(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114201119 A
(43)申请公布日 2022.03.18

(21)申请号 202210145523.X

(22)申请日 2022.02.17

(71)申请人 天津市天河计算机技术有限公司
地址 300457 天津市滨海新区天津经济技术开发区信环西路19号5号楼5102

(72)发明人 庞晓磊 李长松 张婷 刘嘉琦 赵欣婷 徐斌 夏梓峻 张健 孙福兴 贾子傲 王普 杨晶

(74)专利代理机构 天津盛理知识产权代理有限公司 12209
代理人 王利文

(51)Int.Cl.
G06F 3/06(2006.01)

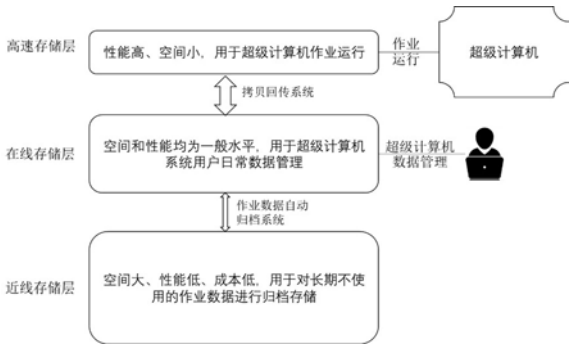
权利要求书2页 说明书5页 附图4页

(54)发明名称

一种超级计算机作业数据分层存储系统及方法

(57)摘要

本发明涉及一种超级计算机作业数据分层存储系统及方法，通过将存储层分为包括高速存储层、在线存储层和近线存储层的三层存储系统，同时基于三层存储系统分别构建了作业数据拷贝回传机制、原始数据存留时间计算机制和作业数据自动归档和换回机制，并将超级计算机系统与三层存储系统进行融合，实现了在控制存储系统整体设备成本的前提下，解决了超级计算机系统用户作业数据存储空间、存取性能和设备成本之间的矛盾，在保持存储设备低成本的同时，提高存储系统总可用空间和存储服务IO性能，提高数据总可用存储空间，降低存储系统设备平均成本。



CN 114201119 A

1.一种超级计算机作业数据分层存储系统,其特征在于:包括设置在超级计算机中的三层存储系统,三层存储系统分别为高速存储层、在线存储层和近线存储层,所述高速存储层挂载一般计算节点;在线存储层挂载登陆节点和小微作业计算节点,在线存储层用于用户登陆、作业数据操作管理和小作业任务的处理。

2.根据权利要求1所述的一种超级计算机作业数据分层存储系统,其特征在于:所述高速存储层选用利于超级计算机作业运行的存储资源;所述在线存储层选用空间和性能利于超级计算机系统用户日常数据管理的存储资源;所述近线存储层选用利于长期不使用的作业数据进行归档存储的存储资源。

3.一种基于权利要求1至2任一项所述的超级计算机作业数据分层存储系统的存储方法,其特征在于,包括以下步骤:

步骤1、构建作业数据拷贝回传机制;

步骤2、构建原始数据存留时间计算机制;

步骤3、构建作业数据自动归档和换回机制;

步骤4、在高速存储层、在线存储层和近线存储层的三层存储系统中植入步骤1的作业数据拷贝回传机制;采用步骤2的原始数据存留时间计算机制和步骤3的作业数据自动归档和换回机制实现超级计算机作业数据分层存储。

4.根据权利要求3所述的一种超级计算机作业数据分层存储系统的存储方法,其特征在于:所述步骤1中作业数据拷贝回传机制为:作业数据存于在线存储层中,当用于进行提交作业数据时,将作业数据自动从在线存储层拷贝到高速存储层进行计算,在作业数据计算完成后,自动将作业数据以及作业数据的计算结果从高速存储层回传至在线存储层。

5.根据权利要求3所述的一种超级计算机作业数据分层存储系统的存储方法,其特征在于:所述步骤1的具体实现方法为:将yhrun交互式提交作业命令和yhbatch批处理式提交作业命令进行重写,在实际执行提交作业之前,先获取作业提交脚本中的作业文件路径,生成在高速存储层中对应的路径,将作业数据拷贝到高速存储层中;然后再实际执行yhrun交互式提交作业命令或yhbatch批处理式提交作业命令提交作业;提交成功后,获取jobid提交作业的ID,通过该提交作业的ID设置触发器,监控作业运行状态,作业运行完毕后,自动将结果数据回传。

6.根据权利要求3所述的一种超级计算机作业数据分层存储系统的存储方法,其特征在于:所述步骤2中原始数据存留时间计算机制为:作业数据第一次提交高速存储层并计算运行完毕后,保留作业数据的预设时间。

7.根据权利要求6所述的一种超级计算机作业数据分层存储系统的存储方法,其特征在于:所述保留作业数据的预设时间的计算方法为:

$$T_{留存时间} = D_{作业数据量}/S_{拷贝速度}/R_{作业运行时间}*(0.1*W_{警告} + E_{错误} + 10*KE_{关键错误})*(1-U_{空间使用率})^{2}*\delta_{常量系数}$$

其中,$T_{留存时间}$为作业原始数据留存时间;$D_{作业数据量}$为作业原始数据总大小,$S_{拷贝速度}$为作业原始数据拷贝速度,$R_{作业运行时间}$为作业运行时间,$W_{警告}$为作业运行日志警告数量,$E_{错误}$为作业运行日志错误数量,$KE_{关键错误}$为作业运行日志关键错误数量,$U_{空间使用率}$为高速存储层当前空间使用率,$\delta_{常量系数}$为常量系数值。

8.根据权利要求3所述的一种超级计算机作业数据分层存储系统的存储方法,其特征在于:所述步骤3中作业数据自动归档和换回机制为:定期扫描在线存储层,若在线存储层

存在超过阈值时间未访问的文件,则将其移动到近线存储层,然后在原有的位置创建一个软连接,指向文件被移动到的位置。

一种超级计算机作业数据分层存储系统及方法

技术领域

[0001] 本发明属于超级计算机存储领域，尤其是一种超级计算机作业数据分层存储系统及方法。

背景技术

[0002] 随着高性能计算技术的不断发展，超级计算机所能提供的计算性能越来越强，有越来越多的科研人员开始使用超级计算机来运行自己的作业，以降低作业运算时间，提高科研工作效率。然而，在同一时期，超级计算机底层存储系统的发展速度则较为平缓，随着超级计算机计算性能的不断提高，与之配套的存储系统开始力不从心，在运行高IO的计算作业时，存储系统开始成为整个高性能计算系统的瓶颈，制约高性能计算系统性能的进一步提高。

[0003] 当前，为了解决超级计算机中的存储系统性能瓶颈问题，一般会使用高性能存储设备来搭建存储集群，以提供较高的存储系统IO性能，从而满足计算作业对底层存储系统的性能需求。而对于超级计算机来说，整个存储系统的数据量非常大，考虑到成本因素，不可能将整个存储集群都使用高性能存储设备来搭建。因此，只能搭建一个小的专有集群，在一定程度上解决问题。当前尚没有一个较通用的方法，能够从整体上来解决这一问题。

[0004] 另一方面，由于作业所产生的数据量的持续增加，整个存储系统长期处于高空间使用率的状态，再叠加上超级计算机作业运行，给其带来的IO压力，导致整个存储系统的稳定性变差，也因此而导致其上层超级计算机系统的运行不稳定。当前为了解决这一问题，只能采用存储系统扩容或者督促用户及时删除无用数据的方法，前者增加了存储系统的硬件成本，而后者则降低了用户对超级计算机的使用满意度。

发明内容

[0005] 本发明的目的在于克服现有技术的不足，提出一种超级计算机作业数据分层存储系统及方法，能够在控制存储系统整体设备成本的前提下，解决超级计算机系统运行存储性能瓶颈的问题，并提高存储系统乃至整个超级计算机的运行稳定性。

[0006] 本发明解决其技术问题是采取以下技术方案实现的：

一种超级计算机作业数据分层存储系统，包括设置在超级计算机中的三层存储系统，三层存储系统分别为高速存储层、在线存储层和近线存储层，所述高速存储层挂载一般计算节点；在线存储层挂载登陆节点和小微作业计算节点，在线存储层用于用户登陆、作业数据操作管理和小作业任务的处理。

[0007] 而且，所述高速存储层选用利于超级计算机作业运行的存储资源；所述在线存储层选用空间和性能利于超级计算机系统用户日常数据管理的存储资源；所述近线存储层选用利于长期不使用的作业数据进行归档存储的存储资源。

[0008] 一种基于超级计算机作业数据分层存储系统的存储方法，包括以下步骤：

步骤1、构建作业数据拷贝回传机制；

步骤2、构建原始数据存留时间计算机制；

步骤3、构建作业数据自动归档和换回机制；

步骤4、在高速存储层、在线存储层和近线存储层的三层存储系统中植入步骤1的作业数据拷贝回传机制；采用步骤2的原始数据存留时间计算机制和步骤3的作业数据自动归档和换回机制实现超级计算机作业数据分层存储。

[0009] 而且，所述步骤1中作业数据拷贝回传机制为：作业数据存于在线存储层中，当用于进行提交作业数据时，将作业数据自动从在线存储层拷贝到高速存储层进行计算，在作业数据计算完成后，自动将作业数据以及作业数据的计算结果从高速存储层回传至在线存储层。

[0010] 而且，所述步骤1的具体实现方法为：将yhrun交互式提交作业命令和yhbatch批处理式提交作业命令进行重写，在实际执行提交作业之前，先获取作业提交脚本中的作业文件路径，生成在高速存储层中对应的路径，将作业数据拷贝到高速存储层中；然后再实际执行yhrun交互式提交作业命令或yhbatch批处理式提交作业命令提交作业；提交成功后，获取jobid提交作业的ID，通过该提交作业的ID设置触发器，监控作业运行状态，作业运行完毕后，自动将结果数据回传。

[0011] 而且，所述步骤2中原始数据存留时间计算机制为：作业数据第一次提交高速存储层并计算运行完毕后，保留作业数据的预设时间。

[0012] 而且，所述保留作业数据的预设时间的计算方法为：

$$T_{留存时间} = D_{作业数据量}/S_{拷贝速度}/R_{作业运行时间}*(0.1*W_{警告} + E_{错误} + 10*KE_{关键错误})*(1-U_{空间使用率})^{2} *\delta_{常量系数}$$

其中，$T_{留存时间}$为作业原始数据留存时间；$D_{作业数据量}$为作业原始数据总大小，$S_{拷贝速度}$为作业原始数据拷贝速度，$R_{作业运行时间}$为作业运行时间，$W_{警告}$为作业运行日志警告数量，$E_{错误}$为作业运行日志错误数量，$KE_{关键错误}$为作业运行日志关键错误数量，$U_{空间使用率}$为高速存储层当前空间使用率，$\delta_{常量系数}$为常量系数值。

[0013] 而且，所述步骤3中作业数据自动归档和换回机制为：定期扫描在线存储层，若在线存储层存在超过阈值时间未访问的文件，则将其移动到近线存储层，然后在原有的位置创建一个软连接，指向文件被移动到的位置。

[0014] 本发明的优点和积极效果是：

1、本发明通过将存储层分为包括高速存储层、在线存储层和近线存储层的三层存储系统，同时基于三层存储系统分别构建了作业数据拷贝回传机制、原始数据存留时间计算机制和作业数据自动归档和换回机制，并将超级计算机系统与三层存储系统进行融合，实现了在控制存储系统整体设备成本的前提下，解决了超级计算机系统用户作业数据存储空间、存取性能和设备成本之间的矛盾，在保持存储设备低成本的同时，提高存储系统总可用空间和存储服务IO性能，提高数据总可用存储空间，降低存储系统设备平均成本。

[0015] 2、本发明通过构建作业数据拷贝回传机制和作业数据自动归档和换回机制，使作业数据在线存储层和高速存储层进行传输操作，解决了超级计算机运行高IO作业时的存储系统性能瓶颈问题。

[0016] 3、本发明通过构建原始数据存留时间计算机制，能够保证需多次提交的作业，其原始数据能够留存在高速存储层中，减少作业原始数据的拷贝成本；同时使高速存储层保

持较低的存储空间使用率，从而解决了超级计算机运行高IO作业时的存储系统性能瓶颈问题，并提高存储系统的稳定性，进而提高整个超级计算机系统的稳定性。

附图说明

[0017] 图1为本发明超级计算机作业数据分层存储系统；

图2为本发明超级计算机系统与分层存储系统融合方法；

图3为本发明超级计算机系统用户提交作业处理流程；

图4为本发明作业数据自动归档整体处理流程；

图5为为本发明作业数据自动换回整体处理流程。

具体实施方式

[0018] 以下结合附图对本发明做进一步详述。

[0019] 一种超级计算机作业数据分层存储系统，包括设置在超级计算机中的三层存储系统，三层存储系统分别为高速存储层、在线存储层和近线存储层。如图2所示，由于超级计算机用户一般通过登录节点进行作业数据管理，而作业任务的运行则通过计算节点，因此高速存储层挂载一般计算节点；在线存储层挂载登陆节点和小微作业计算节点。用户的作业数据存储在在线存储层，用户可以通过登录节点，对作业数据进行管理；当有作业需要运行时，通过“作业拷贝回传机制”，将作业数据从在线存储层拷贝到高速存储层。同时，对于较小的作业任务，可以直接提交到小微作业计算节点上，避免了作业数据拷贝的时间开销。

[0020] 分层存储即是将多种不同性能级别的存储系统融合在一起，性能最高的存储系统放在第一层，用于满足高速数据存取能力的需求；随着存储性能的降低，层级依次向下；性能最低的存储系统放在最后一层，用于满足海量存储空间的需求。各层存储系统发挥各自的特点优势，共同对外提供数据存储服务。如图1所示，高速存储层选用性能高、空间小，且利于超级计算机作业运行的存储资源；所述在线存储层选用空间和性能均为一般水平，且利于超级计算机系统用户日常数据管理的存储资源；所述近线存储层选用空间大、性能低、成本低，且利于长期不使用的作业数据进行归档存储的存储资源。

[0021] 一种基于超级计算机作业数据分层存储系统的存储方法，包括以下步骤：

步骤1、构建作业数据拷贝回传机制。

[0022] 由于构建的高速存储层的存储空间较小，因此仅用于存放超级计算机当前运行中作业的数据，及作业运行完毕后数据的短期留存。在线存储层有较高的存储空间，用于存储超级计算机系统用户在超级计算机上的日常作业数据，这就要求，作业数据能够在高速存储层和在线存储层之间进行自动流转。如图3所示，构建作业拷贝回传机制：对超级计算机作业管理系统进行改造，使作业提交时，将作业原始数据从在线存储层拷贝到高速存储层；作业运行完毕后，将结果数据从高速存储层拷贝到在线存储层。

[0023] 作业拷贝回传机制的具体事项方法维持：将yhrun交互式提交作业命令和yhbatch批处理式提交作业命令进行重写，在实际执行提交作业之前，先获取作业提交脚本中的作业文件路径，生成在高速存储层中对应的路径，将作业数据拷贝到高速存储层中；然后再实际执行yhrun交互式提交作业命令或yhbatch批处理式提交作业命令提交作业；提交成功后，获取jobid提交作业的ID，通过该提交作业的ID设置触发器，监控作业运行状态，作业运

行完毕后，自动将结果数据回传。

[0024] 通过分层存储系统与作业拷贝回传机制相结合，即可实现使用小规模的高速存储系统，支持超级计算机大规模作业运行，满足其对底层存储系统超高IO性能的要求。同时，通过作业拷贝回传装置，使作业数据只有需要被超级计算机访问时，才会被拷贝到高速存储层中，访问完毕后即被移出高速存储层，这一机制可以使高速存储层保持较低的存储空间使用率，从而增加存储系统的稳定性，进而增加整个超级计算机系统的稳定性。

[0025] 步骤2、构建原始数据存留时间计算机制。

[0026] 基于用户对超级计算机的使用习惯，一般作业在被提交运行一次后，短期内有可能进行少量修改并再次提交。因此，为了使作业多次提交时，降低作业数据的拷贝成本，作业第一次提交并运行完毕后，不会立刻将高速存储层中的作业数据删除，而是保留的预设时间。当作业算法或数据进行了少量修改并再次被提交时，仅需将修改部分的数据重新拷贝即可。

[0027] 作业数据在高速存储层中的留存时间既不能太长，也不能太短。如果留存时间太长，会导致高速存储层中的无用数据太多，浪费存储空间；如果留存时间太短，可能导致作业多次提交时，仍需进行完整的数据拷贝。

[0028] 本发明根据作业运行时间、作业日志警告数量、作业日志报错数量、作业日志关键报错数量、作业数据量大小、存储系统总体空间使用率等多个因素，来计算作业数据在高速存储层中的留存时间。如作业的数据量较大，则考虑再次拷贝时的成本较高，因此增加作业数据的留存时间；如作业运行时间较长，则考虑其数据拷贝操作时间占比较小，甚至可以忽略不计，因此降低作业数据的留存时间；另外，如果作业运行时间极短，则作业运行可能已经发生错误，因此该作业被重新提交的可能性极大，应增加作业数据留存时间；当作业运行日志中包含警告或者错误，则其被重新提交的可能性较大，增加作业数据留存时间；当整体的存储系统空间使用率较低时，能够容纳的数据较多，因此增加作业数据留存时间，反之则降低作业数据留存时间。保留作业数据的预设时间的计算方法为：

$$T_{留存时间} = D_{作业数据量}/S_{拷贝速度}/R_{作业运行时间}*(0.1*W_{警告} + E_{错误} + 10*KE_{关键错误})*(1-U_{空间使用率})^2 *\delta_{常量系数}$$

其中，$T_{留存时间}$为作业原始数据留存时间；$D_{作业数据量}$为作业原始数据总大小，$S_{拷贝速度}$为作业原始数据拷贝速度，$R_{作业运行时间}$为作业运行时间，$W_{警告}$为作业运行日志警告数量，$E_{错误}$为作业运行日志错误数量，$KE_{关键错误}$为作业运行日志关键错误数量，$U_{空间使用率}$为高速存储层当前空间使用率，$\delta_{常量系数}$为常量系数值。

[0029] 步骤3、构建作业数据自动归档和换回机制。

[0030] 如图4所示，作业数据自动归档用于将长期不使用的文件从在线存储层移出到近线存储层。作业数据自动归档装置是一个守护进程，每天定期扫描在线存储层，若在线存储层存在超过阈值时间未访问的文件，则将其移动到近线存储层，然后在原有的位置创建一个软连接，指向文件被移动到的位置。

[0031] 如图5所示，由于有软连接，因此即是文件被换出到近线存储层，用户依然可以访问到该文件，只是读写性能会相对差一些。同时使用作业数据自动换回自机制，监控用户对文件的访问，当用户访问文件时，自动将文件换回到在线存储层，保证用户的文件的正常使用。

[0032] 步骤4、在高速存储层、在线存储层和近线存储层的三层存储系统中植入步骤1的作业数据拷贝回传机制;采用步骤2的原始数据存留时间计算机制和步骤3的作业数据自动归档和换回机制实现超级计算机作业数据分层存储。

[0033] 需要强调的是,本发明所述的实施例是说明性的,而不是限定性的,因此本发明包括并不限于具体实施方式中所述的实施例,凡是由本领域技术人员根据本发明的技术方案得出的其他实施方式,同样属于本发明保护的范围。

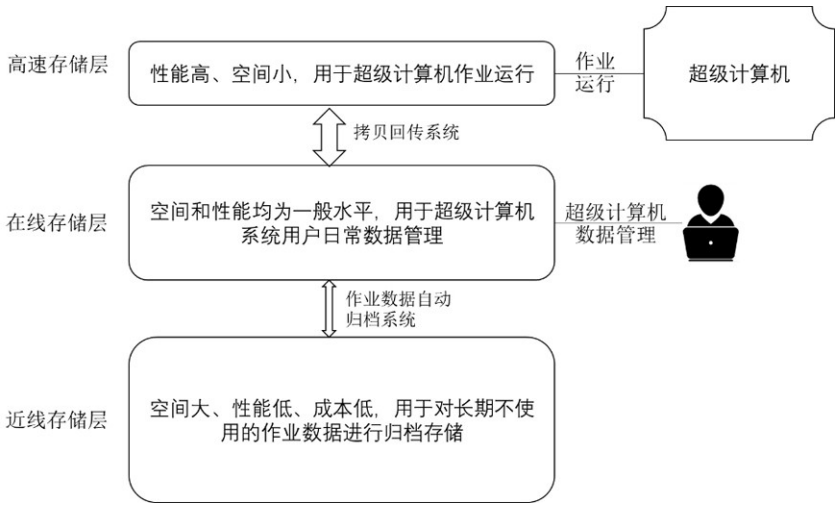

高速存储层
性能高、空间小，用于超级计算机作业运行
作业
运行
超级计算机
拷贝回传系统
在线存储层
空间和性能均为一般水平，用于超级计算机
系统用户日常数据管理
超级计算机
数据管理
作业数据自动
归档系统
近线存储层
空间大、性能低、成本低，用于对长期不使
用的作业数据进行归档存储

图1

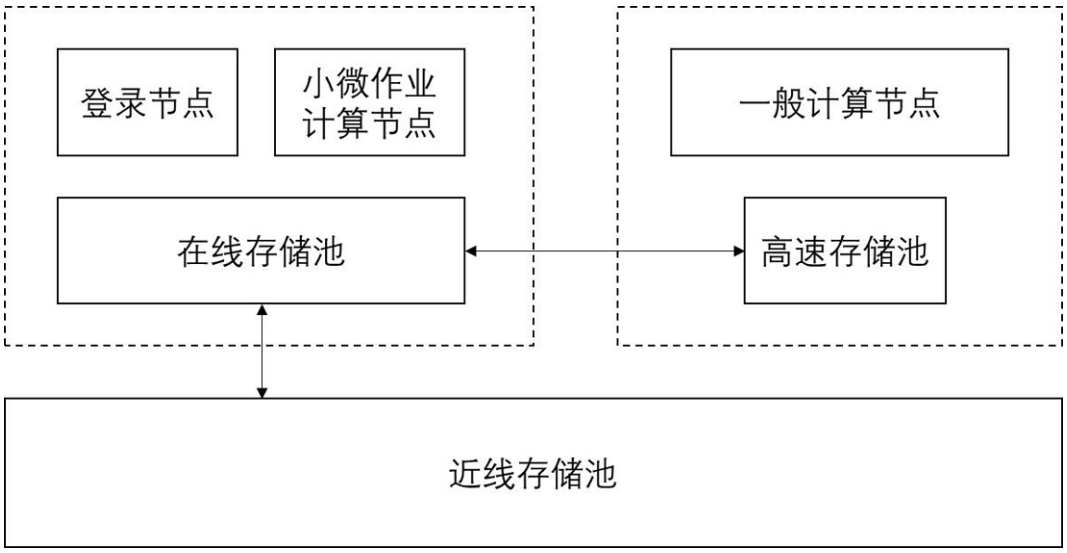

登录节点
小微作业
计算节点
一般计算节点
在线存储池
高速存储池
近线存储池

图2

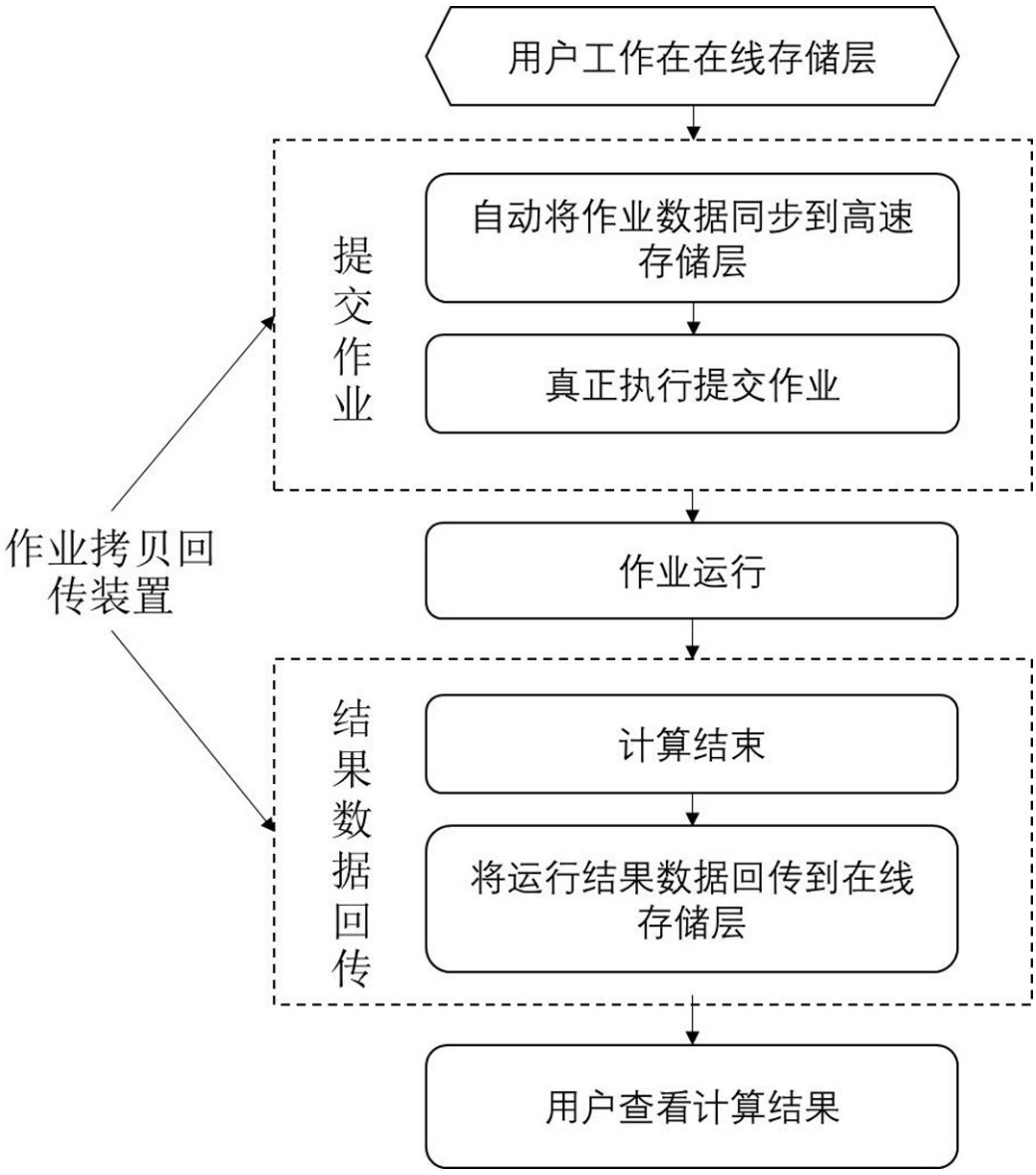

用户工作在在线存储层
提交作业
自动将作业数据同步到高速存储层
真正执行提交作业
作业运行
结果数据回传
计算结束
将运行结果数据回传到在线存储层
用户查看计算结果
作业拷贝回传装置

图3

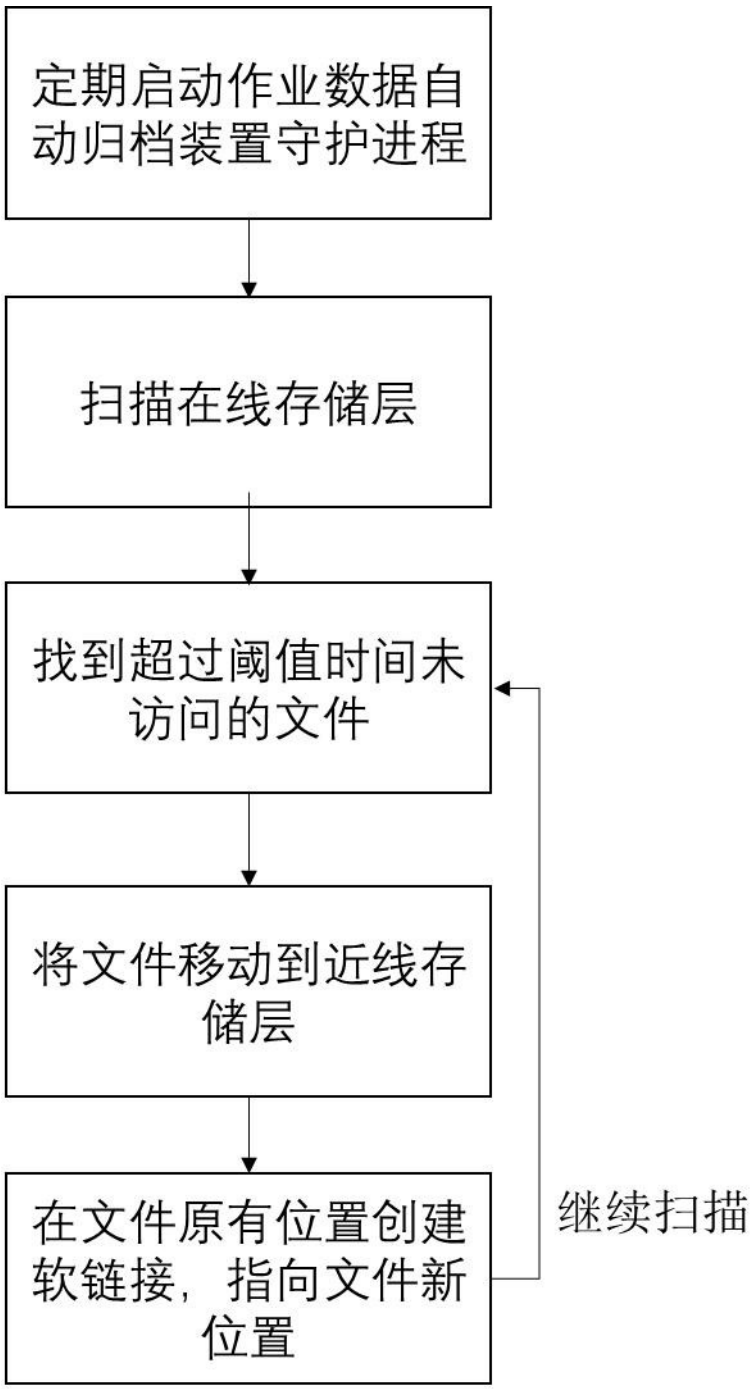

定期启动作业数据自动归档装置守护进程
扫描在线存储层
找到超过阈值时间未访问的文件
将文件移动到近线存储层
在文件原有位置创建软链接，指向文件新位置
继续扫描

图4

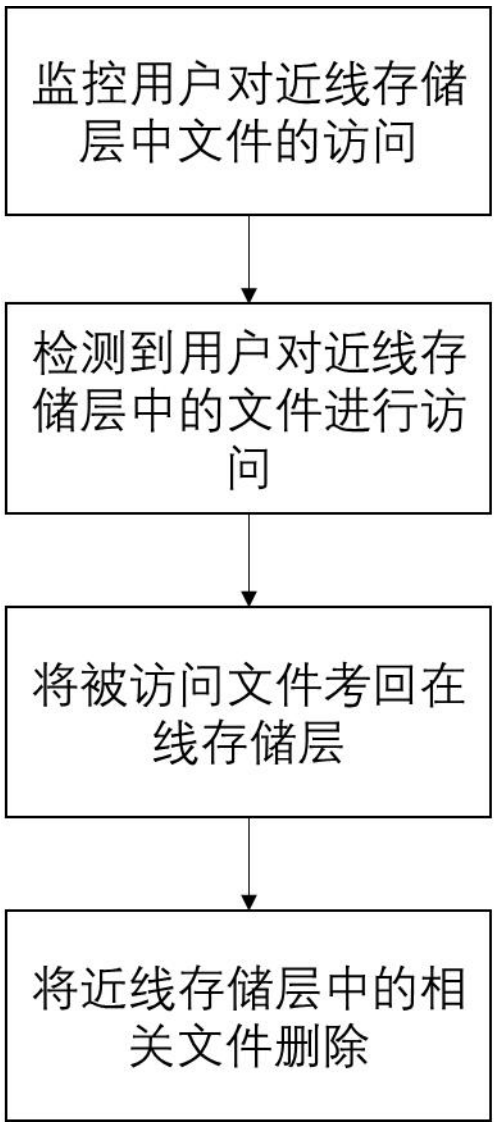

监控用户对近线存储层中文件的访问
检测到用户对近线存储层中的文件进行访问
将被访问文件考回在线存储层
将近线存储层中的相关文件删除

图5