(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103019890 B
(45)授权公告日 2015.04.29

(21)申请号 201210568036.0

(22)申请日 2012.12.24

(73)专利权人 清华大学
地址 100084 北京市海淀区清华园 1 号

(72)发明人 汪东升 王占业

(74)专利代理机构 北京清亦华知识产权代理事务所（普通合伙） 11201
代理人 廖元秋

(51)Int.Cl.
G06F 11/14(2006.01)
G06F 21/60(2013.01)

(56)对比文件
CN 101697134 A,2010.04.21, 全文.
US 2011/0060722 A1,2011.03.10,1-2.
申远南等. 一种基于磁盘块的持续数据保护系统. 《微型电脑应用》. 2011, 第 27 卷（第 4 期）, 30-33.

审查员 万洋

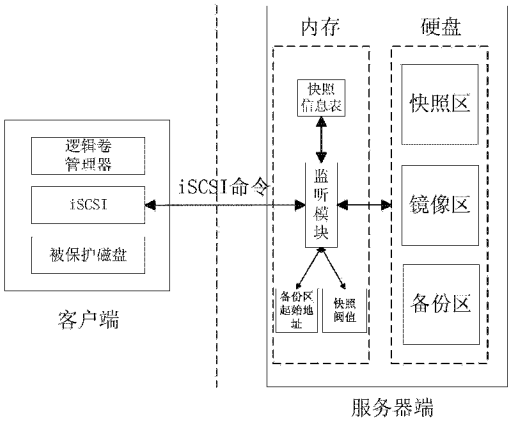
权利要求书2页 说明书5页 附图3页

(54)发明名称

一种块级别的磁盘数据保护系统及其方法

(57)摘要

一种块级别的磁盘数据保护系统及其方法，属于计算机数据存储和保护领域，该系统基于客户端/服务器架构，被保护磁盘位于客户端，服务器端存放备份数据；服务器端分为镜像区、备份区和快照区，被保护磁盘与镜像区做成实时镜像，用户所有对被保护磁盘的操作都会被同步到镜像区；服务器端监听模块将写入到镜像区的数据组织成备份数据单元，依次写入备份区；当备份区内新增数据备份单元数量大于用户设定的阀值时，服务器端对镜像区数据做快照，并将快照数据写入快照区；若用户提出恢复请求，服务器端根据恢复时间点先从快照区将快照数据写入镜像区，再从备份区找出相应备份数据单元写入镜像区，最后返回到被保护磁盘。



CN 103019890 B

1. 一种块级别的磁盘数据保护系统，其特征在于：该系统基于客户端 / 服务器架构，被保护磁盘位于客户端，服务器端存放备份数据；

该服务器端包括内存及磁盘，该磁盘划分为三个逻辑分区，分别是镜像区、快照区和备份区，其中镜像区与客户端中的被保护磁盘互为实时镜像，当有数据写入被保护磁盘时，相同的数据也会同步写入服务器端的镜像区；服务器端内存中存有快照信息表、备份区中的扇区起始地址、监听模块以及快照阀值；快照信息表中包含快照标示符、快照时间和快照地址三列，快照标示符是快照的标识，快照时间是做快照时的系统时间，快照地址为做快照时备份区的扇区起始地址；监听模块用于监听并判断对于镜像区的 iSCSI 命令是进行写操作还是数据恢复操作；

客户端包括逻辑卷管理器、iSCSI 和被保护磁盘三部分，iSCSI 用于将服务器端镜像区挂载到客户端，逻辑卷管理器用于将被保护磁盘与服务器端镜像区做成实时镜像，保证当有数据写入被保护磁盘时，相同的数据同步写入服务器端的镜像区。

2. 一种采用如权利要求 1 所述系统的块级别的磁盘数据保护方法，其特征在于，该方法包括磁盘数据备份和磁盘数据恢复两部分；该磁盘数据备份包括以下步骤：

11) 对服务器端进行初始化，包括对快照阀值赋值，以及对镜像区数据进行一次快照作为快照数据，将快照数据存入快照区，向快照信息表中添加一条记录，该记录的快照时间列为当前系统时间，该记录的快照地址列为当前备份区扇区起始地址，该记录的快照标示符列为一个全局随机数，用来标识该次快照；

12) 服务器端监听镜像区，并判断对于镜像区的 iSCSI 命令是进行写操作还是数据恢复操作，若是写操作，则将本次写操作暂停；

13) 服务器端将本次写操作的写入地址、数据长度、当前系统时间以及数据内容组织成一个备份数据单元；

14) 服务器端从内存中读取备份区扇区起始地址，并以备份区扇区起始地址为目标地址，将备份数据单元写入到备份区中；

15) 更新备份区扇区起始地址，新扇区起始地址为原扇区起始地址加上备份数据单元的长度；

16) 恢复本次写操作，使本次写操作写入镜像区；

17) 自上一次对镜像区做快照起，若备份区内增加的备份数据单元的个数大于快照阀值，对镜像区数据进行一次快照作为快照数据，将快照数据存入快照区，向快照信息表中添加一条记录，该记录的快照时间列为当前系统时间，该记录的快照地址列为当前备份区扇区起始地址，该记录的快照标示符列为一个全局随机数，用来标识该次快照；

该磁盘数据恢复包括以下步骤：

21) 服务器端自动监听镜像区，并判断对于镜像区的 iSCSI 命令是进行写操作还是数据恢复操作，若是数据恢复操作，则服务器端根据接收客户端提交的恢复时间点，查询快照信息表，找出所有快照时间中最临近的早于恢复时间点的一条快照记录，取出该条记录的快照标示符和快照地址；

22) 服务器端根据快照标示符从快照区取出快照数据，并将快照数据写入镜像区；

23) 服务器端以快照地址为起始地址，从备份区中依次向后取出备份数据单元，直到备份数据单元内的系统时间晚于客户端提交的恢复时间点，根据已取出的各个数据单元的扇

区起始地址、数据长度、数据内容将备份数据写入镜像区；

24）服务器端将镜像区数据返回到客户端中的被保护磁盘。

一种块级别的磁盘数据保护系统及其方法

技术领域

[0001] 本发明属于计算机网络技术和计算机数据存储技术领域，特别涉及一种块级别的磁盘数据保护系统及其方法。

背景技术

[0002] 当今世界，政府与各种企业的数据正经历着爆炸性的增长。由于互联网、电子邮件、以及社交网络的出现，以及越来越多、占用存储介质越来越多的各种应用软件所致，数据量呈现巨大的增长态势。来自 10 年 EMC 与 IDC 联合发布的数据，预计在 2020 年，全球的数据量会是 09 年数据量总和的 44 倍，其中个人数据（包括音频、视频、图形文件）占相当大的比重；而对于企业来说，大量数据的管理问题也是不容小觑的，IBM 在 12 年 5 月发布的报告说，企业数据正在以 55% 的速度逐年增长，如今只需两天就能创造出自文明诞生以来到 2003 年所产生的数据总量。因此对于任何组织和个人来说，数据丢失都会带来严重的后果。

[0003] 针对可能发生的数据丢失，数据保护技术应运而生，简单来说，就是提前把用户的数据备份到别处，若被保护数据发生损坏或丢失，再将备份数据写回到用户的设备中。目前，市场主流的存储系统都支持持续数据保护技术，该技术可以监听到用户对被保护数据的每次操作，并将变化的数据与发生变化的时间点保存在服务器上，若被保护数据发生丢失或损坏时，可根据服务器上之前保存的数据和对应的时间点快速恢复被保护数据。这类典型的系统如 linux 平台上多版本文件系统 EXT3COW 和 windows 平台上容灾备份系统 Hyperstor 等。但该类系统存在的问题如下：

[0004] 1、实现记录数据操作的监听模块，都是基于修改被保护数据所在操作系统内核实现的，这会提升该操作系统的不稳定性，同时加大监听模块程序开发和测试的难度。

[0005] 2、针对不同的操作系统平台（例如 linux 和 windows），由于内核 API 差别极大，需要开发出不同版本的监听模块；即使是同一种操作系统，由于内核升级导致内核 API 也会有较大变化，所以，针对特定版本内核设计的监听模块若要移植到另一个版本（例如从 linux 2.4 系统移植到 linux 2.6 系统），同样需要对监听模块的代码进行大量修改。

[0006] 3、数据恢复操作非常耗时，主要原因是备份软件在数据重组时需要扫描和读取备份在服务器上的大量的变化数据和他们对应的时间点信息，引起大量磁盘 I/O 操作。

发明内容

[0007] 本发明的目的是为克服已有技术的不足之处，提出一种块级别的磁盘数据保护系统及其方法，该系统通用并且稳定，系统开发和测试的难度低，该方法减少从备份区读取备份数据单位的数量，且节省恢复的时间。

[0008] 1、本发明提出的一种块级别的磁盘数据保护系统，其特征在于：该系统基于客户端 / 服务器架构，被保护磁盘位于客户端，服务器端存放备份数据；

[0009] 该服务器端包括内存及磁盘，该磁盘划分为三个逻辑分区，分别是镜像区、快照区和备份区，其中镜像区与客户端被保护磁盘互为实时镜像，当有数据写入被保护磁盘时，相

同的数据也会同步写入服务器端的镜像区；服务器端内存中存有快照信息表、备份区中的扇区起始地址、监听模块以及快照阀值；快照信息表中包含快照标示符、快照时间和快照地址三列，快照标示符是快照的标识，快照时间是做快照时的系统时间，快照地址为做快照时备份区的扇区起始地址；监听模块用于监听并判断对于镜像区的 iSCSI 命令是进行写操作还是数据恢复操作；

[0010] 客户端包括逻辑卷管理器、iSCSI 和被保护磁盘三部分，iSCSI 用于将服务器端镜像区挂载到客户端，逻辑卷管理器用于将被保护磁盘与服务器端镜像区做成实时镜像，保证当有数据写入被保护磁盘时，相同的数据同步写入服务器端的镜像区。

[0011] 本发明还提出采用如上述系统的块级别的磁盘数据保护方法，其特征在于，该方法包括磁盘数据备份和磁盘数据恢复两部分；该磁盘数据备份包括以下步骤：

[0012] 11）对服务器端进行初始化，包括对快照阀值赋值（快照阀值可由用户预先设置），以及对镜像区数据进行一次快照作为快照数据，将快照数据存入快照区，向快照信息表中添加一条记录，该记录的快照时间列为当前系统时间，该记录的快照地址列为当前备份区扇区起始地址，该记录的快照标示符列为一个全局随机数，用来标识该次快照；

[0013] 12）服务器端监听镜像区，并判断对于镜像区的 iSCSI 命令是进行写操作还是数据恢复操作，若是写操作，则将本次写操作暂停；

[0014] 13）服务器端将本次写操作的写入地址、数据长度（以扇区为单位）、当前系统时间以及数据内容组织成一个备份数据单元；

[0015] 14）服务器端从内存中读取备份区扇区起始地址，并以备份区扇区起始地址为目标地址，将备份数据单元写入到备份区中；

[0016] 15）更新备份区扇区起始地址，新扇区起始地址为原扇区起始地址加上备份数据单元的长度；

[0017] 16）恢复本次写操作，使本次写操作写入镜像区；

[0018] 17）自上一次对镜像区做快照起，若备份区内增加的备份数据单元的个数大于快照阀值，对镜像区数据进行一次快照作为快照数据，将快照数据存入快照区，向快照信息表中添加一条记录，该记录的快照时间列为当前系统时间，该记录的快照地址列为当前备份区扇区起始地址，该记录的快照标示符列为一个全局随机数，用来标识该次快照。

[0019] 该磁盘数据恢复包括以下步骤：

[0020] 21）服务器端自动监听镜像区，并判断对于镜像区的 iSCSI 命令是进行写操作还是数据恢复操作，若是数据恢复操作，则服务器端根据接收客户端提交的恢复时间点，查询快照信息表，找出所有快照时间中最临近的早于恢复时间点的一条快照记录，取出该条记录的快照标示符和快照地址；

[0021] 22）服务器端根据快照标示符从快照区取出快照数据，并将快照数据写入镜像区；

[0022] 23）服务器端以快照地址为起始地址，从备份区中依次向后取出备份数据单元，直到备份数据单元内的系统时间晚于客户端提交的恢复时间点，根据已取出的各个数据单元的扇区起始地址、数据长度、数据内容将备份数据写入镜像区；

[0023] 24）服务器端将镜像区数据返回到客户端被保护磁盘。

[0024] 本发明提出的一种块级别的磁盘数据保护系统及其方法，其优点是：

[0025] 1、平台通用：磁盘数据保护系统适用于多种操作系统平台，或者同一种操作系统的不同版本。

[0026] 2、系统稳定：磁盘数据保护系统的实现不用修改操作系统内核，保证了被保护数据所在操作系统的稳定性，同时也降低系统开发和测试的难度。

[0027] 3、快速恢复：数据恢复时，磁盘数据保护系统先从快照区读取相应的快照数据，减少从备份区读取备份数据单位的数量，节省恢复的时间。

附图说明

[0028] 图 1 是本发明磁盘数据保护系统组成示意图。

[0029] 图 2 是本发明磁盘数据保护系统快照信息表实施例结构示意图。

[0030] 图 3 是本发明的磁盘数据保护系统数据备份实施例流程框图。

[0031] 图 4 是本发明的磁盘数据保护系统备份数据单元实施例结构示意图。

[0032] 图 5 是本发明的磁盘数据保护系统数据恢复实施例流程框图。

具体实施方式

[0033] 本发明提出一种块级别的磁盘数据保护系统及其方法，结合附图及实施例详细说明如下：

[0034] 本发明提出一种块级别的磁盘数据保护系统，如图 1 所示，数据保护首先需要解决的问题是如何在系统中截取到每次写操作。如果在被保护的业务系统上截取，可能会影响业务系统自身的读写性能，本发明使用远程镜像技术来保障写操作的截取不会影像到业务系统自身的运作。

[0035] 本系统基于客户端 / 服务器架构，被保护磁盘位于客户端，服务器端存放备份数据。

[0036] 服务器端磁盘划分为三个逻辑分区，分别是镜像区、快照区和备份区，其中镜像区与客户端被保护磁盘互为实时镜像，当有数据写入被保护磁盘时，相同的数据也会同步写入服务器端的镜像区；服务器端内存中存有快照信息表、备份区中的扇区起始地址、监听模块以及快照阀值；快照信息表中包含快照标示符、快照时间和快照地址三列，快照标示符是快照的标识，快照时间是做快照时的系统时间，快照地址等于做快照时备份区的扇区起始地址；服务器端内存中设置有快照阀值，该值可由用户预先设置。

[0037] 客户端包括逻辑卷管理器、iSCSI 和被保护磁盘三部分，客户端使用逻辑卷管理器和 iSCSI 来实现被保护磁盘和服务器端镜像区的镜像。逻辑卷管理器用于实现磁盘镜像，但前提是互为镜像的两块磁盘都必须是属于本地操作系统的。iSCSI 可以将一个远程主机的磁盘空间，当作一个块设备来使来挂载到本地操作系统上。所以，本发明的做法是，将服务器端镜像区通过 iSCSI 挂载到客户端的操作系统下，再用逻辑卷管理器将被保护磁盘和镜像区做成镜像，写操作截取的动作在服务器端完成，不影响客户端系统自身的性能和稳定性。

[0038] 服务器端通过监听模块监控针对到镜像区的写操作，并将写入镜像区的写入地址、数据长度(扇区为单位)、当前系统时间及数据内容组织成一个备份数据单元，以内存中的备份区扇区起始地址为目标地址将备份数据单元写入到备份区中。在 linux 系统中，每

个存储设备都会在系统内核中注册一个 make_request 函数用来处理针对该设备的读写请求，当用户对该设备存储设备进行读写请求后，系统都会交给内核中该设备对应的驱动函数 make_request 进行读写处理，其中，make_request 函数中有一个参数叫做 bio，该参数内包含本次操作类型(读或写)、设备目标地址、读写长度以及要读写的数据内容。在监听模块中，本发明实现了拥有监控写功能的 make_request 函数，并将该函数注册给镜像区。当有针对镜像区的写操作时，make_request 函数首先暂停该写操作，并分析传入的 bio 参数，并解析出将写入镜像区的写入地址、数据长度(以扇区为单位)、当前系统时间及数据内容组织成一个备份数据单元，根据内存中的备份区扇区起始地址将备份数据单元写入到备份区中，之后更新备份区扇区起始地址，新的备份区扇区起始地址等于原扇区起始地址加上数据单元的长度，完成以上操作后，恢复该写入镜像区的操作。

[0039] 在数据保护的过程中，监听模块会不间断的监听镜像区的写操作，并将其组成备份数据单元存储至备份区。但当数据单元个数过多时会产生两个问题：

[0040] 1. 如果用户请求的恢复时间点与当前时间较远，服务器端需要取出这两个时间点间的大量备份数据单元，并将这些数据单元写回镜像区，这一过程会十分耗时；

[0041] 2. 若恢复时间点与当前时间之间的某个备份数据单元发生损坏，则该数据单元之前的备份数据都无意义，造成数据丢失。

[0042] 本发明采用快照技术来解决以上问题。服务器端初始化时会对镜像区数据做一次快照，将快照数据存入快照区，向快照信息表中添加一条记录，该记录的快照时间列为当前系统时间，该记录的快照地址列为当前备份区扇区起始地址，该记录的快照标示符列为一个全局随机数，用来标识该次快照，快照信息表结构如图 2 所示，该表中显示有 4 条快照记录，每条记录都描述了一次快照的相关信息，包括快照标示符、快照时间和快照地址。

[0043] 在服务器端运行过程中，监听模块会监测从镜像区上一次快照起备份区内备份数据单元新增的个数，若新增数大于快照阀值时，对镜像区数据做一次快照，将快照数据存入快照区，向快照信息表中添加一条记录，该记录的快照时间列为当前系统时间，该记录的快照地址列为当前备份区扇区起始地址，该记录的快照标示符列为一个全局随机数，用来标识该次快照。

[0044] 本发明的客户端和服务器端的实施例均采用普通商用计算机，客户端支持 windows 平台和 linux 平台，服务器端基于 linux 平台开发。

[0045] 客户端逻辑卷管理器采用 windows 和 linux 平台自身提供的磁盘镜像工具，其中 windows 平台上的磁盘镜像工具是磁盘管理控制台 (Windows Disk Management console)，linux 平台上的磁盘镜像工具是分布式块设备复制器(Distributed Replicated BlockDevice)。iSCSI 采用 windows 和 linux 平台自带的 iSCSI 软件程序。

[0046] 服务器端的监听模块采用 linux 内核模块编程技术(linux kernel moduleprogramming)实现，基于 C 语言开发。快照阀值的默认值是 1024，快照阀值可由用户更改。服务器端使用 linux 提供的 lvcreate 命令对镜像区数据做快照，并将快照数据存入快照区。

[0047] 本发明采用如上述系统的块级别的磁盘数据保护方法，其特征在于，该方法包括磁盘数据备份和磁盘数据恢复两部分；该磁盘数据备份，如图 3 所示包括以下流程：

[0048] 11)对服务器端进行初始化，包括对快照阀值赋值，以及对镜像区数据进行一次快

照作为快照数据，将快照数据存入快照区，向快照信息表中添加一条记录，该记录的快照时间列为当前系统时间，该记录的快照地址列为当前备份区扇区起始地址，该记录的快照标示符列为一个全局随机数，用来标识该次快照；

[0049] 12）服务器端监听镜像区，并判断对于镜像区的 iSCSI 命令是进行写操作还是数据恢复操作，若是写操作，则将本次写操作暂停；

[0050] 13）服务器端将本次写操作的写入地址、数据长度、当前系统时间以及数据内容组织成一个备份数据单元；

[0051] 14）服务器端从内存中读取备份区扇区起始地址，并以备份区扇区起始地址为目标地址，将备份数据单元写入到备份区中；

[0052] 15）更新备份区扇区起始地址，新扇区起始地址为原扇区起始地址加上备份数据单元的长度；

[0053] 16）恢复本次写操作，使本次写操作写入镜像区；

[0054] 17）自上一次对镜像区做快照起，若备份区内增加的备份数据单元的个数大于快照阀值，对镜像区数据进行一次快照作为快照数据，将快照数据存入快照区，向快照信息表中添加一条记录，该记录的快照时间列为当前系统时间，该记录的快照地址列为当前备份区扇区起始地址，该记录的快照标示符列为一个全局随机数，用来标识该次快照。

[0055] 本实施例的备份数据单元结构如图 4 所示，备份区内存放多个备份数据单元，每个备份数据单元包含写入地址、数据长度、系统时间数据内容 4 种信息。

[0056] 该磁盘数据恢复，如图 5 所示，包括以下流程：

[0057] 21）服务器端自动监听镜像区，并判断对于镜像区的 iSCSI 命令是进行写操作还是数据恢复操作，若是数据恢复操作，则服务器端根据接收客户端提交的恢复时间点，查询快照信息表，找出所有快照时间中最临近的早于恢复时间点的一条快照记录，取出该条记录的快照标示符和快照地址；

[0058] 22）服务器端根据快照标示符从快照区取出快照数据，并将快照数据写入镜像区；

[0059] 23）服务器端以快照地址为起始地址，从备份区中依次向后取出备份数据单元，直到备份数据单元内的系统时间晚于客户端提交的恢复时间点，根据已取出的各个数据单元的扇区起始地址、数据长度、数据内容将备份数据写入镜像区；

[0060] 24）服务器端将镜像区数据返回到客户端被保护磁盘。

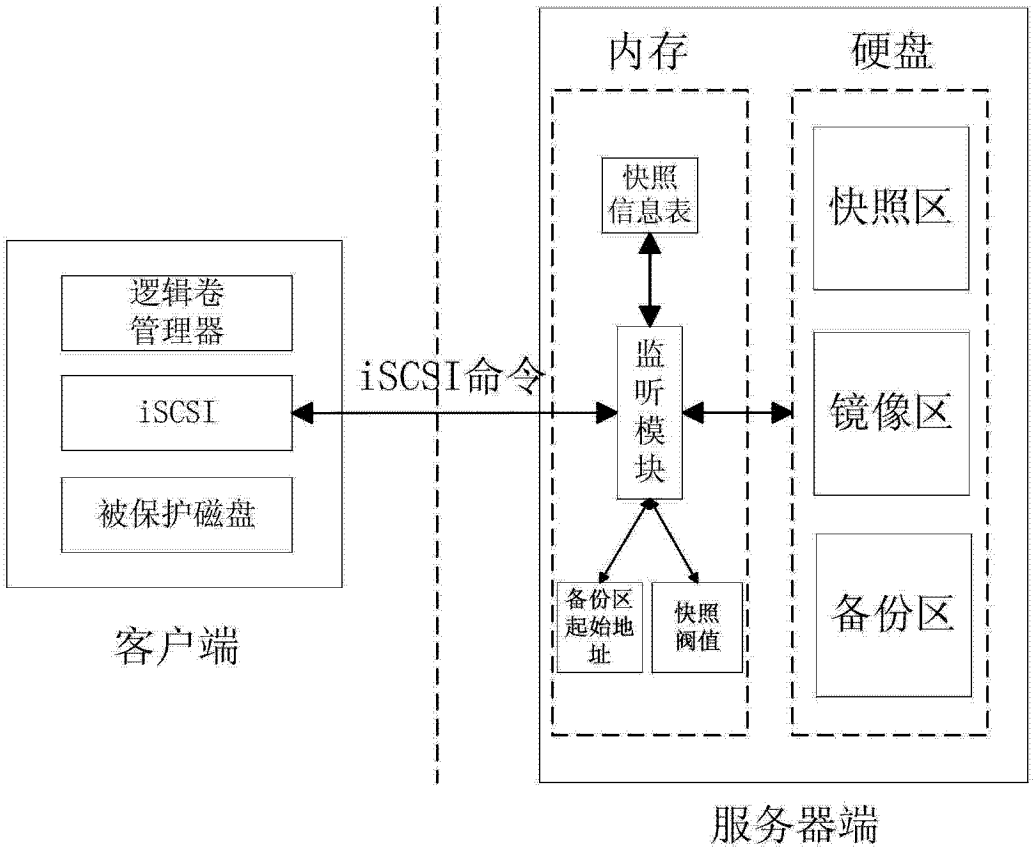


图 1

快照标示符	快照时间	快照地址
120822	20121013 17:23:00	1606500
23840E	20121013 19:02:10	2236358
352B34	20121013 23:31:36	2844145
6A8255	20121014 04:47:52	3410822
. . .		

图 2

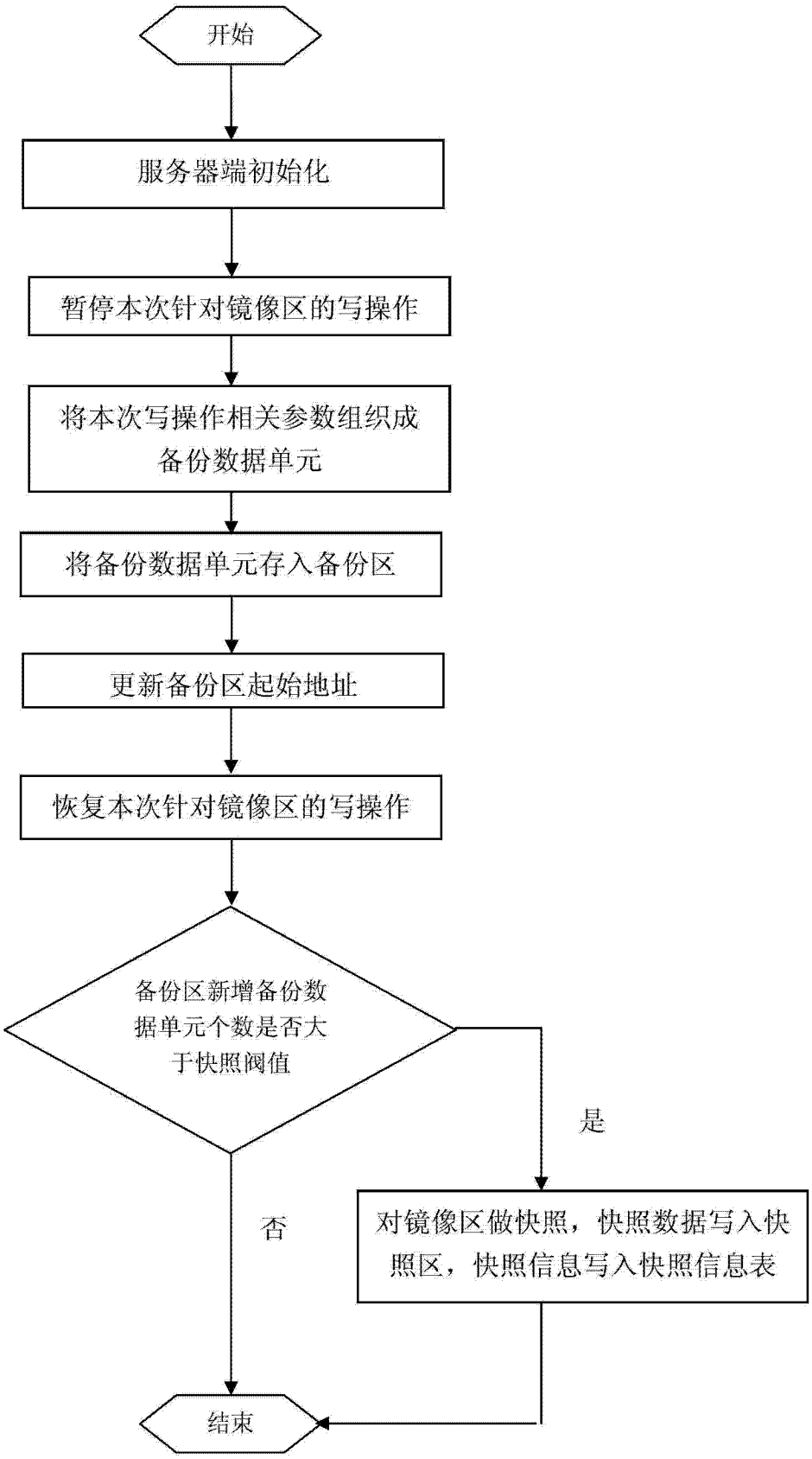

开始
服务器端初始化
暂停本次针对镜像区的写操作
将本次写操作相关参数组织成备份数据单元
将备份数据单元存入备份区
更新备份区起始地址
恢复本次针对镜像区的写操作
备份区新增备份数据单元个数是否大于快照阀值
是
否
对镜像区做快照，快照数据写入快照区，快照信息写入快照信息表
结束

图 3

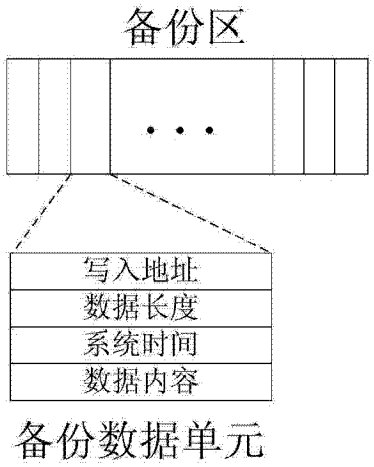

备份区
写入地址
数据长度
系统时间
数据内容
备份数据单元

图 4

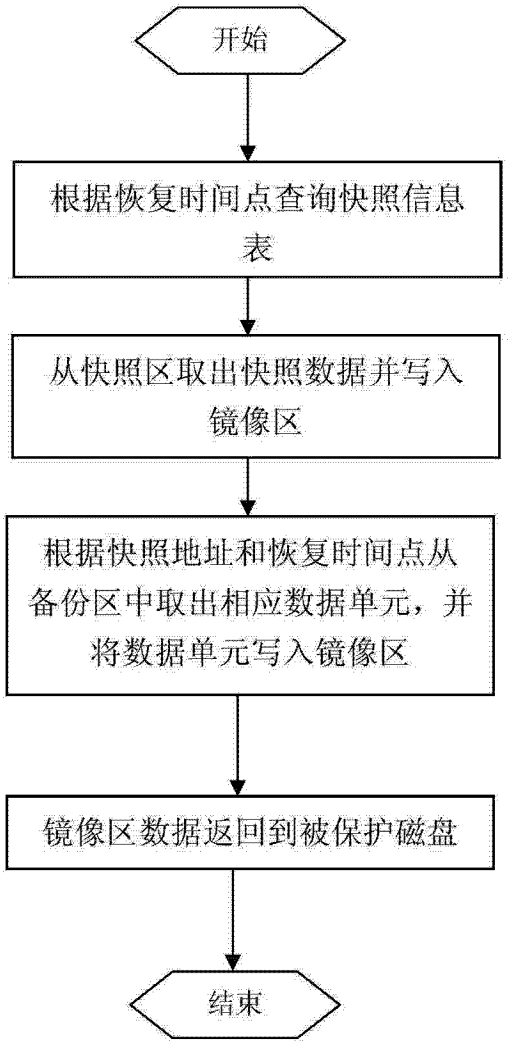

开始
根据恢复时间点查询快照信息表
从快照区取出快照数据并写入镜像区
根据快照地址和恢复时间点从备份区中取出相应数据单元，并将数据单元写入镜像区
镜像区数据返回到被保护磁盘
结束

图 5