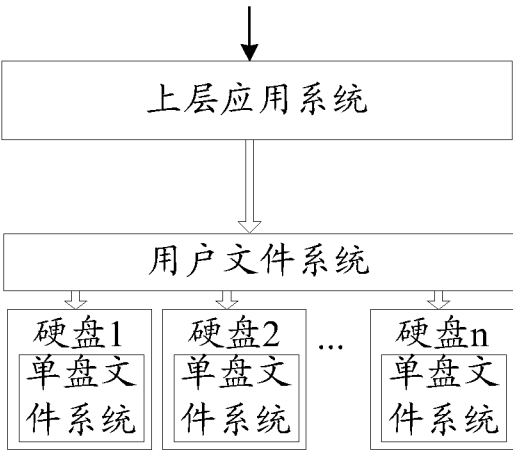


图 4

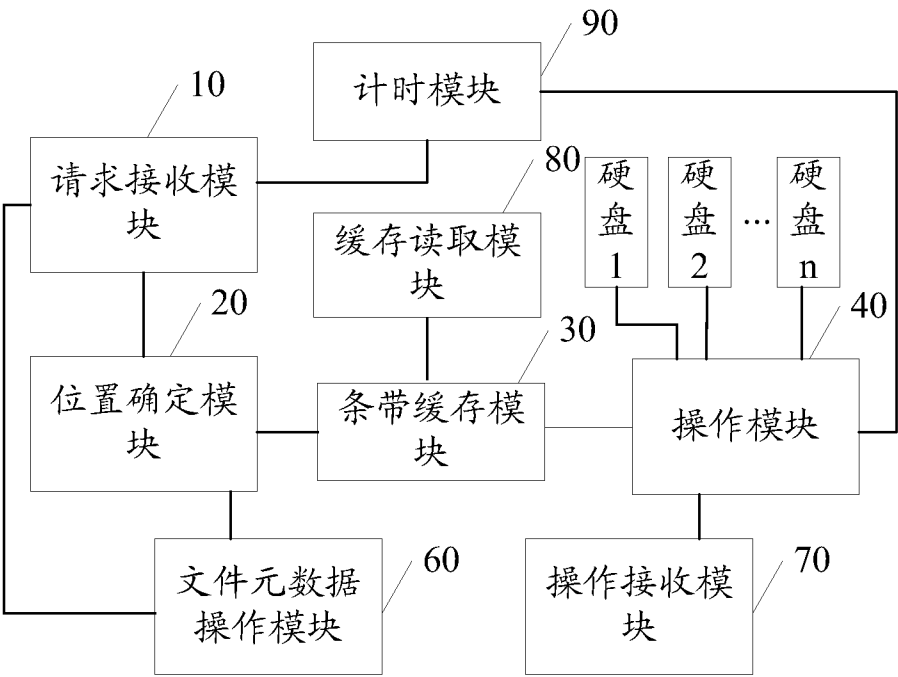


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2012/072107**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)

IPC: G06F 17/-

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: HUAWEI, data stripe, read/write stripe, redundance, RAID, read, write, document, file, data, strip+, hard disk, disk, redundant access independent disk, cache, metadata

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 1737745 A (HUAWEI TECHNOLOGIES CO., LTD.), 22 February 2006 (22.02.2006), the whole document	1-14
A	CN 101916173 A (HANGZHOU H3C TECHNOLOGIES CO., LTD.), 15 December 2010 (15.12.2010), the whole document	1-14

☐ 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
07 June 2012 (07.06.2012)Date of mailing of the international search report
21 June 2012 (21.06.2012)Name and mailing address of the ISA/CN:
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

LIU, YingyingTelephone No.: (86-10) **62411674**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2012/072107

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 1737745 A	22.02.2006	CN 1306381 C	21.03.2007
CN 101916173 A	15.12.2010	None	

A. 主题的分类		
G06F17/30(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)		
IPC: G06F17/-		
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:华为, 读, 写, 文件, 数据条带, 读写条带, 条带, 硬盘, 磁盘, 冗余, RAID, 缓存, 元数据, read, write, document, file, data, strip+, hard disk, disk, redundant access independent disk, cache, metadata		
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 1737745 A (华为技术有限公司) 22. 2 月 2006 (22.02.2006) 全文	1-14
A	CN 101916173 A(杭州华三通信技术有限公司)15. 12 月 2010(15.12.2010) 全文	1-14
☐ 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 07.6 月 2012 (07.06.2012)		国际检索报告邮寄日期 21.6 月 2012 (21.06.2012)
ISA/CN 的名称和邮寄地址: 中华人民共和国国家知识产权局 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 刘莹莹 电话号码: (86-10) 62411674

国际检索报告

关于同族专利的信息

国际申请号

PCT/CN2012/072107

检索报告中引用的 专利文件	公布日期	同族专利	公布日期
CN 1737745 A	22.02.2006	CN 1306381 C	21.03.2007
CN 101916173 A	15.12.2010	无	