(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 103827806 B
(45)授权公告日 2016.03.30

(21)申请号 201280046541.7

(22)申请日 2012.08.11

(30)优先权数据
13/208,094 2011.08.11 US

(85)PCT国际申请进入国家阶段日
2014.03.25

(86)PCT国际申请的申请数据
PCT/US2012/050490 2012.08.11

(87)PCT国际申请的公布数据
WO2013/023200 EN 2013.02.14

(73)专利权人 净睿存储股份有限公司
地址 美国加利福尼亚

(72)发明人 J·科尔格洛夫 J·海斯 E·米勒 王锋

(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 叶勇

(51)Int.Cl.
G06F 3/06(2006.01)
G06F 11/14(2006.01)

(56)对比文件
US 2002/0087544 A1,2002.07.04,
US 7031971 B1,2006.04.18,
US 2010/0153620 A1,2010.06.17,

审查员 武晓冬

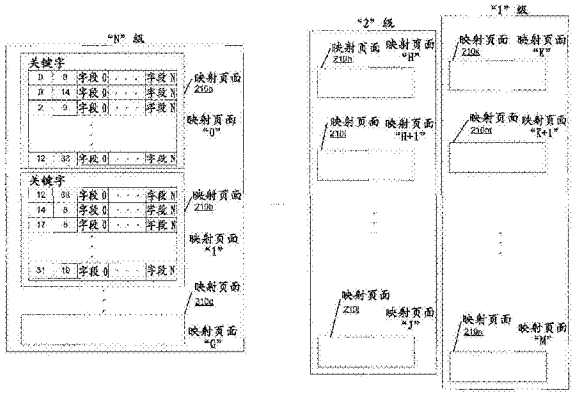
权利要求书3页 说明书23页 附图16页

(54)发明名称

闪存阵列中的逻辑扇区映射

(57)摘要

对包括多台固态存储设备的存储系统中存储的数据高效地执行用户存储器虚拟化的系统和方法。数据存储子系统支持多个映射表。在映射表内的记录被安排在多个等级中。每个等级都存储着关键字值与指针值的配对。等级按时间分选。新记录被插入在创建的最新的(最年轻的)等级中。不执行原地编辑。全部等级除最年轻等级外可以是只读的。系统可以进一步包括覆盖表,它标识映射表内无效的关键字。



CN 103827806 B

1. 一种计算机系统，包括：

数据存储介质；

数据存储控制器，连接到数据存储介质；以及

映射表单元，存储有映射表，该映射表包括多个条目，每个映射表条目都包含元组，该元组内含关键字；

其中，映射表单元把映射表组织为包括多个按时间排序的等级的层次，使得所述多个按时间排序的等级中的较年轻的等级与所述多个按时间排序的等级中的较老的等级相比在所述层次中更高地出现，每个等级都包括一个或多个映射表条目；

其中，数据存储控制器被配置成：

识别具有重复关键字的两个或更多个等级的组；并且

对所述两个或更多个等级的组产生要被加入到所述多个按时间排序的等级的新的单一等级，该新的单一等级包括所述两个或更多个等级的组内最年轻的记录。

2. 根据权利要求 1 的计算机系统，进一步包括高速缓存，被配置为存储映射表的至少一部分的高速缓存副本。

3. 根据权利要求 1 的计算机系统，其中，响应于收到请求，数据存储控制器被进一步配置为：

以与请求对应的至少给定关键字访问映射表；

识别一个或多个条目，它们对应于给定关键字；

判断一个或多个条目中哪一个对应于按时间排序等级中最年轻的等级；以及

返回与按时间排序等级中最年轻的等级对应的条目以满足请求。

4. 根据权利要求 3 的计算机系统，其中，控制器被配置为根据元组的关键字值分选每个等级。

5. 根据权利要求 4 的计算机系统，其中，存储介质中存储的映射表的条目被分组为若干页面，以及查询的结果被用于检索这些页面中的特定页面。

6. 根据权利要求 5 的计算机系统，其中，响应于收到特定页面，控制器被配置为使用给定关键字识别特定页面内的映射，该映射包括在存储介质中存储的与给定关键字对应的数据项的位置的标识。

7. 根据权利要求 1 的计算机系统，其中，数据存储控制器被进一步配置为：

把一个或多个新条目插入到新等级中。

8. 根据权利要求 1 的计算机系统，其中，每个映射表条目进一步包括关于与给定关键字对应的用户数据在数据存储介质上的存储位置的指示。

9. 根据权利要求 8 的计算机系统，其中，数据存储介质包括一台或多台固态存储设备。

10. 根据权利要求 1 的计算机系统，其中，所述多个按时间排序等级中除最年轻等级之外的全部等级都是只读的。

11. 根据权利要求 1 的计算机系统，进一步包括存储有覆盖映射表的覆盖表的覆盖表单元，其中，覆盖表和映射表由不同的关键字索引。

12. 根据权利要求 11 的计算机系统，其中，覆盖表标识映射表中无效的一个或多个元组。

13. 根据权利要求 11 的计算机系统，其中，覆盖表包含与某数值范围对应的一个或多

个条目。

14. 根据权利要求 11 的计算机系统，其中，覆盖表中的条目能够被用于修改从映射表返回的元组中的一个或多个字段以响应查询。

15. 根据权利要求 11 的计算机系统，其中，索引覆盖表所用的关键字从访问映射表所用的关键字的字段中产生。

16. 根据权利要求 11 的计算机系统，其中，索引覆盖表所用的关键字从访问映射表产生的元组中的字段中产生。

17. 根据权利要求 11 的计算机系统，进一步包括高速缓存，被配置为存储覆盖表的至少一部分的高速缓存副本。

18. 根据权利要求 11 的计算机系统，其中，覆盖表被组织为多个按时间排序的等级，每个等级都包括一个或多个覆盖表条目。

19. 一种在存储系统中使用的方法，所述方法包括 ：

存储映射表，该映射表包括多个条目，每个映射表条目都包含元组，该元组内含关键字 ；

存储映射表索引，该映射表索引包括多个条目 ；

其中，映射表被组织为包括多个按时间排序的等级的层次，使得所述多个按时间排序的等级中的较年轻的等级与所述多个按时间排序的等级中的较老的等级相比在所述层次中更高地出现，每个等级都包括一个或多个映射表条目 ；

识别具有重复关键字的两个或更多个等级的组 ；以及

对所述两个或更多个等级的组产生要被加入到所述多个按时间排序的等级的新的单一等级，该新的单一等级包括所述两个或更多个等级的组内最年轻的记录。

20. 根据权利要求 19 的方法，其中，响应于收到特定页面，该方法进一步包括使用访问映射表所用的给定关键字识别特定页面内的映射，该映射包括在存储介质中存储的与给定关键字对应的数据项的位置的标识。

21. 根据权利要求 20 的方法，进一步包括存储映射表的至少一部分的高速缓存副本。

22. 根据权利要求 20 的方法，其中，响应于收到请求，方法进一步包括 ：

以与请求对应的至少给定关键字访问映射表 ；

识别一个或多个条目，它们对应于给定关键字 ；

判断一个或多个条目中哪一个对应于按时间排序等级中最年轻的等级 ；以及

返回与按时间排序等级中最年轻的等级对应的条目以满足请求。

23. 根据权利要求 22 的方法，其中，进一步包括根据元组的关键字值分选每个等级。

24. 根据权利要求 23 的方法，其中，存储介质中存储的映射表的条目被分组为若干页面，以及查询的结果被用于检索这些页面中的特定页面。

25. 根据权利要求 19 的方法，进一步包括 ：

把一个或多个新条目插入到新等级中。

26. 根据权利要求 19 的方法，其中，每个映射表条目进一步包括关于与给定关键字对应的用户数据在数据存储介质上的存储位置的指示。

27. 根据权利要求 26 的方法，其中，数据存储介质包括一台或多台固态存储设备。

28. 根据权利要求 19 的方法，其中，所述多个按时间排序等级中除最年轻等级之外的

全部等级都是只读的。

29. 根据权利要求 19 的方法，进一步包括覆盖映射表的覆盖表，其中，覆盖表和映射表由不同的关键字索引。

30. 根据权利要求 29 的方法，其中，覆盖表标识映射表中无效的一个或多个元组。

31. 根据权利要求 29 的方法，其中，覆盖表包含与某数值范围对应的一个或多个条目。

32. 根据权利要求 29 的方法，进一步包括使用覆盖表中的条目修改从映射表返回的元组中的一个或多个字段以响应查询。

33. 根据权利要求 29 的方法，其中，索引覆盖表所用的关键字从访问映射表所用的关键字的字段中产生。

34. 根据权利要求 29 的方法，其中，索引覆盖表所用的关键字从访问映射表产生的元组中的字段中产生。

35. 一种在存储系统中使用的设备，所述设备包括 ：

用于存储映射表的装置，该映射表包括多个条目，每个映射表条目都包含元组，该元组内含关键字 ；

用于存储映射表索引的装置，该映射表索引包括多个条目 ；

其中，映射表被组织为包括多个按时间排序的等级的层次，使得所述多个按时间排序的等级中的较年轻的等级与所述多个按时间排序的等级中的较老的等级相比在所述层次中更高地出现，每个等级都包括多个映射表条目 ；

用于识别具有重复关键字的两个或更多个等级的组的装置 ；以及

用于对所述两个或更多个等级的组产生要被加入到所述多个按时间排序的等级的新的单一等级的装置，该新的单一等级包括所述两个或更多个等级的组内最年轻的记录。

36. 根据权利要求 35 的设备，进一步包括用于在高速缓存中存储映射表索引的至少一部分的高速缓存副本的装置。

37. 根据权利要求 35 的设备，进一步包括用于根据元组的关键字值分选高速缓存副本的装置。

闪存阵列中的逻辑扇区映射

技术领域

[0001] 本发明涉及计算机网络，更确切地说，涉及对于多台固态存储设备当中存储的数据高效地执行用户存储器虚拟化。

背景技术

[0002] 随着计算机内存存储和数据带宽的增加，商家日常管理的数据的量和复杂度也在增加。大规模的分布式存储系统，比如数据中心，典型情况下运行许多商务操作。数据中心也可以被称为服务器机房，是中心化的储存库，或者是物理的或者是虚拟的，用于对属于一个或多个商家的数据进行存储、管理和传播。分布式存储系统可以连接到由一个或多个网络互连的客户计算机。如果分布式存储系统的任何部分性能不佳，公司运转就可能受损。所以分布式存储系统对于数据可用性和高性能功能保持高标准。

[0003] 分布式存储系统包括若干物理卷，它们可以是硬盘、固态设备、使用另一种存储技术的存储设备或存储设备的分区。软件应用程序，比如逻辑卷管理器或磁盘阵列管理器，提供了对海量存储阵列分配空间的工具。此外，这种软件允许系统管理员创建存储组的单元，包括逻辑卷。存储器虚拟化实现了从物理存储器抽象（分离）出逻辑存储器，以便访问逻辑存储器而终端用户不必识别物理存储器。

[0004] 为了支持存储器虚拟化，卷管理器执行输入 / 输出（I/O）重定向，方式为把终端用户使用逻辑地址送来的 I/O 请求转换为使用与存储设备中物理位置相关联的地址的新请求。由于某些存储设备可以包括附加的地址转换机制，比如可用于固态存储设备的地址转换层，所以上述从逻辑地址到另一种地址的转换可能不表示唯一或最后的地址转换。重定向利用了在一个或多个映射表中存储的元数据。此外，在一个或多个映射表中存储的信息可用于存储器去重复以及把特定快照级的虚拟扇区映射到物理位置。卷管理器可以保持虚拟化存储器的映射信息视图一致。不过，所支持的地址空间可能受限于保持映射表所用的存储容量。

[0005] 与选中的存储磁盘相关联的技术和机制确定了卷管理器使用的方法。例如，提供硬盘、硬盘分区或外部存储设备的逻辑单元号（LUN）粒度级映射的卷管理器受限于大块数据的重定向、定位、去除重复数据等。另一种类型存储盘的一个实例是固态盘（SSD）。SSD 可以仿真 HDD 接口，但是 SSD 利用固态内存存储永久数据而不是 HDD 中发现的机电设备。例如，SSD 可以包括闪存的存储体。所以，在包括 SSD 存储器同时使用为 HDD 开发的映射表分配算法的系统中可能实现不了由一个或多个映射表支持的大地址空间。

[0006] 考虑以上情况，期望的是对多台固态存储设备当中存储的数据高效地执行存储器虚拟化的系统和方法。

发明内容

[0007] 公开了对多台固态存储设备当中存储的数据高效地执行用户存储器虚拟化的计算机系统和方法的多个实施例。

[0008] 在一个实施例中，连接到网络的数据存储子系统在网络上接收来自客户计算机的读写请求。数据存储子系统包括多个数据存储位置，在包含多台存储设备的设备组上。数据存储子系统进一步包括至少一个映射表，包括按时间分选的多个等级。在一个实施例中，每个等级都存储一个或多个元组，每个元组都包括一个或多个值，它们可用作查找关键字。此外，每个元组都可以包括若干数据值，与这些关键字值相关联。在一个实施例中，映射表是虚拟至物理地址转换表。在另一个实施例中，映射表是去重复表。数据存储子系统进一步包括数据存储控制器，被配置为响应检测到把一个或多个新元组插入到映射表中的条件，创建要向多个等级添加的新的最高等级(最年轻等级)。每个元组都可以被存储在映射表内分开的记录即条目中。这些记录可以按关键字值分选。

[0009] 在也预期的实施例中，系统包括一个或多个“覆盖”表。覆盖表可用于修改由映射表回答的对若干查询的响应。修改后的响应可以包括标注响应无效或者修改响应对映射表的查询而提供的元组中若干字段值。对覆盖表的访问可以以相对快的方式判定给定关键字不是有效的。在也预期的实施例中，映射表的全部等级除了最年轻的都是只读的。

[0010] 在考虑了以下说明和附图后，这些和其他实施例将变得显而易见。

[0011] 附图简要说明

[0012] 图 1 是广义框图，展示了网络架构的一个实施例；

[0013] 图 2 是映射表的一个实施例的广义框图；

[0014] 图 3A 是访问映射表所用的初级索引的一个实施例的广义框图；

[0015] 图 3B 是访问映射表所用的初级索引的另一个实施例的广义框图；

[0016] 图 4 是初级索引和映射表的另一个实施例的广义框图；

[0017] 图 5A 是广义流程图，展示了执行读访问的方法的一个实施例；

[0018] 图 5B 是广义流程图，展示了执行写操作的方法的一个实施例；

[0019] 图 6 是具有共享映射表的多节点网络的一个实施例的广义框图；

[0020] 图 7 是访问映射表所用的次级索引的一个实施例的广义框图；

[0021] 图 8 是访问映射表的三级索引的一个实施例的广义框图；

[0022] 图 9 展示了利用覆盖表方法的一个实施例；

[0023] 图 10 是对映射表内若干等级进行展平操作的一个实施例的广义框图；

[0024] 图 11 是对映射表内若干等级进行展平操作的另一个实施例的广义框图；

[0025] 图 12 是广义流程图，展示了映射表内等级展平方法的一个实施例；

[0026] 图 13 是广义流程图，展示了映射表内批量阵列任务高效处理方法的一个实施例；

[0027] 图 14 是广义框图，展示了存储设备内数据布局架构的实施例。

[0028] 虽然本发明允许多种修改和替代形式，但是在附图中举例显示并且本文详细介绍了若干特定实施例。不过应当理解，附图及其详细说明并非旨在使本发明限于所公开的具体形式，而是相反，本发明要覆盖落入由附带权利要求书定义的本发明的精神和范围内的一切修改、等价和替代。

具体实施方式

[0029] 为了彻底理解本发明，在以下说明中阐述了众多特定细节。不过，本领域的普通技术人员应当认识到，没有这些特定的细节，也能够实施本发明。在某些实例中，为了避免模

糊本发明，未详细显示众所周知的电路、结构、信号、计算机程序指令和技术。

[0030] 参考图 1，显示了网络架构 100 的一个实施例的广义框图。正如以下进一步介绍，网络架构 100 的一个实施例包括客户计算机系统 110a-110b，通过网络 180 彼此互连并连接到数据存储器阵列 120a-120b。网络 180 可以通过交换机 140 连接到第二个网络 190。客户计算机系统 100c 经由网络 190 连接到客户计算机系统 110a-110b 和数据存储器阵列 120a-120b。此外，网络 190 还可以通过交换机 150 连接到因特网 160 或网络之外别处。

[0031] 应当在意，在替代实施例中，客户计算机和服务器、交换机、网络、数据存储器阵列以及数据存储设备的数量和类型不限于图 1 所示的这些。在多个时间一台或多台客户机可以离线运行。此外，在运行期间，随着用户连接、断开以及重新连接到网络架构 100，各个客户计算机连接类型可以改变。另外，虽然本说明一般地讨论了网络附属存储，但是本文介绍的系统和方法也适用于直接附属存储系统，并且可以包括被配置为执行所介绍方法的一个或多个方面的主机操作系统。众多这样的替代是可能的并被预期。简短地提供了图 1 所示的每个组件的进一步说明。首先介绍由数据存储器阵列 120a-120b 提供的某些特征的概况。

[0032] 在网络架构 100 中，每个数据存储器阵列 120a-120b 都可用于共享不同服务器和计算机当中的数据，比如客户计算机系统 110a-110c。此外，数据存储器阵列 120a-120b 可用于进行磁盘镜像、备份和恢复、存档数据的存档和检索以及从一台存储设备到另一台存储设备的数据迁移。在替代实施例中，一个或多个客户计算机系统 110a-110c 可以通过快速局域网（LAN）彼此链接以便形成集群。这样的客户机可以共享存储器资源，比如在数据存储器阵列 120a-120b 之一内驻留的集群共享卷。

[0033] 数据存储器阵列 120a-120b 的每一个都包括存储子系统 170 用于数据存储。存储子系统 170 可以包括多台存储设备 176a-176m。这些存储设备 176a-176m 可以向客户计算机系统 110a-110c 提供数据存储服务。存储设备 176a-176m 的每一台都使用具体的技术和机构执行数据存储。每台存储设备 176a-176m 内所用技术和机构的类型可以至少部分地被用于确定若干算法，用于控制和调度进出每台存储设备 176a-176m 的读写操作。例如，这些算法可以定位这些操作对应的具体物理位置。此外，这些算法还可以执行用于这些操作的输入 / 输出（I/O）重定向、去除存储子系统 170 中的重复数据以及支持地址重定向和去重复的一个或多个映射表。

[0034] 在以上算法中所用的逻辑可以被包括在基本操作系统（OS）132、卷管理器 134、存储子系统控制器 174 内、每台存储设备 176a-176m 内的控制逻辑或其他方面的一个或多个中。此外，本文介绍的逻辑、算法和控制机构可以包括硬件和 / 或软件。

[0035] 每台存储设备 176a-176m 都可以被配置为接收读写请求并包括多个数据存储位置，每个数据存储位置都可寻址为阵列中的行和列。在一个实施例中，在存储设备 176a-176m 内的数据存储位置可以安排为逻辑的冗余存储容器或 RAID 阵列（非昂贵 / 独立磁盘冗余阵列）。

[0036] 在某些实施例中，每台存储设备 176a-176m 都可以利用与常规硬盘驱动器（HDD）不同的数据存储技术。例如，一台或多台存储设备 176a-176m 可以包括或进一步连接到含有固态内存的存储器以存储永久数据。在其他实施例中，一台或多台存储设备 176a-176m 可以包括或进一步连接到使用其他技术的存储器，比如自旋力矩转移技术、磁阻性随机存

取存储器(MRAM)技术、遮板式磁盘、memristors、相变存储器或其他存储器技术。这些不同的存储器技术可以导致存储设备之间不同的 I/O 特征。

[0037] 在一个实施例中,含有的固态内存包括固态驱动器(SSD)技术。HDD 技术与 SDD 技术之间技术和机制的差异可以导致数据存储设备 176a-176m 的输入 / 输出(I/O)特征的差异。固态盘(SSD)也可以称为固态驱动器。SSD 没有移动部件或机械延迟,可以具有比 HDD 更短的读取访问时间和等待时间。不过,SSD 的写入性能一般低于读出性能,并且可能由 SSD 内空闲可编程块的可用性显著地影响。

[0038] 通过在存储设备 176a-176m 内的用户存储器与物理位置之间创建存储器虚拟化层可以改进存储器阵列效率。在一个实施例中,卷管理器的虚拟层被放置在操作系统(OS)的设备 - 驱动器栈中,而不是在存储设备内或网络中。许多存储器阵列以粗粒度级执行存储器虚拟化以允许完全在内存中存储虚拟至物理的映射表。不过,这样的存储器阵列不能够集成诸如数据压缩、去重复和修改后复制操作的特征。许多文件系统支持细粒度的虚拟至物理的映射表,但是它们不支持大的存储器阵列,比如设备组 173a-173m。而是使用卷管理器或磁盘阵列管理器支持设备组 173a-173m。

[0039] 在一个实施例中,一个或多个映射表可以存储在存储设备 176a-176m,而不是内存,比如 RAM172、内存介质 130 或处理器 122 内的高速缓存。存储设备 176a-176m 可以是利用闪存的 SSD。SSD 的短读取访问时间和等待时间可以允许在服务于来自客户计算机的存储器访问请求的同时发生少量的相关读取操作。在存储器访问请求的服务期间,这些相关读取操作可以用于访问一个或多个索引、一个或多个映射表以及用户数据。

[0040] 在一个实施例中,通过相关读取操作可以执行 I/O 重定向。在另一个实施例中,通过相关读取操作可以执行内联去重复。在又一个实施例中,批量阵列任务比如大的复制、移动或清零操作可以整体在映射表内执行,而不是访问持有用户数据的存储器位置。这样的直接映射操作可以大幅降低 I/O 流量和存储设备 176a-176m 内的数据移动。既服务于存储器访问请求又执行从 SSD 的相关读取操作的合并时间可以少于服务于来自自旋 HDD 的存储器访问请求。

[0041] 此外,映射表内的信息可以压缩。可以选择具体的压缩算法以允许对各个成分的识别,比如多个记录当中某记录内的关键字。所以,可以发生对多个压缩记录当中给定关键字的搜索。如果发现了匹配,可以仅仅对匹配的记录解压缩。对映射表记录内的元组进行压缩可以进一步实现细粒度级的映射。这种细粒度级的映射可以允许直接映射操作,作为普通批量阵列任务的替代。以下将讨论涉及高效存储器虚拟化的进一步细节。

[0042] 同样如图所示,网络架构 100 包括客户计算机系统 110a-110c,通过网络 180 和 190 彼此互连并连接到数据存储器阵列 120a-120b。网络 180 和 190 可以包括各种各样的技术,包括无线连接、直接局域网(LAN)连接、广域网(WAN)连接,比如因特网、路由器、存储器区域网络、以太网和其他。网络 180 和 190 可以包括一个或多个 LAN,它们也可以是无线的。网络 180 和 190 可以进一步包括远程直接存储器存取(RDMA)硬件和 / 或软件、传输控制协议 / 因特网协议(TCP/IP)硬件和 / 或软件、路由器、中继器、交换机、网格和 / 或其他。在网络 180 和 190 中可以使用诸如光纤信道、以太网光纤信道(FCoE)、iSCSI 等协议。交换机 140 可以利用与网络 180 和 190 都相关联的协议。网络 190 可以与因特网 160 所用的一组通信协议接合,比如传输控制协议(TCP)和因特网协议(IP)即 TCP/IP。交换机 150 可以

是 TCP/IP 交换机。

[0043] 客户计算机系统 110a-110c 是任何数量的固定或移动计算机的代表，比如台式个人计算机(PC)、服务器、服务器群、工作站、便携式电脑、手持计算机、服务器、个人数字助理(PDA)、智能电话等。一般来说，客户计算机系统 110a-110c 包括一个或多个处理器，内含一个或多个处理器核。每个处理器核都包括电路，根据预定义的通用指令集执行指令。例如可以选择 x86 指令集架构。作为替代，可以选择 **Alpha®、PowerPC®、SPARC®** 或任何其他通用指令集架构。处理器核可以访问高速缓存内存子系统存取数据和计算机程序指令。高速缓存子系统可以连接到存储器层次，内含随机存取存储器(RAM)和存储设备。

[0044] 客户计算机系统内的每个处理器核和存储器层次都可以连接到网络接口。除了硬件组件，每个客户计算机系统 110a-110c 还可以包括在存储器层次内存储的基本操作系统(OS)。基本 OS 可以各种各样操作系统的任何一种的代表，例如 **MS-DOS®、MS-WINDOWS®、 OS/2®、UNIX®、Linux®、Solaris®、AIX®、** DART 或其他。因此，基本 OS 可以可用于向终端用户提供多种服务以及提供可用于支持多个程序执行的软件框架。此外，每个客户计算机系统 110a-110c 都可以包括系统管理程序，用于支持虚拟机(VM)。正如本领域技术人员熟知，虚拟化可用于台式计算机和服务器中以便完全或部分地把软件比如 OS 与系统的硬件分离。虚拟化可以向终端用户提供多个 OS 运行在同一机器上的错觉，每个都具有其自己的资源和对每个数据存储器阵列 120a-120b 内的存储设备 176a-176m 上创建的逻辑存储实体(如 LUN)的访问权限。

[0045] 每个数据存储器阵列 120a-120b 都可用于在不同的服务器当中共享数据，比如客户计算机系统 110a-110c。每个数据存储器阵列 120a-120b 都包括用于数据存储的存储子系统 170。存储子系统 170 可以包括多台存储设备 176a-176m。这些存储设备 176a-176m 的每一台都可以是 SSD。控制器 174 可以包括用于处理收到的读 / 写请求的逻辑。随机存取存储器(RAM)可用于批量操作，比如收到的写请求。在多个实施例中，进行批量写操作时(或其他操作)可以使用非易失性存储器(如 NVRAM)。

[0046] 内存介质 130 中存储的基本 OS132、卷管理器 134 (或磁盘阵列管理器 134)、任何 OS 驱动程序(未显示)和其他软件可以提供对文件的访问的功能以及对这些功能的管理。基本 OS132 可以是存储器操作系统比如 NetApp Data **ONTAP®**或其他操作系统。基本 OS132 和 OS 驱动程序可以包括在内存介质 130 上存储的并由处理器 122 可执行的程序指令以执行存储子系统 170 中的一项或多项内存访问操作，它们对应于收到的请求。图 1 所示的系统通常可以包括一个或多个文件服务器和 / 或块服务器。

[0047] 每个数据存储器阵列 120a-120b 都可以使用网络接口 124 连接到网络 180。类似于客户计算机系统 110a-110c，在一个实施例中，网络接口 124 的功能可以包括在网络适配器卡上。实施网络接口 124 的功能时可以硬件和软件都使用。随机存取存储器(RAM)和只读存储器(ROM)都可以包括在网络接口 124 的网卡实施上。可以使用一个或多个专用集成电路(ASIC)提供网络接口 124 的功能。

[0048] 除了以上情况，数据存储器阵列 120a-120b 内的每个存储器控制器 174 都可以支持若干存储器阵列功能，比如快照、复制和高可用性。此外，每个存储器控制器 174 都可以支持虚拟机环境，包括多个卷，每个卷都包括多个快照。在一个实例中，存储器控制器 174

可以支持几十万个卷，其中每个卷都包括几千个快照。在一个实施例中，卷可以以固定尺寸的扇区映射，比如在存储设备 176a-176m 内 4 千字节（KB）页面。在另一个实施例中，卷可以以可变尺寸的扇区映射，比如用于写请求。可以使用卷 ID、快照 ID 和扇区号标识给定的卷。

[0049] 地址转换表可以包括多个条目，其中每个条目都保持着对应数据成分的虚拟至物理的映射。这个映射表可用于把来自每个客户计算机系统 110a-110c 的逻辑读 / 写请求映射到存储设备 176a-176m 中的物理位置。在收到的读 / 写请求对应的查找操作期间，可以从映射表中读取“物理”指针值。然后可以使用这个物理指针值定位存储设备 176a-176m 内的物理位置。应当注意，可以使用物理指针值访问存储设备 176a-176m 的给定存储设备内的另一个映射表。从而在物理指针值与目标存储位置之间可以存在一个或多个等级的间接。

[0050] 在另一个实施例中，映射表可以包括去重复所用的信息（去重复表相关信息）。去重复表中存储的信息可以包括对于给定数据成分算出的一个或多个哈希值与到保持着给定数据成分的存储设备 176a-176m 之一中物理位置的物理指针之间的映射。此外，给定数据成分的长度和对应条目的状态信息也可以存储在去重复表中。

[0051] 现在转向图 2，显示了映射表的一个实施例的广义框图。如早先讨论，一个或多个映射表可以用于 I/O 重定向或变换、用户数据的重复副本的去重复、卷快照映射等。映射表可以存储在存储设备 176a-176m 中。图 2 所示的框图表示映射表的组织和存储的一个实施例的逻辑表达。所示的每级都可以包括不同时段对应的映射表条目。例如，“1”级包括的信息可以早于“2”级中存储的信息。同样，“2”级包括的信息可以早于“3”级中存储的信息。在图 2 所示的记录、页面和级中存储的信息可以在存储设备 176a-176m 内以随机存取的方式存储。另外，给定映射表条目的部分或全部副本可以存储在 RAM172 中、控制器 174 内的缓冲区中、内存介质 130 中以及处理器 122 内或连接的一个或多个高速缓存中。在多个实施例中，对应的索引可以包括在每级映射中，它们是该级的一部分（如后面图 4 的描述）。这样的索引可以包括该级内映射表条目的标识以及它们存储之处的标识（如该页面的标识）。在其他实施例中，与映射表条目相关联的索引可以是截然不同的实体或若干实体，逻辑上它们不是级本身的一部分。

[0052] 一般来说，每个映射表都包括一组行和列。单个记录可以存储在映射表中作为行。记录也可以被称为条目。在一个实施例中，记录存储了至少一个元组，包括关键字。元组还可以（或可以不）包括数据字段，包括若干数据，比如识别或定位在存储子系统 170 中存储的数据成分所用的指针。应当注意，在多个实施例中，存储子系统可以包括具有内部映射机构的存储设备（如 SSD）。在这样的实施例中，元组中的指针本质上可以不是实际的物理地址。相反，指针可以是该存储设备映射到该设备内物理位置的逻辑地址。逻辑地址与物理位置之间的这种内部映射可以随着时间而改变。在其他实施例中，映射表中的记录可以仅仅包含关键字字段，没有另外的相关联的数据字段。与给定记录所对应数据成分相关联的属性可以存储在表中的若干列即字段中。状态信息比如有效指示符、数据年龄、数据尺寸等都可以存储在字段中，比如图 2 所示的字段 0 至字段 N。在多个实施例中，每列都存储着给定类型所对应的信息。在某些实施例中，可以对选定的字段使用压缩技术，在某些情况下可以产生其压缩后表达为零位长度的字段。

[0053] 关键字是映射表中的实体，可以区分数据的一行与另一行。每行也可以称为条目或记录。关键字可以是单列，它也可以由识别记录所用的一组列组成。在一个实例中，地址转换映射表可以利用包括卷标识符（ID）、逻辑或虚拟地址、快照 ID、扇区号等的关键字。给定收到的读 / 写存储访问请求可以识别具体卷、扇区和长度。扇区可以是卷中存储数据的逻辑块。在不同卷上扇区可以具有不同的尺寸。地址转换映射表可以以扇区尺寸为单位映射卷。

[0054] 卷标识符（ID）可以用于访问卷表格，它传达了卷 ID 和对应的当前快照 ID。这条信息连同收到的扇区号可用于访问地址转换映射表。所以，在这样的实施例中，用于访问地址转换映射表的关键字值是卷 ID、快照 ID 和收到的扇区号的结合。在一个实施例中，分选地址转换映射表内的记录时，依据卷 ID，继之以扇区号然后是快照 ID。这种排序可以把不同快照中的数据成分的不同版本组合在一起。所以，在查找存储器访问读请求期间，以对存储设备 176a-176m 更少的读取操作就可以找到对应的数据成分。

[0055] 地址转换映射表可以传达物理指针值，它指示了存储着收到的数据存储器访问请求所对应的数据成分的数据存储子系统 170 内的位置。关键字值可以与映射表中存储的一个或多个关键字值进行比较。在展示实例中，为了易于展示，显示了更简单的关键字值，比如“0”、“2”、“12”等。物理指针值可以存储在对应记录的一个或多个字段中。

[0056] 物理指针值可以包括区段标识符（ID）和标识存储器位置的地址。区段可以是在每台存储设备 176a-176m 中分配的基本单元。区段可以具有独立冗余磁盘阵列（RAID）级和数据类型。在分配期间，区段可以具有为对应的存储所选择的一台或多台存储设备 176a-176m。在一个实施例中，在存储设备 176a-176m 中一台或多台选定的存储设备的每台上可以给区段分配等量的存储空间。数据存储访问请求可以对应于多个扇区，它们可能引起多个并行的查找。写请求可以被放置在 NVRAM 缓冲区，比如 RAM172，并且可以把写完成应答发送到客户计算机 110a-110c 中对应的客户计算机。在随后时间，异步过程可以把缓冲的写请求刷新到存储设备 176a-176m。

[0057] 在另一个实例中，图 2 所示的映射表可以是去重复表。去重复表利用的关键字可以包括从与存储访问请求相关联的数据成分确定的哈希值。去重复操作的初始步骤可以与其他操作并发地执行，比如读 / 写请求、垃圾收集操作、剪裁操作等。对于给定的写请求，从客户计算机系统 110a-110c 之一发送的数据可以是数据流，比如字节流。正如本领域技术人员熟知，可以把数据流划分为固定长度或可变长度组块的序列。组块算法可以执行把数据流划分为离散数据成分，它们可以被称为“组块”。组块可以是亚文件、内容可寻址的数据单位。在多个实施例中，表格或其他结构可以用于确定具体组块算法用于给定文件类型或数据类型。确定文件的类型时可以通过参考其文件名扩展名、分开的标识信息、数据自身的内容或其他方面。产生的组块然后可以存储在数据存储器阵列 120a-120b 之一中，以便允许这些组块的共享。这样的组块可以分开存储或以多种方式分组在一起。

[0058] 在多个实施例中，组块可以由数据结构表示，它允许从其组块重构更大的数据成分（如根据所存储数据的一个或多个更小组块可以重构某具体文件）。对应的数据结构可以记录其对应的组块，包括算出的相关联哈希值、指向在数据存储器阵列 120a-120b 之一中其位置的指针以及其长度。对于每个数据成分，都可以使用去重复应用程序计算对应的哈希值。例如，哈希函数比如 Message-Digest 算法 5 （MD5）、Secure Hash 算法（SHA）或其他

算法可用于计算对应的哈希值。为了知晓所收到的写请求所对应的给定数据成分是否已经存储在数据存储器阵列 120a-120b 之一中，对给定的数据成分算出的哈希值的若干位（或哈希值若干位的子集）可以与在一个或多个数据存储器阵列 120a-120b 中存储的数据成分的哈希值中的若干位进行对比。

[0059] 映射表可以包括一个或多个等级，如图 2 所示。映射表可以包括 16 至 64 个等级，尽管在映射表内支持的另一数量的等级有可能并被预期。在图 2 中，为了易于展示，显示了标注为“1”级、“2”级和“N”级的三个等级。映射表内的每个等级都可以包括一个或多个分区。在一个实施例中，每个分区是 4 千字节（KB）的页面。例如，“N”级被显示为包括页面 210a 至 210g、“2”级包括页面 210h 至 210j 而“1”级包括页面 210k 至 210n。有可能并预期其他分区尺寸也可以选择为用于映射表内的每个等级。此外，有可能一个或多个等级具有单一分区，它是该等级自身。

[0060] 在一个实施例中，映射表内的多个等级按时间分选。例如，在图 2 中，“1”级可以早于“2”级。同样，“2”级可以早于“N”级。在一个实施例中，检测到在映射表中插入一个或多个新记录的条件时，就可以创建新的等级。在多个实施例中，创建新等级时，给予新等级的编号 / 标记大于给予时间上先于新等级的等级的编号。例如，如果最近创建的等级被分配了值 8，那么新创建的等级可以被分配值 9。以这种方式，可以建立或确定等级之间的时间关系。正如可以认识到，数字值不必严格为顺序的。此外，替代实施例可以颠倒该编号模式，使得更新等级具有更小的数字标记。另外，其他实施例可以利用非数字标记在等级之间区分。许多这样的实施例是可能的并被预期。每个下一个更早的等级具有从先前更新等级的标签整数值中减去一的标签。未显示的分开表格可以用于在逻辑上描述该映射表。例如，分开表格的每个条目都可以包括给定的等级 ID 和在给定的等级 ID 内存储的页面 ID 的列表。

[0061] 通过为插入新记录而创建新的最高等级，通过添加新记录更新了映射表。在一个实施例中，单一等级被创建为新的最高等级并且每个新记录都被插入到单一等级中。在另一个实施例中，先对新记录搜索重复关键字，再插入到映射表中。单一等级可以被创建为新的最高等级。找到存储着重复关键字的给定记录时，在给定记录前面已缓冲的每个记录都可以被插入到单一等级中。新记录可以以保持内存次序的方式被缓冲，比如以请求完成的次序。然后可以创建另一个单一等级并可以把新记录的剩余部分插入到这另一个单一等级中，除非找到了存储着重复关键字的另一个记录。如果找到了这样的记录，那么重复这些步骤。把相同关键字值作为新记录之一存储的映射表内的现有记录不被新记录的插入在原地被编辑或盖写。

[0062] 尽管等级尺寸被展示为更低等级的增加大于更新等级，但是更高等级可能在比相邻等级更大或更小之间交替。插入到映射表中的更新记录的数量可以随着时间而改变并创建了波动的等级尺寸。由于更低等级的波动，更低等级可以大于更新等级。检测到若干具体条件时，两个或更多更低的等级可以被展平为单一等级。后面提供了进一步的细节。

[0063] 由于对映射表中存储的记录没有原地编辑，所以在更高等级中放置的更新记录可以推翻位于更低等级中存储着同样关键字值的记录。例如，通过给定关键字值访问映射表时，可以找到一个或多个等级存储的记录保持着与该给定关键字值匹配的关键字值。在这样的情况下，可以选择这一个或多个等级中的最高等级把其对应的记录中存储的信息提供

为访问的结果。后面提供了进一步的细节。此外，后面提供了关于把一个或多个新记录插入到映射表中和存储信息的检测条件的进一步细节。

[0064] 在一个实施例中，给定页面内的条目可以按关键字分选。例如，可以根据条目中包括的关键字以上升的次序对条目分选。此外，在多个实施例中，等级内的页面可以根据期望的任何分选次序分选。在多个实施例中，也可以对等级内的页面分选(如根据关键字或其他方面)。在图 2 的实例中，N 级的页面 210a 包括根据上升次序的关键字值所分选的记录。在多个实施例中，可以使用一列或多列存储关键字值。在图 2 的实例中，在用于存储关键字值的每个元组中显示了两列或字段。利用这样的关键字值，就可以以期望次序分选记录。进行分选时可以根据记录的任何关键字值或记录关键字值的任何组合。在所示实例中，第一个记录存储的关键字值包括在两列中存储的 0 和 8。而最后一个记录存储的关键字值包括 12 和 33。在这个展示实例中，页面 210a 中第一个与最后一个记录之间的每个分选后记录都在第一列中存储了 0 与 12 之间的关键字值，并且以(至少部分地)根据对第一列从 0 至 12 上升次序存储关键字值的方式安排这些记录。同样，页面 210b 包括分选后的记录，其中第一个记录存储了关键字值 12 和 39，而最后一个记录存储了关键字值 31 和 19。在这个展示实例中，页面 210b 中第一个与最后一个记录之间的每个分选后记录都在第一列中存储了 12 与 31 之间的关键字值，并且以从 12 至 31 上升次序存储关键字值的方式安排这些记录。

[0065] 除了以上情况，N 级内页面根据期望的次序分选。在多个实施例中，级内页面分选的方式可以反映页面内条目被分选的次序。例如，级内页面可以根据关键字值上升次序分选。由于页面 210b 中的第一个关键字值大于页面 210a 中的最后一个关键字，所以在分选次序中页面 210b 跟随页面 210a。页面 210g 然后将可能包括其关键字值大于在页面 210a 至 210f (未显示)中包括的关键字值的条目。以这种方式，级内全部条目都根据共同模式分选。条目被简单地细分为页面或其他的尺寸单元。正如可以认识到，可以按期望使用其他分选模式。

[0066] 现在参考图 3A，显示了访问映射表所用的初级索引的一个实施例的广义框图。关键字发生器 304 可以接收一个或多个请求者数据输入 302。在一个实施例中，映射表是地址转换目录表。给出的所收到读 / 写请求可以识别具体的卷、扇区和长度。关键字发生器 304 可以产生查询关键字值 306，包括卷标识符(ID)、逻辑即虚拟地址、快照 ID 和扇区号。其他组合是可能的并且也可以利用其他或附加值。查询关键字值 306 的不同部分可以与映射表内可以相邻也可以不相邻的列中存储的值进行对比。在所示实例中，为了易于展示，使用关键字值“22”。

[0067] 正如早先介绍，与关键字发生器 304 相关联的组块算法和 / 或分段算法都可以接收存储访问请求对应的数据 302。这些算法可以产生一个或多个数据成分并选择哈希函数计算每个数据成分的对应哈希值或查询关键字值 306。可以使用产生的哈希值检索去重复表。

[0068] 正如图 3A 所示，初级索引 310 可以对存储设备 176a-176m 中存储的数据提供位置识别信息。例如，再次参考图 2，对应的初级索引 310 (或其部分)可以被逻辑地包括在“1”级、“2”级和“N”级的每级中。同样，每级和每个对应的初级索引都可以以随机存取方式被物理地存储在存储设备 176a-176m 内。

[0069] 在一个实施例中，初级索引 310 可以被划分为若干分区，比如分区 312a 至 312b。

在一个实施例中，分区的尺寸可以从 4 千字节（KB）页面到 256KB，尽管其他尺寸是可能的并被预期。初级索引 310 的每个条目都可以存储关键字值。此外，每个条目都可以存储该关键字值所对应的对应唯一虚拟页面标识符（ID）以及等级 ID。每个条目都可以存储对应的状态信息比如有效性信息。以查询关键字值访问初级索引 310 时，可以在索引 310 内条目中搜索与该关键字值匹配或以其他方式对应的一个或多个条目。然后可以使用匹配条目的信息定位和检索某映射，它标识的存储位置是收到的读写请求的目标。换言之，索引 310 标识映射的位置。在一个实施例中，索引中的击中提供的对应页面 ID 标识了在存储设备 176a-176m 内既存储着关键字值又存储着对应物理指针值的页面。可以用该关键字值搜索由对应页面 ID 标识的页面以找到物理指针值。

[0070] 在图 3A 的实例中，收到的请求对应于关键字“22”。这个关键字然后被用于访问索引 310。对索引 310 的搜索产生了对分区 312b 内某条目的击中。匹配的条目在这种情况下包括诸如页面 20 和 3 级的信息。根据这个结果，在映射表 3 级内被标识为页面 28 的页面中找到了对该请求所期望的映射。使用这条信息，就可以对映射表进行访问以检索所期望的映射。如果对初级索引 310 的访问要求对存储器的访问，那么为了获得所期望的映射将需要至少两次存储器存取。所以，如以下介绍在多个实施例中，初级索引的若干部分可以高速缓存或以其他方式存储在相对快速存取的内存中，以便消除对存储设备的一次存取。在多个实施例中，映射表的全部初级索引被高速缓存。在某些实施例中，在初级索引已经变得大到无法将其全部高速缓存或者在其他方面大于所期望时，可以在高速缓存中使用次级、三级或其他索引部分以减小其尺寸。下面讨论了次级类型索引。除了以上情况，在多个实施例中，最近击中所对应的映射页面至少在某段时间也被高速缓存。以这种方式，展现了以时间局部性访问的过程能够更快地得到服务（即最近访问的位置将使其映射被高速缓存并易于获得）。

[0071] 现在参考图 3B，显示了访问映射表所用的被高速缓存的初级索引的一个实施例的广义框图。图 3A 的电路和逻辑部分的对应者一致地编号。被高速缓存的初级索引 314 可以包括在映射表中的多个等级的每个初级索引 310 中存储的信息的副本。初级索引 314 可以存储在 RAM172、控制器 174 内的缓冲区、内存介质 130 和处理器 122 内的高速缓存的一个或多个中。在一个实施例中，初级索引 314 可以按关键字值分选，尽管以其他方式分选是可能的。初级索引 314 还可以被划分为若干分区，比如分区 316a 至 316b。在一个实施例中，分区 316a 至 316b 的尺寸可以与初级索引 310 内分区 312a 至 312b 的尺寸相同。

[0072] 类似于初级索引 310，初级索引 314 的每个条目都可以存储关键字值、对应的唯一虚拟页面标识符（ID）、关键字值所对应的等级 ID 以及诸如有效信息的状态信息中的一个或多个。以查询关键字值访问初级索引 314 时，它可以传达的对应页面 ID 标识在存储设备 176a-176m 内既存储着关键字值又存储着对应指针值的页面。由对应的页面 ID 所标识的页面可以用关键字值搜索以找到该指针值。如图所示，初级索引 314 可以具有多个记录存储着相同的关键字值。所以，多次击中可以从搜索给定关键字值中产生。在一个实施例中，可以选择具有等级 ID（或者用于识别最年轻等级或最新近条目的无论何种指示器）的最高值的击中。这种从多次击中里选择一个击中可以由这里未显示的合并逻辑实施。后面提供了对合并逻辑的进一步说明。

[0073] 现在转向图 4，显示了映射表和访问映射表所用的初级索引的另一个实施例的广

义框图。图 3A 的电路和逻辑部分的对应者一致地编号。映射表 340 可以具有与图 2 所示的映射表类似的结构。不过，现在显示了每个等级的对应初级索引 310 的存储。初级索引 310a 至 310i 的一个或多个的副本可以包括在索引副本 330 中(如高速缓存的副本)。副本 330 一般可以对应于图 3B 中描述的高速缓存中的索引。索引副本 330 中的信息可以存储在 RAM172、控制器 174 内的缓冲区、内存介质 130 和处理器 122 的高速缓存中。在所示的实施例中，初级索引 310a 至 310i 中的信息可以用映射页面存储在存储设备 176a-176m 中。同样所示的是次级索引 320，可以用于访问初级索引，比如以图表所示的初级索引 310i。同样，访问和更新映射表 340 可以如早先介绍的那样发生。

[0074] 映射表 340 包括多个等级，比如“1”级至“N”级。在展示的实例中，每级都包括多个页面。“N”级被显示为包括页面“0”至“D”，“N-1”级包括页面“E”至“G”，等等。同样映射表 340 内的等级可以按时间分选。“N”级可以年轻于“N-1”级，等等。映射表 340 可以按至少某关键字值访问。在展示的实例中，映射表 340 按关键字值“27”和页面 ID“32”访问。例如，在一个实施例中，等级 ID“8”可用于标识映射表 340 中要搜索的特定等级(或“子表”)。识别了所期望的子表，然后页面 ID 可以用于识别子表内的期望页面。最后，可以使用关键字识别期望页面内的期望条目。

[0075] 正如以上讨论，对高速缓存中索引 330 的访问可以产生多次击中。在一个实施例中，这多次击中的结果被提供给了合并逻辑 350，它识别出哪次击中被用于访问映射表 340。合并逻辑 350 可以表示存储控制器内包括的硬件和 / 或软件。在一个实施例中，合并逻辑 350 被配置为识别最新近(最新的)映射所对应的击中。这样的识别有可能根据对某条目的对应等级的识别或其他方面。在所示实例中，收到了 8 级、页面 32、关键字 27 所对应的查询。响应该查询，访问了 8 级的页面 32。如果在页面 32 内找到了关键字 27 (击中)，就返回对应结果(如在所示实例中的指针 xF3209B24)。如果在页面 32 内未找到关键字 27，就返回未击中的指示。这个物理指针值可以从映射表 340 输出以服务于关键字值“27”所对应的存储访问请求。

[0076] 在一个实施例中，映射表 340 支持内联映射。例如，被检测为具有足够小的目标的映射可能被表示为在存储设备 176a-176m 内没有存储用户数据的实际物理扇区。一个实例可以是用户数据内的重复模式。不是在存储设备 176a-176m 内实际存储重复模式(如一串零)的多个副本作为用户数据，对应的映射可以具有以状态信息标注的指示，比如在映射表中字段 0 至字段 N 的字段之一内，它指示了对于读取请求要返回什么数据。不过，在存储设备 176a-176m 内的目标位置没有实际存储这个用户数据。此外，指示可以存储在初级索引 310 以及可以使用的任何附加索引(未显示)的状态信息内。

[0077] 除了以上情况，在多个实施例中，存储系统可以同时支持数据组织、存储模式等的多个版本。例如随着系统硬件和软件演变，新的特征可以加入或以其他方式提供。(例如)更新的数据、索引和映射可以利用这些新特征。在图 4 的实例中，新的 N 级可以对应于系统的一个版本，而旧些的 N-1 级可以对应于先前版本。为了容纳这些不同版本，可以把元数据与每个等级相关联地存储，它表明哪个版本、哪些特征、压缩模式等被该等级使用。这个元数据可以存储为索引的以部分、页面自身或双方。进行访问时，这个元数据就指明数据应当如何恰当地处理。此外，新模式和特征可以动态地应用而不需要停顿系统。以这种方式，系统升级更灵活并且没有必要重建陈旧的数据以反映更新的模式和方法。

[0078] 现在转向图5A，显示了服务读访问的方法的一个实施例。一般来说，以上介绍的在网络架构100和映射表340中实施的组件可以根据方法500运行。为了讨论目的，在这个实施例中的步骤以连续的次序显示。不过，某些步骤可以以不同于所示的次序出现，某些步骤可以并发地进行，某些步骤可以与其他步骤结合，而某些步骤在另一个实施例中可以不出现。

[0079] 读和存储(写)请求可以从客户机110a-110c之一传达到数据存储器阵列120a-120b之一。在所示实例中，收到了读请求500，并且在方框502中可以产生对应的查询关键字值。在某些实施例中，请求自身可以包括访问该索引所用的关键字并且不要求“产生”关键字502。正如早先介绍，查询关键字值可以是虚拟地址索引，包括卷ID、与收到请求相关联的逻辑地址即虚拟地址、快照ID、扇区号等。在用于去重复的实施例中，查询关键字值可以使用哈希函数或其他函数产生。用于查询关键字值的其他值也是可能的并被预期，它被用于访问映射表。

[0080] 在方框504中，查询关键字值可用于访问高速缓存中的一个或多个索引以识别映射表的一个或多个部分，它们可能存储着该关键字值对应的映射。此外，也可以搜索已经被高速缓存的最近使用的映射。如果检测到在高速缓存中映射上的击中(方框505)，就可以使用高速缓存中映射执行所请求的访问(方框512)。如果在高速缓存中映射上没有击中，对在高速缓存中索引上是否存在着击中可以做出判断(方框506)。如果是这样，就使用该击中所对应的结果识别和访问该映射表(方框508)。例如，利用初级索引310，存储查询关键字值的条目也可以存储唯一的虚拟页面ID，它标识映射表内特定的单一页面。查询关键字值和相关联的物理指针值都可以存储在这个特定单一页面。在方框508中，可以使用该查询关键字值访问映射表的被识别部分并执行搜索。然后可以返回映射表结果(方框510)并用于执行存储器访问(方框512)，它对应于原始读请求的目标位置。

[0081] 在某些实施例中，响应读请求的索引查询可能导致未击中。这样的未击中可能由于只有一部分索引在高速缓存中或由于错误条件(如对不存在位置的读取访问、地址讹误等)。在这样的情况下，可以执行对所存储的索引的访问。如果对所存储的索引的访问产生击中(方框520)，那么可以返回结果(方框522)，它被用于访问映射表(方框508)。相反，如果对所存储的索引的访问导致未击中，就可以检测错误条件。可以以各种各样期望方式的任何一种完成错误条件的处理。在一个实施例中，可以产生异常(方框524)然后按照期望进行处理。在一个实施例中，在方框510中返回了一部分映射表。在多个实施例中，这个部分是页面，它可以为4KB页面或其他尺寸。正如先前讨论，可以分选页面内的记录以便利对其中包括的内容进行更快的搜索。

[0082] 在一个实施例中，映射表把传统数据库系统的方法用于每个页面中的信息存储。例如，映射表内的每个记录(或行或条目)一个紧跟一个地存储。这种方式可用在面向行的或行存储的数据库中并且另外用于相关数据库。这些类型的数据库利用了基于值的存储结果。基于值的存储(VBS)架构存储唯一数据值仅仅一次而且自动生成的索引系统保持着全部值的上下文。在多个实施例中，数据可以按行存储并且可以在行内的若干列(字段)上使用压缩。在某些实施例中，所用的技术可以包括存储基值并具有更小的字段尺寸用于偏移量，以及/或者具有一组基值，而行中的列包括基值选择器和与这个基值的偏移量。在两种情况下，压缩信息都可以存储在分区内(如在开始处)

[0083] 在某些实施例中，映射表对于每个页面中的信息存储利用了面向列的数据库系统(列存储)方法。列存储分开地存储数据库表的每列。此外，属于同列的属性值可以相邻地存储、压缩和稠密地包装。所以，读取表中若干列的子集，比如在页面内，可以相对快地执行。列数据可以具有一致的类型并可以允许使用存储器尺寸最优化，在面向行的数据中可能不可用。某些压缩模式，比如 Lempel-Ziv-Welch (LZ) 和行程编码(RLE)，利用了相邻数据检测的相似性进行压缩。可以选择的压缩算法允许对页面内的各个记录进行识别和索引。压缩映射表内的记录可以实现细粒度的映射。在多个实施例中，特定数据部分所用的压缩类型可以与数据相关联地存储。例如，压缩类型有可能存储在索引中，作为与压缩后数据同一页面的一部分(如在某种类型的头部)或其他方式。以这种方式，在存储系统内可以并行地使用多种压缩技术和算法。此外，在多个实施例中，在存储数据之时可以动态地确定用于存储页面数据的压缩类型。在一个实施例中，至少部分地根据被压缩数据的性质和类型，可以选择各种各样压缩技术之一。在某些实施例中，将执行多种压缩技术，然后展示最佳压缩的一种将被选为在压缩该数据时使用。许多这样的方式都是可能的并被预期。

[0084] 如果在映射表的任何等级中找到了查询关键字值 306 的匹配(方框 508)，那么在方框 510，可以向合并逻辑 350 传达某击中的一个或多个指示。例如，可以把一个或多个击中指示从“1”级传达到“J”级，如图 4 所示。合并逻辑 350 可以选择传达击中指示的“1”级至“J”级中的最高等级，它也可以是最年轻等级。选中的等级可以把对应记录中存储的信息提供为访问的结果。

[0085] 在方框 512 中，为了处理对应的请求，可以读取被选中页面的匹配记录内的一个或多个对应字段。在一个实施例中，当该页面内的数据以压缩格式存储时，该页面被解压缩并且对应的物理指针值被读出。在另一个实施例中，只有匹配的记录被解压缩并且对应的物理指针值被读出。在一个实施例中，完整的物理指针值可以在映射表与对应的目标物理位置之间分离。所以，为了完成数据存储访问请求，可以访问存储用户数据的多个物理位置。

[0086] 现在转向图 5B，显示了收到的写请求对应的方法的一个实施例。响应收到的写请求(方框 530)，可以创建该请求所对应的新的映射表条目(方框 532)。在一个实施例中，新的虚拟至物理地址映射可以被添加(方框 534)到映射表，它使写请求的虚拟地址与存储对应数据成分的物理位置配对。在多个实施例中，新映射可以与其他新映射一起高速缓存，并且被添加到映射表条目的新的最高等级。然后可以执行写入到永久存储器的操作(方框 536)。在多个实施例中，可以到被认为更高效的稍后时间点才执行把新的映射表条目写入到永久存储器中的映射表(方框 538)。正如先前讨论，在使用固态存储设备的存储系统中，写入到存储器比从存储器读出慢得多。所以，调度写入到存储器的方式为使它们对整个系统性能的影响最小。在某些实施例中，把新记录插入到映射表中可以与其他更大的数据更新合并。以这种方式合并更新可以提供更高效的写操作。应当注意，以图 5B 的方法，正如以本文介绍的每一种方法，为了易于讨论若干操作被描述为以特定次序出现。不过，这些操作事实上可以以不同次序出现，并且在某些情况下这些操作的多个操作可以同时出现。一切这样的实施例都被预期。

[0087] 除了以上情况，在某些实施例中可以使用去重复。图 5B 描述的操作 550 一般可以对应于去重复的系统和方法。在所示实例中，可以产生所收到的写请求对应的哈希值(方

框 540)，用于访问去重复表(方框 542)。如果在去重复表中存在着击中(方框 544)(即该数据的副本已经存在于系统内)，那么可以把新条目添加到去重复表(方框 548)以反映新的写入。在这样的情况下，无需把数据本身写入到存储器并且可以丢弃所收到的写入数据。作为替代，如果在去重复表中存在着未击中，那么为新数据创建新条目并存储在去重复表中(方框 546)。此外，执行向存储器写入该数据(方框 536)。另外，在索引中可以创建新条目以反映新数据(方框 538)。在某些实施例中，如果在内联去重复操作期间出现未击中，在此时就不进行去重复表中的插入。相反，在内联去重复操作期间，对整个去重复表的仅仅一部分(如去重复表的高速缓存中部分)才可能出现以哈希值的查询。如果出现未击中，可以创建新条目并存储在高速缓存中。随后，在去重复操作的后处理期间，比如收集垃圾期间出现的操作，对整个去重复表可能出现以哈希值的查询。未击中可以表明该哈希值是唯一的哈希值。所以，新条目比如哈希至物理指针的映射可以被插入到去重复表中。作为替代，如果在去重复的后处理期间检测到击中(即检测到重复)，可以执行去重复以消除一个或多个检测到的副本。

[0088] 现在参考图 6，显示了具有共享映射表的多节点网络一个实施例的广义框图。在所示实例中，使用三个节点 236a 至 360c 形成映射节点的集群。在一个实施例中，每个节点 236a 至 360c 都可以负责一个或多个逻辑单元号(LUN)。在描述的实施例中，显示了许多映射表等级，1 至 N 级。1 级可以对应于最老的等级，而 N 级可以对应于最新的等级。对于由特定节点管理的 LUN 的映射表条目，这个特定节点自身就可以具有更新的条目存储在节点自身上。例如，节点 360a 被显示为存储了映射子表 362a 和 364a。这些子表 362a 和 364a 可以对应于节点 360a 通常负责的若干 LUN。同样，节点 360b 包括子表 362b 和 364b，它们可以对应于由该节点管理的若干 LUN，而节点 360c 包括子表 362c 和 364c，它们可以对应于由该节点管理的若干 LUN。在这样的实施例中，这些“更新”等级的映射表条目仅仅由其对应的管理节点保存，并且通常在其他节点上找不到。

[0089] 与以上讨论的相对更新的等级相反，更老的等级(即 N-2 级下至 1 级)表示的映射表条目在节点 360a 至 360c 中任何节点都可以存储着这些条目的副本的意义上可以由全部节点 360a 至 360c 所共享。在所示实例中，这些更老等级 370、372 和 374 被共同标识为共享表 380。此外，正如先前讨论，在多个实施例中，这些更老等级是静态的——除了后面讨论的合并或类似操作以外。一般来说，静态层是不承受修改的层(即它是“固定的”)。假设这样的等级在这个意义上是固定的，可以对这些更低等级的任何副本进行访问而不必考虑另一个副本是否已经或正在被修改。因此，任何节点都可以安全地存储共享表 380 的副本，并且以能够恰当地服务以下请求的置信度，对这些表的请求进行服务。使共享表 380 的若干副本存储在多个节点 360 上，当执行查找和以其他请求服务请求时可以允许使用多个负载平衡模式。

[0090] 除了以上情况，在多个实施例中，可以被共享的等级 380 的组织方式可以反映若干节点 360 自身。例如，节点 360a 可以对 LUN1 和 2 负责，节点 360b 可以对 LUN3 和 4 负责，而节点 360c 可以对 LUN5 和 6 负责。在多个实施例中，映射表条目可以包括若干元组，它们本身标识对应的 LUN。在这样的实施例中，根据关键字值、存储空间的绝对宽度或量或者其他方面可以分选共享的映射表 380。如果等级 380 中映射表条目的分选部分基于 LUN，那么条目 370a 可以对应于 LUN1 和 2，条目 370b 可以对应于 LUN3 和 4，而条目 370c 可以对应于

LUN5 和 6。这样的组织可以加速由给定节点对目标为特定 LUN 的请求的查找，方式为有效地降低需要被搜索的数据量、允许协调程序直接选择对特定 LUN 负责的节点作为请求的目标。这些和其他组织和分选模式是可能的并被预期。此外，如果期望把对某 LUN 的责任从一个节点移动到另一个节点，对该节点的原始节点映射可以被刷新到共享的等级(如以及合并)。然后对该 LUN 的责任被传递到新的节点，然后它开始服务该 LUN。

[0091] 现在参考图 7，显示了访问映射表所用的次级索引的一个实施例的广义框图。正如早先介绍，请求者数据输入 302 可以由关键字发生器 304 接收，它产生查询关键字值 306。查询关键字值 306 用于访问映射表。在某些实施例中，图 3 所示的初级索引 310 可能大到无法存储在 RAM172 或内存介质 130 中(或大于期望规模)。例如，索引的更老等级由于后面图 10 和图 11 中介绍的合并和展平操作可能增长到非常庞大。所以对于至少一部分初级索引，可以高速缓存次级索引 320，从而代替初级索引 310 的对应部分。次级索引 320 对存储设备 176a-176m 中存储的数据的位置标识可以提供更粗略的粒度等级。所以，次级索引 320 可以小于它对应的初级索引 310 的部分。所以，次级索引 320 可以存储在 RAM172 或内存介质 130 中。

[0092] 在一个实施例中，次级索引 320 被划分为若干分区，比如分区 322a 至 322b。此外，次级索引可以根据等级组织，更新近的等级首先出现。在一个实施例中，更老等级具有更低的号码而更年轻等级具有更高的号码(如等级 ID 可以随每个新等级增加)。次级索引 320 的每个条目都可以标识关键字值的某个范围。例如实例中所示的第一个条目可以标识 22 级中关键字值从 0 至 12 的范围。这些关键字值可以对应于与初级索引 310 的给定页面内第一个记录和最后一个记录相关联的关键字值。换言之，次级索引中的条目可以仅仅存储关键字 0 的标识和关键字 12 的标识，以表明对应的页面包括该范围内的若干条目。再次参考图 3A，分区 312a 可以是页面而其第一个记录及其最后一个记录的关键字值分别是 0 和 12。所以，次级索引 320 内某条目存储着范围 0 至 12，如图 7 所示。由于在映射表内若干等级中保持着再映射，所以关键字值的范围可以对应于多个页面和相关联的等级。次级索引 320 内的若干字段可以存储这种信息，如图 7 所示。每个条目都可以存储该关键字值范围对应的一个或多个对应的唯一虚拟页面标识符(ID)和相关联的等级 ID。每个条目还可以存储对应的状态信息比如有效性信息。保持的页面 ID 和相关联的等级 ID 的列表可以指示给定查询关键字值有可能被存储之处，但是不确认该关键字值在该页面和等级中出现。次级索引 320 小于初级索引 310，但是也具有存储设备 176a-176m 中存储的数据的位置标识的粗等级的粒度。次级索引 320 可以小到足以存储在 RAM172 中或内存介质 130 中。

[0093] 以查询关键字值 306 访问次级索引 320 时，它可以传达一个或多个对应的页面 ID 和相关联的等级 ID。这些结果然后被用于访问和检索已存储的初级索引的若干部分。然后可以用查询关键字值搜索一个或多个已识别的页面以发现物理指针值。在一个实施例中，等级 ID 可以用于判定已识别的也存储着查询关键字值 306 的一个或多个等级中最年轻的等级。然后可以检索对应页面内的记录并且可以读取物理指针值，用于处理存储访问请求。在展示的实例中，查询关键字值 27 在关键字 16 至 31 的范围内。对应的条目中存储的页面 ID 和等级 ID 以查询关键字值向映射表传达。

[0094] 现在参考图 8，显示了访问映射表所用的三级索引的一个实施例的广义框图。图 4 的电路和逻辑部分的对应者一致地编号。正如早先介绍，图 3 所示的初级索引 310 可能大

到无法存储在 RAM172 或内存介质 130 中。此外，随着映射表 340 增长，次级索引 320 也可能变得大到无法存储在这些内存中。所以，可以在次级索引 320 前访问三级索引 330，这仍然可以比访问初级索引 310 更快。

[0095] 对于存储设备 176a-176m 中存储的数据的位置标识，三级索引 330 可以提供比次级索引 320 更粗等级的粒度。所以，三级索引 330 可以小于它对应的次级索引 320 的部分。应当注意，初级索引 310、次级索引 320、三级索引 330 等的每一个都可以以压缩格式存储。选择的压缩格式可以与在映射表 340 内存储信息所用的压缩格式相同。

[0096] 在一个实施例中，三级索引 330 可以包括多个分区，比如分区 332a、332b 等。可以用查询关键字值 306 访问三级索引 330。在展示的实例中，在从 0 至 78 的关键字值范围之间发现了“27”的查询关键字值 306。三级索引 330 中的第一个条目对应于这个关键字值范围。三级索引 330 中的列可以指示在次级索引 320 内要访问哪个分区。在展示的实例中，0 至 78 的关键字值范围对应于次级索引 320 内的分区 0。

[0097] 还应当注意，可以访问某过滤器（未显示）以判断查询关键字值是否在索引 310 至 330 的任何一个之内。这个过滤器可以是判断某元素是不是集合成员的概率性数据结构。假阳性是有可能的，但是假阴性是不可能的。这样的过滤器的一个实例是 Bloom 过滤器。如果这样的过滤器的访问判定某特定值不在全索引 142 中，那么不向存储器发送查询。如果过滤器的访问判定查询关键字值在对应的索引中，那么可能不知道对应的物理指针值是否被存储在存储设备 176a-176m 中。

[0098] 除了以上情况，在多个实施例中，一个或多个覆盖表可以用于修改或取消由映射表响应查询所提供的元组。这样的覆盖表可以用于施加过滤条件，在响应对映射表的访问时或创建新等级时的展平操作期间使用。在多个实施例中，其他硬件和 / 或软件可以用于施加过滤条件。在某些实施例中，以类似于以上介绍的映射表的方式，覆盖表可以被组织为按时间排序的等级。在其他实施例中，它们以不同方式组织。用于覆盖表的关键字不必与用于基本映射表的关键字匹配。例如，覆盖表可以包含单一条目，声明某特定卷已经被删除或因其他原因不可访问（如不存在查询这个元组的自然访问路径），以及对引用该卷标识符的元组对应的查询的的响应转而是无效的。在另一个实例中，覆盖表中的条目可以指示某存储位置已经被释放，所以引用该存储位置的任何元组都是无效的，从而使查找的结果而不是由映射表所用的关键字无效。在某些实施例中，覆盖表可以修改若干字段以响应对基本映射表的查询。在某些实施例中，关键字范围（关键字值的范围）可以用于高效地标识要应用相同操作（取消或修改）的多个值。以这种方式，通过在覆盖表中创建“取消”条目而不修改映射表，可以（高效地）从映射表中“删除”若干元组。在这种情况下，覆盖表可以包括与非关键字数据字段不相关联的关键字。

[0099] 现在转向图 9，显示了在包括映射和覆盖表的系统中处理读请求的方法的一个实施例。为了响应正被接收的读请求（方框 900），产生该请求对应的映射表的关键字（方框 908）和第一个覆盖表的关键字（方框 902）。在这个实例中，对覆盖和映射表的访问被显示为并发地出现。不过，在其他实施例中，对这些表的访问可以以任何期望的次序非并发地（如顺序地或以其他方式在时间上分开地）执行。使用为映射表产生的关键字，可以从映射表中检索对应的元组（方框 910）。如果第一个覆盖表包含覆盖表关键字对应的“取消”条目（条件框 906），在映射表中发现的任何元组都被视为无效并且可以向请求者返回对这个结果的

指示。相反，如果该覆盖表包含覆盖表关键字对应的“修改”条目(条件框 912)，在第一个覆盖表条目中的值可以用于修改从映射表中检索的元组中的一个或多个字段(方框 922)。一旦完成了这个过程，便根据来自映射表的元组(无论是否修改过)产生第二个覆盖表的关键字(方框 914)，并且在第二个覆盖表中完成第二次查找(方框 916)，它可以是与第一个覆盖表相同的表也可以是不同的表。如果在第二个覆盖表中发现了“取消”条目(条件框 920)，来自映射表的元组就被视为无效(方框 918)。如果在第二个覆盖表中发现了“修改”条目(条件框 924)，来自映射表的元组的一个或多个字段可以被修改(方框 926)。这样的修改可以包括撤消元组、规格化元组或其他。修改后的元组然后可以返回给请求者。如果第二个覆盖表不包含修改条目(条件框 924)，可以不修改地向请求者返回元组。在某些实施例中，覆盖表的至少某些部分可以被高速缓存以提供对其内容更快的存取。在多个实施例中，在第一个覆盖表中检测出的取消条目可以用来使任何其他对应的查找(如方框 914、916 等)短路。在其他实施例中，可以并行和“竞争”地执行访问。众多这样的实施例都是可能并被预期。

[0100] 现在转向图 10，显示了对映射表内若干等级的展平操作的一个实施例的广义框图。在多个实施例中，响应检测到一个或多个条件可以执行展平操作。例如，随着时间过去，由于插入新记录，映射表 340 在增长并积聚若干等级，对查询关键字值搜索更多等级的成本可能变得不合意地高。为了限制要搜索的等级数量，可以把多个等级展平为单一新等级。例如，在逻辑上接近或在时间顺序上相邻的两个或更多等级可以选为进行展平操作。在两个或更多记录对应于同一关键字值时，最年轻的记录可以保留而其他记录不包括在新的“展平的”等级中。在这样的实施例中，对于给定关键字的搜索，新展平的等级将返回与由对应的多个等级的搜索提供的相同的结果。由于在新展平的等级中搜索的结果与其取代的两个或更多等级相比不改变，所以展平操作不需要与对映射表的更新操作同步。换言之，对表的展平操作可以关于对该表的更新异步地执行。

[0101] 正如先前注意，更老的等级在其映射不被修改的意义上是固定的(即从 A 至 B 的映射保持不变)。结果，对正在被展平等级的修改不进行(如由于用户的写入)并且不要求对这些等级的同步锁定。此外，在基于节点的集群环境中，其中每个节点都可以存储索引的更老等级的副本(如关于图 6 的讨论)，可以在一个节点上采取展平操作而不需要锁定其他节点中的对应等级。结果，处理可以在全部节点中继续同时展平以异步方式发生在任何节点上。在后面的时间点，其他节点可以展平等级或使用已经展平的等级。在一个实施例中，已经被用于形成展平后等级的两个或更多等级可以被保留用于故障恢复、镜像法或其他目的。除了以上情况，在多个实施例中，已经被取消的记录不可以被重新插入到新等级中。例如可以执行以上介绍的展平以响应检测到映射表中的等级数量已经达到了给定阈值。作为替代，可以执行展平以响应检测到一个或多个等级的尺寸已经超过了某阈值。可以考虑的又一个条件是系统上的负载。是否展平这些等级的决策除单独考虑这些条件外还可以考虑它们的组合。是否展平的决策也可以对该条件的当前值以及该条件将来的预测值都进行考虑。可以执行展平的其他条件也是可能的并被预期。

[0102] 在展示的实例中，若干记录被简单显示为关键字和指针对。为了易于展示，页面被显示为包括四个记录。“F”级及其下一个相邻的逻辑邻居，“F-1”级可以被考虑用于展平操作。“F”级可以年轻于“F-1”级。尽管这里显示了两个等级要被展平，但是有可能并预期了

可以选择三个或更多等级进行展平。在所示实例中“F-1”级具有的记录可以存储与“F”级中发现的关键字值相同。使用双向箭头标识出存储着横跨两个相邻等级的相同关键字值的记录。

[0103] 新的“新 F”级包括在“F”级和“F-1”级中发现的重复关键字值所对应的关键字。此外，新的“新 F”级包括存储着重复关键字值的若干记录中最年轻(或在这种情况下更年轻)记录所对应的指针值。例如，“F”级和“F-1”级每级都包括存储着关键字值 4 的记录。更年轻的记录在“F”级中并且这个记录还存储了指针值 512。所以，“F”级包括的记录存储着关键字值 4 也存储着指针值 512，而不是在更老的“F-1”级中发现的指针值 656。此外，新的“新 F”级包括具有唯一在“F”级和“F-1”级之间发现的关键字值的记录。例如，新的“新 F”级包括的记录具有在“F”级中发现的关键字和指针对 6 和 246，以及在“F-1”级中发现的关键字和指针对 2 和 398。如图所示，等级内的每个页面按关键字值分选。

[0104] 正如以上注意，在多个实施例中，覆盖表可以用于修改或取消基本映射表中若干关键字值所对应的元组。这样的覆盖表可以以类似于映射表的方式管理。例如，可以展平覆盖表并且把相邻条目合并在一起以节省空间。作为替代，管理覆盖表的方式可以与管理映射表所用的方式不同。在某些实施例中，覆盖表可以包含单一条目，它引用覆盖表关键字的某范围。以这种方式，能够限制覆盖表的尺寸。例如，如果覆盖表包含 k 个有效条目，覆盖表(展平后)需要包含不多于 k+1 个条目，把若干范围标注为无效，对应于映射表中有效条目之间的间隙。所以，覆盖表可以用于以相对高效的方式标识出可以从映射表撤消的元组。除了以上情况，虽然先前的讨论介绍了使用覆盖表取消或修改对来自映射表的若干请求的响应，但是覆盖表也可用于在映射表的展平操作期间取消或修改若干值。所以，在映射表的展平操作期间创建新等级时，可以取消否则有可能被插入到新等级中的关键字值。作为替代，可以先修改某值再插入到新等级中。这样的修改可以引起映射表中给定范围关键字值所对应的单一记录，(在新等级中)以多个记录取代——每个都对应原始记录的一个子范围。此外，某记录可以用对应于更小范围的新记录取代，多个记录也有可能由单一记录取代，其范围覆盖了原始记录的全部范围。一切这样的实施例都被预期。

[0105] 现在参考图 11，显示了对映射表内若干等级的展平操作实施例的广义框图。正如先前讨论，等级可以按时间排序。在展示的实例中，“F”级包括一个或多个索引并且对应的映射在逻辑上位于更老的“F-1”级之上。同样，“F”级在逻辑上位于更年轻的“F+1”级之下。同样，“F-2”级在逻辑上位于更年轻的“F-1”级之下而“F+2”级在逻辑上位于更老的“F+1”级之上。在一个实例中，“F”和“F-1”级可以被视为用于展平操作。使用双向箭头展示有若干记录存储着横跨两个相邻等级的相同关键字值。

[0106] 正如早先介绍，新的“新 F”级包括在“F”级和“F-1”级中发现的重复关键字值所对应的关键字值。此外，新的“新 F”级包括存储着重复关键字值的若干记录中最年轻(或在这种情况下更年轻)的记录所对应的指针值。在完成展平操作后，“F”级和“F-1”级可能尚未从映射表中去除。同样，在基于节点的集群中，每个节点都可以验证其准备好采用新的单一等级，比如“新 F”级。并且不再使用它取代的两个或更多的等级(比如“F”级和“F-1”级)。可以先进行这种验证再使新等级变为取代。在一个实施例中，两个或更多的被取代的等级比如“F”级和“F-1”级可以被保留在存储器中，用于故障恢复、镜像法或其他目的。为了保持等级及其映射的时间顺序，新展平的 F 级在逻辑上被放置更年轻的等级(如 F+1 级)

之下以及它取代的原始等级(如“F”级和“F-1”级)之上。

[0107] 现在转向图12,显示了展平映射表内若干等级的方法1000的一个实施例。在以上介绍的网络架构100和映射表340中实施的组件一般可以根据方法1000操作。为了讨论目的,在这个实施例中的步骤以连续的次序显示。不过,某些步骤可以以不同于所示的次序出现,某些步骤可以并发地进行,某些步骤可以与其他步骤结合,而某些步骤在另一个实施例中可以不出现。

[0108] 在方框1002,为映射表和对应的索引分配了存储空间。在方框1004,为展平映射表内两个或更多等级判断一个或多个条件。例如,搜索映射表内当前数量的等级的成本可能大于执行展平操作的成本。此外,成本可以基于以下各项至少其一:要展平的结构中等级的当前(或预测)数量、一个或多个等级中条目的数量、会被取消或修改的映射条目的数量以及系统上的负载。成本还可以包括执行对应操作的时间、一条或多条总线的占用、对应操作期间使用的存储空间、一组等级中重复条目的数量已经达到了某个阈值等。此外,每个等级内许多记录的计数可用于估算对两个相邻等级执行的展平操作何时可以产生记录数量等于前一个等级内的记录数量两倍的新的单一等级。单独地或以任何组合地取得的这些条件以及其他条件是可能的并被预期。

[0109] 在方框1006,由于存储了数据并发现了新映射,所以访问并更新了索引和映射表。映射表内的许多等级随着把新记录插入到映射表中而增加。如果检测出展平映射表内两个或更多等级的条件(条件框1008),那么在方框1010,标识出要展平的一组或多组等级。一组等级可以包括两个或更多等级。在一个实施例中,两个或更多等级是相邻的等级。尽管最低等级即最老等级可以是进行展平的最优候选,也可以选择更年轻的组。

[0110] 在方框1012,对每组产生新的单一等级,包括对应的组内最新的记录。在更早的实例中,新的单一“新F”级包括“F”级和“F+1”级当中最年轻的记录。在方框1014,在基于节点的集群中,可以从集群内每个节点中请求应答以表明各自节点准备好采用由展平操作所产生的新等级。当每个节点应答它可以采用新等级时,在方框1016,用新等级取代已标识组内的当前等级。在其他实施例中,不需要横跨节点的同步。在这样的实施例中,某些节点可以先于其他节点开始使用新等级。另外,即使在新展平的等级可用之后,某些节点也可以继续使用原始等级。例如,某特定节点可以使原始等级数据被高速缓存并使用,而不使用新展平等级的非高速缓存的数据。众多这样的实施例是可能的并被预期。

[0111] 现在转向图13,显示了高效处理映射表内批量阵列任务的方法1100的一个实施例。类似于介绍的其他方法,在以上介绍的网络架构100和映射表340中实施的组件一般可以根据方法1100操作。此外,在这个实施例中的步骤以连续的次序显示。不过,某些步骤可以以不同于所示的次序出现,某些步骤可以并发地进行,某些步骤可以与其他步骤结合,而某些步骤在另一个实施例中可以不出现。

[0112] 在映射表内以压缩格式存储信息可以实现细粒度映射,它可以允许直接操作映射表内的映射信息,作为普通的批量阵列任务的替代。直接映射操作可以降低I/O网络和总线的通信量。正如早先介绍,闪存的“查找时间”短,这允许许多相关读操作以短于自旋磁盘单次操作的时间发生。这些相关读取可以用于执行联机的细粒度映射以集成节省空间的特征象压缩和去重复。此外,这些相关读取操作可以允许存储器控制器174完全在映射表内执行批量阵列任务,而不是访问(读取和写入)存储设备176a-176m内存储的用户数据。

[0113] 在方框 1102，接收大型即批量阵列任务。例如，批量复制或移动请求可以对应于几十个或几百个虚拟机的备份，再加上由这些虚拟机执行和更新的企业应用数据。收到的请求与全部这种数据的移动、分支、克隆或复制相关联，与收到的请求相关联的数据量可能大到 16 吉字节（GB）或更大。要是为了处理这个请求而访问用户数据，大量处理时间可能要花费在该请求上因而系统性能要降低。此外，典型情况下，虚拟化环境具有的输入 / 输出（I/O）总资源小于物理环境。

[0114] 在方框 1104，存储器控制器 174 可以存储所收到请求所对应的指示，它使新关键字的范围与老关键字的范围相关，其中关键字的两个范围都对应于收到的请求。例如，如果收到请求是 16GB 数据的复制，就可以存储 16GB 数据所对应的开始关键字值和结束关键字值。同样，开始和结束关键字值的每一个都可以包括卷 ID、收到请求内的逻辑或虚拟地址、快照 ID、扇区号等。在一个实施例中，这种信息可以与索引中存储的信息分开存储，比如初级索引 310、次级索引 320、三级索引 330 等。不过，当处理后面请求期间访问索引时可以访问这种信息。

[0115] 在方框 1106，数据存储器控制器 174 可以向客户计算机系统 110a-110c 的对应客户机传达响应，表明完成了收到的请求而不预先访问用户数据。所以，存储器控制器 174 处理收到的请求时处理器 122 上可以停机时间很短或没有以及没有负载。

[0116] 在方框 1108，存储器控制器 174 可以设置条件、指示或旗标，或者缓冲区更新操作，用于更新映射表中的一个或多个记录，对应于取代映射表中老关键字的新关键字。对于移动请求和复制请求，都可以把新关键字所对应的一个或多个新记录插入映射表。关键字可以被插入到所创建的新的最高等级，正如早先介绍。对于移动请求，在新记录已经被插入到映射表后可以从映射表中去除对应的一个或多个老记录。映射表中的记录或者立即或者在更晚时间实际地更新。

[0117] 对于清零或擦除请求，可以存储某范围的关键字值现在对应于一串二进制零的指示。另外，如以上讨论，覆盖表可以用于标识不是（或不再）有效的关键字值。用户数据可以不被盖写。对于擦除请求，用户数据可以在更晚时间被盖写，此时“释放的”存储位置被分配给后续存储（写入）请求的新数据。对于外面导向的碎片整理请求，可以选择相邻的地址用于扇区重新组织，它可以有助于在客户计算机系统 110a-110c 的客户机上执行的应用程序。

[0118] 如果存储器控制器 174 收到了新关键字之一所对应的数据存储访问请求（方框 1110），并且新关键字已经被插入映射表（条件框 1112），那么在方框 1114，可以用新关键字访问索引和映射表。例如，可以用新关键字访问初级索引 310、次级索引 320 或三级索引 330 中任一个。当映射表的一个或多个页面由索引识别出时，然后就访问这些已识别的页面。在方框 1116，可以用映射表中找到的与新关键字相关联的物理指针值服务存储访问请求。

[0119] 如果存储器控制器 174 收到了新关键字之一所对应的数据存储访问请求（方框 1110），并且新关键字尚未被插入映射表（条件框 1112），那么在方框 1118，可以用对应的老关键字访问索引和映射表。可以访问保持着老关键字范围和新关键字范围的存储器以确定对应的老关键字值。当映射表的一个或多个页面由索引识别出时，然后就访问这些已识别的页面。在方框 1120，可以用映射表中找到的与老关键字相关联的物理指针值服务存储访问请求。

[0120] 现在转向图 14，显示了存储设备内数据布局架构实施例的广义框图。在一个实施例中，存储设备 176a-176m 内的数据存储位置可以被安排到独立冗余磁盘阵列（RAID）中。如图所示，不同类型的数据可以根据数据布局架构存储在存储设备 176a-176k 中。在一个实施例中，存储设备 176a-176k 的每台都是 SSD。SSD 内的分配单位可以包括 SSD 内的一个或多个擦除块。

[0121] 用户数据 1230 可以存储在一台或多台存储设备 176a-176k 内包括的一个或多个页面内。在 RAID 条带与存储设备 176a-176k 之一的每个交集内，存储的信息可以被格式化为一系列逻辑页面。每个逻辑页面又可以包括页面头以及对页面中数据的校验和。发出读取时它可以为了一个或多个逻辑页面并且在每个页面中的数据都可以用校验和验证。由于每个逻辑页面可以包括页面头，它包含用于该页面的校验和（它可以被称为“介质”校验和），所以用于数据的实际页面尺寸可能小于一个逻辑页面。在某些实施例中，对于存储设备间恢复数据 1250 比如 RAID 奇偶校验信息的页面，页面头可以更小，使得奇偶校验页面保护该数据页面中的页面校验和。在其他实施例中，可以计算存储着设备间恢复数据 1250 的奇偶校验页面中的校验和，使得该数据页面校验和的校验和与覆盖对应数据页面的奇偶校验页面的校验和相同。在这样的实施例中，奇偶校验页面头不需要小于数据页面头。

[0122] 设备间 ECC 数据 1250 可以是从保留用户数据的其他存储设备上的一个或多个页面产生的奇偶校验信息。例如，设备间 ECC 数据 1250 可以是在 RAID 数据布局架构中所用的奇偶校验信息。尽管存储的信息被显示为存储设备 176a-176k 中的相邻逻辑页面，但是在本领域众所周知逻辑页面可以以随机次序安排，其中每台存储设备 176a-176k 都是 SSD。

[0123] 设备内 ECC 数据 1240 可以包括由设备内冗余模式使用的信息。设备内冗余模式利用了给定存储设备内的 ECC 信息，比如奇偶校验信息。这种设备内冗余模式及其 ECC 信息对应于给定设备并可以被保留在给定设备之内，但是与可以被内部产生并由设备自身保留的 ECC 截然不同。一般来说，设备的内部产生并保留的 ECC 对于包括该设备在内的系统是不可见的。

[0124] 设备内 ECC 数据 1240 还可以被称为设备内故障恢复数据 1240。设备内故障恢复数据 1240 可以用于保护给定存储设备免于隐藏的扇区故障（LSE）。LSE 是访问给定扇区时才被发现的故障。所以，先前存储在给定扇区中的任何数据都可能丢失。在存储设备故障后 RAID 重建期间遇到时，单一 LSE 就可能导致数据丢失。术语“扇区”典型情况下是指 HDD 上的存储器基本单位，比如磁盘上给定磁轨内的区段。这里，术语“扇区”也可以指在 SSD 上分配的基本单位。当存储设备内的给定扇区或其他存储单位不可访问时就发生了隐藏的扇区故障（LSE）。对于该给定扇区的读或写操作可能无法完成。此外，可能存在着无法纠正的纠错码（EDD）故障。

[0125] 给定存储设备内包括的设备内故障恢复数据 1240 可以用于提高给定存储设备内的数据存储可靠性。设备内故障恢复数据 1240 是增加到另一台存储设备内可以包括的其他 ECC 信息，比如在 RAID 数据布局架构中采用的奇偶校验信息。

[0126] 在每台存储设备内，设备内故障恢复数据 1240 可以被存储在一个或多个页面中。正如本领域技术人员熟知，通过对用户数据 1230 内的选定信息位执行某函数可以得到设备内故障恢复数据 1240。基于 XOR 的运算可以用于导出奇偶校验信息以存储在设备内故障恢复数据 1240 中。设备内冗余模式的其他实例包括单奇偶校验检查（SPC）、最大距离可分

(MDS)擦除码、交错奇偶校验检查码(IPC)、混合 SPC 和 MDS 码(MDS+SPC),以及列对角奇偶校验(CDP)。这些模式根据传递的可靠性而改变并且计算了依赖于模式的开销。

[0127] 除了以上介绍的故障恢复信息以外,系统还可以被配置为对设备上区域计算校验和值。例如,向设备写入信息时可以计算校验和。这种校验和由系统存储。从设备读回信息时,系统可以再次计算校验和并将其与原始被存储的值进行对比。如果两个校验和不同,该信息不是恰当地读取,所以系统可以使用其他模式恢复数据。校验和函数的实例包括循环冗余检查(CRC)、MD5 和 SHA-1。

[0128] SSD 内的擦除块可以包括几个页面。一个页面可以包括 4KB 的数据存储空间。擦除块可以包括 64 个页面即 256KB。在其他实施例中,擦除块可以大到 1 兆字节(MB)并包括 256 个页面。为了降低跟踪分配单位的开销,可以以提供尺寸足够大而数量又相对少的单位的方式选择分配单位的尺寸。在一个实施例中,一个或多个状态表可以保持分配单元的状态(已分配、空闲、已擦除、有故障)、磨损等级以及在分配单元内已经发生的(可纠正和 / 或不可纠正)故障数量的计数。在一个实施例中,分配单位与 SSD 的总存储容量相比相对小。用于页面、擦除块和其他单位配置的数据存储空间的其他量也是可能的并被预期。

[0129] 元数据 1260 可以包括页面头信息、RAID 条带识别信息、用于一个或多个 RAID 条带的日志数据等。在多个实施例中,每个条带开始处的单一元数据页面都可以从其他条带头重建。作为替代,这个页面可以处于奇偶校验碎片中的不同偏移量处,使得数据能够由设备间的奇偶校验保护。在一个实施例中,元数据 1260 可以存储着表明这份数据不要去重复的特定旗标值或与其相关联。

[0130] 除了设备间的奇偶校验保护和设备内的奇偶校验保护以外,存储设备 176a-176k 中的每个页面都可以包括附加保护,比如每个给定页面内存储的校验和。校验和(8 字节、4 字节或其他)可以被放置在页面内部,页面头后以及对应的数据前,它可以被压缩。对于又一个等级的保护,数据位置信息可以被包括在校验和值中。每个页面中的数据可以包括这种信息。虚拟地址和物理地址都可以包括在这种信息中。扇区号、数据块和偏移量号、磁轨号、平面号等也可以包括在这种信息中。如果地址转换映射表的内容丢失,这种映射信息也可以用于重建该表。

[0131] 在一个实施例中,存储设备 176a-176k 中的每个页面都存储着特定类型的数据,比如数据类型 1230-1260。作为替代,页面可以存储不止一种类型的数据。页面头可以存储对于对应页面的数据类型进行识别的信息。在一个实施例中,设备内冗余模式把设备划分为用于存储用户数据的位置组。例如,划分可以是 RAID 布局内的条带对应的设备内的一组位置。在所示实例中,为了易于展示,仅仅显示了两个条带 1270a 和 1270b。

[0132] 在一个实施例中,存储器控制器 174 内的 RAID 引擎可以判断对存储设备 176a-176k 使用的保护等级。例如,RAID 引擎可以判定对存储设备 176a-176k 采用 RAID 双重奇偶校验。设备间冗余数据 1250 可以表示从对应的用户数据产生的 RAID 双重奇偶校验值。在一个实施例中,存储设备 176j 和 176k 可以存储双重奇偶校验信息。应当理解,其他等级的 RAID 奇偶校验保护也是可能的并被预期。此外,在其他实施例中,双重奇偶校验信息的存储可以在若干存储设备之间循环而不是被存储在每个 RAID 条带的存储设备 176j 和 176k 之内。为了易于展示和说明,双重奇偶校验信息的存储被显示为存储在存储设备 176j 和 176k 中。尽管每台存储设备 176a-176k 都包括多个页面,但是为了易于展示仅仅标注了

页面 1212 和页面 1220。

[0133] 应当注意，以上介绍的实施例可以包括软件。在这样的实施例中，实施本方法和 / 或机制的程序指令可以在计算机可读介质上传送或存储。被配置为存储程序指令的众多类型的介质是可用的并包括硬盘、软盘、CDROM、DVD、闪存、可编程 ROM （PROM）、随机存取存储器（RAM）以及多种其他形式的易失性或非易失性存储器。

[0134] 在多个实施例中，本文介绍的方法和机制的一个或多个部分可以形成云计算环境的一部分。在这样的实施例中，若干资源可以在因特网上提供为根据一个或多个不同模型的服务。这样的模型可以包括基础设施即服务（IaaS）、平台即服务（PaaS）以及软件即服务（SaaS）。在 IaaS 中，计算机基础设施被交付为服务。在这样的情况下，计算装备一般由服务供应商拥有和运行。在 PaaS 模型中，由开发商开发软件解决方案所用的软件工具和基本装备可以提供为服务并由服务供应商主办。典型情况下，SaaS 包括服务供应商把软件特许为按需服务。以上模型的众多组合是可能的并被预期。

[0135] 尽管已经相当详细地介绍了上面的若干实施例，但是一旦以上公开内容被全面认识到，众多变种和修改对于本领域技术人员将变得显而易见。以下权利要求书意在被解释为包含一切这样的变种和修改。

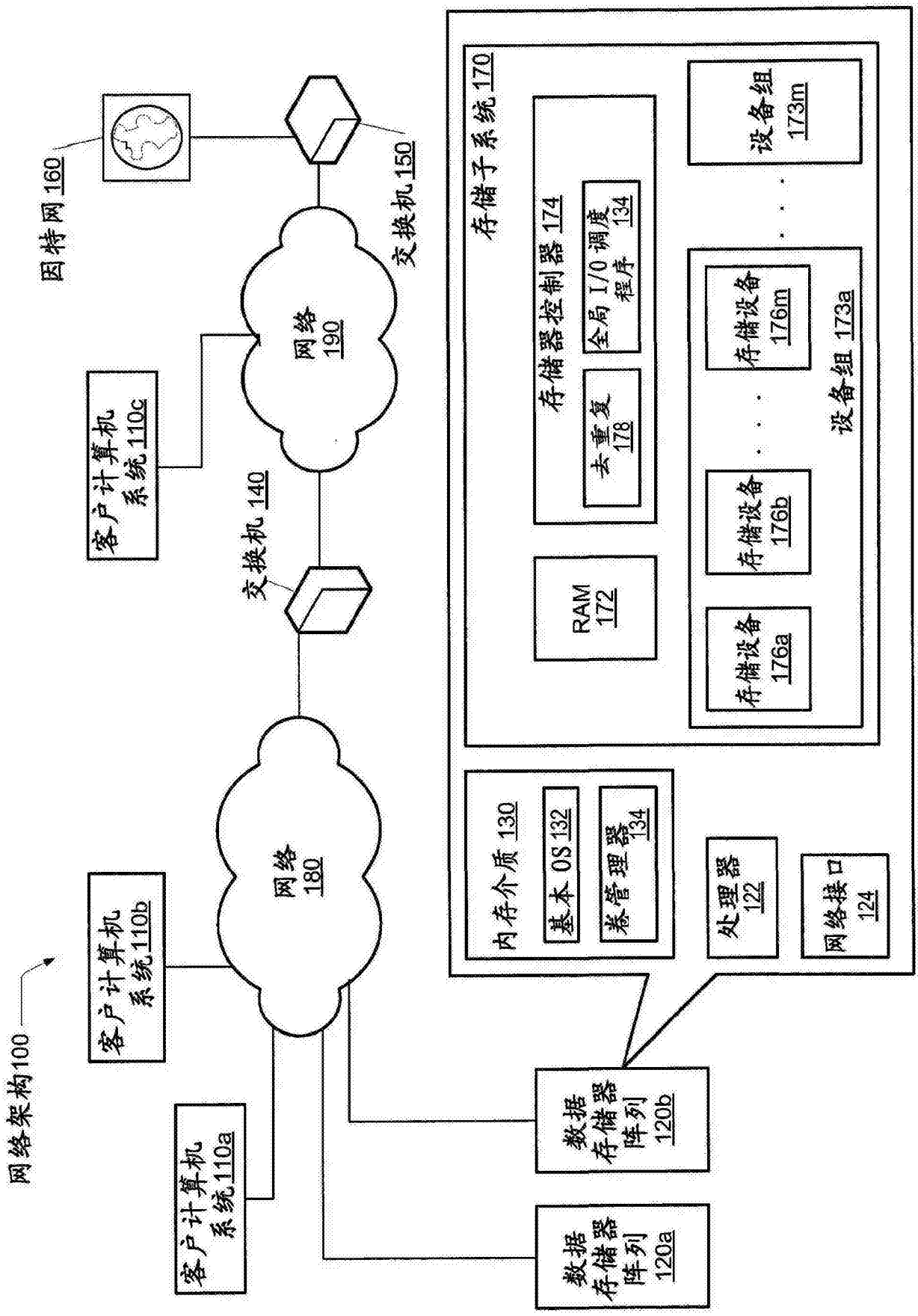

网络架构100
客户计算机系统110a
客户计算机系统110b
客户计算机系统110c
因特网160
网络180
网络190
交换机140
交换机150
数据存储器阵列120a
数据存储器阵列120b
存储子系统170
内存介质130
基本OS132
卷管理器134
处理器122
网络接口124
存储器控制器174
全局I/O调度程序134
去重复178
RAM172
设备组173a
存储设备176a
存储设备176b
存储设备176m
设备组173m

图 1

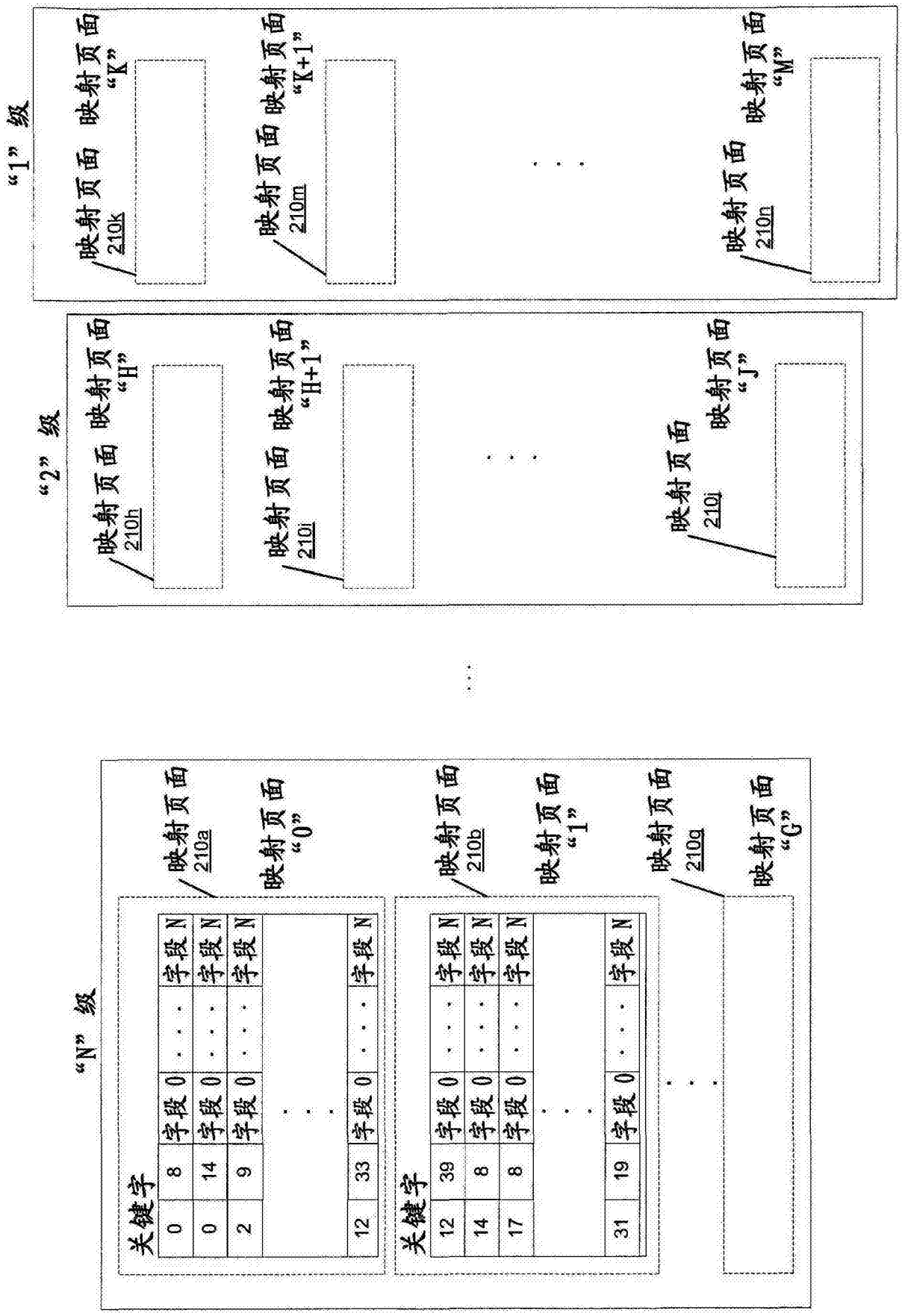

“1”级
映射页面 210k 映射页面“K”
映射页面 210m 映射页面“K+1”
映射页面 210n 映射页面“M”
“2”级
映射页面 210h 映射页面“H”
映射页面 210i 映射页面“H+1”
映射页面 210j 映射页面“J”
“N”级
映射页面 210a 映射页面“0”
关键字
0 8 字段0 ... 字段N
0 14 字段0 ... 字段N
2 9 字段0 ... 字段N
12 33 字段0 ... 字段N
映射页面 210b 映射页面“1”
关键字
12 39 字段0 ... 字段N
14 8 字段0 ... 字段N
17 8 字段0 ... 字段N
31 19 字段0 ... 字段N
映射页面 210q 映射页面“G”

图 2

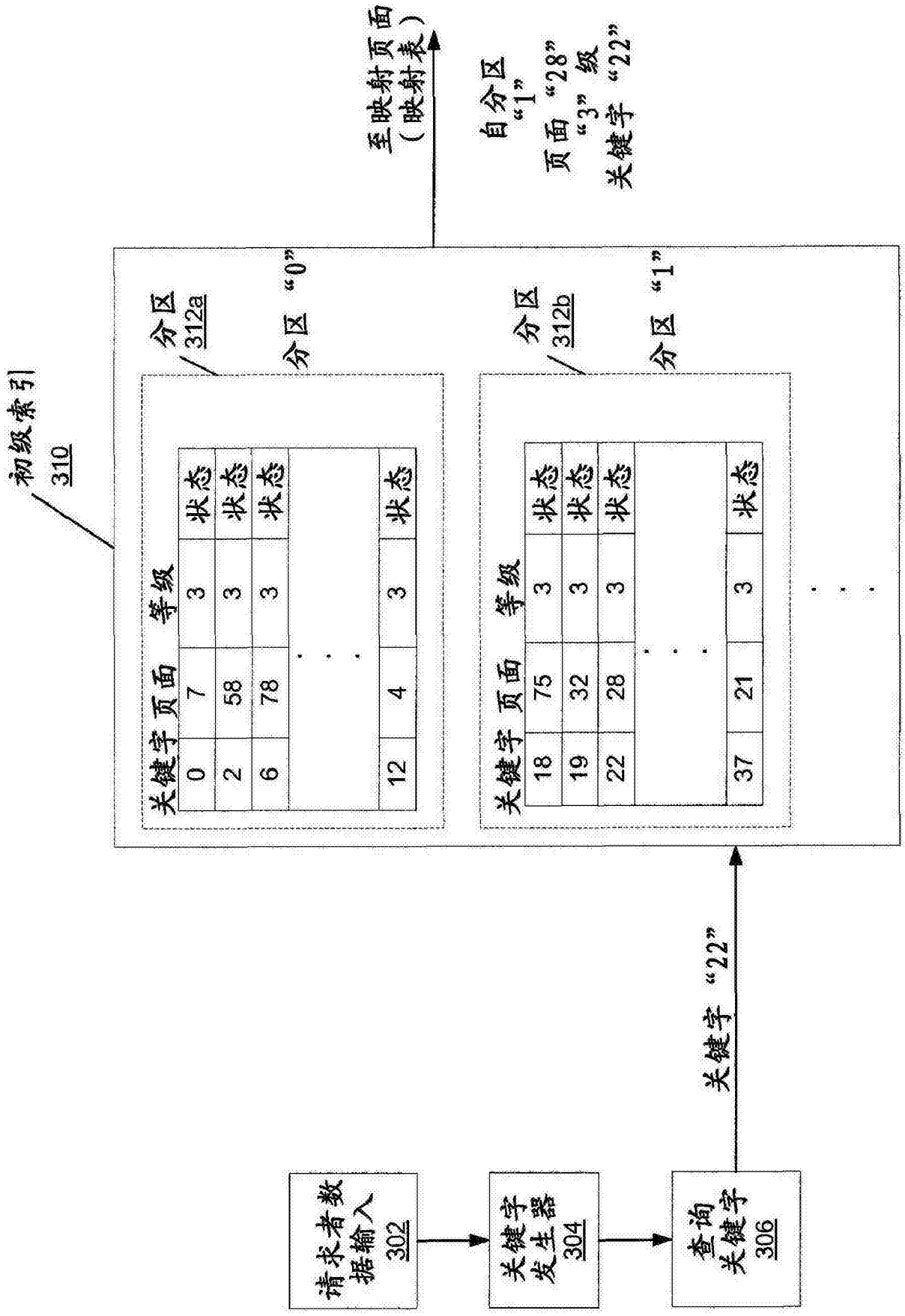

请求者数据输入 302
关键字发生器 304
查询关键字 306
关键字 “22”
初级索引 310
分区 312a
分区 “0”
关键字 页面 等级
0 7 3 状态
2 58 3 状态
6 78 3 状态
12 4 3 状态
分区 312b
分区 “1”
18 75 3 状态
19 32 3 状态
22 28 3 状态
37 21 3 状态
至映射页面（映射表）
自分区 “1”
页面 “28”
“3” 级
关键字 “22”

图 3A

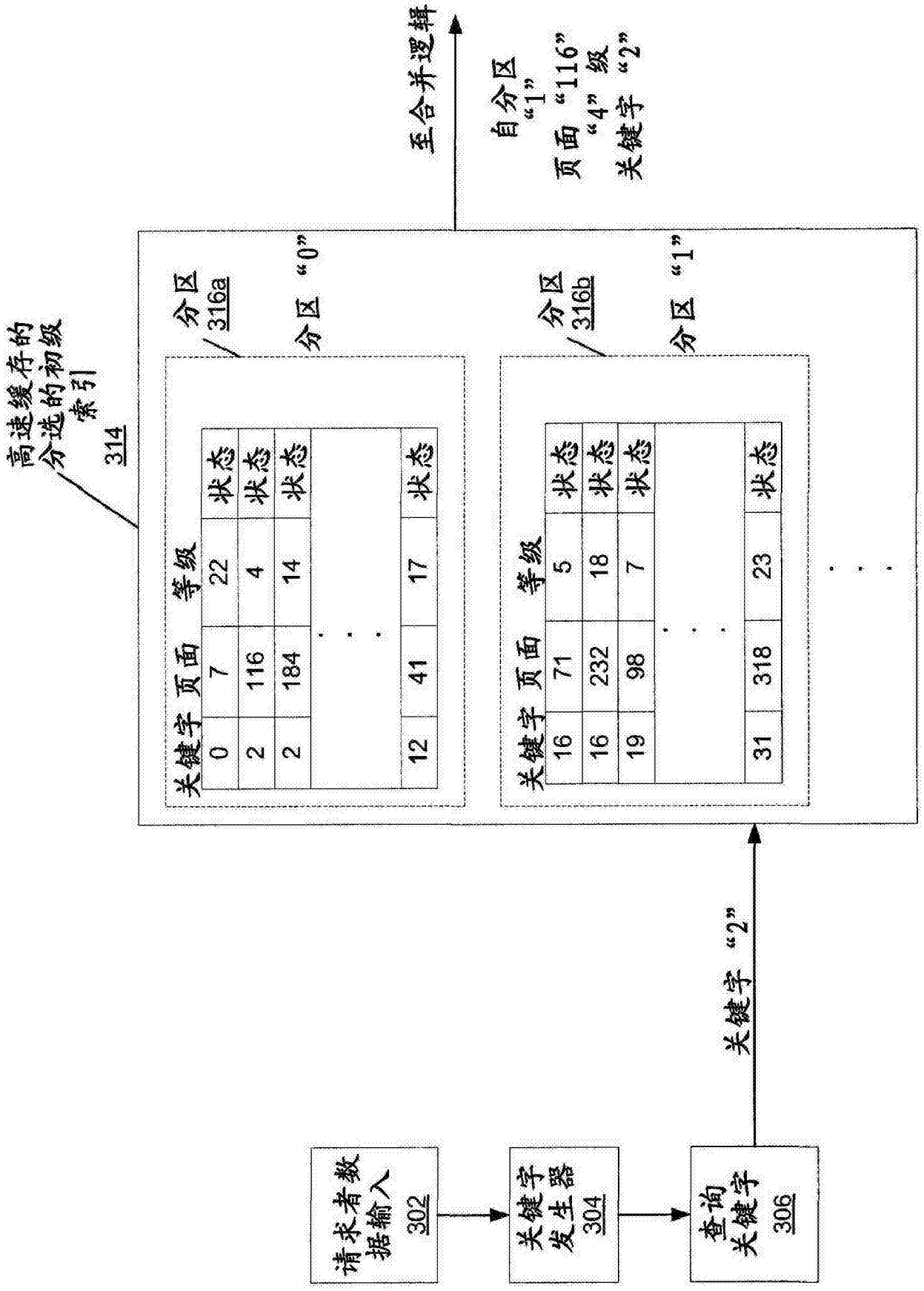

至合并逻辑
自分区 “1” 页面 “116” “4” 级 关键字 “2”
高速缓存的分选的初级索引 314
分区 316a
分区 “0”
关键字 页面 等级 状态
0 7 22
2 116 4
2 184 14
12 41 17
分区 316b
分区 “1”
关键字 页面 等级 状态
16 71 5
16 232 18
19 98 7
31 318 23
关键字 “2”
请求者数据输入 302
关键字发生器 304
查询关键字 306

图 3B

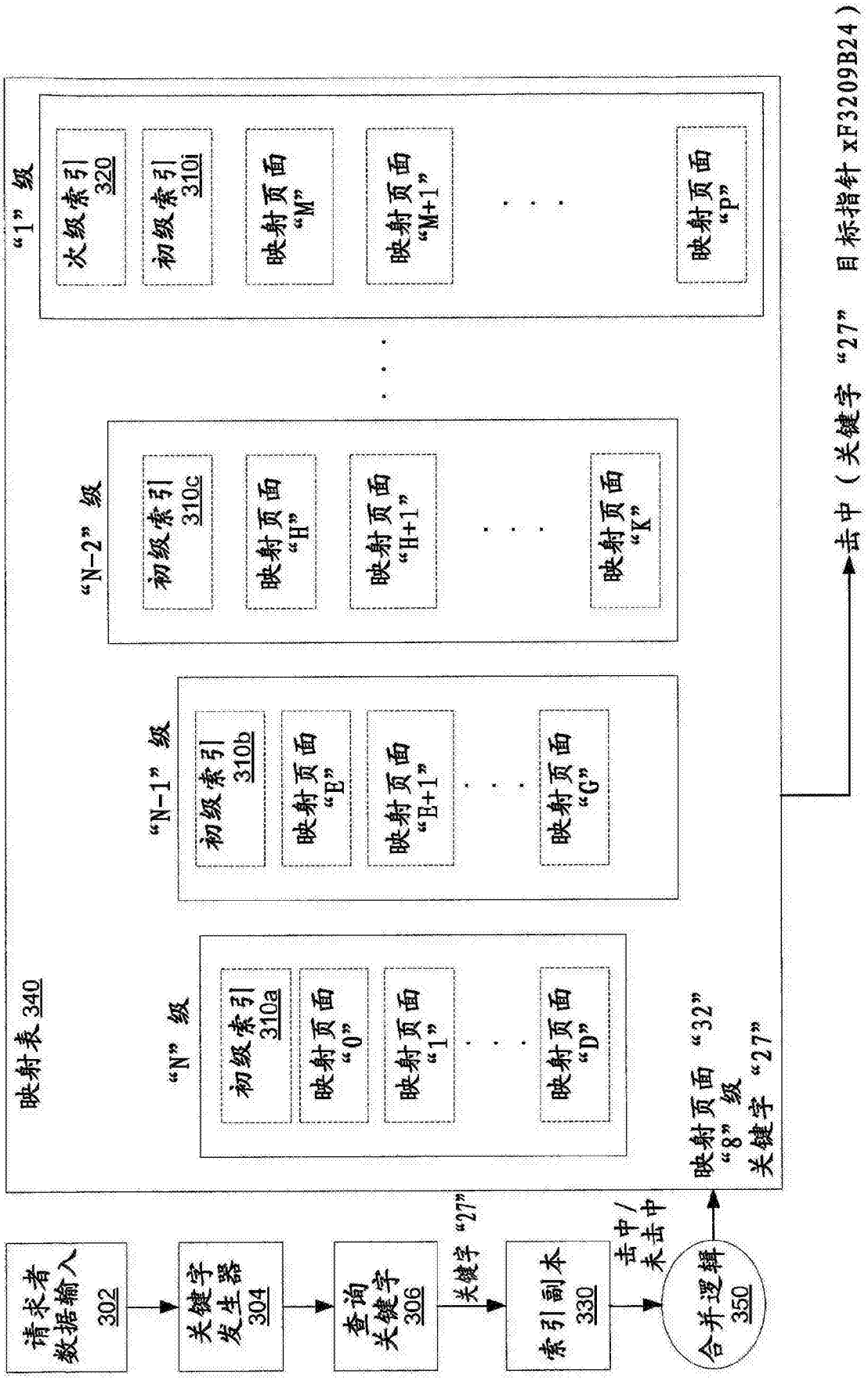

请求者数据输入 302
关键字发生器 304
查询关键字 306
关键字 "27"
索引副本 330
击中/未击中
合并逻辑 350
映射表 340
"N" 级
初级索引 310a
映射页面 "0"
映射页面 "1"
映射页面 "D"
"N-1" 级
初级索引 310b
映射页面 "E"
映射页面 "E+1"
映射页面 "G"
"N-2" 级
初级索引 310c
映射页面 "H"
映射页面 "H+1"
映射页面 "K"
"1" 级
次级索引 320
初级索引 310i
映射页面 "M"
映射页面 "M+1"
映射页面 "P"
映射页面 "32"
"8" 级
关键字 "27"
击中（关键字 "27" 目标指针 xF3209B24）

图 4

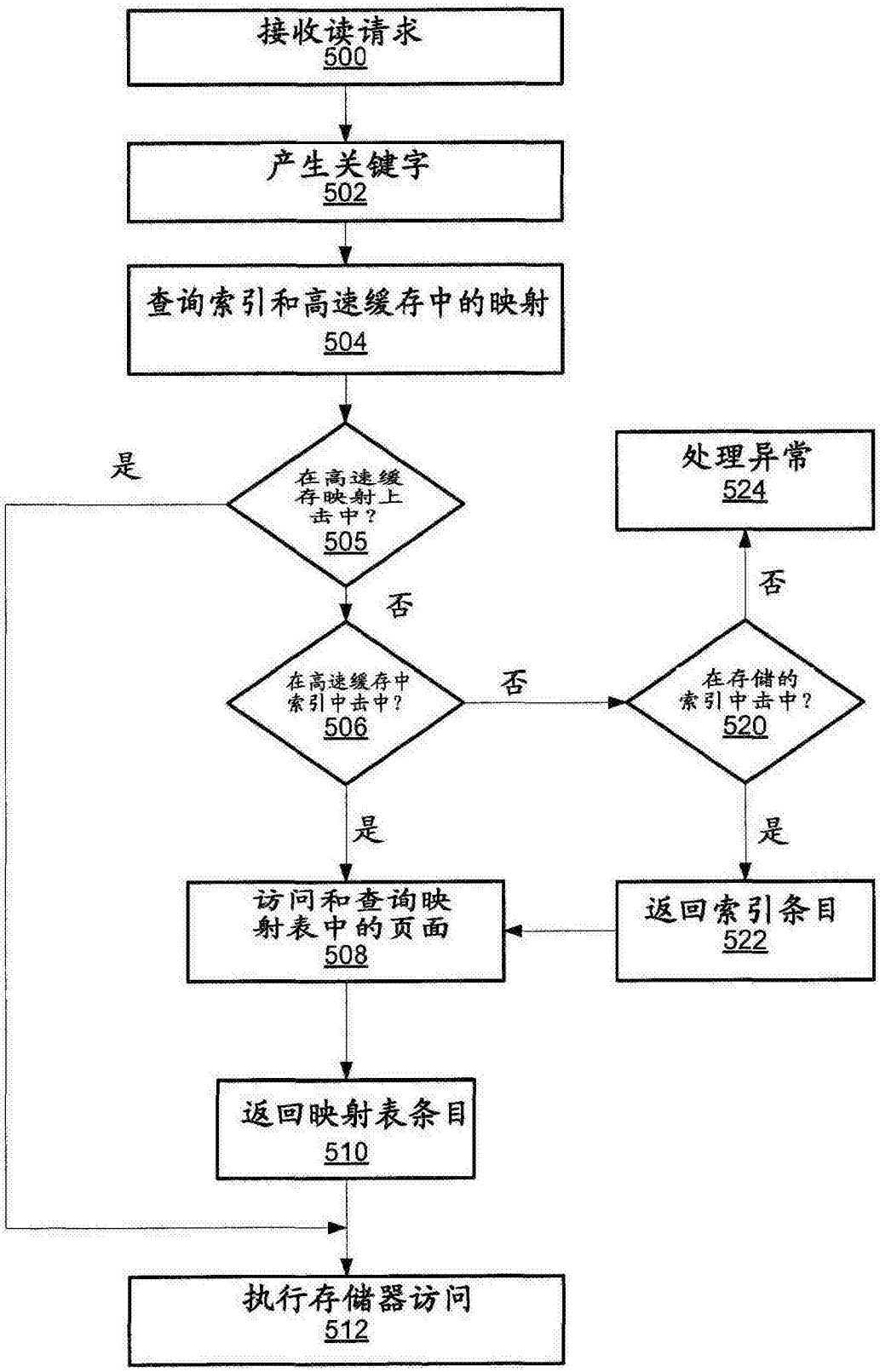

接收读请求
500
产生关键字
502
查询索引和高速缓存中的映射
504
在高速缓存映射上击中？
505
是
否
在高速缓存中索引中击中？
506
否
处理异常
524
否
在存储的索引中击中？
520
是
是
访问和查询映射表中的页面
508
返回索引条目
522
返回映射表条目
510
执行存储器访问
512

图 5A

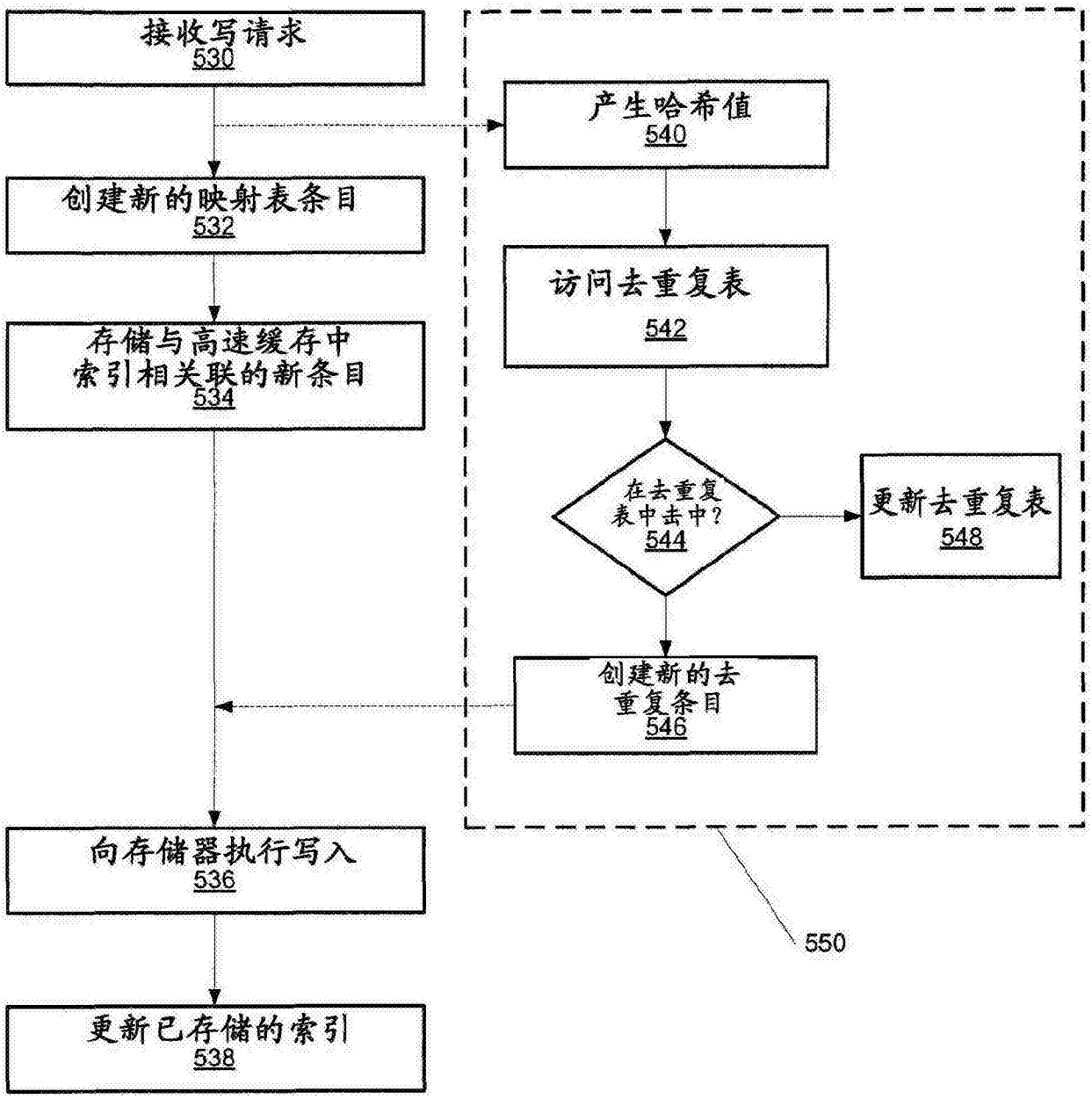

接收写请求
530
产生哈希值
540
创建新的映射表条目
532
访问去重复表
542
存储与高速缓存中
索引相关联的新条目
534
在去重复
表中击中?
544
更新去重复表
548
创建新的去
重复条目
546
向存储器执行写入
536
更新已存储的索引
538
550

图 5B

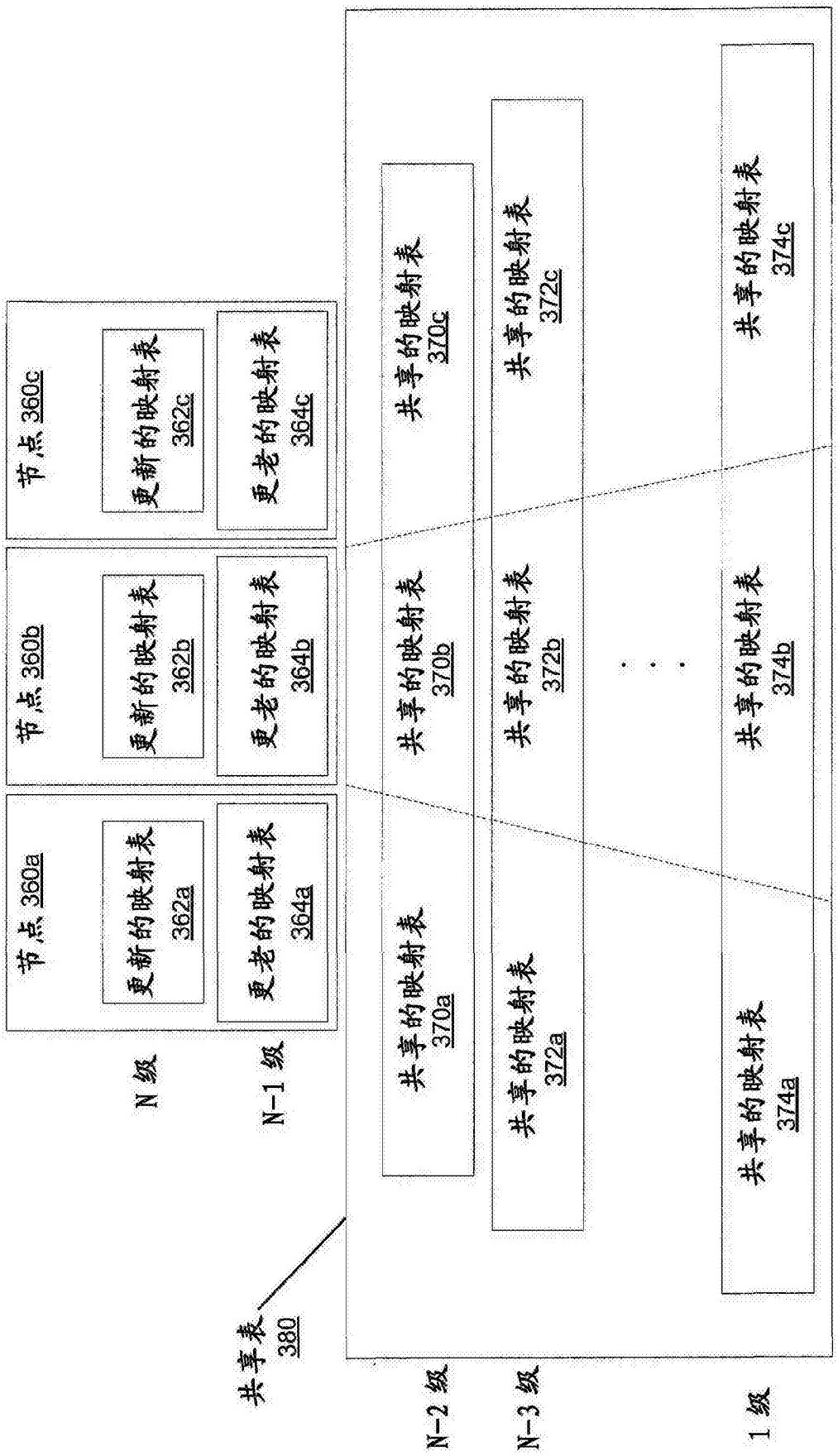

节点 360a
更新的映射表 362a
更老的映射表 364a
节点 360b
更新的映射表 362b
更老的映射表 364b
节点 360c
更新的映射表 362c
更老的映射表 364c
N 级
N-1 级
共享表 380
共享的映射表 370a
共享的映射表 370b
共享的映射表 370c
共享的映射表 372a
共享的映射表 372b
共享的映射表 372c
共享的映射表 374a
共享的映射表 374b
共享的映射表 374c
N-2 级
N-3 级
1 级

图 6

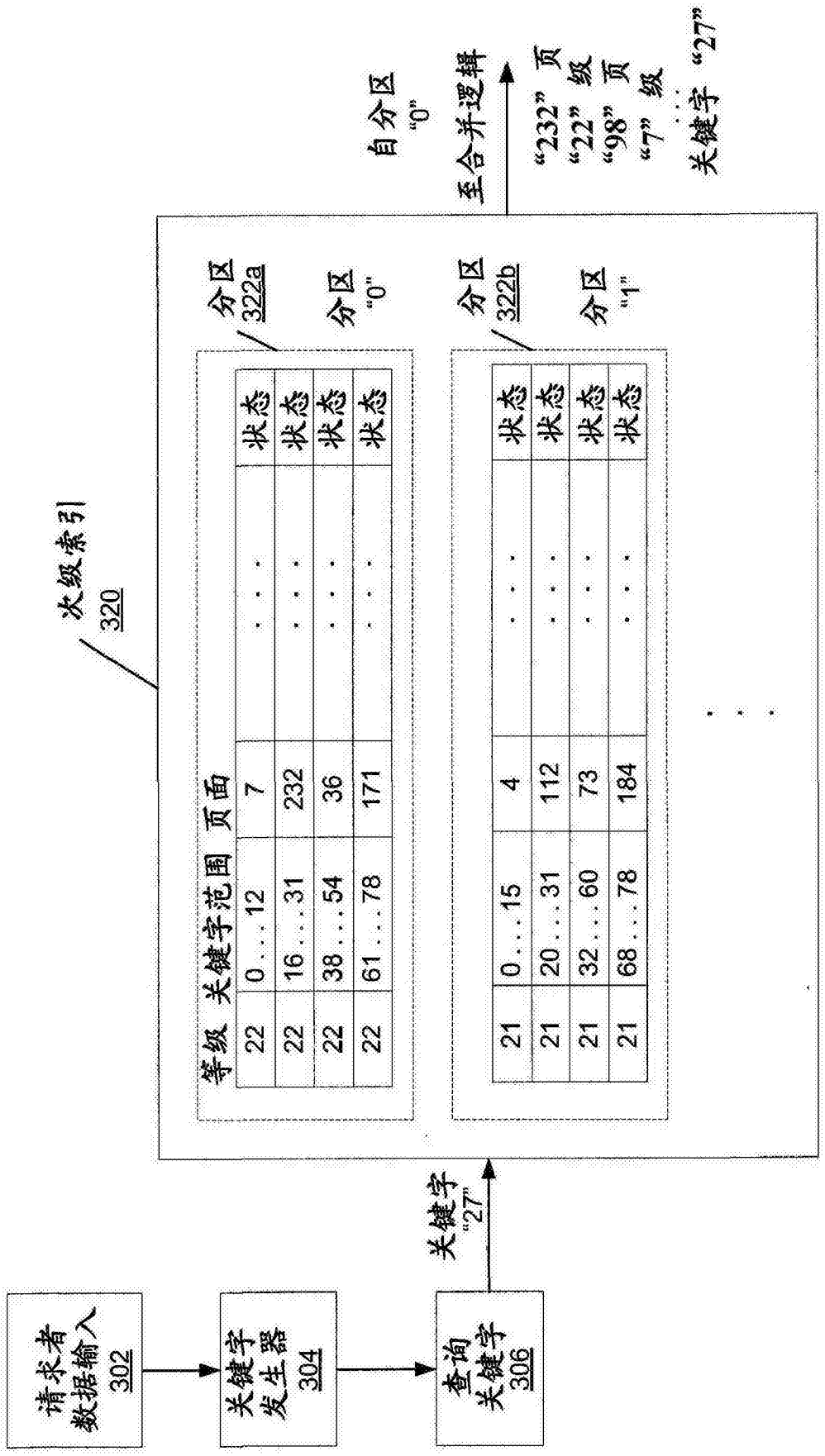

请求者数据输入 302
关键字发生器 304
查询关键字 306
关键字 "27"
次级索引 320
分区 322a
分区 "0"
分区 322b
分区 "1"
等级 关键字范围 页面
22 0...12 7 状态
22 16...31 232 状态
22 38...54 36 状态
22 61...78 171 状态
21 0...15 4 状态
21 20...31 112 状态
21 32...60 73 状态
21 68...78 184 状态
自分区 "0"
至合并逻辑
"232" 页
"22" 级
"98" 页
"7" 级
关键字 "27"

图 7

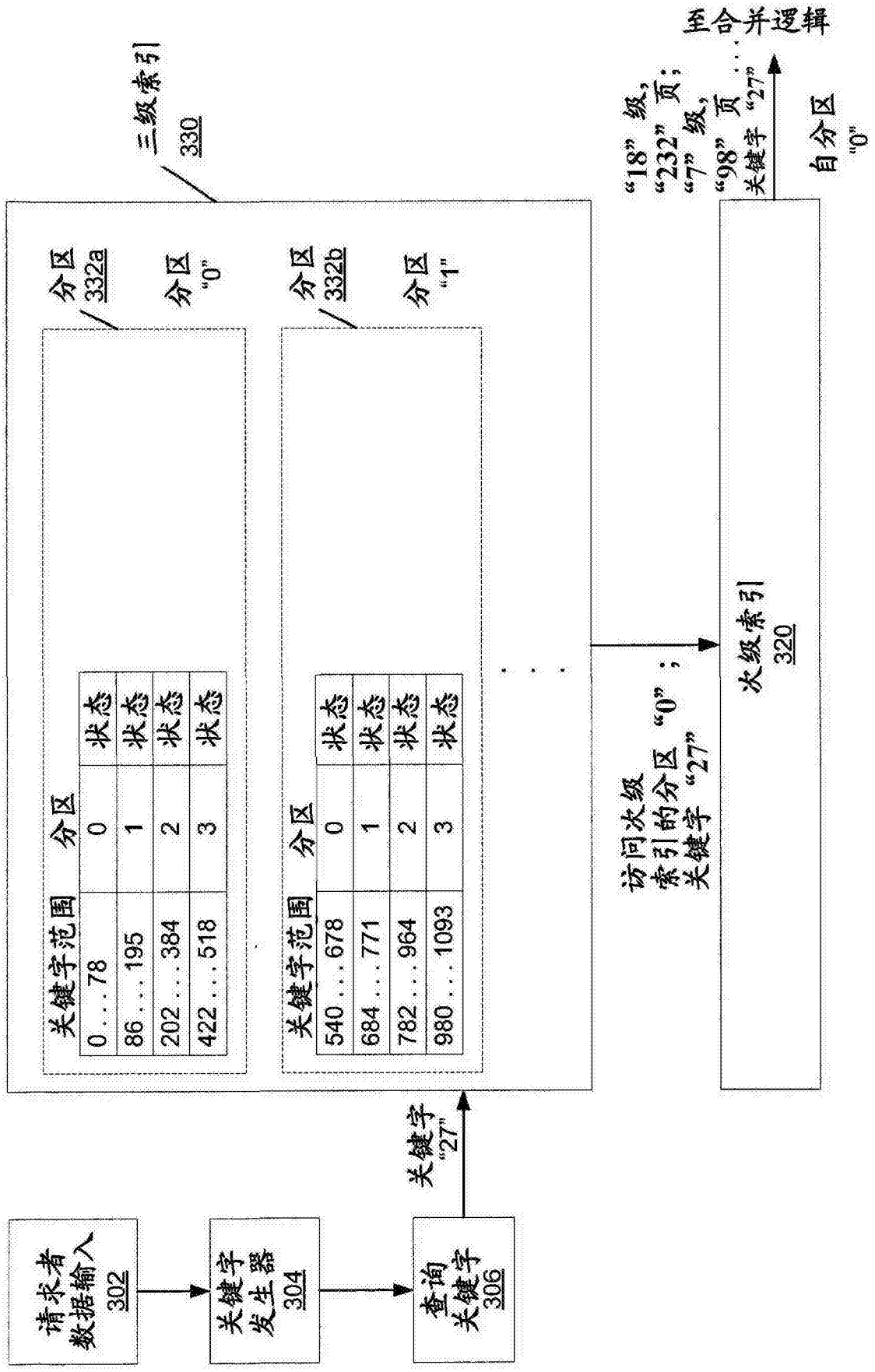

至合并逻辑
三级索引
330
分区
332a
分区
"0"
分区
332b
分区
"1"
关键字范围 分区
0 . . . 78 0 状态
86 . . . 195 1 状态
202 . . . 384 2 状态
422 . . . 518 3 状态
关键字范围 分区
540 . . . 678 0 状态
684 . . . 771 1 状态
782 . . . 964 2 状态
980 . . . 1093 3 状态
访问次级
索引的分区 "0" ;
关键字 "27"
次级索引
320
"18" 级,
"232" 页;
"7" 级,
"98" 页 . . .
关键字 "27"
自分区
"0"
请求者
数据输入
302
关键字
发生器
304
查询
关键字
306
关键字
"27"

图 8

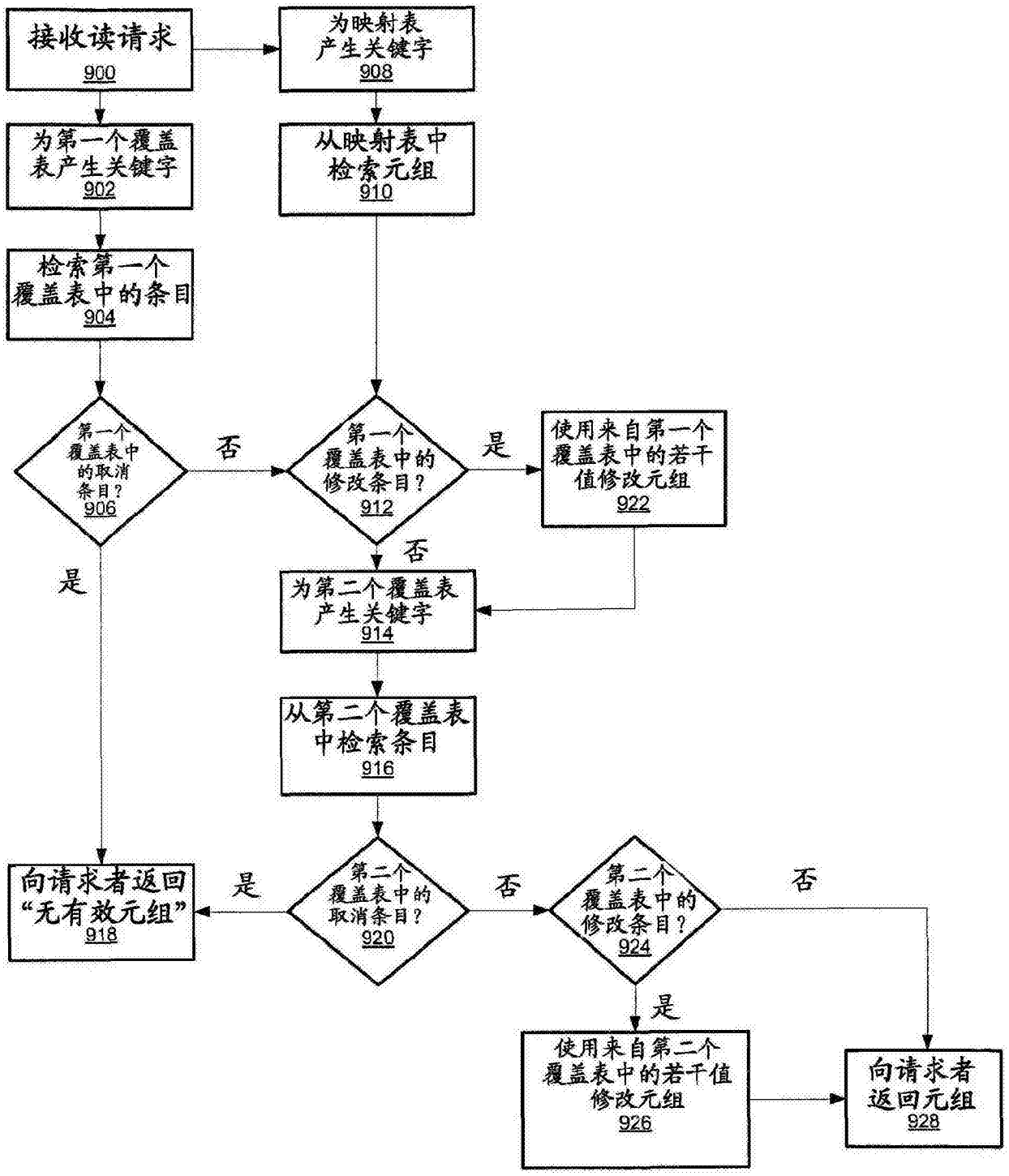

接收读请求
900
为第一个覆盖表产生关键字
902
检索第一个覆盖表中的条目
904
第一个覆盖表中的取消条目？
906
否
是
向请求者返回“无有效元组”
918
为映射表产生关键字
908
从映射表中检索元组
910
第一个覆盖表中的修改条目？
912
是
使用来自第一个覆盖表中的若干值修改元组
922
否
为第二个覆盖表产生关键字
914
从第二个覆盖表中检索条目
916
第二个覆盖表中的取消条目？
920
是
否
第二个覆盖表中的修改条目？
924
否
是
使用来自第二个覆盖表中的若干值修改元组
926
向请求者返回元组
928

图 9

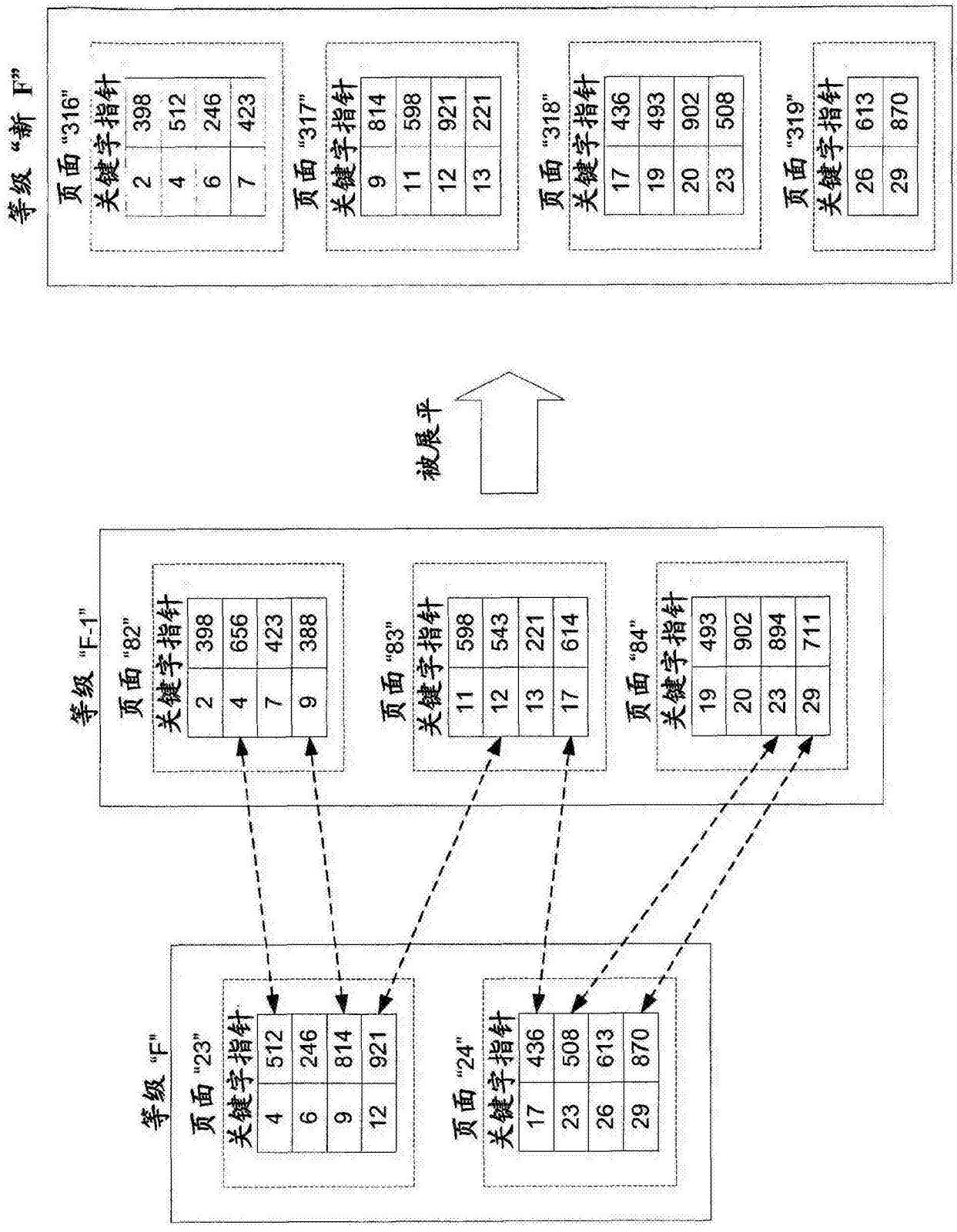

等级 "新 F"
页面 "316"
关键字指针
2 398
4 512
6 246
7 423
页面 "317"
关键字指针
9 814
11 598
12 921
13 221
页面 "318"
关键字指针
17 436
19 493
20 902
23 508
页面 "319"
关键字指针
26 613
29 870
被展平
等级 "F-1"
页面 "82"
关键字指针
2 398
4 656
7 423
9 388
页面 "83"
关键字指针
11 598
12 543
13 221
17 614
页面 "84"
关键字指针
19 493
20 902
23 894
29 711
等级 "F"
页面 "23"
关键字指针
4 512
6 246
9 814
12 921
页面 "24"
关键字指针
17 436
23 508
26 613
29 870

图 10

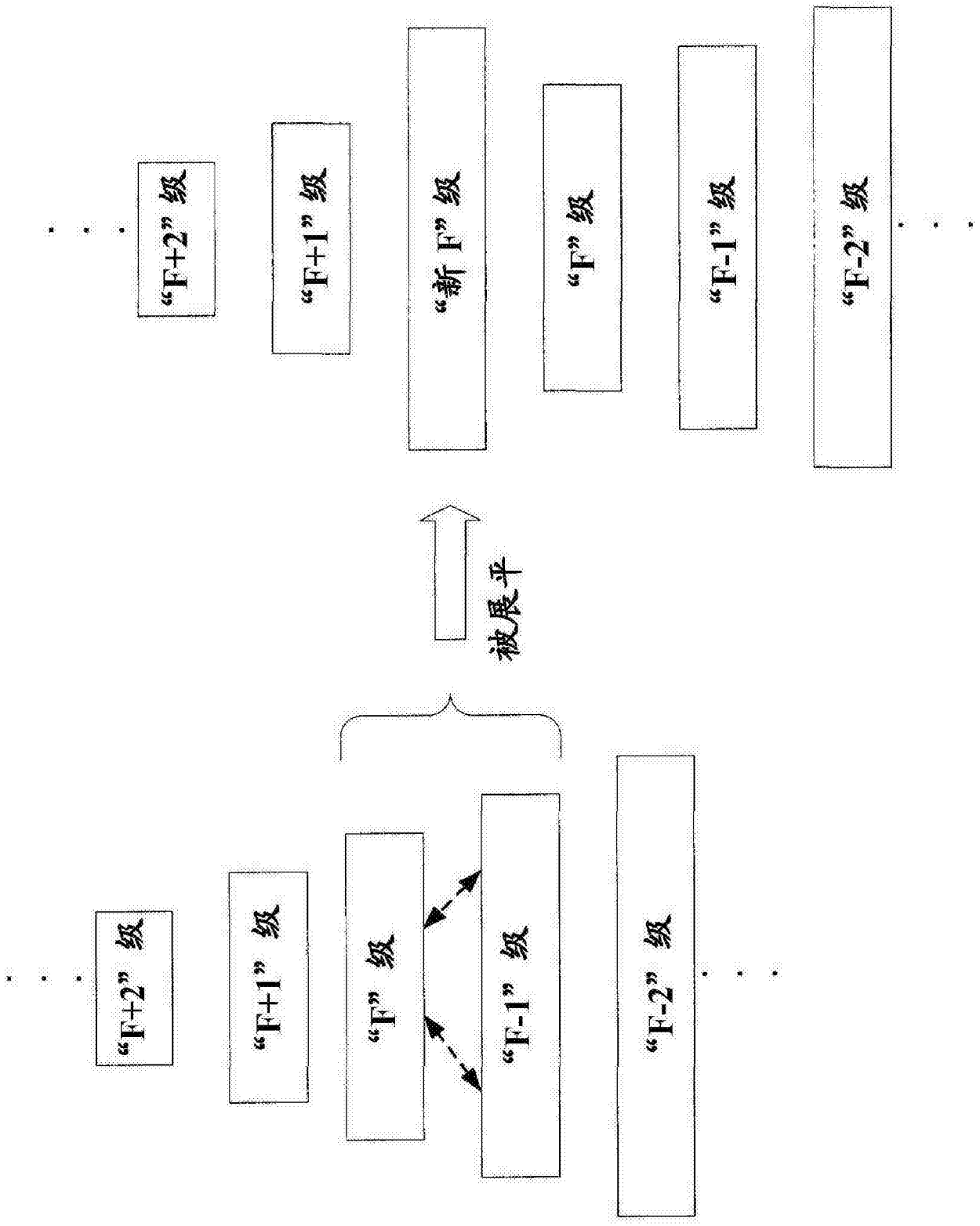


图 11

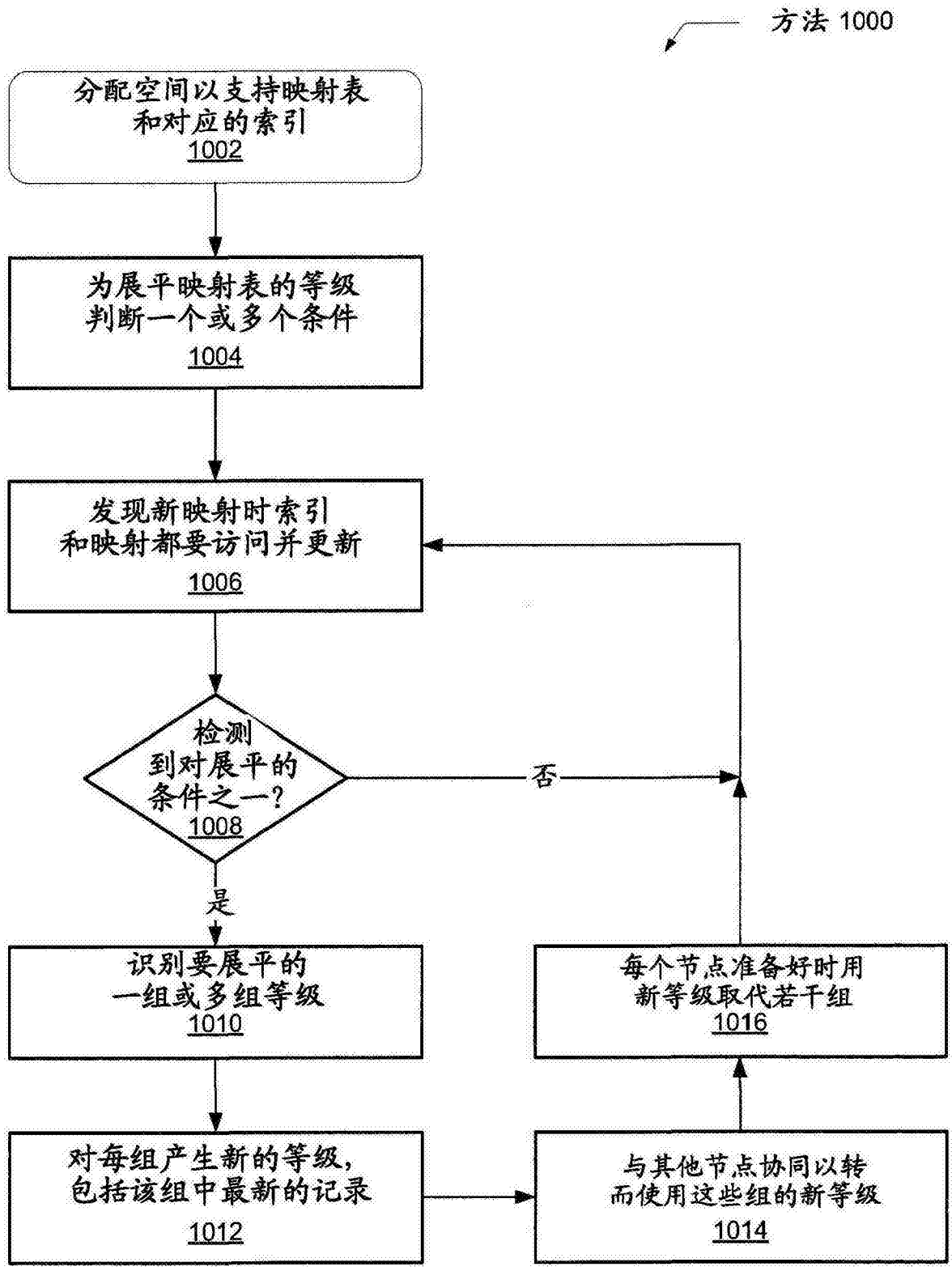

方法 1000
分配空间以支持映射表和对应的索引
1002
为展平映射表的等级判断一个或多个条件
1004
发现新映射时索引和映射都要访问并更新
1006
检测到对展平的条件之一?
1008
否
是
识别要展平的一组或多组等级
1010
对每组产生新的等级，包括该组中最新的记录
1012
与其他节点协同以转而使用这些组的新等级
1014
每个节点准备好时用新等级取代若干组
1016

图 12

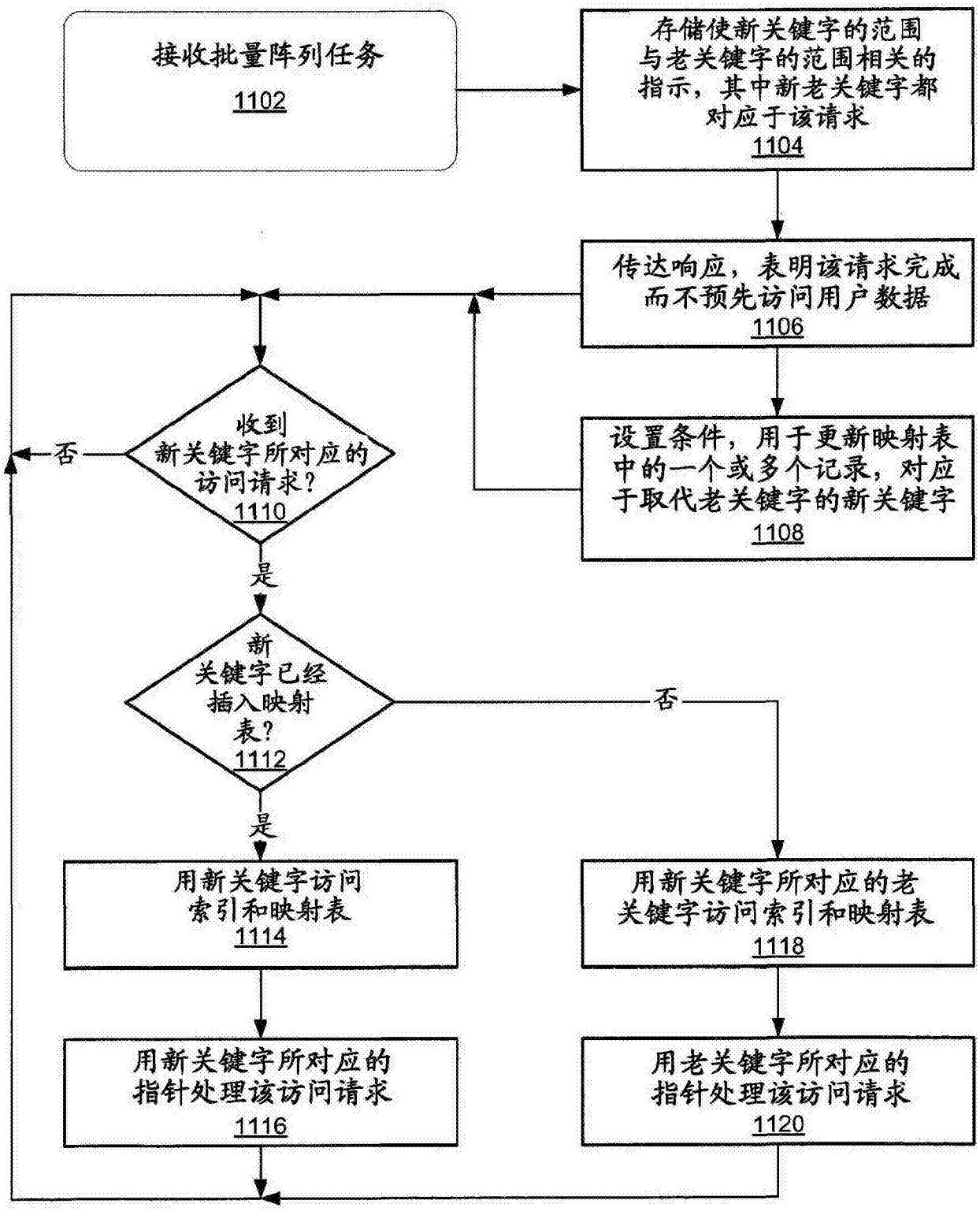

接收批量阵列任务
1102
存储使新关键字的范围与老关键字的范围相关的指示，其中新老关键字都对应于该请求
1104
传达响应，表明该请求完成而不预先访问用户数据
1106
设置条件，用于更新映射表中的一个或多个记录，对应于取代老关键字的新关键字
1108
收到新关键字所对应的访问请求？
1110
否
是
新关键字已经插入映射表？
1112
否
是
用新关键字访问索引和映射表
1114
用新关键字所对应的指针处理该访问请求
1116
用新关键字所对应的老关键字访问索引和映射表
1118
用老关键字所对应的指针处理该访问请求
1120

图 13

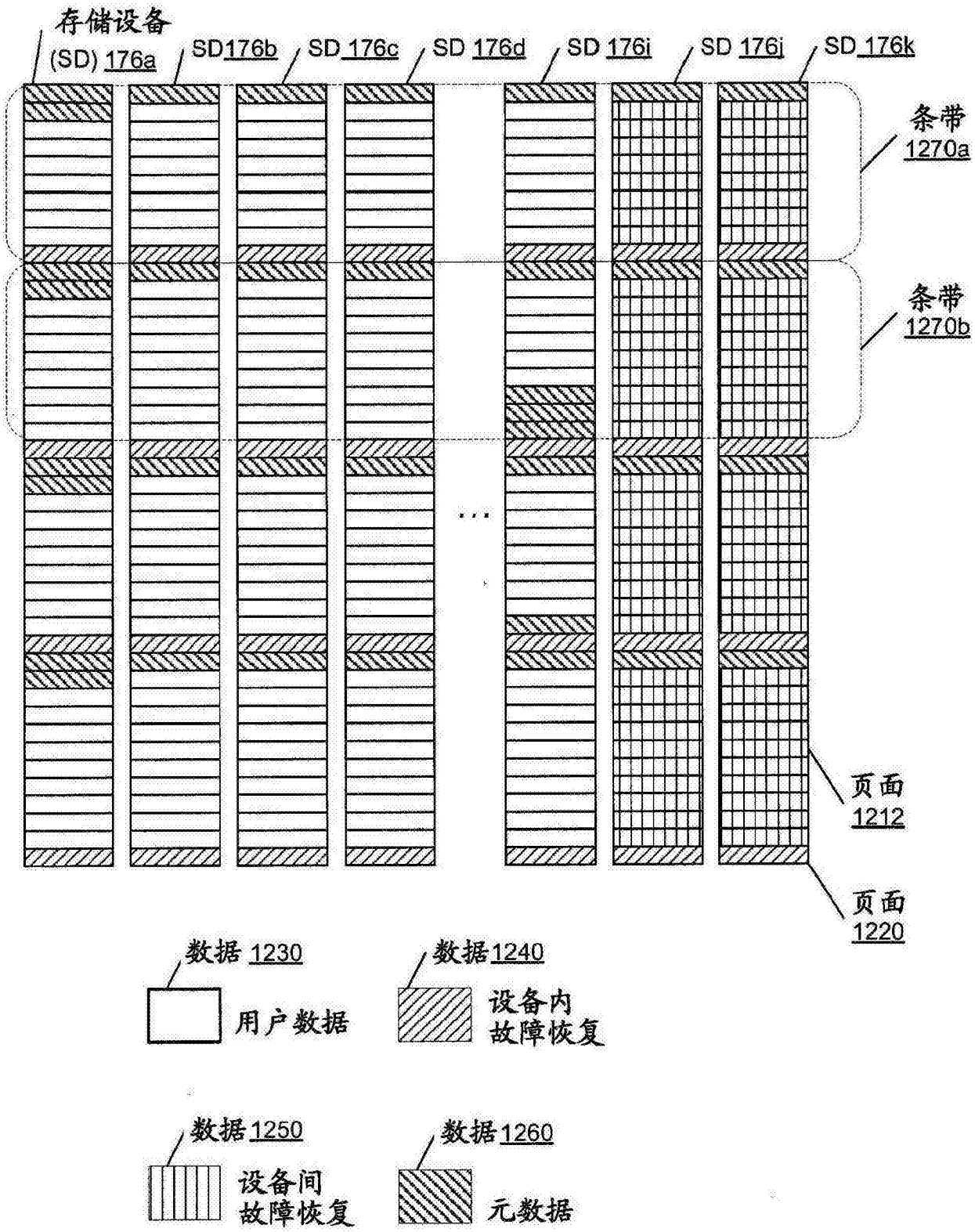

存储设备
(SD) 176a
SD176b
SD 176c
SD 176d
SD 176i
SD 176j
SD 176k
条带
1270a
条带
1270b
...
页面
1212
页面
1220
数据 1230
用户数据
数据1240
设备内
故障恢复
数据1250
设备间
故障恢复
数据1260
元数据

图 14