(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 103365926 A
(43)申请公布日 2013.10.23

(21)申请号 201210103128.1

(22)申请日 2012.03.30

(71)申请人 伊姆西公司
地址 美国马萨诸塞州

(72)发明人 赵军平 谢纲 杨加林 齐巍 胡凤华

(74)专利代理机构 北京市金杜律师事务所 11256
代理人 王茂华

(51)Int.Cl.
G06F 17/30(2006.01)

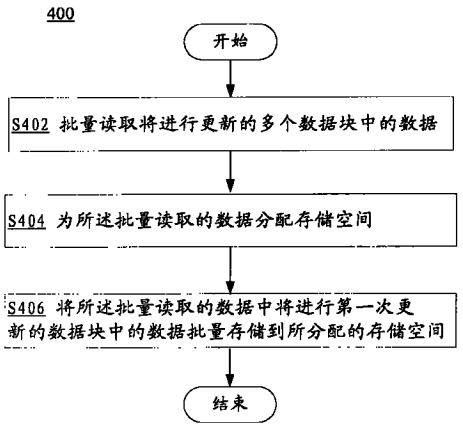
权利要求书3页 说明书10页 附图5页

(54)发明名称

在文件系统中用于保存快照的方法和装置

(57)摘要

本发明的实施方式涉及一种在文件系统中保存快照的方法和装置。该方法可以包括：批量读取将进行更新的多个数据块中的数据；为所述批量读取的数据分配存储空间；以及将所述批量读取的数据中将进行第一次更新的数据块中的数据批量存储到所分配的存储空间。根据本发明实施方式的方法，可以例如通过一次IO操作批量地保存数据快照，从而提高文件系统的效率。



CN 103365926 A

1. 一种在文件系统中保存快照的方法，包括：

批量读取将进行更新的多个数据块中的数据；

为所述批量读取的数据分配存储空间；以及

将所述批量读取的数据中将进行第一次更新的数据块中的数据批量存储到所分配的存储空间。

2. 根据权利要求 1 所述的方法，进一步包括：

在所述将进行第一次更新的数据块数目达到第一值、和 / 或最长相邻的所述将进行第一次更新的数据块数目达到第二值时，执行所述保存快照的方法，

其中，所述第一值和所述第二值为预定值或可在运行中调整。

3. 根据权利要求 2 所述的方法，其中所述第一值为使得所述将进行第一次更新的数据块数目占将进行更新的所述多个数据块数目的比例达到第一阈值的值；所述第二值为使得所述将进行第一次更新的数据块数目占将进行更新的所述多个数据块数目的比例达到第二阈值的值，并且所述第一阈值和所述第二阈值为预定值或可在运行中调整。

4. 根据权利要求 1-3 之任一所述的方法，进一步包括：

批量更新映射表，所述映射表记录了所述将进行第一次更新的数据块和所述将进行第一次更新的数据块中的数据在所述存储空间中的位置之间的对应关系。

5. 根据权利要求 4 所述的方法，其中，所述更新映射表包括仅通过一次读写操作更新所述映射表。

6. 根据权利要求 1-3 之任一所述的方法，进一步包括通过查找在存储器中存储的位图来获得所述将进行第一次更新的数据块的信息，其中所述信息包括所述将进行第一次更新的数据块的数目、分布、以及首尾位置中的一个或者多个。

7. 根据权利要求 1-3 之任一所述的方法，其中，将进行更新的所述多个数据块为连续分布，并且所述批量读取将进行更新的所述多个数据块中的数据包括通过一次读取操作读取连续分布的所述多个数据块中的数据。

8. 根据权利要求 1-3 之任一所述的方法，其中，所述批量读取的起始位置为将进行更新的所述多个数据块中第一个将进行第一次更新的数据块；和 / 或所述批量读取的结束位置为将进行更新的所述多个数据块中最后一个将进行第一次更新的数据块。

9. 根据权利要求 1-3 之任一所述的方法，其中，为所述批量读取的数据分配的所述存储空间包括一次性分配的连续存储空间，并且所述批量存储包括仅通过一次写入操作进行存储。

10. 根据权利要求 1-3 之任一所述的方法，其中，为所述批量读取的数据分配的所述存储空间大小相应于存储所述将进行第一次更新的数据块中的数据所需的存储空间大小，并且所述保存快照的方法进一步包括：

仅存储所述批量读取的数据中所述将进行第一次更新的数据块中的数据。

11. 根据权利要求 1-3 之任一所述的方法，其中，所述批量存储采用完全条带写的方式。

12. 根据权利要求 1-3 之任一所述的方法，其中，所述存储空间包括磁盘上的专用存储空间，并且所述批量存储的数据为快照。

13. 一种在文件系统中保存快照的方法，其中所述文件系统具有将进行第一次更新的

一段或多段连续的数据块，所述方法包括：

批量读取每段所述连续数据块中的数据；

为所述批量读取的数据分配存储空间；以及

将所述批量读取的数据批量存储到所分配的存储空间。

14. 根据权利要求 13 所述的方法，其中，并行地针对所述将进行第一次更新的一段或多段连续数据块，执行所述读取步骤、所述分配步骤以及所述存储步骤。

15. 一种在文件系统中保存快照的装置，包括：

读取装置，用于批量读取将进行更新的多个数据块中的数据；

分配装置，用于为所述批量读取的数据分配存储空间；以及

存储控制装置，用于将所述批量读取的数据中将进行第一次更新的数据块中的数据批量存储到所分配的存储空间。

16. 根据权利要求 15 所述的装置，进一步包括：

触发装置，用于在所述将进行第一次更新的数据块数目达到第一值、和 / 或最长相邻的所述将进行第一次更新的数据块数目达到第二值时，触发所述保存快照的装置的执行，

其中，所述第一值和所述第二值为预定值或可在运行中调整。

17. 根据权利要求 16 所述的装置，其中所述第一值为使得所述将进行第一次更新的数据块数目占将进行更新的所述多个数据块数目的比例达到第一阈值的值；所述第二值为使得所述将进行第一次更新的数据块数目占将进行更新的所述多个数据块数目的比例达到第二阈值的值，并且所述第一阈值和所述第二阈值为预定值或可在运行中调整。

18. 根据权利要求 15-17 之任一所述的装置，进一步包括：

更新装置，用于批量更新映射表，所述映射表记录了所述将进行第一次更新的数据块和所述将进行第一次更新的数据块中的数据在所述存储空间中的位置之间的对应关系。

19. 根据权利要求 18 所述的装置，其中，所述更新映射表包括仅通过一次读写操作更新所述映射表。

20. 根据权利要求 15-17 之任一所述的装置，进一步包括信息获取装置，用于通过查找在存储器中存储的位图来获得所述将进行第一次更新的数据块的信息，其中所述信息包括所述将进行第一次更新的数据块的数目、分布、以及首尾位置中的一个或者多个。

21. 根据权利要求 15-17 之任一所述的装置，其中，将进行更新的所述多个数据块为连续分布，并且所述批量读取将进行更新的所述多个数据块中的数据包括通过一次读取操作读取连续分布的所述多个数据块中的数据。

22. 根据权利要求 15-17 之任一所述的装置，其中，所述批量读取的起始位置为将进行更新的所述多个数据块中第一个将进行第一次更新的数据块；和 / 或所述批量读取的结束位置为将进行更新的所述多个数据块中最后一个将进行第一次更新的数据块。

23. 根据权利要求 15-17 之任一所述的装置，其中，为所述批量读取的数据分配的所述存储空间包括一次性分配的连续存储空间，并且所述批量存储包括仅通过一次写入操作进行存储。

24. 根据权利要求 15-17 之任一所述的装置，其中，为所述批量读取的数据分配的所述存储空间大小相应于存储所述将进行第一次更新的数据块中的数据所需的存储空间大小，并且所述存储控制装置进一步用于：

仅存储所述批量读取的数据中所述将进行第一次更新的数据块中的数据。

25. 根据权利要求 15-17 之任一所述的装置，其中，所述批量存储采用完全条带写的方式。

26. 根据权利要求 15-17 之任一所述的装置，其中，所述存储空间包括磁盘上的专用存储空间，并且所述批量存储的数据为快照。

27. 一种在文件系统中保存快照的装置，其中所述文件系统具有将进行第一次更新的一段或多段连续的数据块，所述装置包括：

读取装置，用于批量读取每段所述连续数据块中的数据；

分配装置，用于为所述批量读取的数据分配存储空间；以及

存储控制装置，用于将所述批量读取的数据批量存储到所分配的存储空间。

28. 根据权利要求 27 所述的装置，其中，并行地针对所述将进行第一次更新的一段或多段连续的数据块，启动所述读取装置、所述分配装置以及所述存储控制装置。

在文件系统中用于保存快照的方法和装置

技术领域

[0001] 本发明总体上涉及文件系统领域，更具体地，涉及在文件系统中用于保存快照的方法和装置。

背景技术

[0002] 快照 (Snapshot) 通常指生产文件系统 (Production File System，以下称为 PFS) 关于指定数据集合的一个完全可用拷贝，该拷贝包括相应数据在某个时间点（拷贝开始的时间点）的映像，其能够在存储设备发生逻辑错误或文件损坏的情况下进行快速的数据恢复，例如将数据恢复到某个可用的时间点的状态。快照的应用非常广泛，例如作为备份的源、作为数据挖掘的源、作为保存应用程序状态的检查点，甚至仅作为单纯的数据复制的一种手段等。创建快照的技术也有很多种，例如包括镜像分离、指针重映射、日志文件等等，其中，用于保护文件系统的较为经典的快照技术包括第一次写时复制快照。

[0003] 第一次写时复制 (copy on first write，以下称为 COFW) 是指在对数据块进行第一次写操作之前将原始内容复制并存储到专用存储中，并且更新追踪表以维护相应的映射。这种技术一般会在块级完成。如后文将详细描述的，在现有的 COFW 技术中，每个第一次更新（写）的数据块均必须各自经历一个完整的 COFW 周期，包括：读原始内容、分配新的存储区域、写入原始内容并最终更新追踪表（通常在硬盘上持久保存）。如后文将进一步说明的，这将产生大量琐碎的硬盘 I/O 并消耗大量的计算资源，并导致极大的性能负担并最终高度影响到正常的 PFS 操作。

[0004] 因此在本领域中，极需一种更为有效、开销更小的快照创建方案。案。

发明内容

[0005] 为了缓解现有技术中 COFW 快照存在的上述缺陷，本发明的实施方式提供一种改进、高效的、在文件系统中用于保存快照的方法和装置。

[0006] 根据本发明的一个实施方式，提供一种在文件系统中保存快照的方法，该方法可以包括：批量读取将进行更新的多个数据块中的数据；为所述批量读取的数据分配存储空间；以及将所述批量读取的数据中将进行第一次更新的数据块中的数据批量存储到所分配的存储空间。

[0007] 在本发明的可选实施方式中，所述方法进一步可以包括：在所述将进行第一次更新的数据块数目达到第一值、和 / 或最长相邻的所述将进行第一次更新的数据块数目达到第二值时，执行所述保存快照的方法，其中，所述第一值和所述第二值为预定值或可在运行中调整。

[0008] 在本发明的可选实施方式中，其中所述第一值为使得所述将进行第一次更新的数据块数目占将进行更新的所述多个数据块数目的比例达到第一阈值的值；所述第二值为使得所述将进行第一次更新的数据块数目占将进行更新的所述多个数据块数目的比例达到第二阈值的值，并且所述第一阈值和所述第二阈值为预定值或可在运行中调整。

[0009] 在本发明的可选实施方式中，所述方法进一步可以包括：批量更新映射表，所述映射表记录了所述将进行第一次更新的数据块和所述将进行第一次更新的数据块中的数据在所述存储空间中的位置之间的对应关系。

[0010] 在本发明的可选实施方式中，所述更新映射表包括仅通过一次读写操作更新所述映射表。

[0011] 在本发明的可选实施方式中，所述方法进一步包括通过查找在存存储器中存储的位图来获得所述将进行第一次更新的数据块的信息，其中所述信息包括所述将进行第一次更新的数据块的数目、分布、以及首尾位置中的一个或者多个。

[0012] 在本发明的可选实施方式中，将进行更新的所述多个数据块为连续分布，并且所述批量读取将进行更新的所述多个数据块中的数据包括通过一次读取操作读取连续分布的所述多个数据块中的数据。

[0013] 在本发明的可选实施方式中，所述批量读取的起始位置为将进行更新的所述多个数据块中第一个将进行第一次更新的数据块；和/或所述批量读取的结束位置为将进行更新的所述多个数据块中最后一个将进行第一次更新的数据块。

[0014] 在本发明的可选实施方式中，为所述批量读取的数据分配的所述存储空间包括一次性分配的连续存储空间，并且所述批量存储包括仅通过一次写入操作进行存储。

[0015] 在本发明的可选实施方式中，为所述批量读取的数据分配的所述存储空间大小相应于存储所述将进行第一次更新的数据块中的数据所需的存储空间大小，并且所述保存快照的方法进一步包括：仅存储所述批量读取的数据中所述将进行第一次更新的数据块中的数据之外的数据。

[0016] 在本发明的可选实施方式中，所述批量存储采用完全条带写的方式。

[0017] 在本发明的可选实施方式中，所述存储空间包括磁盘上的专用存储空间，并且所述批量存储的数据为快照。

[0018] 根据本发明的另一实施方式，提供一种在文件系统中保存快照的方法，其中所述文件系统具有将进行第一次更新的一段或多段连续的数据块，所述方法可以包括：批量读取每段所述连续数据块中的数据；为所述批量读取的数据分配存储空间；以及将所述批量读取的数据批量存储到所分配的存储空间。

[0019] 在本发明的可选实施方式中，并行地针对所述将进行第一次更新新的一段或多段连续数据块，执行所述读取步骤、所述分配步骤以及所述存储步骤。

[0020] 根据本发明的又一实施方式，提供一种在文件系统中保存快照的装置，包括：读取装置，用于批量读取将进行更新的多个数据块中的数据；分配装置，用于为所述批量读取的数据分配存储空间；以及存储控制装置，用于将所述批量读取的数据中将进行第一次更新的数据块中的数据批量存储到所分配的存储空间。

[0021] 在本发明的可选实施方式中，所述装置进一步包括：触发装置，用于在所述将进行第一次更新的数据块数目达到第一值、和/或最长相邻的所述将进行第一次更新的数据块数目达到第二值时，触发所述保存快照的装置的执行，其中，所述第一值和所述第二值为预定值或可在运行中调整。

[0022] 在本发明的可选实施方式中，所述第一值为使得所述将进行第一次更新的数据块数目占将进行更新的所述多个数据块数目的比例达到第一阈值的值；所述第二值为使得所

述将进行第一次更新的数据块数目占将进行更新的所述多个数据块数目的比例达到第二阈值的值，并且所述第一阈值和所述第二阈值为预定值或可在运行中调整。

[0023] 在本发明的可选实施方式中，所述装置进一步包括：更新装置，用于批量更新映射表，所述映射表记录了所述将进行第一次更新的数据块和所述将进行第一次更新的数据块中的数据在所述存储空间中的位置之间的对应关系。

[0024] 在本发明的可选实施方式中，所述更新映射表包括仅通过一次读写操作更新所述映射表。

[0025] 在本发明的可选实施方式中，所述装置进一步包括信息获取装置，用于通过查找在存储器中存储的位图来获得所述将进行第一次更新的数据块的信息，其中所述信息包括所述将进行第一次更新的数据块的数目、分布、以及首尾位置中的一个或者多个。

[0026] 在本发明的可选实施方式中，将进行更新的所述多个数据块为连连续分布，并且所述批量读取将进行更新的所述多个数据块中的数据包括通过一次读取操作读取连续分布的所述多个数据块中的数据。

[0027] 在本发明的可选实施方式中，所述批量读取的起始位置为将进行更新的所述多个数据块中第一个将进行第一次更新的数据块；和 / 或所述批量读取的结束位置为将进行更新的所述多个数据块中最后一个将进行第一次更新的数据块。

[0028] 在本发明的可选实施方式中，为所述批量读取的数据分配的所述存储空间包括一次性分配的连续存储空间，并且所述批量存储包括仅通过一次写入操作进行存储。

[0029] 在本发明的可选实施方式中，为所述批量读取的数据分配的所述存储空间大小相应于存储所述将进行第一次更新的数据块中的数据所需的存储空间大小，并且所述存储控制装置进一步用于：仅存储所述批量读取的数据中所述将进行第一次更新的数据块中的数据。

[0030] 在本发明的可选实施方式中，所述批量存储采用完全条带写的方式。

[0031] 在本发明的可选实施方式中，所述存储空间包括磁盘上的专用存储空间，并且所述批量存储的数据为快照。

[0032] 根据本发明的另一实施方式，提供一种在文件系统中保存快照的装置，其中所述文件系统具有将进行第一次更新的一段或多段连续的数据块，所述装置可以包括：读取装置，用于批量读取每段所述连续数据块中的数据；分配装置，用于为所述批量读取的数据分配存储空间；以及存储控制装置，用于将所述批量读取的数据批量存储到所分配的存储空间。

[0033] 在本发明的可选实施方式中，并行地针对所述将进行第一次更新的一段或多段连续的数据块，启动所述读取装置、所述分配装置以及所述存储控制装置。

附图说明

[0034] 通过参考附图阅读下文的详细描述，本发明实施方式的上述以及其他目的、特征和优点将变得明显。在附图中，以示例性而非限制性的方式示出了本发明的若干实施方式，其中相同的参考标号表示相同或相似的元素。

[0035] 图 1 示出了本发明可以实施于其中的虚拟文件系统架构的示例性图示；

[0036] 图 2 示出了根据现有技术的 COFW 快照过程的示意性图示；

[0037] 图 3A 和 3B 示出了根据现有技术的 COFW 快照过程的示意性缺陷图示；

[0038] 图 4 示出了根据本发明一个实施方式的在文件系统中保存数据快照的方法的流程图。

[0039] 图 5 示出了根据本发明另一实施方式的在文件系统中保存数据快照的方法的流程图。

[0040] 图 6A-6D 示出了根据本发明优选实施方式的在文件系统中保存数据快照的具体示例。

[0041] 图 7 示出了根据本发明又一实施方式的在文件系统中保存数据快照的装置的框图。

[0042] 图 8 示出了根据本发明另一实施方式的在文件系统中保存数据快照的装置的框图。

具体实施方式

[0043] 为更好地理解本发明，在此对本发明所可能采用的术语进行简要说明。要注意的是，在此的说明仅出于更全面地理解本发明而示出，并不作为对本发明任何方面的限制。本发明所称的“生产文件系统”(PFS) 指代用于生产环境并由快照保护的文件系统，既可读也可写。要注意的是，虽然本发明出于示例性目的示出 PFS，但本领域技术人员可以理解，根据本发明各个方面的方法和装置也可以应用于其他类型的文件系统。

[0044] 图 1 示出了本发明可以实施于其中的虚拟文件系统架构的示例性图示 100。如图所示，虚拟文件系统 101 总体而言包括 PFS 102 部分和快照 106 部分。PFS 102 通常由硬盘 104 上组织成的各 PFS 卷 (volume)103 构建而成。而快照 106 存储在快照存储空间 105 中，该存储空间 105 可以是诸如 SavVol 之类专门用于容纳快照的专用存储。在 PFS 卷 103 与快照存储空间 105 这一级之间发生 COFW 快照过程。

[0045] 如所已知的，图 1 所例示的文件系统通常以“高速缓存 (cached)”模式装配，其支持页高速缓存和缓冲器高速缓存机制以获得低延时和高带宽。在所述“高速缓存”模式中，脏数据块由文件系统 IO 以例如异步 (asyc) 方式进行冲刷。这里的术语“文件系统 IO”包括但不限于诸如 List IO 之类的非阻塞 IO，List IO 由 FSBN 排序，其是允许其他处理在传输完成之前继续的一种输入 / 输出处理形式。在上述冲刷期间，连续的脏数据块合并 (merge) 成例如由 FSBN 所排序的盘区 (extent)。本领域技术人员将理解，术语“盘区”指代文件系统中由 FSBN 所排序的连续范围的脏数据块，其中数据块的数目是可变的，一般由具体的文件系统一次 IO 所能访问的最大数据块的数目所决定。典型的，盘区中数据块的数目例如可以为 32 个。本领域技术人员还将理解，这里所称的“盘区”仅是出于更好地描述本发明的目的而示出，并不作为对文件系统的任何限制，用于实现本发明各个方面的文件系统完全可以不包括该“盘区”。

[0046] 为后文更好地理解本发明的各种优点，在此示出现有技术中典型的数据更新过程。在接收文件的更新 (写) 请求 ；之后，如果满足预定条件 (诸如已聚集 32 个脏数据块，或者到达存储器水印)，则生产文件系统开始处理写操作。此时数据块被合并，并且由 FSBN 排序，继而这些连续的脏数据块由例如 List IO 进行冲刷，在所述冲刷期间如有必要，则进行日志记录 (journallog)。在上述过程期间，快照监控 PFS 上的每一次写操作，对于 List

IO 中的每个数据块，进行如图 2 所示的根据现有技术的 COFW 快照过程 200 ：

[0047] 在步骤 S201，首先检查数据块是否是第一次更新，并且如果是，则进行步骤 S202，从 PFS 卷中读取原始内容。接下来，步骤 S203 为所读取的原始内容分配存储空间。作为示例，可以是从诸如 SavVol 之类的用于容纳快照的专用存储中分配例如槽 (slot) 来存储被快照的块。接着，步骤 S204，将原始内容写入所分配的例如 SavVol 的槽中。接下来步骤 S205，更新映射表以维护映射，所述映射表例如记录了被快照的块与其在诸如 SavVol 之类的快照存储空间上的位置之间的映射关系。作为示例，该映射表典型地为块映射表 (BlockMap)，并且采用 B+ 树 (B 树) 的方式进行组织。同样作为示例，该映射表也可以存储在诸如 SavVol 之类的快照存储空间中。此后，继续进行 PFS 写操作。最后，步骤 S206 周期性地将映射表冲刷到硬盘。

[0048] 如上所述，现有的数据更新过程将脏数据块组织成盘区内的可变向量，并且提供起始块地址 (例如假定为 N)，而快照筛选器将按每块 (例如 8KB) 切断了这些连续的块，这会导致性能问题。例如，图 3A 和 3B 详细示出了根据现有技术的 COFW 快照过程的示意性缺陷图示。如图 3A 所示，假定 PFS 文件系统的一次 List IO 涉及从块号为 N 至 N+31 的总共 32 个脏数据块，参考图 3B，现有技术的 COFW 快照过程尤其在密集型写和快照工作负载下将遭受如下巨大的性能问题，包括 ：

[0049] 1、存在大量小型 (例如 8KB) 块读取和块写入 IO 操作，这使得对硬盘或 SAN 存储产生巨大的压力，而另一方面，小型 IO 通常很难被优化，诸如写操作对于合并或级化 (staging) 而言要求更多的资源，并且由于存储一般为共享，从而预取可能也不能有效工作。

[0050] 2、频繁地分配诸如槽之类的存储快照的空间并且因而频繁地产生中断。

[0051] 3、频繁的映射表 (诸如块映射表) 更新，而众所周知由于用于数据一致性的内部锁机制，B 树的更新开销非常大。

[0052] 4、多个快照流 (或线程) 之间存在锁竞争，这是因为在系统中存在用于数据完整性和快速恢复的同步点，因此并行性劣化。

[0053] 有鉴于此，本发明提出一种改进的在文件系统中保存数据快照的方法和装置。图 4 示出了根据本发明一个实施方式的在文件系统中保存数据快照的方法的流程图 400。为更为清楚地理解本发明，以下结合图 6A- 图 6D 详细描述如图 4 所述的方法的流程图。图 6A- 图 6D 示出了根据本发明优选实施方式的在文件系统中保存数据快照的具体示例。在图 6A- 图 6D 中，假定盘区的长度为 10。本领域技术人员应理解，所述的具体示例仅为更为清楚地理解本发明而示出，并不作为对本发明的任何限制。

[0054] 如方法 400 所示出的，本发明开始之后，进行步骤 S402，批量读取将进行更新的多个数据块中的数据。所述读取操作典型地包括通过批量的、诸如 List IO 之类的 IO 操作来从 PFS 中读取。

[0055] 根据本发明的优选实施方式，将进行更新的所述多个数据块为连续分布，并且批量读取将进行更新的所述多个数据块中的数据包括通过一次读取操作读取连续分布的所述多个数据块 (例如 32 个连续数据块) 中的数据。

[0056] 根据本发明的优选实施方式，所述批量读取的起始位置为将进行更新的所述多个数据块中第一个将进行第一次更新的数据块，也即第一个将被快照的数据块。在图 6B 的示

例中，批量读取的起始位置例如可以是 PFS 中的块 #2。根据本发明的优选实施方式，所述批量读取的结束位置为将进行更新的所述多个数据块中最后一个将进行第一次更新的数据块，也即最后一个将被快照的数据块。在图 6B 的示例中例如为 PFS 中的块 #10，在图 6C 所示的示例中例如为 PFS 中的块 #9。在上述优选实施方式中，所读取的数据块的数目可以大于将进行第一次更新的数据块的数目。例如，在图 6B 的示例中，所读取的数据块可以优选地为从 PFS 中的块 #2 至块 #10，或者在图 6C 的示例中，所读取的数据块可以优选地为从 PFS 中的块 #2 至块 #9。

[0057] 接下来，过程前进至步骤 S404，为所述批量读取的数据块分配存储空间。根据本发明的优选实施方式，为所述批量读取的数据分配的所述存储空间包括一次性分配的连续存储空间。此外，根据本发明的优选实施方式，所述存储空间包括磁盘上的专用存储空间，诸如 SavVol 中的槽。根据本发明的另一优选实施方式，为所述批量读取的数据分配的所述存储空间大小相应于（例如等于）存储所述将进行第一次更新的数据块中的数据所需的存储空间大小。例如，在图 6A 的示例中，需要分配 8 个槽；而图 6B 的示例中，需要分配 7 个槽。所分配的存储空间（或槽）的地址可以例如保持在存储空间表（例如槽表）中供后续参考。所述存储空间表（例如槽表）可以例如采用链表实现。

[0058] 接着，过程前进至步骤 S406，将所述批量读取的数据块中将进行第一次更新的数据块中的数据批量存储到所分配的存储空间。根据本发明的优选实施方式，该步骤 S406 中的批量存储包括仅通过一次写入操作进行存储。例如，所述写入操作可以包括通过一个 List IO 之类的 IO 操作进行。要注意的是，如上文关于步骤 S402 描述的那样，有时所读取的数据块数目会大于第一次更新的数据块的数目，这时，根据本发明的优选实施方式，步骤 S406 中的批量存储将包括仅存储所述批量读取的数据中所述将进行第一次更新的数据块中的数据，而对其余的数据不进行存储。例如，在图 6A 的示例中，不写入块 #2 和块 #9 所对应的数据。此外，考虑到在后端存储中通常以 RAID 方式进行的配置，根据本发明的优选实施方式，所述批量存储可以采用完全条带写 (full stripe write) 的方式。根据本发明的另一优选实施方式，所述批量存储的数据为快照。至此，过程结束。

[0059] 要注意的是，根据本发明的各个方面在文件系统中保存快照的方法与现有技术中的保存快照的方法是兼容的，它们也可以共存并在运行中切换。因此，如果进一步考虑到快照性能和存储器开销之间的性能平衡，优选的，可以在进行如图 4 所示出方法 400 之前查看需要进行快照块的状态，据此判断采用本发明所述的在文件系统中保存数据快照的方法是否具有所要求的性能，从而决定是采用根据本发明的保存快照的方法还是传统的保存快照的方法。下面结合图 6A 和 6B 详细说明如何查看状态和进行所述判断。本领域技术人员员应知，如下详细说明的所述查看状态和所述判断均是为进一步优化本发明性能而作出的可选示例。其不是必须的。根据本发明的优选实施方式，可以通过查找在存储器中存储的位图来获得所述将进行第一次更新的数据块的信息，所述信息包括（例如盘区中）所述将进行第一次更新的数据块的数目、分布、以及首尾位置中的一个或多个。由于位图通常位于存储器中，因此查询速度将会非常快。根据本发明的另一优选实施方式，所述判断包括在步骤 S402 之前，进一步确定在所述将进行第一次更新的数据块数目达到第一值、和 / 或确定最长相邻的所述将进行第一次更新的数据块数目达到第二值时，执行如图 4 中方法 400 的步骤 S402-S406 所述的过程。根据本发明的优选实施方式，所述第一值和所述第二值为预定

值或可在运行中调整。

[0060] 作为进一步补充的又一优选实施方式，所述第一值为使得所述将进行第一次更新的数据块数目占将进行更新的所述多个数据块数目的比例达到第一阈值的值；所述第二值为使得所述将进行第一次更新的数据块数目占将进行更新的所述多个数据块数目的比例达到第二阈值的值。根据本发明的优选实施方式，所述第一阈值和所述第二阈值为预定值或可在运行中调整。

[0061] 作为本发明的优选实施方式，所述第一阈值可以为 80％。作为本发明的又一优选实施方式，所述第二阈值可以为 50％。例如，在图 6A 的示例中，盘区中存在 8 个第一次更新的数据块，其占盘区的总数据块比例达 80％，则根据本发明一个方面的策略，可以采用根据本发明的在文件系统中保存数据快照的方法。否则，如果仅有 20％的数据块是第一次更新的数据块，则作为一种选择，可以采用传统保存数据快照的方法。在图 6B 的示例中，盘区中存在 7 个第一次更新的数据块，则根据本发明一个方面的策略，即考虑将进行第一次更新的数据块数目所占盘区比例的这一策略，由于该比例未达 80％，因而需要采用传统保存数据快照的方法；然而，根据本发明的另一方面的策略，即考虑最长相邻的所述将进行第一次更新的数据块数目所占盘区比例的这一策略，由于该比例达到 60％，可以采用根据本发明的在文件系统中保存数据快照的方法。

[0062] 一旦确定使用根据本发明的用于保存文件快照的方法，本发明还可优选地提取快照表供后续参考，所述快照表例如可以采用链表的形式记录下（例如盘区中的）哪些数据块需要被快照的信息，这些信息例如可以是 PFS 块号。根据本发明进一步的优选实施方式，快照表中内的 PFS 块号可以依据 B 树的节点按升序排列（但可不相邻）。例如对于图 6A 所示的示例，在快照表中可以例如标注块 #1、块 #3- 块 #8 以及块 #10 作为将被快照的块。

[0063] 根据本发明进一步的优选实施方式，在图 4 中方法 400 所示的步骤 S406 之后，还可以包括批量更新映射表的步骤，所述映射表记录了所述将进行第一次更新的数据块和所述将进行第一次更新的数据块中的数据在所述存储空间中的位置之间的对应关系。作为本发明进一步的优选实施方式，可以基于所述快照表和所述存储空间表（槽表）来更新所述映射表。

[0064] 根据本发明的优选实施方式，所述更新映射表包括仅通过一次读写操作更新所述映射表。根据本发明的又一优选实施方式，所述映射表的更新可以遵循现有的锁机制。要注意的是，在所述映射表中的映射或关键字仍然在块级进行保持，因此根据本发明的方面的保存文件快照的方法可以与当前的设计相兼容。

[0065] 图 5 示出了根据本发明另一实施方式的在文件系统中保持数据快照的方法的流程图 500。

[0066] 与图 4 所示出的方法 400 的不同之处在于，图 5 所示的方法 500 针对的是文件系统中将进行第一次更新的一段或多段连续的数据块。针对这些数据块，在方法开始之后，步骤 S502，批量读取每段连续数据块中的数据。所述批量读取的具体方法例如如前文参照图 4 的方法 400 中步骤 S402 中所示出的，读取操作典型地包括通过批量的、诸如 List IO 之类的 IO 操作（例如一次 IO 操作）来从 PFS 中读取。结合图 6D 进行说明，在图 6D 所示的示例中，针对（例如盘区中）将进行第一次更新的连续数据块 #2- 块 #4，以及块 #7- 块 #10，分别进行根据本发明一个实施方式的如图 5 所示的方法 500。

[0067] 接下来，步骤前进到步骤 S504，为所述批量读取的数据分配存储空间。与图 4 的方法 400 中步骤 S404 的分配类似，步骤 S504 中所分配的存储空间包括磁盘上的专用存储空间，诸如 SavVol 中的槽。

[0068] 继而，步骤前进至步骤 S506，将所述批量读取的数据批量存储到所分配的存储空间。与前述步骤 S406 类似，步骤 S506 中的批量存储也包括仅通过一次写入操作（例如通过一个 List IO）进行存储。至此，过程结束。

[0069] 需要注意的是，根据本发明的优选实施方式，如图 5 所示的方法可以针对将进行第一次更新的一段或多段连续数据块并行地进行。

[0070] 本领域技术人员应理解，以上所描述的说明性流程图的每个框以及流程图中框的组合可以由计算机程序指令来执行。这些程序指令可以被提供至处理器以生产机器，从而使得所述指令在处理器上执行时创建用于实现一个或多个流程图框中所指定操作的装置。所述计算机程序指令可以由处理器执行以使得所述处理器执行一系列操作步骤来产生计算机实施的处理，以使得在处理器上执行的指令提供用于实现一个或多个流程图框中所指定操作的装置。所述计算机程序指令还可以使得流程图框中所示出的至少一些操作步骤并行执行。此外，诸如可能在多处理器计算机系统中出现的某些步骤还可以跨多个的处理器执行。此外，流程图图示中的一个或多个框或框的组合还可以在不背离本发明的范围或精神的情况下与其它框或框的组合同时执行，或者甚至以不同于所图示的顺序来执行。

[0071] 图 7 示出了根据本发明一个实施方式的在文件系统中保存数据快照的装置的框图。

[0072] 如图所示，装置 700 包括读取装置 701，配置用于批量读取将进行更新的多个数据块中的数据；分配装置 702，配置用于为所述批量读取的数据分配存储空间；以及存储控制装置 703，配置用于将所述批量读取的数据中将进行第一次更新的数据块中的数据批量存储到所分配的存储空间。

[0073] 在本发明的可选实施方式中，装置 700 进一步包括：触发装置 704，配置用于在所述将进行第一次更新的数据块数目达到第一值、和 / 或最长相邻的所述将进行第一次更新的数据块数目达到第二值时，触发所述保存快照的装置的执行。其中，所述第一值和所述第二值为预定值或可在运行中调整。

[0074] 在本发明的可选实施方式中，所述第一值为使得所述将进行第一次更新的数据块数目占将进行更新的所述多个数据块数目的比例达到第一阈值的值；所述第二值为使得所述将进行第一次更新的数据块数目占将进行更新的所述多个数据块数目的比例达到第二阈值的值，并且所述第一阈值和所述第二阈值为预定值或可在运行中调整。

[0075] 在本发明的可选实施方式中，所述装置 700 进一步包括更新装置 705，配置用于批量更新映射表，所述映射表记录了所述将进行第一次更新的数据块和所述将进行第一次更新的数据块中的数据在所述存储空间中的位置之间的对应关系。

[0076] 在本发明的可选实施方式中，所述更新映射表包括仅通过一次读写操作更新所述映射表。

[0077] 在本发明的可选实施方式中，所述装置 700 进一步包括信息获取装置 706，配置用于通过查找在存储器中存储的位图来获得所述将进行第一次更新的数据块的信息，其中所述信息包括所述将进行第一次更新的数据块的数目、分布、以及首尾位置中的一个或多个。

[0078] 在本发明的可选实施方式中，将进行更新的所述多个数据块为连续分布，并且所述批量读取将进行更新的所述多个数据块中的数据包括通过一次读取操作读取连续分布的所述多个数据块中的数据。

[0079] 在本发明的可选实施方式中，所述批量读取的起始位置为将进行更新的所述多个数据块中第一个将进行第一次更新的数据块；和/或所述批量读取的结束位置为将进行更新的所述多个数据块中最后一个将进行第一次更新的数据块。

[0080] 在本发明的可选实施方式中，为所述批量读取的数据分配的所述存储空间包括一次性分配的连续存储空间，并且所述批量存储包括仅通过一次写入操作进行存储。

[0081] 在本发明的可选实施方式中，为所述批量读取的数据分配的所述存储空间大小相应于存储所述将进行第一次更新的数据块中的数据所需的存储空间大小，并且所述存储控制装置进一步配置用于：仅存储所述批量读取的数据中所述将进行第一次更新的数据块中的数据。

[0082] 在本发明的可选实施方式中，所述批量存储采用完全条带写的方式。

[0083] 在本发明的可选实施方式中，所述存储空间包括磁盘上的专用存储空间，并且所述批量存储的数据为快照。

[0084] 图 8 示出了根据本发明另一实施方式的在文件系统中保存数据快照的装置的框图。

[0085] 如图所示，装置 800 配置为针对文件系统中将进行第一次更新的一段或多段连续的数据块，包括读取装置 801，配置用于批量读取每段所述连续数据块中的数据；分配装置 802，配置用于为所述批量读取的数据分配存储空间；以及存储控制装置 803，配置用于将所述批量读取的数据批量存储到所分配的存储空间。

[0086] 在本发明的可选实施方式中，并行地针对所述将进行第一次更新的一段或多段连续数据块，启动所述读取装置 801、所述分配装置 802 以及所述存储控制装置 803。

[0087] 应当注意，尽管在上文详细描述中提及了设备的若干装置或子装置，但是这种划分仅仅并非强制性的。实际上，根据本发明的实施方式，上文描述的两个或更多装置的特征和功能可以在一个装置中具体化。反之，上文描述的一个装置的特征和功能可以进一步划分为由多个装置来具体化。

[0088] 特别地，除硬件实施方式之外，本发明的实施方式还可以通过计算机程序产品的形式实现。例如，参考图 4 和图 5 描述的方法 400 和 500 可以通过计算机程序产品来实现。该计算机程序产品可以存储在例如 RAM、ROM、硬盘和/或任何适当的存储介质中，或者通过网络从适当的位置下载到计算机系统上。计算机程序产品可以包括计算机代码部分，其包括可由适当的处理设备（例如，中央处理单元 CPU）执行的程序指令。所述程序指令至少可以包括：用于批量读取将进行更新的多个数据块中的数据的指令，为所述批量读取的数据分配存储空间的指令，以及将所述批量读取的数据中将进行第一次更新的数据块中的数据批量存储到所分配的存储空间的指令。

[0089] 上文已经结合若干具体实施方式阐释了本发明的精神和原理。以下将结合 COFW 的特点，对根据本发明的各种实施方式的在文件系统中保存快照的方法的诸多优点加以描述。

[0090] 由于 COFW 通常采用高速缓存模式和 List IO，PFS 默认按照高速缓存模式进行更

新（高速缓存写操作一般为NAS服务器的默认模式），而在高速缓存模式中数据块往往在硬盘上为连续的，这样，利用根据本发明的各种实施方式将会将会易于结合批量I/O操作并获得非常良好的性能。同样，由于COFW的本地规则，即通常本地地分配和修改相邻块，高速缓存模式也将有助于进行随机更新。

[0091] 更为具体而言，根据本发明的实施方式，将PFS读和SavVol写分别合并为（例如单个）批量IO操作，这显著地降低了从硬盘读取或向硬盘写入的IO数量（典型地为32:1，具体取决于文件系统的设置，例如取决于文件系统一次IO所能访问的数据块的数目），并且通过适当地支持提前读和完全条带写而提高了IO性能。与此同时，显著地降低了快照存储空间分配和映射表更新的函数调用（多达32:1）。并在多个快照服务线程之间的锁竞争以及因而支持更多写入流（即向不同文件）并行运行。特别的，本发明尤其对于顺序写和快照而言能够获得更好的多的快照性能。同时，使用检测更新模式和自动切换快照的方法的策略更为灵活地在性能和存储器消耗之间进行平衡。而且，本发明与当前的快照可以兼容，实际上，本发明的快照方式与现有技术中的快照方式可以共存并在运行中切换。此外，由于几乎所有变化都在存储器结构/逻辑中发生，因此本发明还易于实现。

[0092] 虽然已经参考若干具体实施方式描述了本发明，但是应该理解，本发明并不限于所公开的具体实施方式。本发明旨在涵盖所附权利要求的精神和范围内所包括的各种修改和等同布置。所附权利要求的范围符合最宽泛的解释，从而包含所有这样的修改及等同结构和功能。

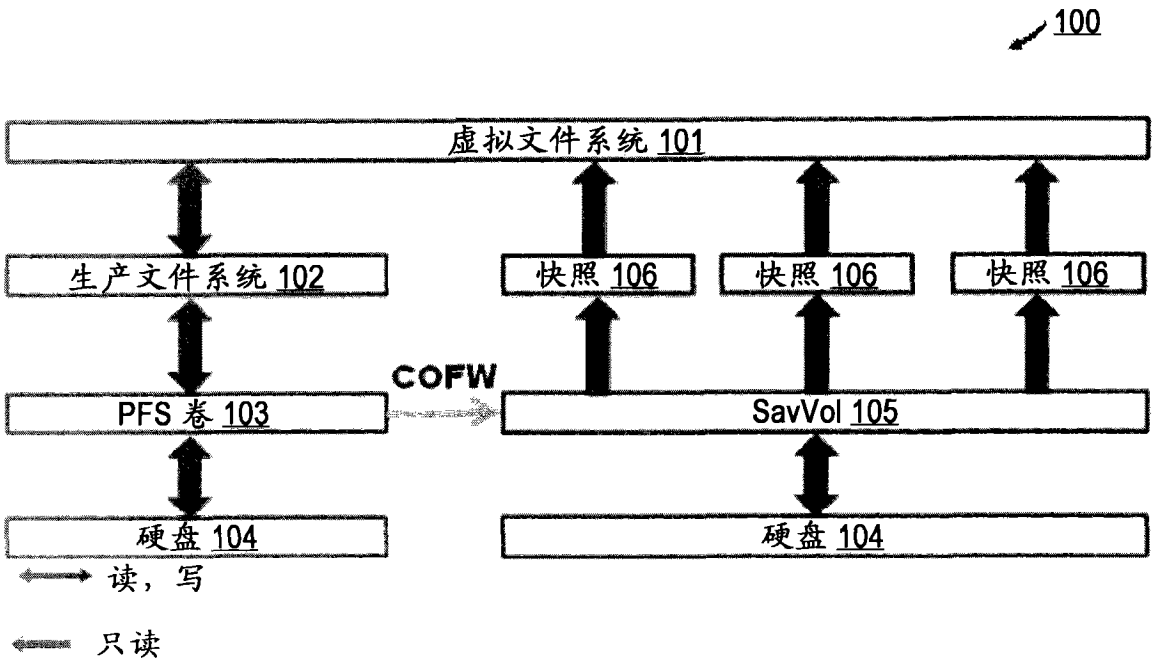


图 1

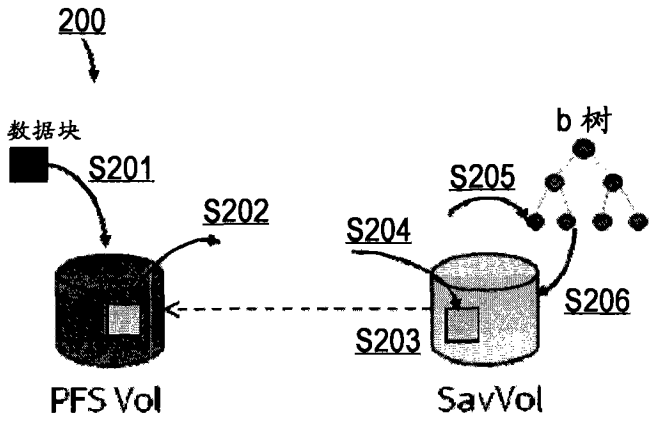
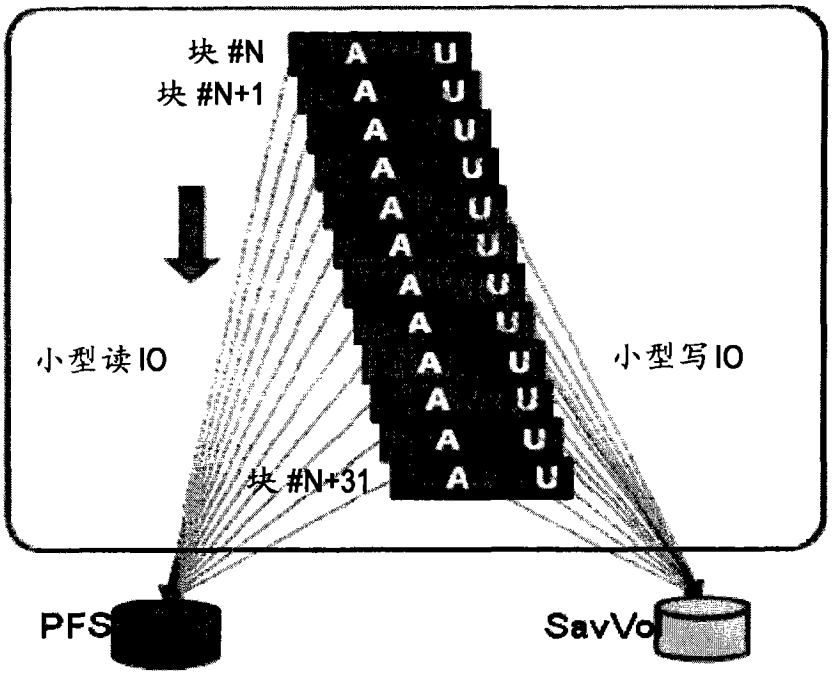


图 2

块 #N	N+1	N+2	N+3	...	N+30	N+31

图 3A



R: 从 PFS 卷中读原始拷贝
A: 从 SavVol 中分配新槽
W: 将原始拷贝写入槽中
U: 更新映射表，以及更新位图

图 3B

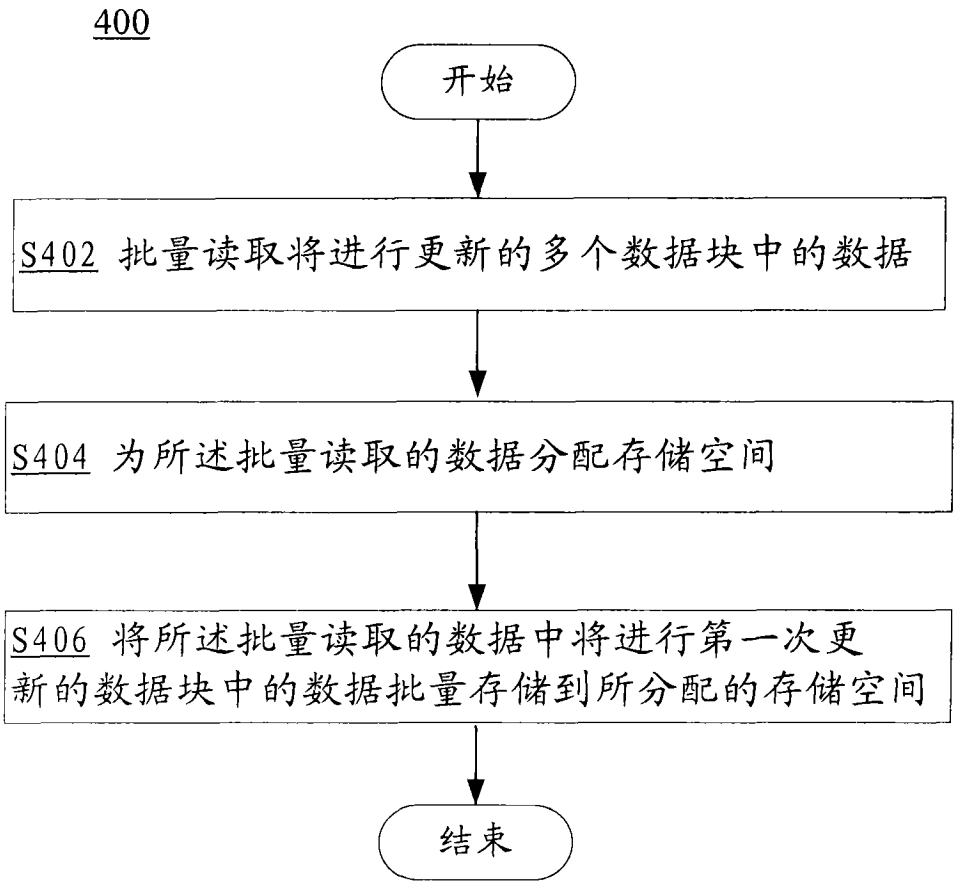

400
开始
S402 批量读取将进行更新的多个数据块中的数据
S404 为所述批量读取的数据分配存储空间
S406 将所述批量读取的数据中将进行第一次更新的数据块中的数据批量存储到所分配的存储空间
结束

图 4

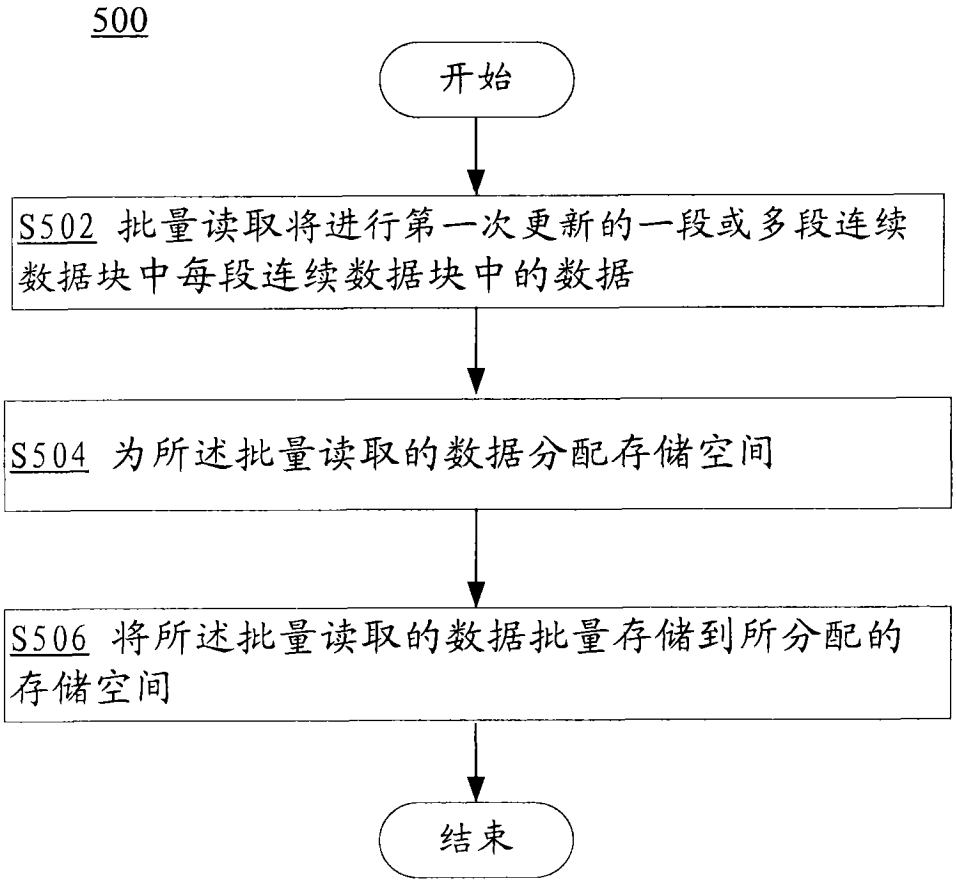


图 5

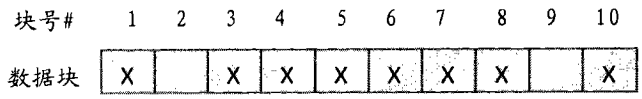


图 6A

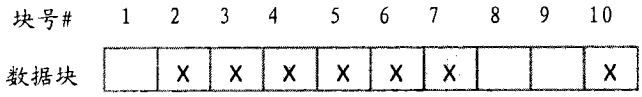


图 6B

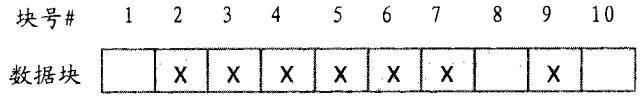


图 6C

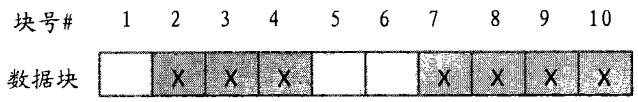

块号# 1 2 3 4 5 6 7 8 9 10
数据块 X X X X X X X X

图 6D

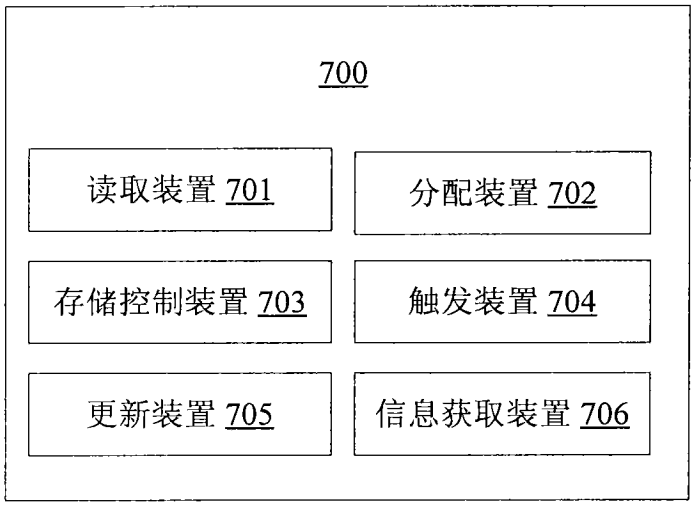

700
读取装置 701
分配装置 702
存储控制装置 703
触发装置 704
更新装置 705
信息获取装置 706

图 7

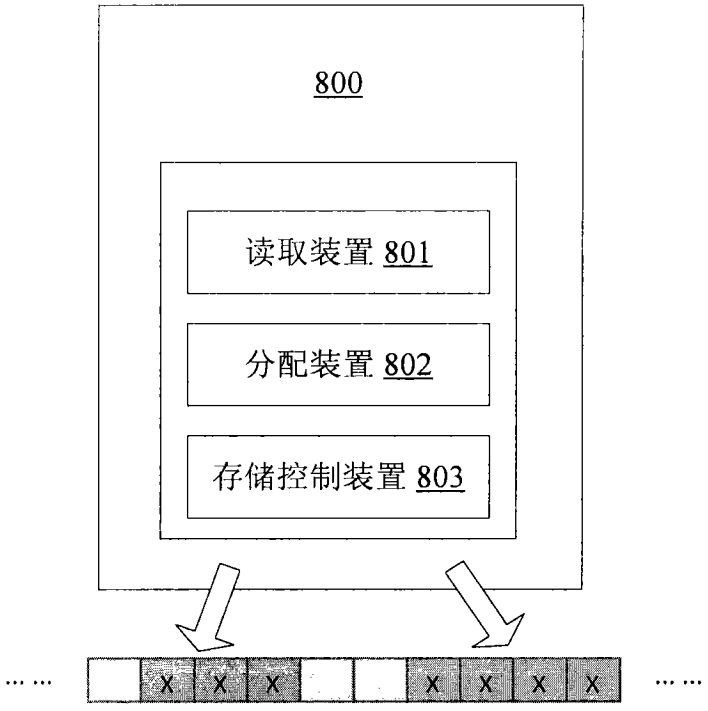

800
读取装置 801
分配装置 802
存储控制装置 803
…… X X X X X X X X ……

图 8