



# [12] 发明专利申请公开说明书

[21] 申请号 02816665.5

[43] 公开日 2004 年 11 月 17 日

[11] 公开号 CN 1547714A

[22] 申请日 2002.8.2 [21] 申请号 02816665.5

[30] 优先权

[32] 2001. 8. 3 [33] US [31] 60/309,803

[32] 2001. 11. 9 [33] US [31] 10/007,003

[86] 国际申请 PCT/US2002/024728 2002.8.2

[87] 国际公布 WO2003/012699 英 2003.2.13

[85] 进入国家阶段日期 2004.2.25

[71] 申请人 易斯龙系统公司

地址 美国华盛顿州

[72] 发明人 舒亚 M·帕特尔

保罗 A·米克塞尔

达雷恩 P·沙克

[74] 专利代理机构 北京英赛嘉华知识产权代理有限公司

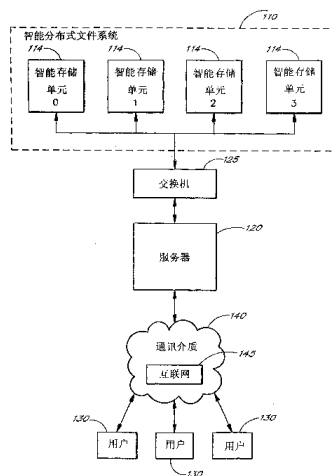
代理人 葛强 方挺

权利要求书 6 页 说明书 32 页 附图 18 页

[54] 发明名称 提供用于在存储设备的分布式文件系统中进行信息追踪的元数据的系统和方法

[57] 摘要

智能分布式文件系统(110)能够将文件数据存储存储在可像单一文件系统一样得到访问的多个智能存储单元中。该智能分布式文件系统(110)利用元数据数据结构来跟踪和管理每个文件的详细信息,包括如文件数据块的块位置和设备,从而可以允许单一文件系统内有不同级别的复制和/或冗余,便于冗余参数的改变,提供高级别的元数据保护,以及实时地复制和移动数据等等。



1. 一种与多个智能存储设备通信的分布式文件系统，其中所述分布式文件系统包括报文通信系统并被配置为可存储和管理与存储于所述分布式文件系统的文件和目录有关的元数据，以使所述元数据包括元数据数据块的位置、内容数据块和奇偶校验数据块。

2. 根据权利要求1所述的分布式文件系统，被进一步配置成可在所述分布式文件系统中提供不同级别的复制和冗余，其中，所述用于文件或目录的所述复制和冗余信息与文件或目录的相应元数据保存在一起。

3. 根据权利要求2所述的分布式文件系统，其特征在于，所述冗余信息包括奇偶校验信息。

4. 根据权利要求3所述的分布式文件系统，其特征在于，所述奇偶校验信息包括奇偶校验数据块和奇偶校验数据块的位置，并且所述元数据存储有与所述奇偶校验数据块的位置有关的信息。

5. 根据权利要求1所述的分布式文件系统，被进一步配置成可改变文件或目录的冗余参数，其中所述文件或目录的相应元数据被更新以存储与所述冗余数据的位置有关的信息。

6. 根据权利要求5所述的分布式文件系统，被进一步配置成可遍及多个智能存储设备的子集而实时地存储文件或目录数据的拷贝，并可在文件或目录的相应元数据中存储所述复制的文件或目录数据的位置。

7. 根据权利要求1所述的分布式文件系统，被进一步配置成可实时地移动数据，其中所述被移动的数据的位置与所述数据的相应元数据保存在一起。

8. 根据权利要求1所述的分布式文件系统, 被进一步配置成可实时地复制数据, 其中所述被复制的数据的位置与所述数据的相应元数据保存在一起。

5        9. 根据权利要求8所述的分布式文件系统, 被进一步配置成可响应高容量的数据请求而复制数据。

10. 根据权利要求8所述的分布式文件系统, 被进一步配置成可响应存储数据的硬件的高利用率而复制数据。

10

11. 根据权利要求8所述的分布式文件系统, 被进一步配置成可处理比写请求更多的读请求。

12. 根据权利要求8所述的分布式文件系统, 被进一步配置可处理块  
15 业务。

13. 一种用于在多个模块化存储单元中存储数据文件的虚拟文件系统, 所述虚拟文件系统包括:

多个存储单元, 其被配置成:

20

存储数据块;

接收数据文件读请求;

检索与所述被请求的数据文件相应的位置数据信息, 其中所述位置数据信息包括与和所述被请求的数据文件相应的数据块有关的存储位置信息;

25

检索本地存储的数据块;

从所述多个存储单元请求远程存储的数据块;

接收来自所述多个存储单元的远程存储的数据块的拷贝; 以及

返回所述被请求的数据文件。

14. 根据权利要求 13 所述的虚拟文件系统，其特征在于，所述存储单元包括存储设备和处理模块，其中所述本地存储的数据被保存于所述存储设备，而所述处理模块则从所述存储设备中检索所述本地存储的数据块。

5

15. 根据权利要求 13 所述的虚拟文件系统，进一步包括与交换部件通信的写模块，其中，所述写模块被配置成可接收数据文件写请求并确定与所述数据文件写请求相应的多个数据块的存储位置。

10 16. 根据权利要求 15 所述的虚拟文件系统，其特征在于，所述写模块被进一步配置成可将所述多个数据块分布在所述多个存储单元的至少两个存储单元中。

15 17. 根据权利要求 16 所述的虚拟文件系统，其特征在于，所述数据文件写请求包括镜像保护信息。

18. 根据权利要求 17 所述的虚拟文件系统，其特征在于，所述写模块被进一步配置成可将与所述数据文件写请求相应并符合所述镜像保护信息的镜像数据分布在所述多个存储单元中的至少两个存储单元中。

20

19. 根据权利要求 16 所述的虚拟文件系统，其特征在于，所述数据文件写请求包括奇偶校验保护信息。

20 20. 根据权利要求 19 所述的虚拟文件系统，其特征在于，所述写模块被进一步配置成可将与所述数据文件写请求相应并符合所述奇偶校验保护信息的奇偶校验数据分布在所述多个存储单元中的至少两个存储单元中。

21. 一种模块化存储单元，其配置成可与多个模块化存储单元通信以  
30 提供分布式文件存储，所述模块化存储单元包括：

存储模块，其被配置成可存储数据块；以及  
处理模块，其被配置成：  
接收对数据文件的请求；  
检索与所述被请求的数据文件相应的文件位置数据结构，其中所述  
5 文件位置数据结构包括本地存储的数据块和远程存储的数据块的位置的  
列表；  
从所述存储模块检索所述本地存储的数据块；  
从所述多个模块化存储单元中的至少一个请求远程存储的数据块；  
接收所述被请求的远程存储的数据块；以及  
10 返回所述被请求的数据文件。

22. 根据权利要求 21 所述的模块化存储单元，其特征在于，所述数  
据块至少有两种不同的大小。

15 23. 根据权利要求 21 所述的模块化存储单元，其特征在于，所述数  
据块具有相同的大小。

24. 根据权利要求 21 所述的模块化存储单元，其特征在于，所述存  
储模块包括非易失性存储设备和易失性存储设备。

20

25. 根据权利要求 24 所述的模块化存储单元，其特征在于，所述检  
索本地存储的数据块包括：

查询用于所述被请求的本地存储的数据块的所述易失性存储设备；  
及

25 如果没有找到所述被请求的数据块，则查询用于所述被请求的本地  
存储的数据块的非易失性存储设备。

26. 根据权利要求 25 所述的模块化存储单元，其特征在于，所述检  
索本地存储的数据块包括在所述易失性存储设备中保存所述被检索的本  
地存储的数据。  
30

27. 根据权利要求 25 所述的模块化存储单元, 其特征在于, 所述处理模块被进一步配置成可将所述检索的本地存储的数据块和所述接收的远程存储的数据块合并在一起, 以形成所述被请求的数据文件。

5

28. 根据权利要求 22 所述的模块化存储单元, 其特征在于, 所述处理模块被进一步配置成可接收用于在所述存储模块中存储的数据块。

29. 一种用于存储与数据文件和目录有关的信息的位置数据结构, 其中所述数据文件和目录使用不同的冗余和奇偶校验保护, 所述位置数据结构包括:

元数据部分, 其包括与相应数据文件或数据目录有关的元数据; 以及

内容部分, 其包括与数据文件内容的位置有关的位置信息或者与目录的子项有关的位置数据结构信息。

15

30. 根据权利要求 29 所述的位置数据结构, 其特征在于, 所述元数据包括名称、地址、类型、大小、安全数据、镜像数据、奇偶校验数据、数据大小、块大小以及位置信息地址中的至少一个。

20

31. 根据权利要求 29 所述的位置数据结构, 其特征在于, 所述位置数据结构保存有与目录有关的信息, 并且所述元数据包括镜像数据和奇偶校验数据。

25

32. 根据权利要求 31 所述的位置数据结构, 其特征在于, 所述镜像数据是从所述目录的任何子项的镜像数据中确定的。

33. 根据权利要求 32 所述的位置数据结构, 其特征在于, 所述镜像数据至少与所述目录的任何子项的最大镜像相匹配。

30

34. 根据权利要求 31 所述的位置数据结构, 其特征在于, 所述奇偶校验数据是从所述目录的任何子项的奇偶校验数据中确定的。

5 35. 根据权利要求 34 所述的位置数据结构, 其特征在于, 所述奇偶校验数据至少与所述目录的任何子项的最大奇偶校验保护相匹配。

36. 根据权利要求 31 所述的位置数据结构, 其特征在于, 所述位置信息包括与所述目录的子项相应的位置数据结构的位置。

10 37. 根据权利要求 29 所述的位置数据结构, 其特征在于, 所述位置数据结构保存了与数据文件有关的信息。

38. 根据权利要求 37 所述的位置数据结构, 其特征在于, 所述位置信息包括文件的数据块的位置。

15

39. 一种智能存储设备, 其被配置成可与其它的智能存储设备通信并处理对文件的请求, 其中所述被请求的文件的数据被分布在所述智能存储单元的子集中。

20 40. 根据权利要求 39 所述的智能存储设备, 其特征在于, 所述数据包括元数据和内容数据。

41. 根据权利要求 39 所述的智能存储设备, 其特征在于, 所述数据包括奇偶校验数据。

25

42. 根据权利要求 39 所述的智能存储设备, 其特征在于, 所述数据包括镜像数据。

提供用于在存储设备的分布式文件系统中进行信息追踪的元数据的系统  
和方法

5

发明领域

本发明的系统和方法一般涉及分布式文件存储领域，尤其涉及智能分布式文件管理。

10 

技术背景

互联网的爆炸式成长带来了新的领域，在该领域中，信息被持续地交换和访问。响应于这种增长，共享数据的规模也在增大。用户所要求的比标准 HTML 文档更多，他们希望访问多种数据，例如音频数据、视频数据、图像数据和编程数据。因此，存在着对在提供快速而可靠的数据访问的同时能够存储大的数据组的数据存储器的需要。

一种响应是采用可以存储大量的数据但很难提供高流量的单一的存储设备。随着数据容量的增加，访问数据的时间量也随之增加。虽然处理速度和能力已经得到提高，但是磁盘 I/O（输入/输出）操作性能没有以相同的速率提高，致使 I/O 操作效率低下，特别是在大数据文件方面。

20 

另一种响应是允许多个服务器利用诸如存储区域网络（SAN）解决方案的系统架构来访问共享磁盘，但是这种系统价格昂贵，并需要复杂的技术来建立并控制数据完整性（integrity）。而且还需要高速适配器来处理大容量的数据请求。

25 

常规方法的一个问题在于，它们受到其可扩展性（scalability）的限制。因而，随着数据容量的增加，系统需要也相应增长，但是扩充费用高昂而且具有高的破坏性。

常规方法的另一共同问题在于它们的灵活性有限。这些系统通常被配置为使用预先确定的纠错控制。例如，可使用 RAID（磁盘阵列）系统在物理磁盘级别上提供数据文件的冗余和镜像，因而在确定数据在哪里



存储或应使用的冗余参数类型方面给管理员很少的灵活性或根本没有灵活性。

### 发明内容

5 智能分布式文件系统能够很好地将文件数据存储在不像单一文件系统一样得到访问的一组智能 (smart) 存储单元中。智能分布式文件系统很好地利用元数据 (metadata) 数据结构来跟踪和管理每个文件的详细信息, 包括如文件数据块的块位置和设备, 以允许单一文件系统内不同级别的复制和/或冗余, 方便冗余参数的改变, 为元数据提供高级别的保护, 10 实时地复制和移动数据, 等等。

本发明内容的目的、特定方面、优点和新颖的特征在本文中得到了描述。应该理解, 未必所有这些优点都在本发明的任何特定的实施例中得到实现。因此, 例如, 本领域技术人员应意识到, 可以以只实现本文中所教导的一个优点或一组优点而无需实现本文教导或建议的其他优点 15 的方式来实现或完成本发明。

### 附图的简要说明

图 1 是本发明一个实施例的高层方框图;  
图 2 示出了图 1 所示的组成部件之间的示例数据流;  
20 图 3 示出了示例性的智能存储单元的高层方框图;  
图 4 示出了示例文件目录;  
图 5 示出了元数据数据结构的一个实施例;  
图 6A 示出了数据位置表结构的一个实施例;  
图 6B 示出了数据位置表结构的另一个实施例;  
25 图 6C 示出了数据位置表结构的又一个实施例;  
图 6D 示出了数据位置表结构的再一个实施例;  
图 7A 示出了用于目录的元数据数据结构的一个实施例;  
图 7B 示出了用于文件的元数据数据结构的一个实施例;  
图 8A 示出了数据位置表的一个实施例;  
30 图 8B 示出了数据位置表的另一个实施例;

图 8C 示出了数据位置表的又一个实施例;

图 9 示出了带有相应的示例数据的文件的示例元数据数据结构;

图 10 示出了用于检索数据的流程图的一个实施例;

图 11 示出了的用于实现名称解析的流程图的一个实施例;

5 图 12 示出了用于检索文件的流程图的一个实施例;

图 13 示出了用于创建奇偶校验信息的流程图的一个实施例;

图 14 示出了用于实现纠错的流程图的一个实施例。

### 具体实施方式

10 以下将参照附图对代表了本发明的一个实施例和示例应用的系统和方法进行说明。对该系统和方法的变换所代表的其它实施例也将得到描述。

为说明的目的,一些实施例将在互联网内容-传送和网络托管(web hosting)的背景中描述。发明人期望本发明不受使用本系统和方法的环境类型的限制,本系统和方法可使用在其他环境中,诸如,举例来说,互联网、万维网、医院专用网络、政府机构广播网、合作企业互连网、企业内部互联网,局域网,广域网等等。然而,在涉及本发明的实施例的说明书和附图中,背景环境是互联网内容-传送和网络托管。也应意识到,在其他的实施例中,所述系统和方法可作为单一的模块实现并且/或者可以其他各种模块等协同实现。而且,本文所描述的特定的实现方式  
15 为说明的目的而提出的,它并不是对本发明的限制。本发明的范围由所附权利要求来定义。

现在将参照上面简要描述的附图来描述这些和其它的特征。附图和相关的说明被提供用来阐明本发明的实施例,但其不限制本发明的范围。  
25 在全部附图中,参考标号可被重复使用以指明所参考部件间的对应关系。另外,每个参考标号的第一个数字一般都表示该部件第一次出现的图。

### I. 概述

30 本发明的系统和方法提供了智能分布式文件系统,其能够在一组可作为单一文件系统得到访问的智能存储单元中存储数据。该智能分布式

文件系统跟踪和管理每个文件的详细的元数据。元数据可以包括涉及和/或描述文件的任何数据，例如，包括设备和块位置信息在内的文件数据块的位置、元数据和/或数据块（如果有的话）的冗余备份的位置、纠错信息、访问信息、文件名、文件的大小、文件类型，等等。另外，对该  
5 由文件系统管理的不同文件和/或数据块，智能分布式文件系统允许进行不同级别的复制和/或冗余，从而方便了系统激活时冗余参数的改变，并能够实现元数据和数据的实时复制和移动。进一步，通过从智能存储单元组中定位和收集文件数据，各智能存储单元都可响应文件请求。

在存在大量的读（READ）请求的情况下，特别是其与写（WRITE）  
10 请求成比例的情况下，智能分布式文件系统可以很好地提供对数据的访问。这是由于增加了锁定智能的智能存储单元组以及对单个智能存储单元的浏览的复杂性从而保证了一致性。此外，在对大的数据块的请求很普遍的期间，智能分布式文件系统能够很好地对块的交互进行处理。

一些实施例的一个好处是，用于文件和目录的元数据由智能分布式  
15 文件系统管理和访问。元数据可指出用于目录或文件的元数据被定位在哪里、内容数据存储在哪里、元数据和/或内容数据的镜像备份存储在哪里、以及与系统相关的奇偶校验或其他纠错信息存储在哪里。可使用如设备和块位置信息来存储数据位置信息。因而，智能分布式文件系统可通过利用分布和存储在智能存储单元组中的元数据来定位和检索被请求  
20 的内容数据。另外，因为智能分布式文件系统可以访问元数据，因而智能分布式文件系统可被用于选择数据被存储的位置并根据请求来移动、复制和/或改变数据而不对智能存储单元组产生破坏。

一些实施例的另一个好处是，用于每个文件的数据可跨越几个智能  
存储单元存储并以时序的方式访问。用于每个文件的数据块可能分布在  
25 智能存储单元的子集中，这样数据存取时间就可减少。进一步，不同的文件可能跨越不同数量的智能存储单元以及跨越不同智能存储单元组而分布。这种架构使智能分布式文件系统能够根据诸如文件的大小、重要度、预期访问速率，以及可用的存储容量、CPU 利用率和每个智能存储单元的网络利用率等因素而智能地存储数据块。

一些实施例的另外的益处是，该系统和方法可被用于对数据块或文件提供不同的保护方案，例如纠错、冗余和镜像，以使存储在智能存储单元之中的不同的数据块或文件可以有不同的保护类型。例如，一些目录或文件可以被镜像，其他的目录和文件可利用不同的错误或丢失纠正方案的错误和/或丢失（loss）纠正数据而得到保护，以及其它较不重要的目录或文件可能不使用任何保护方案。

一些实施例的进一步的好处是该系统和方法可实时地增加、删除和/或修改智能存储单元而不破坏或中断正在进行的数据请求。因而，在需要更多的存储器时，附加的智能存储单元可实时地加到智能存储单元组并且并入到智能分布式文件系统中，而不打断文件请求或使现有的智能存储单元离线。在现有文件的数据块或新文件被智能分布式文件系统跨越现在已包括了新的智能存储单元的智能存储单元组分布的同时，现有的智能存储单元可处理对文件的请求。

一些实施例的另外的益处是，通过将数据块复制在一个或多个智能存储单元上，该系统和方法可实现对这些块的存储的实时修改，因而为任何单独的数据块创造了多点访问。这种复制有助于减少已被观测的频繁访问模式下用于文件或文件组的单个智能存储单元的CPU利用率和网络资源需求。这些访问模式被智能存储单元监测，并且智能分布式文件系统为智能存储单元提供了在智能分布式文件系统依旧工作的同时复制这类数据的灵活性。

## II. 示例操作

为说明的目的，现在讨论一个其中使用了正在运行的智能分布式文件系统的示例场景。在该示例场景中，智能分布式文件系统由一家通过互联网网站提供电影下载的公司使用。该公司可使用智能分布式文件系统来存储和管理由消费者经网站访问的可下载的电影及电影预告片、广告和消费者信息的拷贝。数据可以按照不同的保护级别得到存储，并且可以跨越多个智能存储单元存储以进行快速访问。

例如，该公司可能想要在智能分布式文件系统中跨越几个智能存储单元地存储客户调查电子邮件以提供对这些电子邮件的快速访问。然而，

该公司可以保留全部电子邮件的备份带并可能感到对客户调查立即恢复并不是很重要的。该公司可能指示智能分布式文件系统不对客户调查电子邮件使用纠错或镜像保护。因此，如果一个或更多的智能存储单元变得不可访问，该公司可能感到将这些智能存储单元上的访问客户调查电子邮件延迟到电子邮件可从备份带中恢复为止是可接受的。

为进行广告，该公司可能指示智能分布式文件系统使用高纠错参数，这样如果一个或更多的智能存储单元出现故障，则智能分布式文件系统可以恢复数据而不打断广告的进行。例如，该公司可能依据不同的容错测量结果以帮助确定对特定的文件需要提供多少保护。对于重要的信息，该公司可能想确保容错水平 X，而对重要性较低的信息，公司想要确保容错水平 Y，其中  $X > Y$ 。应意识到，容错是被用于说明可靠性的一种测量方法，但可以使用附加于或代替容错的其他方法。因此，即使一个或更多的智能存储单元出现故障，该公司也可以向其广告客户可靠地保证广告的可用性。

对于顶级电影下载，该公司可有利地建立智能分布式文件系统以自动存储电影数据的多重拷贝，从而使更多的客户能够访问数据并保证如果一个或更多的智能存储单元出现故障，那么丢失的数据可从其他单元重建或恢复。另外，如果请求的数量增加并且/或者一个或更多的智能存储单元开始充满对驻留在智能存储单元上的数据的请求，则可创建顶级电影下载的另外拷贝并将其存储在智能存储单元之中。

该公司可选择提供其他不是如此流行的电影，并可能由于较少的请求而指示智能分布式文件系统存储较少量的拷贝。进一步，随着“顶级下载电影”变得不那么流行，该公司可有利地建立智能分布式文件系统以从存储电影的智能存储单元上删除电影的额外拷贝，并将“较不流行的”电影移动至性能较低的智能存储单元（例如，那些可用磁盘空间较少的智能存储单元）。智能分布式文件系统可被设定以利用智能存储单元自动地照顾这些任务。

另外，随着该公司获得更多的电影，该公司可能为智能分布式文件系统增加额外的智能存储单元。然后，该公司可能使用新的智能存储单元以存储更多的电影、存储现有电影的更多拷贝并且/或者重新分配现有

的电影数据以改善响应时间。增加的智能存储单元被合并到智能分布式文件系统中，这样即使智能分布式文件系统管理和存储多组智能存储单元中的数据，智能分布式文件系统也作为单一的文件系统出现。

本例中，智能分布式文件系统为该公司提供了对顶级电影下载的可靠的和快速的访问、对较不流行的电影的快速访问以及客户调查电子邮件访问的能力。对每个文件，该公司可能设置错误和/或丢失纠正参数并可选择应存储多少文件的附加拷贝。在某些情况下，该公司可能手工地选择应存储的数据拷贝数并确定存储数据的地方。在其他的情况下，该公司可能依赖智能分布式文件系统的特性以选择应存储的数据拷贝数、  
10 将使用的错误和/或丢失纠正方案（如果有的话），以及/或者数据应存储的位置。因而，该公司能有效地使用其存储空间而更好地响应用户的请求。存储空间不被浪费在稀少的请求文件上，并且不为不重要的文件生成和存储纠错信息。

尽管上面的例子涉及提供用于下载的电影的公司，但应该意识到，  
15 该例子只是用于说明智能分布式文件系统的的一个实施例的特性。进一步，智能分布式文件系统也可在其他的环境中使用，并可能使用其他类型的数据和/或其结合，包括例如声音文件、音频文件、图形文件、多媒体文件、数字相片、可执行文件，等等。

### 20 III. 智能分布式文件系统

图 1 示出了智能分布式文件系统 110 的一个实施例，其与网络服务器 120 通信以提供远程文件访问。智能分布式文件系统 110 可使用多种协议，例如 NFS（网络文件系统）或 CIFS，与网络服务器 120 通信。用户 130 经通信介质 140（例如互联网 145）与网络服务器 120 交互，以请求被智能分布式文件系统 110 管理的文件。示例的智能分布式文件系统  
25 110 使用了交换部件 125，交换部件 125 与一组智能存储单元 114 以及网络服务器 120 通信。智能分布式文件系统 110 使单独文件的数据块能够跨越多个智能存储单元 114 分布。数据被存储以使对数据的访问可以比数据被存储在单一的设备上提供较高的流量。另外，智能分布式文件系统  
30 110 可被用于存储利用各种保护方案得到存储的的各种的数据文件。

示例的智能分布式文件系统 110 在一组智能存储单元 114 中存储数据。为更详细描述智能存储单元 114, 请参考下面标题为“智能存储单元”的部分。

5 示例的智能分布式文件系统使用诸如负载平衡交换机的交换部件 125, 交换部件 125 为请求指明能处理被请求的数据类型的应用服务器。使用高速技术将引入的请求转发到适当的应用服务器以使延迟最小, 从而保证数据的完整性。

应该意识到可以使用不同的负载平衡交换机 125, 例如, 1000Base-T (铜) 千兆比特负载平衡以太网交换机、Extreme Networks Summit7I、  
10 Foundry Fast Iron II、北电网络的 Alteon ACE 交换机 180、F5 Big - Ip) 以及标准以太网交换机或其他的负载平衡交换机。智能分布式文件系统使用支持大帧尺寸 (例如 “jumbo” 以太网帧) 的交换机。另外, 可使用 Foundry 网络公司的 SERVER IRON 交换机、Asante 公司的 InstraSwitch 6200 交换机, Asante 公司的 HotStack、CISCO 公司的 Catalyst 交换机及  
15 其他的商业产品和/或私有的产品来实现负载平衡交换机 125。然而, 本领域的普通技术人员将意识到, 也可以使用很大范围的交换部件 125, 或使用其他技术。此外, 应该意识到, 也可配置交换机部件 125 以传输不同大小的网络帧。

高度重要的文件可能用高纠错参数存储, 在磁盘、主板、CPU、操作  
20 系统或其它硬件或软件出现故障而阻止了对一个或更多的智能存储单元的访问时, 该纠错参数为数据提供了高的恢复率。如果数据丢失或缺少, 智能存储单元 114 可使用元数据中的冗余信息或镜像信息以从另外的单元获得数据或重建数据。高需求的文件可实时地跨越另外的智能存储单元 114 而被镜像, 从而提供平稳的更高的流量。

25 在智能分布式文件系统 110 的一个实施例, 元数据数据结构至少受到与它引用的数据 (包括与元数据数据结构相应的目录的任何下属内容) 一样的保护。由于如果没有元数据数据结构则很难恢复数据, 因此元数据数据结构中的数据丢失会危害智能分布式文件系统 110。在智能分布式文件系统 110 中, 元数据数据结构的替代拷贝可根据需要被镜像于  
30 很多单元中以提供必需的保护。因而, 带有奇偶校验保护的文件可能使

它的元数据数据结构以至少与奇偶校验保护相同或更好的方式得到存储，并且被镜像两次的文件可能使它的元数据结构至少被镜像在两个单元中。

尽管图 1 示出了智能分布式文件系统 110 的一个实施例，应该意识到，也可以采用其他的实施例。例如，可以采用另外的服务器，如可与交换部件 125 通信的应用服务器。这些应用服务器可能包括如音频流服务器、视频流服务器、图像处理服务器、数据库服务器，等等。此外，可能有另外的设备，例如与交换部件 125 通信的工作站。另外，尽管图 1 示出的智能分布式文件系统 110 与四个智能存储单元 114 协同工作，应该意识到，智能分布式文件系统 110 也可与不同数目的智能存储单元 114 协同工作。

还应意识到术语“远程”可包括非本地存储（也就是不能通过本地总线访问）的设备、部件和/或模块。因而，远程设备可包括物理上位于同一房间并经诸如交换机或局域网的设备连接的设备。在其他情况下，远程设备也可位于分离的地理范围内，例如在不同的位置、国家等等。

也应该意识到，可使用智能分布式文件系统 110 存储多种类型的数据。例如，智能分布式文件系统 110 可与大的文件应用程序一起使用，例如，视频点播（video-on-demand）、在线音乐系统、网站镜像、大的数据库、大的图形文件、CAD/CAM 设计、软件更新、公司介绍（corporate presentation）、保险请求文件、医疗成像文件、公司文件存储等等。

图 2 示出了一种示例环境，其中网站用户 130 提交了观看数字视频点播的请求。在事件 A 中，用户 130 通过互联网 145 向网站发送请求，请求观看电影 mymovie.movie 的拷贝。该请求被网站服务器 120 接收，并且服务器 120 确定该文件位于 movies\comedy\mymovie.movie。在事件 B 中，智能分布式文件系统 110 的交换部件 125 看到该请求而连接到智能分布式文件系统 110，并使用标准负载平衡技术将该请求转发到可用的智能存储单元 114，如智能存储单元 0。在事件 C 中，智能存储单元 0 接收对文件/DFSR/movies/comedy/mymovie.movie 的请求，并从它的根元数据数据结构（用于根目录/DFSR）确定出用于子目录 movies（电影）的元数据数据结构被存储在智能存储单元 2。在事件 D 中，智能存储单元 0 发



送请求到智能存储单元 2，请求用于子目录 comedy（喜剧）的元数据数据结构的位置。在事件 E 中，智能存储单元 0 收到信息，用于子目录 comedy（喜剧）的元数据数据结构被存储在智能存储单元 3。在事件 F 中，智能存储单元 0 发送请求到智能存储单元 3，请求用于文件 mymovie.movie 的元数据数据结构的位置。在事件 G 中，智能存储单元 0 收到这样的信息：用于文件 mymovie.movie 的元数据数据结构被存储在智能存储单元 0 内。智能存储单元 0 随后从本地存储器中检索出用于文件 mymovie.movie 的元数据数据结构。从该元数据数据结构中，智能存储单元 0 检索出用于文件 mymovie.movie 的数据位置表（data location table），该数据位置表中保存有文件中的各个数据块的位置。智能存储单元 0 随后使用数据位置表信息以开始检索本地存储的块并发送对存储在其他智能存储单元上的数据的请求。

在文件的数据或数据的一部分被检索之后，文件数据被发送到请求服务器 120 以被转发给请求的用户 130。在一个例子中，文件数据可能被发送到视频流服务器，其可调节数据在何时以何方式被发送到用户 130。应该意识到，在一些实施例，利用预读（read ahead）技术以检索更多的数据，然后被请求从而减少等待时间是有利的。

#### IV. 智能文件系统结构

表 1 示出了一组示例的文件系统层的一个实施例，通过该文件系统层，文件请求被处理以访问物理存储设备。示例的文件系统层包括用户层、虚拟文件系统层、本地文件系统层、本地文件存储层和存储设备层。

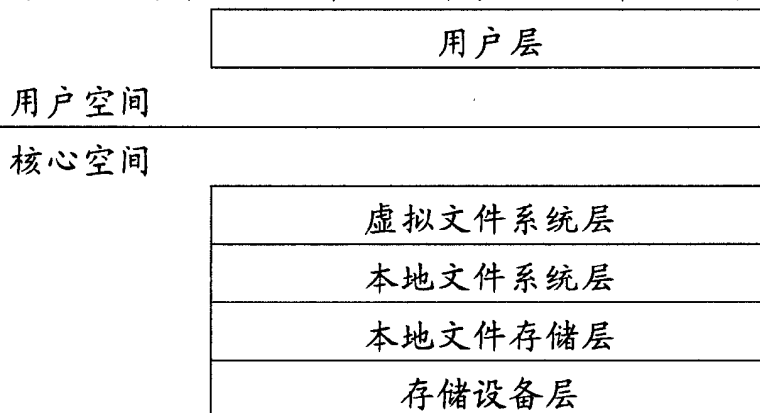


表 1

在一类文件请求中，请求通过用于文件共享的用户层协议应用程序被接收，用户层协议如 HTTPD（Apache 全球网服务）、FTPD 或在 Unix 上使用的实现了微软公司的 Windows 文件共享服务器协议的一个版本的 SMBD。用户层协议应用程序通过例如对 libc（C 运行时期库）的函数调用来实现核心层的操作，如打开、读、搜索、写或关闭系统调用。

系统调用被传送到虚拟文件系统层（“VFS”），虚拟文件系统层则维持了缓冲存储器。缓冲存储器可以是如最近最少使用的（“LRU”）高速缓冲存储器，其用于存储从较低的文件系统层接收的数据或元数据数据结构。

下一层是本地文件系统层，该层维护文件系统的等级命名系统（hierarchical naming system）并发送目录和文件名请求到下层，即本地文件存储层。本地文件系统层处理元数据数据结构的查找和管理。例如，在一些系统中，例如在基于 Unix 的文件系统中，元数据数据结构是文件的概括，包括与文件访问许可、数据块位置和参考计数有关的信息。一旦利用其名字打开了文件，其它的文件操作利用唯一的标识符来引用文件，该标识符标识出了用于特定文件的元数据结构。这种方法的益处是，单个文件可能有许多不同的名字，单个文件可通过不同的路径访问，并且可复制新文件以覆盖 VFS 命名空间中的旧文件，而不会经标准 UNIX 用户层应用程序（例如，‘MV’ 命令）而覆盖实际的文件数据。在诸如内容-传送和网络托管的环境中，这些益处可能更有利，因为内容可在适当的位置更新而不会破坏当前的内容服务。元数据数据结构内的参考计数使系统只在打开的文件柄全部被关闭时才能够无效该数据块。

第四层是本地文件存储层，其控制“缓冲器请求到块请求”的翻译以及数据缓冲器请求的管理。例如，本地文件存储层使用块分配方案来改善并最大化用于写和读的流量，以及用于读的块检索方案。

最后一层是存储设备层，其托管（host）文件系统使用的用于磁盘硬件的特定块的设备驱动器，例如，如果物理存储设备是 ATA 磁盘，那么存储设备层托管 ATA 磁盘驱动程序。

## V. 智能存储单元

在一个实施例中，智能存储单元 114 是即插即用 (plug-and-play) 的、高密度的、可架式安装的应用设备，其对高流量数据传送是优选的。智能存储单元可被配置成与各种其他的智能存储单元通信以提供单一的虚拟文件系统。随着所需存储空间的增多或一个或更多的智能存储单元出现故障，另外的智能存储单元可被安装而不会使整个系统关机或导致服务中断。

如本文所使用的，字模块是指嵌入在硬件或固件中的逻辑或由程序设计语言（例如 C 或 C++）编写的、可能具有进入点和退出点的软件指令的集合。软件模块可能被编译并连接为可执行程序，此可执行程序可安装在动态连接库中，或由解释性程序设计语言例如 Basic、Perl 或 Python 编写。应该意识到软件模块可能被其他模块或自身所调用，并且/或者可响应检测到的事件或中断而被调用。软件指令可能被嵌入在固件（例如 EPROM）中。应该进一步意识到，硬件模块可能包含逻辑连接单元，例如门和触发器，和/或可能由可编程单元（例如可编程门阵列或处理器）组成。本文描述的模块优选地由软件模块实现，但也可能在硬件或固件中出现。

图 3 示出了智能存储单元 114 的一个实施例，其包括管理模块 320、处理模块 330、缓冲器 340，堆栈 350 和存储设备 360。示例的智能存储单元 114 可被配置成与交换部件 125 通信，发送和接收如图 1 所示的请求。

### A. 管理模块

在一个实施例中，智能存储单元包括用于执行管理任务的管理模块 320，这些任务例如安装、参数设置、智能分布式文件系统监控，发生在智能分布式文件系统 110 上的事件的记录以及升级。

### B. 处理模块

示例的处理模块 330 可被配置为接收对数据文件的请求，检索本地和/或远程存储的关于被请求的数据文件的元数据，并检索本地和/或远程

存储的被请求数据文件的数据块。另外，在一个或更多的被请求数据块损坏或丢失时，处理模块 330 还可完成数据恢复和纠错。

在一个实施例中，处理模块 330 包括响应文件请求的五个模块，块分配管理器模块 331、块缓冲模块 333、本地块管理模块 335、远程块管理模块 337 和块设备模块 339。

### 1. 块分配管理器模块

块分配管理器 331 模块确定在哪里分配块，响应 READ（读）请求对块进行定位，以及引导设备进行故障恢复。关于在哪里分配块的信息可以由缺省参数所设置的策略、系统管理员利用工具（例如图形用户界面或外壳接口）设置的策略、或这些策略的组合确定。在一个实施例中，块分配管理器 331 驻留在本地文件系统层并与标准网络软件层（例如 TCP/IP 和以太网）配合工作，和/或代替加州大学柏克利分校软件设计通用文件系统（“BSD UFS”）。

示例的块分配管理器 331 包括 3 个子模块，块请求翻译模块、预分配（forward allocator）模块和故障恢复模块。

#### a. 块请求翻译模块

块请求翻译模块接收引入的读请求，完成名称查找，定位适当的设备，并从该设备取出数据以完成请求。如果数据是直接可用的，块请求翻译模块依据数据块是存储在本地存储设备还是存储在其它的智能存储单元的存储设备上，以发送数据请求到本地块管理器模块或到远程块管理器模块。

在一个实施例中，块请求翻译模块包括名称查找过程，其将在哪下面标题为“智能分布式文件系统过程 - 名称查找处理”中得到讨论。

块请求翻译模块也可能对设备故障做出响应。例如，如果设备关机，则块请求翻译模块可使用如奇偶校验信息请求可用于重构数据的本地的和远程的数据块。因而即使 READ 不能被执行，数据也可被生成。另外，块请求翻译模块可与故障恢复模块通信，这样故障的恢复模块可使用奇偶校验或其他的错误或丢失纠正数据来重建数据并跨越智能分布式文件

系统的自由空间地为损失纠正数据重新划分带区 (re-stripe)。在其他的实施例中，块请求翻译模块可能请求出错或丢失数据的干净拷贝。

#### b. 预分配模块

5 预分配模块根据诸如冗余、空间和性能等因素确定应被用于写请求的设备的块。这些参数由系统管理员设置、可以从嵌入在智能分布式文件系统 110 中并作为智能分布式文件系统 110 的逻辑而使用的信息中获得，或可以从其组合中获得。预分配模块 110 接收来自于其他使用智能分布式文件系统的智能存储单元的统计数据，并使用这些统计数据决定  
10 存放新来数据的最好位置。被收集的统计数据包括如所测量的 CPU 利用率、网络利用率和磁盘利用率。

预分配模块还可根据远程智能存储单元的应答时间而从远程块管理器模块接收等待时间信息。如果中间设备的等待时间相对于其他智能存储单元达到高水平，如果可能，则根据冗余设置，分配方案可能被调整  
15 以有利于其他的智能存储单元减少对慢的智能存储单元的使用。在一个有利的例子中，智能分布式文件系统可能已将数据块从一个智能存储单元移动到另外的智能存储单元，并相应地更新了对应的元数据结构。等待时间条件可能被记录系统记录并报告给系统管理员。慢的连接条件的原因可能是诸如劣质的网卡、错误的双工协商或设备数据被相对频繁地  
20 读或写。

可使用多种策略确定在什么地方存储数据。这些策略可依据系统的目的进行调整，例如，遵照系统管理员设置的参数、满足已选择的冗余水平和/或性能改进。下面提供了可被预分配模块采用以存储数据的几个示例性的策略。应该意识到，可采用的策略很多，它们并且可以与下文  
25 讨论的结合使用，也可在其之外。

预分配模块可包括跨越多个智能存储单元的用于划分数据带区的分配方案。对数据划分带区是普通的技术，其通常使用在高端 RAID 存储设备上，但也在有多个磁盘的单用户工作站机器上使用。划分数据带区简单地意味着文件数据的不同部分存在于和/或被存储在不同的存储设备或  
30 磁盘上。划分数据带区的优势是当读请求跨越分配于多个磁盘上的块时，

每个磁盘都分担数据检索的总流量。利用典型的系统，划分数据带区是在软件设备层上完成的。也就是说，文件系统中没有关于划分数据带区的信息。只有文件系统下面的软件层明白该结构。在硬件的一些特定块中，这种划分是在比软件设备层更低、实际是在硬件层上完成的。在智能分布式文件系统 110 中，文件系统自身可处理数据带区的划分。这种实现为划分数据带区配置提供了更大的灵活性。作为示例，典型的 RAID 技术的局限之处在于，所有的磁盘必须是一样大小的并具有相同的性能特征。这些约束对于保证数据跨越设备地均匀分布是必须的。为更详细地讨论 RAID，请参考由 PAUL MASSIGLIA 撰写的“RAID BOOK”，第六版（1997），该文也被引入本文以作为参考。

利用智能分布式文件系统 110，不同的智能存储单元 114 可使用不同的磁盘和不同大小的磁盘并将其加入到文件带区中。预分配模块使用性能度量标准或预置规则，在根元数据数据结构中查寻磁盘设备信息并计算文件数据所跨越的智能存储单元的数目。然后预分配模块可分配文件的数据块到一组智能存储单元中。

预分配模块也可还包括用于奇偶校验或其他错误或丢失纠正保护的分配方案。在大多数 RAID 系统中，当文件带区被使用时，奇偶校验保护也可被使用，以使除一个以外的所有磁盘可被用于数据存储。最后一个磁盘纯粹用于奇偶校验信息。该奇偶校验信息通常利用对跨越所有数据磁盘的每个数据块按位异或（“XOR”）来计算。当磁盘故障发生时，该奇偶校验信息用于实现数据恢复。通过对剩余的磁盘数据块和奇偶校验信息按位 XOR 可重新计算丢失的数据。在典型的 RAID 系统中，数据是不可恢复的，直到替换磁盘被插入到阵列中以重建丢失的数据为止。

利用智能分布式文件系统 110，由于奇偶保护发生在文件系统层而不是软件设备层，因此丢失的数据可被重新计算并重写在其余智能存储单元的其他部分的未用空间中。如果没有足够的未用空间来重写数据，奇偶校验数据可能被重新计算的数据覆盖，并且冗余比最初的水平下降的事实可被记录和/或报告给系统管理员。

预分配模块可能还包括用于镜像数据的分配方案，也就是在不同的智能存储单元上制造可用数据的多个拷贝。预分配模块可能使用一种分

配方案以利用在存储空间、网络利用率、和/或 CPU 利用率方面使用较少的智能存储单元对跨越智能存储单元的数据块的位置进行负载平衡。镜像可提供增加的性能和增加的容错。如果要求对特定的内容块进行镜像，预分配模块分配用于最初数据及镜像数据的空间。如果要求大于 1 的容错水平，预分配器可利用容错计算并产生数据带区的镜像，以逻辑地划分智能存储单元或智能存储单元的子集。例如，如果在智能分布式文件系统 110 中有 10 个智能存储单元 114，并且要求容错为 2，那么预分配器可以逻辑地将智能分布式文件系统分为两部分，每部分带有 5 个智能存储单元，使数据带区在每个部分中跨越四个智能存储单元，并使用每个部分中的第 5 个智能存储单元作为奇偶校验磁盘。智能存储单元的这种划分可被称为阵列镜像切分。

### c. 故障恢复模块

故障恢复模块实时地重新配置智能分布式文件系统 110，以恢复由于设备故障而不再有效的数据。在维护性能的同时，故障恢复模块可实现重新配置而不中断服务，并可在短的时间周期内将数据恢复到所需的冗余水平。

如上面所讨论的，远程块管理器模块 337 检测故障并将这类故障的通知传送到故障恢复模块。对于最初的故障，故障恢复模块定位不能满足系统管理员设置的或智能分布式文件系统 110 设置的冗余参数的任何数据块。

首先，能从奇偶校验信息中重新创建的数据被重新创建，并且请求被发送到预分配模块以分配用于新数据空间。预分配器监测 CPU 和网络利用率并开始积极地操作直到 CPU 和网络利用率达到预定标志为止。该预定标志可能是系统管理员设置的或依据如计算机处理器等因素预先设置的。一旦达到该标志，故障恢复模块可有利地在标志的时间以能达到的速率重新计算数据，以减少对智能存储单元性能的冲击。

如果最近故障设备重新联机，故障恢复模块与被恢复设备的远程块管理器模块 337 通信以检验数据的完整性并修正任何不一致。

智能分布式文件系统 110 也可支持热备用设备 (hot standby device) 的进入。热备用设备是空闲的存储设备, 其在当前不处理任何数据存储, 但在设备故障时其将被投入使用。在这样的情况下, 故障恢复模块可通过与热备用设备的远程块管理器模块 337 进行通信, 从而利用热备用设备重建丢失的数据。

## 2. 块缓冲模块

块缓冲模块 333 管理数据块、名称查找和元数据数据结构的缓冲。在一个实施例中, 块缓冲模块 333 与 BSD 虚拟文件系统缓冲存储器联合工作或代替它工作。

块缓冲模块 333 可使用最近最少使用缓冲算法来高速缓冲数据块和元数据数据块, 但应该意识到还有多种缓冲算法可以使用, 例如, 频率缓冲。块缓冲模块 333 可依据哪个性能最好而确定使用哪种块缓冲算法。而在其他的实施例中, 算法可能被设置为默认的。

最近最少使用缓冲 (“LRU”) 是在大多数系统中使用的典型的缓冲方案。LRU 依据这样的原理工作, 即, 一旦数据被访问, 就很可能被再次访问。因而, 数据按其最后使用的顺序而被存储, 这样最长时间没有被访问的数据被丢弃。

频率缓冲存储被最频繁访问的数据。因为磁盘写是相对地时间集中的操作, 可通过在元数据数据结构中跟踪访问频率来获得附加性能并根据访问频率进行缓冲。

另外, 块缓冲模块 333 可利用特征为所请求的数据比所需要的数据更多的“点播 (on demand)”协议或“预读 (read ahead)”协议。块缓冲模块 333 可发送对一组数据的请求并请求该组数据的一定数量的前置数据。例如, 块缓冲模块 333 可执行预读, 例如一个报文预读, 二个报文预读、十个报文预读、二十报文预读等等。在其他的实施例中, 块缓冲模块 333 可根据请求的等待时间来使用预读技术。例如, 块缓冲模块 333 可执行 K 个报文预读, 其中 K 是利用读速率和链路的等待时间计算出来的。块缓冲模块 333 也可利用其他根据 CPU 和网络利用率的算法来



确定预读数据的大小。此外，块缓冲模块可使用设置缓冲协议 (set caching protocol)，或可改变缓冲协议以响应系统的性能水平。

缓冲器 340 可用由通用多用户操作系统提供的默认大小来实现或更改默认大小以不同程度地增加缓冲块大小，但不能严重地冲击系统性能。这样的调整可通过不同的性能测试来决定，这些性能测试依赖例如被存储的数据的类型，处理速度、在智能分布式文件系统智能存储单元的数量和被使用的保护方案等因素。

### 3. 本地块管理器模块

本地块管理器模块 335 管理本地存储在存储设备 360 上的数据块的分配、存储和检索。本地块管理器 335 可执行零拷贝文件读以将数据从磁盘移动到存储设备 360 的另外的部分 (例如，网卡)，从而改善性能。本地块管理器 335 还可基于所使用的存储设备 360 来实现调整，以提高性能。在一个实施例中，本地块管理器模块 335 驻留在本地文件存储层并可与 FreeBSD 快速文件系统联合使用或替代它使用。

### 4. 远程块管理器模块

远程块管理器模块 337 管理设备间通信，包括例如块请求、块应答和远程设备故障检测。在一个实施例中，远程块管理器模块 337 驻留在本地文件系统层。

在一个实施例中，智能存储单元 114 可能通过远程块管理器 337 与智能分布式文件系统 110 中的其他的智能存储设备 114 连接和/或通信。

远程块管理器模块 337 可使智能存储单元 114 能够经诸如 TCP 的连接而相互对话。在一个实施例中，每个智能存储单元之间至少有二个 TCP 连接，一个用于文件数据传输，一个用于控制报文传输。双通道 TCP 通信架构的优势在于，只要数据块以页大小的倍数形式发送，数据就可经 DMA (直接存储器存取) 传送直接从网络接口卡发送到系统内存，并经 DMA 传送从系统内存发送到系统的其他部分 (可能又是网络接口卡) 而无需将数据从系统内存的一个部分复制到其它部分。这是因为由于该信息是在控制通道传送的，因而其不包含非-数据报头或识别信息，从而无

需使用 CPU 解析数据报文。在高性能服务器和操作系统中，这些从系统内存的一部分到另外部分的内存拷贝是对系统性能的严重限制。

在一个实施例中，远程块管理器模块 337 使用报文通信进行通信，报文通信利用了例如数据块访问报文（如 READ, READ\_RESPONSE, 5 WRITE, 和 WRITE-RESPONSE）、元数据访问报文（例如，GET\_INODE, GET\_INODE\_RESPONSE, SET\_ADDRESS, GET\_ADDRESS, 以及 INVALIDATE\_INODE）、目录报文（如，ADD\_DIR 和 REMOVE\_DIR）、状态报文、以及其他类型的各种报文。

虽然以上讨论了双通道协议，但应该意识到也可使用其他的通信协  
10 议以在智能存储单元 114 之间进行通信。

## 5. 块设备模块

块设备模块 339 托管 (host) 用于被文件系统使用的特定磁盘硬件的设备驱动程序。例如，如果物理存储设备是 ATA 磁盘，那么块设备模块  
15 339 托管 ATA 磁盘驱动程序。

### C. 缓冲器

高速缓冲存储器或高速缓冲器 340 可由本领域中所周知的多种产品实现，例如，1G 的 RAM 高速缓冲。图 3 所示的缓冲器 340 可以存储最近被访问的或在设定的时间量内将被访问的数据块。缓冲器 340 可由高  
20 速存储机构，如静态 RAM 设备、动态 RAM 设备、内部缓存、磁盘高速缓存，以及其它类型的多种设备来实现。通常，从缓冲器 340 访问数据比访问非易失性存储设备的时间要快。缓冲器 340 存储数据，这样如果智能存储单元 114 需要从存储设备 360 访问数据，可首先被检查缓冲器  
25 340，看数据是否已经被检索。因而，使用缓冲 340 可改善智能存储单元在检索数据块时的性能。

### D. 网络堆栈

在一个实施例中，智能存储单元 310 还包括网络堆栈 350，其使用协议（例如 TCP/IP）处理引入的和输出的报文通信量。然而，应该意识到，也可使用其他的协议或数据结构来实现堆栈 350。

#### 5 E. 存储设备

存储设备 360 是可用于存储数据块的非易失性存储设备。存储设备 360 可使用本领域公知的多种产品实现，例如 4 个 1.25 GB 的 ATA100 设备，SCSI 设备，等等。另外，用于智能分布式文件系统 110 中的智能存储单元 114 的存储设备 360 的大小可以相同，或者也可不同于智能存储  
10 单元 114 的大小。

#### F. 系统信息

在一个实施例中，智能存储单元 114 运行在能够使智能存储单元 114 与其他智能存储单元 114 通信的计算机上。该计算机可能是使用一个或多个微处理器的通用目的计算机，例如，奔腾处理器、奔腾 II 处理器、  
15 奔腾 Pro 处理器、奔腾 IV 处理器、xx86 处理器，8051 处理器，MIPS（每秒百万条指令）处理器，强力 PC 处理器、SPARC 处理器、Alpha 处理器等等。

在一个实施例中，处理器单元运行源代码开放的 FreeBSD 操作系统并执行标准的操作系统功能，如打开、读、写和关闭文件。应该意识到，  
20 也可以使用其他的操作系统，例如，微软公司的 Microsoft® Windows® 3.X、Microsoft® Windows 98、Microsoft® Windows® 2000、Microsoft® Windows® NT、Microsoft® Windows® CE、Microsoft® Windows® ME、PALM 公司的 Palm Pilot OS、苹果公司的 Apple® MacOS®、磁盘操作系统（DOS）、  
25 UNIX、IRIX、Solaris、SunOS、FreeBSD、Linux®或 IBM 公司的 IBM® OS/2®操作系统。

在一个实施例中，计算机装备有常规的网络连接，例如，以太网（IEEE 802.3）、令牌环网（IEEE 802.5）、光纤分布式数据连接接口（FDDI）、或异步传输模式（ATM）。进一步，计算机可被配置为支持多种网络协

议，如基于 UDP/TCP 的 NFS v2/v3、Microsoft® CIFS、HTTP（超文本连接协议）1.0、HTTP. 1.1、DAFS、FTP（文件传送协议）等等。

在一个实施例中，智能存储设备 114 包括单或双 CPU 2U 可架式安装的配置、多个 ATA100 接口、以及支持超大（jumbo）的 9K 以太网帧的  
5 1000/100 网络接口卡。然而，应该意识到，也可以使用不同的配置。

## VI. 智能分布式文件系统数据结构

图 4 示出了可与智能分布式文件系统一起使用的示例目录结构。在该例中，根目录（ROOT directory）被命名为“DFSR”并包括多个子目  
10 录 IMPORTANT（重要文件目录）、TEMP（临时目录）和 USER（用户目录）。子目录 IMPORTANT 包括子目录 PASSWORD（口令目录）和 CREDITCARD（信用卡目录）。文件 USER.TXT 和 ADMIN.TXT 存储在 PASSWORD 子目录内。因而，USER.TXT 文件的地址是：

/DFSR/IMPORTANT/PASSWORDS/USER.TXT

15 关于目录和文件的信息或元数据由智能分布式文件系统 110 存储和维护。

### A. 元数据数据结构

图 5 示出了用于存储元数据的示例数据结构 510。该示例数据结构  
20 510 存储了下列各项信息：

字段	描述
Mode (模式)	文件的模式（例如常规文件、特定的块、特殊字符、目录、符号链接、先入先出、插件、淡入淡出（whiteout）、未知）
Owner（所有者）	拥有文件所有权的智能存储单元的帐户
Timestamp（时间标记）	文件的最后修改的时间标记
Size（大小）	元数据文件的大小
Parity Count (奇偶校验计数)	使用的奇偶校验设备的数量

Mirror Count (镜像计数)	使用的镜像设备的数量
Version (版本)	元数据结构的版本
Type (类型)	数据位置表的类型 (例如, 类型 0、类型 1、类型 2、或类型 3)
Data Location Table (数据位置表)	数据位置表地址或实际的数据位置表信息
Reference Count (参考计数)	所引用的元数据结构的数目
Flags (标志)	文件许可 (例如标准的 UNIX 许可)
Parity Map Pointer (奇偶校验映射指针)	指示奇偶校验块信息的指针

应该意识到, 示例数据结构 510 示出了用于存储元数据的数据结构 510 的一个实施例, 可以根据本发明采用多种实现方式。例如, 数据结构 510 可能包括不同的字段, 这些字段可具有不同的类型, 这些字段可被分  
5 组并单独存储等等。

图 6A、6B、6C 和 6D 提供了用于数据位置表的一些类型 (即分别为类型 0、类型 1、类型 2 和类型 3) 的示例数据位置表结构。在图 6A 中, 类型 0 的数据位置表包括 24 个直接块条目, 这意味着数据位置表中的条  
目包括指示数据块存储位置的设备/块数目对。在图 6B 中, 类型 1 数据位  
10 置表包括 15 个直接块条目、三个单独的间接条目、三个二级间接条目和三个三级间接的条目。代表单独间接条目的条目指示出了直接条目的附加数据位置表所存储的位置。代表二级间接条目的条目指示出了包括单独间接条目的数据位置表所存储的位置。代表三级间接条目的条目指出了包括二级间接条目的数据位置表所存储的位置。

15 因为任何块都可以跨越任意数量的设备而被镜像, 所以元数据数据结构 510 是灵活的, 足够代表具有多个位置的块, 并仍可提供来自固定空间内的直接索引的快速访问。因而, 类型可以有利地与元数据数据结

构 510 联系在一起以指示出待要使用的数据位置表的类型。在元数据数据结构 510 的一个实施例中，可能有用于 24 个数据条目的空间 (room)，例如，24 个指针。

数据文件小的时候可以使用类型 0；数据位置地址被作为直接条目  
5 存储。因而，类型 0 元数据数据结构包括 24 个直接条目。类型 1 用于支持至多达两倍的较大的文件和镜像（文件的三个拷贝）。类型 1 使用 15 个直接条目，三个单级间接条目、三个二级间接条目和三个三级间接条目。类型 2 可被用于支持至多达 7 倍的镜像（8 个文件拷贝），并包括八个单级间接条目，八个二级间接条目和八个三级间接条目。类型 3 的数  
10 据位置表可以进一步将全部磁盘地址镜像为三级间接条目。结果，其可存储至多达 24 个完整的文件拷贝。

应该意识到可以使用各种不同的数据位置表，而且图 6A、6B、6C 和 6D 仅示出了示例性的实施例。在其他的实施例中，例如，数据位置表可包括直接和间接条目的不同混合。进一步，在其他的实施例中，数据  
15 位置表可包括指示表中各个条目的条目类型的条目字段。这些类型可包括，例如上面讨论的（例如，直接的、单级间接的、二级间接的，三级间接的）以及其他的（例如，四级间接的等等）。另外，数据位置表可包括至多达 X 水平的更深嵌套的数据位置表，其中 X 是整数。

## 20 1. 目录元数据

图 7A 示出了用于目录 PASSWORD 的一组示例元数据。在图 7A 中，数据结构存储了关于 PASSWORD 目录的信息。该目录被镜像两次（总共三个拷贝）。因为目录结构相对较小（例如，其适于在一个块中），所以只使用了三个直接的指针，每个拷贝使用一个指针。该组示例元数据的  
25 示例包括数据位置表 710，其包括直接条目 720 以及一组未用的块条目 730，直接条目 720 用设备/块号对指出了数据块的位置。

## 2. 文件元数据

图 7B 示出了用于文件 USER.TXT 的一组示例元数据。在图 7B 中，  
30 该数据结构存储了关于文件 USER.TXT 的信息。对用于 USER.TXT 文件

数据的每个数据块都有一个拷贝，并且数据采用 3+1 奇偶校验方案保护。用于 USER.TXT 的内容数据的大小是 45K，块的大小是 8K，因而，有 6 个数据块，其中第 6 个数据块没有完全使用。数据位置表 710 示出了 6 个数据块中的每一个被存储的位置 720，其中，数据块由设备号和块号引用，其中第一个条目与第一数据块相应。进一步，用于内容数据的奇偶校验信息的位置被存储在奇偶校验映射 740 中，其位置由数据结构的最后位置的“奇偶校验映射指针”指出。USER.TXT 文件使用 3 + 1 奇偶校验方案存储，因而，对每三块数据块，存储一个奇偶校验数据块。因为在 3+1 奇偶校验方案中有六个块，因而有两个奇偶校验数据块（6 除以 3 并舍入到最接近的整数）。奇偶校验映射示出了两个奇偶校验数据块的存储位置，其中奇偶校验数据块被设备号和块号引用，并且第一条目与第一奇偶校验数据块相应。

#### B. 数据位置表数据结构

智能分布式文件系统 110 可为各种数据文件提供存储器并可提供如何存储这些数据文件的灵活性。数据文件的冗余和镜像在文件系统级别被完成，从而使智能分布式文件系统 110 能够支持不同文件的不同冗余参数。例如，一些目录可得到镜像、奇偶校验保护，或根本不受保护。

图 8A、8B 和 8C 示出了示例的数据位置表，其可用于存储保护类型和级别不同的数据文件的数据位置信息。图 8A、8B 和 8C 示出了不同的数据位置表，应该意识到，也可以使用不同的格式和/或结构。

图 8A 示出了示例数据位置表 810，其指出了相应文件的每个数据块在哪里存储。虽然应该意识到，数据位置表 810 可能与一组元数据相应，但应注意，图中并未示出与文件（如图 7B 所示的文件）相应的元数据。示例的数据位置表 810 既包括直接条目也包括间接条目。

直接条目包括设备 ID/块对。设备 ID 指出数据被存储的智能存储单元，偏移或块地址则指出了数据在存储数据的设备上的位置。数据位置表上的一个示例条目可以为：

条目	设备	块
1	7	127

它表明数据块 1 被存储在 7 号设备的块 127 上。

示例数据位置表 810 还可包括指向另外的数据位置表的间接条目，以使数据位置表能够跟踪较大的数据组的数据位置。虽然间接条目的级别在理论上无限的，但为改善流量最好还是要限制该级别。例如，数据位置表可被限制为只允许至多二级间接条目或至多三级间接条目。示例数据位置表 810 示出了两个级别的间接条目。

进一步，数据位置表的最后条目可被保留以用于存储奇偶校验映射的地址（如果有）。在其他的例子中，奇偶校验映射的地址可被存储在其他的位置，例如，作为元数据数据结构中的一个条目。如果一组数据不包括奇偶校验保护，则地址值可被设置为标准值，例如 NULL（零）。

图 8B 示出了用于被镜像在两个附加位置的数据的数据位置表。该数据位置表包括设备 ID 以及用于数据的每个拷贝的块或偏移地址。在该示例的数据位置表中，镜像位置被逐块（block-by-block）地选择。应该意识到，也可采用其它的方案，例如，选择一个或更多的智能存储单元来镜像特定的智能存储单元。尽管图 8B 所示的数据位置表只包括直接条目，但应该意识到间接条目也可被使用。

在一个实施例中，用于文件的镜像信息可以被存储在文件的相应元数据结构中。信息可包括例如数据的拷贝数以及每个拷贝的数据位置表的位置。应该意识到，数据位置表可作为单一的数据结构被存储，并且/或者数据位置表的单独拷贝可被存储在不同的位置。

虽然图 8B 所示的带有镜像数据的示例数据位置表不包括奇偶校验保护，但应该意识到该数据位置表也可包括奇偶校验信息。

图 8C 示出了带有奇偶校验映射的数据位置表。在该示例的数据位置表中，数据用 3 + 1 奇偶校验方案保护，也就是说，为每三个数据块创建一组奇偶校验数据。可以使用本领域中用于创建数据的公知技术，例如，对数据一起逐比特地（bit-by-bit）、逐字节地（byte-by-byte）或逐块地（block-by-block）进行 XOR 以创建奇偶校验块。

示例的数据位置表提供了与由 21 个数据块（块 0 到块 20）组成的数据文件有关的信息。因为奇偶校验方案是 3 + 1，所以要为每三个数据



块的组创建奇偶校验块。表 2 示出了图 8C 所示的一些数据块和一些奇偶校验块之间的对应关系。

数据块			奇偶校验块
0 设备 5 块 100	1 设备 9 块 200	2 设备 7 块 306	0 设备 0 块 001
3 设备 5 块 103	4 设备 9 块 203	5 设备 7 块 303	1 设备 8 块 001

表 2

5

该示例数据位置表包括奇偶校验映射或奇偶校验位置表。在示例的奇偶校验映射中，用于创建数据的一组块条目与奇偶校验映射之间存在着一对一的映射。在其他实施例中，奇偶校验映射还可包括大小可变的条目，在由于设备故障而导致数据的任何直接位置都无效的情况下，该大小可变的条目可以利用设备和块号来指明哪些块将被一起受到奇偶校验的 XOR 以再生数据。在其他的实施例中，奇偶校验产生方案是预先设置的，这样可以由智能分布式文件系统 110 确定出奇偶校验数据的位置和对应性而不必指明将被一起进行 XOR 以再生数据的数据块。

10 在一个实施例中，奇偶校验映射由元数据数据结构指明，例如，在元数据数据结构的最后一项条目中被指明，而并不包含在元数据数据结构之中。因为该映射只在智能存储单元 114 出现故障的非正常情况下才需使用，所以其可被指明而不是直接包括在元数据结构中。奇偶校验映射也可使用大小可变的条目来表达奇偶校验重组块，以使智能存储单元 114 能够在单一时间 (single time) 上横移 (traverse) 奇偶校验映射，同时重建数据并在对横移的奇偶校验映射进行解析。在一些情况下，与奇偶校验计算时间相比，检索和解析条目的计算和 I/O 时间是可以忽略的。

20 虽然图 8C 所示出的带有奇偶校验位置信息的示例数据位置表 810 不包括镜像信息或间接条目，但应该意识到，这两者中的一个或两者均可

与奇偶校验位置信息联合使用。进一步，应该意识到，也可使用其他的数据结构，而且上述数据位置表数据结构的用意只是为了示出本发明的一个实施例。

### 5 C. 示例数据

图 9 示出了示例数据位置表 910 和奇偶校验映射 920 以及相应的存储该数据的设备。图 9 的例子示出了数据如何被存储在设备的不同位置上，被存储的数据“带区 (stripes)”跨过了每个设备的不同的偏移地址，并且奇偶校验数据可被存储在不同的设备上，即使对来自同一文件的数据也是一样。在其他的实施例中，数据可被存储在每个设备的同一偏移地址中。

例如，用于第一带区的奇偶校验数据被存储在设备 3 的位置 400 上，并与存储在设备 0 的位置 100 上的数据块 0、存储在设备 1 的位置 200 上的数据块 1 和存储在设备 2 的位置 300 上的数据块 2 相关。用于第二带区的奇偶校验数据被存储在设备 2 的位置 600 上，并与存储在设备 0 的位置 300 上数据块 3、存储在设备 4 的位置 800 上的数据块 4 和存储在设备 1 的位置 700 数据块 5 相关。

在一些实施例中，单个设备确定在哪里和/或怎样将该位置映射到磁盘上的实际位置。例如，如果设备 0 带有 4 个物理硬盘，每个硬盘带有 100 块的存储容量，那么设备 0 将允许从位置 0 到位置 399 的存储。下面是可被用于确定该位置如何映射到磁盘上的块的指南的一组示例：

磁盘号 = (位置/每片磁盘的块数) 的基数 (floor)

磁盘上的块 = 位置 MOD 每片磁盘的块数。

注意，MOD 是模数操作符，其取的是除运算的余数。应该理解，上述指南仅仅是可用于将位置映射到磁盘和磁盘块的指南的一个示例，也可以使用许多其他的指导方针或方案。例如，一个实施例可使用代表每个磁盘的块范围的链接列表并引导列表的横移。链接列表具有允许多种尺寸的磁盘的优势。

由于数据存储和奇偶校验信息的灵活性，在新的智能存储单元被添加时，新的数据可移被存储在新的智能存储单元并且/或者现有的数据可

被移动到新的智能存储单元（例如，通过在删除现有单元上的数据之前制造拷贝）而不破坏系统。另外，响应于高的请求量、磁盘故障、冗余或奇偶校验参数中的变化等等，数据块或全部文件可被实时地移动或复制。

5

## VII. 智能分布式文件系统的处理过程

### A. 检索数据

图 10 示出了用于检索数据（“检索数据过程”）的流程图的一个实施例。可检索多种数据类型，例如，目录元数据、文件元数据、内容数据等等。

10

在起始状态开始，检索数据过程接收数据被存储的位置（块 1010）。在一个实施例中，该位置可由智能存储单元 ID 以及偏移或块地址指明。在其他的实施例中，可使用存储设备 ID，然而在其他的实施例中，也可使用表以将这些 ID 映射到其他 ID 之上，等等。

15

接着，检索数据过程确定数据是否是本地存储的（块 1020）。如果数据是本地存储的，则检索数据过程从本地存储器中检索数据（块 1030）。在一个实施例中，检索数据过程可首先检查高速缓存，如果数据不在那里，则检查存储设备。在其他的实施例，检索数据过程可以只检查存储设备。

20

如果数据不是存储在本地，那么检索数据过程将对数据的请求发送到数据所存储的智能存储单元（块 1040）。在一个实施例中，请求经图 1 所示的交换部件 125 发送。接收数据过程随后接收被请求的数据（块 1050）。

25

检索数据过程收集被请求的数据并返回该数据（块 1060）。在一些实施例中，数据在全部数据被收集之后返回。在其他的实施例中，在数据在本地存储器中检索或从其他的智能存储单元接收时，返回数据组或数据的一部分。该部分可根据文件位置表按照顺序返回或在被检索或接收时返回。在数据被返回之后，检索数据过程进入结束状态。

30

应该意识到，图 10 示出了检索数据过程的一个实施例，也可使用其他的实施例。在另一个例子中，可以同时使用一个以上的检索数据过程，

从而利用诸如并行处理、管道技术或异步 I/O 的技术或这些技术的组合以通过多个检索数据过程对数据进行并行检索。

## B. 处理名称查找

- 5 图 11 示出了用于名称查找过程（“名称查找过程”）的一个实施例。在开始状态开始之后，名称查找过程接收文件名（块 1110），检索根目录的元数据，将根元数据的位置设置为当前位置 CURRENT（块 1120）。在一个实施例中，根目录的数据可存储在数据结构中，例如图 5 所示的数据结构，但应该意识到，也可使用多种数据结构来存储根目录的元数据。此外，在一些实施例中，根目录的元数据可存储在各个智能存储单元 114 上，这样每个智能存储单元 114 都带有相同或相似的根目录元数据的拷贝。在其他的实施例中，根目录的元数据可被存储在智能分布式文件系统 110 中的其他位置上，或与文件请求一起被发送给智能存储单元 114。应该意识到，可以使用公知的用于保证多个拷贝的完整性的技术，  
10 例如，经 mutexes（多用户终端执行程序）和/或 semaphores（信号装置）进行锁定，等等。

- 名称查找过程可以随后检索下一个令牌，该令牌为文件名称的一部分（块 1130）。名称查找过程随后请求从存储用于当前位置（CURRENT）的数据的智能存储单元 114 获取该令牌的元数据的位置地址（块 1140）。  
20 该请求可以是本地的或远程的。名称查找过程可以随后将返回的地址设置为当前地址（块 1150）并确定是否有另外的令牌（块 1160），其中所述令牌表示目录分层结构中的一个单一的层次。如果有另外的令牌，名称查找过程将返回块 1130。如果没有更多的令牌，则名称查找过程返回当前位置的（块 1170）值或参考并进入结束状态。

- 25 应该意识到，也可以使用名称查找过程的其它实施方式。例如，名称查找过程可以检索文件的元数据数据。另外，一旦找到了被请求数据的位置，名称查找过程可以确定数据是存储在本地或存储在其他的智能存储单元上。如果数据存储在本地，名称查找过程可以发送读请求到智能存储单元 114 的本地块管理器模块 335；如果数据存储在另外的智能存

储单元上，则名称查找过程可以发送读请求到远程智能存储单元的 114 的远程块管理器模块 337。

### C. 处理文件请求

5 图 12 示出了用于处理文件请求（“文件请求过程”）的流程图的一个实施例。在开始状态开始后，文件请求过程接收检索文件的请求（块 1210）。在一个实施例中，用文件的全路径名称，包括位置和文件名，来指定文件。在其他的实施例中，路径可以是相对路径和/或其他的数据结构，例如，可被用于存储关于文件的地址信息的表。下一步，文件请  
10 求过程执行名称查找过程，如图 11 所示出的（块 1220）那样，以确定文件的元数据数据结构的位置。

虽然可以使用其他的检索文件过程，但文件请求过程可以随后用如图 10 所示的和上面讨论的检索文件过程来检索文件的元数据（块 1230）。在一个实施例中，文件的元数据可包括数据位置表，其提供对整个智能  
15 分布式文件系统的位置的访问，文件中的每个数据块存储在这些位置中。

然后，对（For）文件中的每个数据块（块 1240，1270），通过在文件的元数据中进行查找，文件请求过程获得数据块的位置（块 1250），并用例如如图 10 所示的和讨论上面的检索文件过程来检索数据块（块  
1260），也可使用其他的检索文件过程。

20 文件请求过程随后返回文件的数据（块 1280）并进入结束状态。在一些实施例中，在全部数据被收集之后返回文件。在其他的实施例中，随着数据的检索而返回一个或更多的数据块。这些部分可依照文件位置表按顺序返回，或在检索或接收时被返回。在一个实施例中，文件请求过程可将数据块按顺序放置，并且/或者其他模块（例如流服务器）可对  
25 数据块进行排序。在数据被返回之后，检索数据过程进入结束状态。

应该意识到，虽然图 12 示出了文件请求过程的一个实施例，但也可使用其他的实施例。例如，文件请求过程可以使用与图 11 所示的不同的名称查找过程来确定文件的位置。在另外的例子中，可以同时使用一个  
一个以上的检索数据过程来检索数据块，从而利用诸如并行处理、管道

技术或异步 I/O 的技术或这些技术的组合以通过多个检索数据过程对数据进行并行检索。

#### D. 奇偶校验产生过程

5 图 13 示出了用于生成奇偶校验信息（“奇偶校验产生过程”）的一个实施例的流程图。在开始状态开始之后，奇偶校验产生过程接收与一组数据相关的奇偶校验方案信息（块 1310）。这组数据可代表文件数据、文件元数据、目录元数据、文件数据的子集，等等。奇偶校验产生过程接收与这组数据相关的数据位置信息（块 1320）。下一步，对每组奇偶  
10 校验数据（块 1330, 1370），奇偶校验产生过程检索一组数据（块 1340）。例如，如果奇偶校验是 3+1，奇偶校验产生过程使用如图 10 所示的数据检索过程检索数据的前三块。下一步，奇偶校验产生过程生成用于该组数据的奇偶校验数据（块 1350），例如，执行基于逐位（bit-by-bit）、逐字节（byte-by-byte）或逐块（block-by-block）的 XOR 数据操作。奇偶  
15 校验产生过程可随后在缓冲器中存储数据并返回到块 1330，直到用于该组数据的奇偶校验信息被生成为止。奇偶校验信息生成之后，奇偶校验产生过程确定在哪里存储奇偶校验数据（块 1380）。奇偶校验产生过程可能使用轮转（rotating）奇偶校验方案，其中用于文件数据的每个连续带区的每个奇偶校验块被轮流地存储在下一个设备上。奇偶校验产生过  
20 程在与保留有用于当前带区的数据的任何设备不同的设备上分配奇偶校验块，以保证如果发生设备故障，奇偶校验信息不会与数据信息一样丢失。奇偶校验产生过程还可考虑其他的因素，例如存储容量、CPU 利用率和网络利用率，以将一些设备排除在用于奇偶校验存储的考虑之外。奇偶校验产生过程随后在分配空间内存储缓冲的数据（块 1390），将奇  
25 偶校验数据的位置记录在奇偶校验映射中（块 1395），并返回到结束状态。

应该意识到，虽然图 13 示出了奇偶校验产生过程的一个实施例，但也可使用其他的实施例。例如，奇偶校验产生可并行地检索数据块并且并行地生成奇偶校验信息，或使公知的管道技术或异步 I/O 技术。进一步，  
30 奇偶校验产生过程可存储奇偶校验信息和奇偶校验信息的位置而不必写

到临时的缓冲器中，或者奇偶校验产生过程可返回奇偶校验数据或奇偶校验数据的指针。

#### E. 数据恢复过程

5 图 14 示出了用于恢复丢失的或损坏的数据（“数据恢复过程”）的流程图的一个实施例。在开始状态开始之后，数据恢复过程接收关于所使用的奇偶校验方案的信息（块 1410）。数据恢复过程随后接收关于故障或损坏的磁盘或数据信息（块 1420）。下一步，数据恢复过程接收用于奇偶校验块组的地址信息，故障或损坏的数据被分配在这些组中（块  
10 1430）。数据恢复过程随后从可用的智能存储单元中检索出数据块（块 1440）。数据可以利用如图 10 所示的检索数据过程被检索。数据恢复过程进行纠错（块 1450），例如，根据奇偶校验方案对块执行 XOR，并将结果存储在缓冲器中（块 1460）。缓冲器中的数据代表丢失的数据。数据恢复过程可以随后返回缓冲器中的数据（块 1470）并进入结束状态。

15 应该意识到，虽然图 14 示出了数据恢复过程的一个实施例，但也可使用其他的实施例。例如，数据恢复过程可返回恢复的数据而不存储它。

### VIII. 结论

20 虽然已经描述了本发明的特定实施例。但这些实施例只是一些示例而已，其意图并不是对本发明范围的限制。因此，本发明的宽度和范围应该根据权利要求及其等价物来定义。

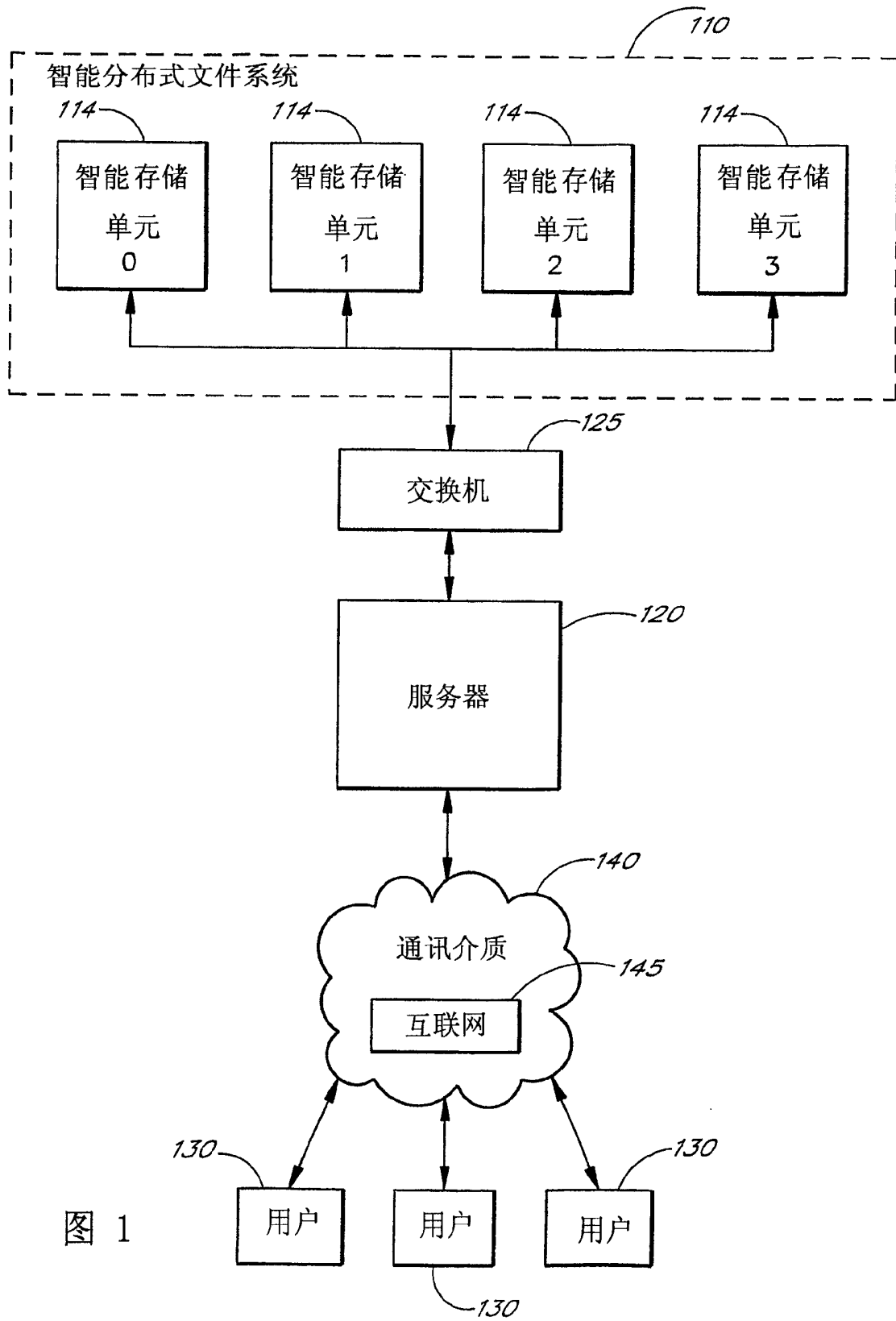


图 1



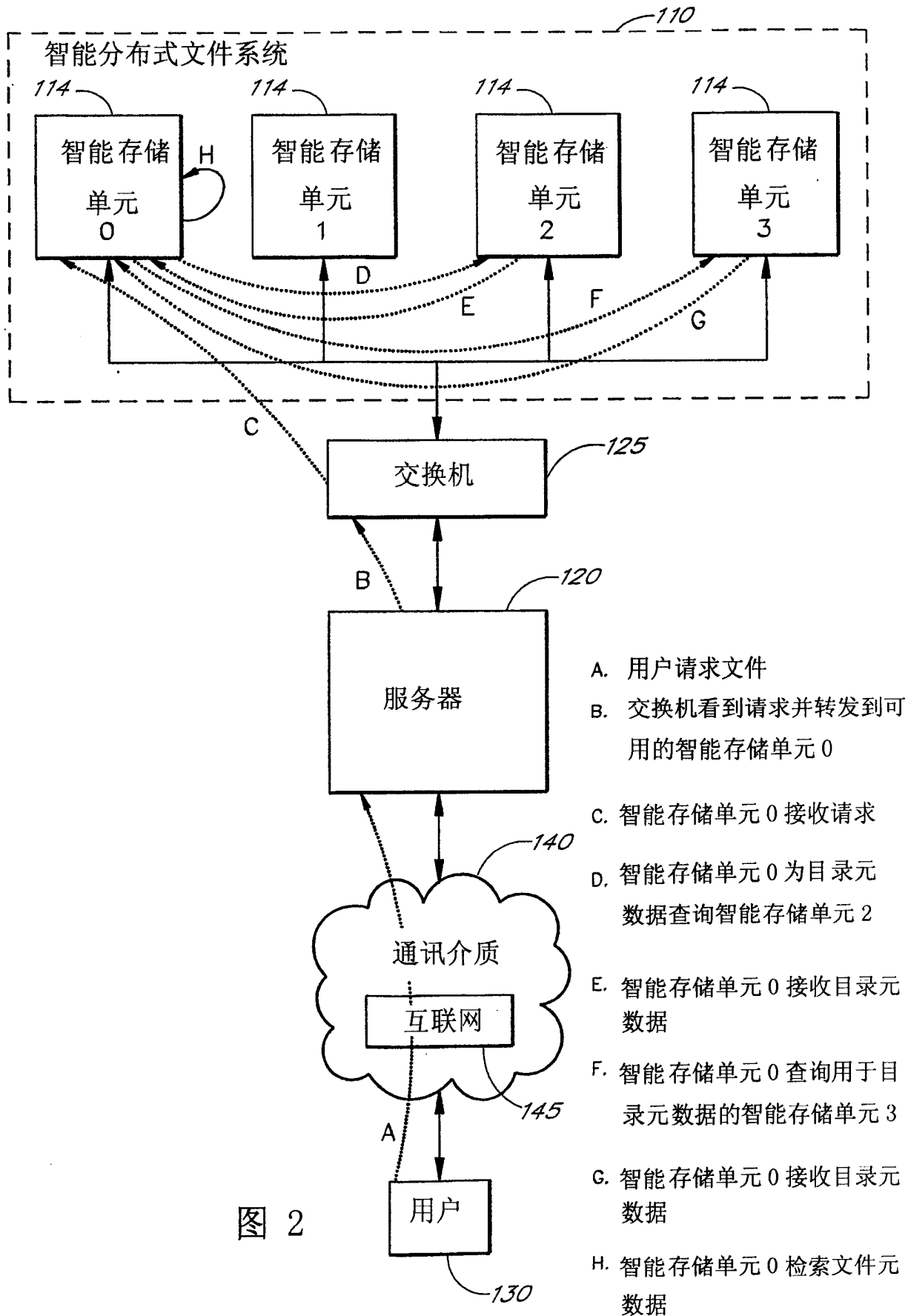


图 2

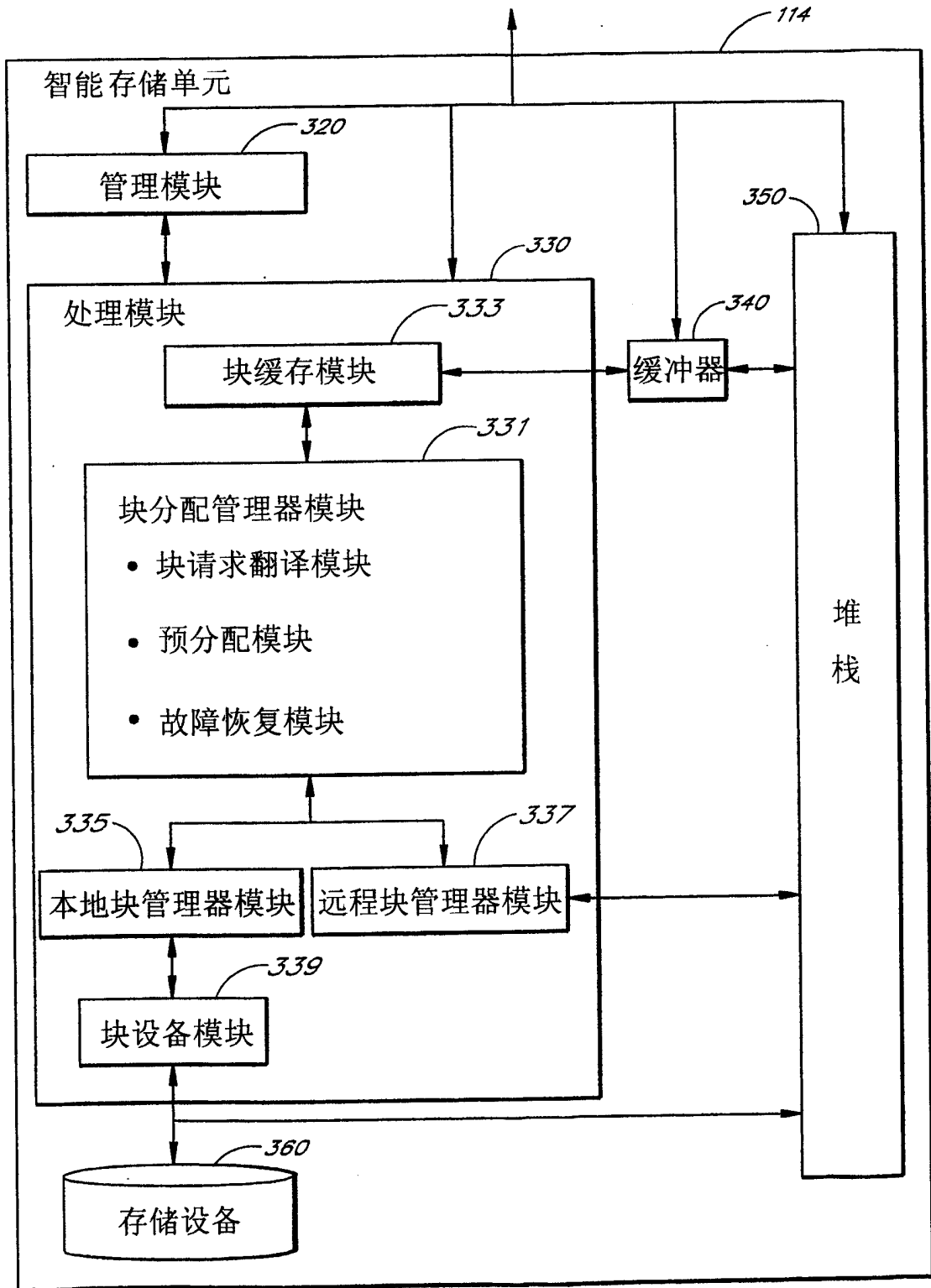


图 3

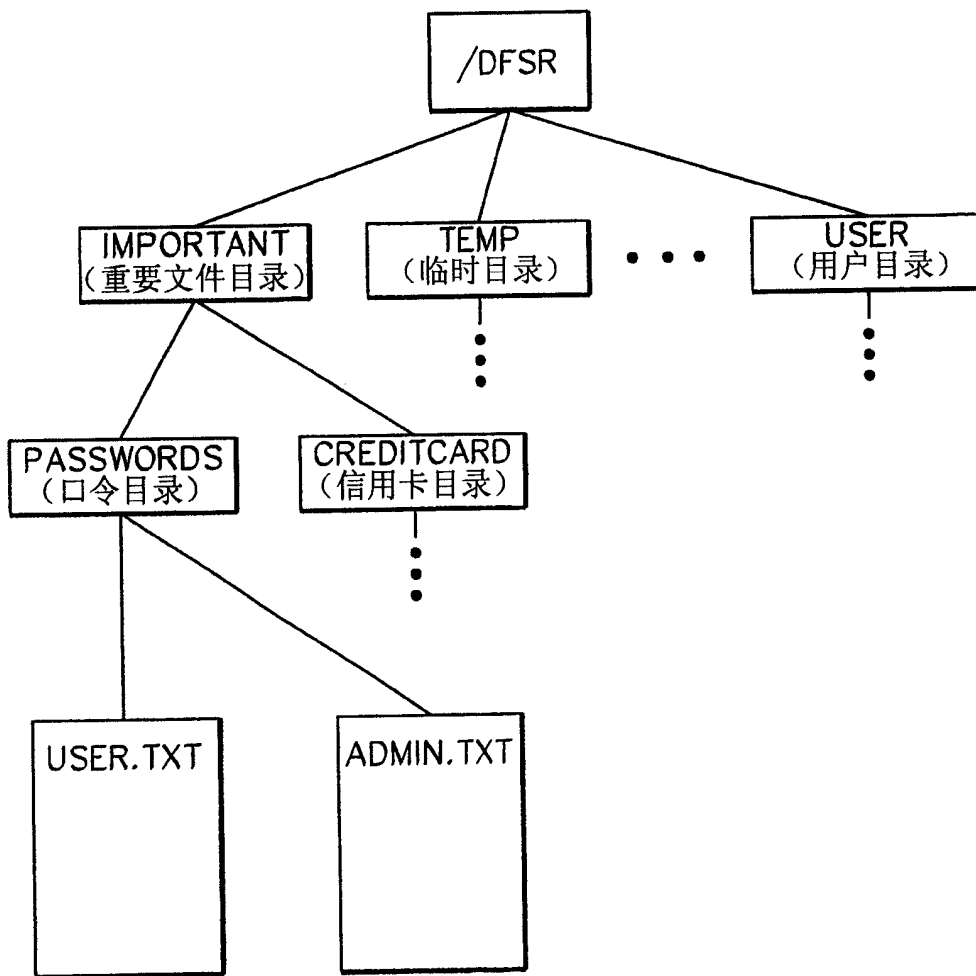


图 4

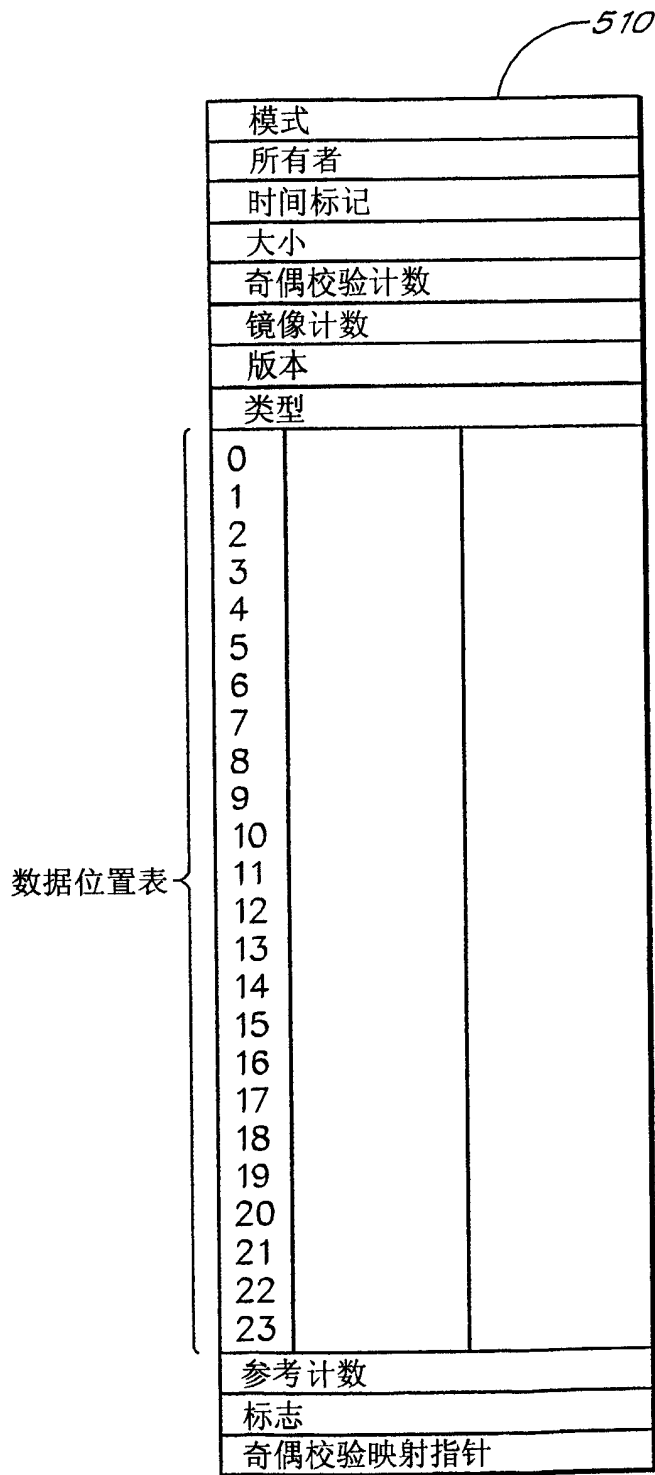


图 5

类型 0

	设备	块	
0			D0
1			D1
2			D2
3			D3
4			D4
5			D5
6			D6
7			D7
8			D8
9			D9
10			D10
11			D11
12			D12
13			D13
14			D14
15			D15
16			D16
17			D17
18			D18
19			D19
20			D20
21			D21
22			D22
23			D23

图 6A

类型 1

	设备	块	
0			D0
1			D1
2			D2
3			D3
4			D4
5			D5
6			D6
7			D7
8			D8
9			D9
10			D10
11			D11
12			D12
13			D13
14			D14
15			S10
16			D10
17			T10
18			S11
19			D11
20			T11
21			S12
22			D12
23			T12

图 6B

类型 2

	设备	块	
0			SI0
1			DI0
2			TI0
3			SI1
4			DI1
5			TI1
6			SI2
7			DI2
8			TI2
9			SI3
10			DI3
11			TI3
12			SI4
13			DI4
14			TI4
15			SI5
16			DI5
17			TI5
18			SI6
19			DI6
20			TI6
21			SI7
22			DI7
23			TI7

图 6C

类型 3

	设备	块	
0			TI0
1			TI1
2			TI2
3			TI3
4			TI4
5			TI5
6			TI6
7			TI7
8			TI8
9			TI9
10			TI10
11			TI11
12			TI12
13			TI13
14			TI14
15			TI15
16			TI16
17			TI17
18			TI18
19			TI19
20			TI20
21			TI21
22			TI22
23			TI23

图 6D

模式	DIRECTORY (目录)				
所有者	ROOT (根)				
时间标记	65536				
大小	345				
奇偶校验计数	0				
镜像计数	3				
版本	1				
类型	1				
710	设备 1	块 11	720		
	设备 2	块 21			
	设备 3	块 31			
730			730		
				参考计数	1
				标记	777
				奇偶校验映射指针	NULL (零)

图 7A

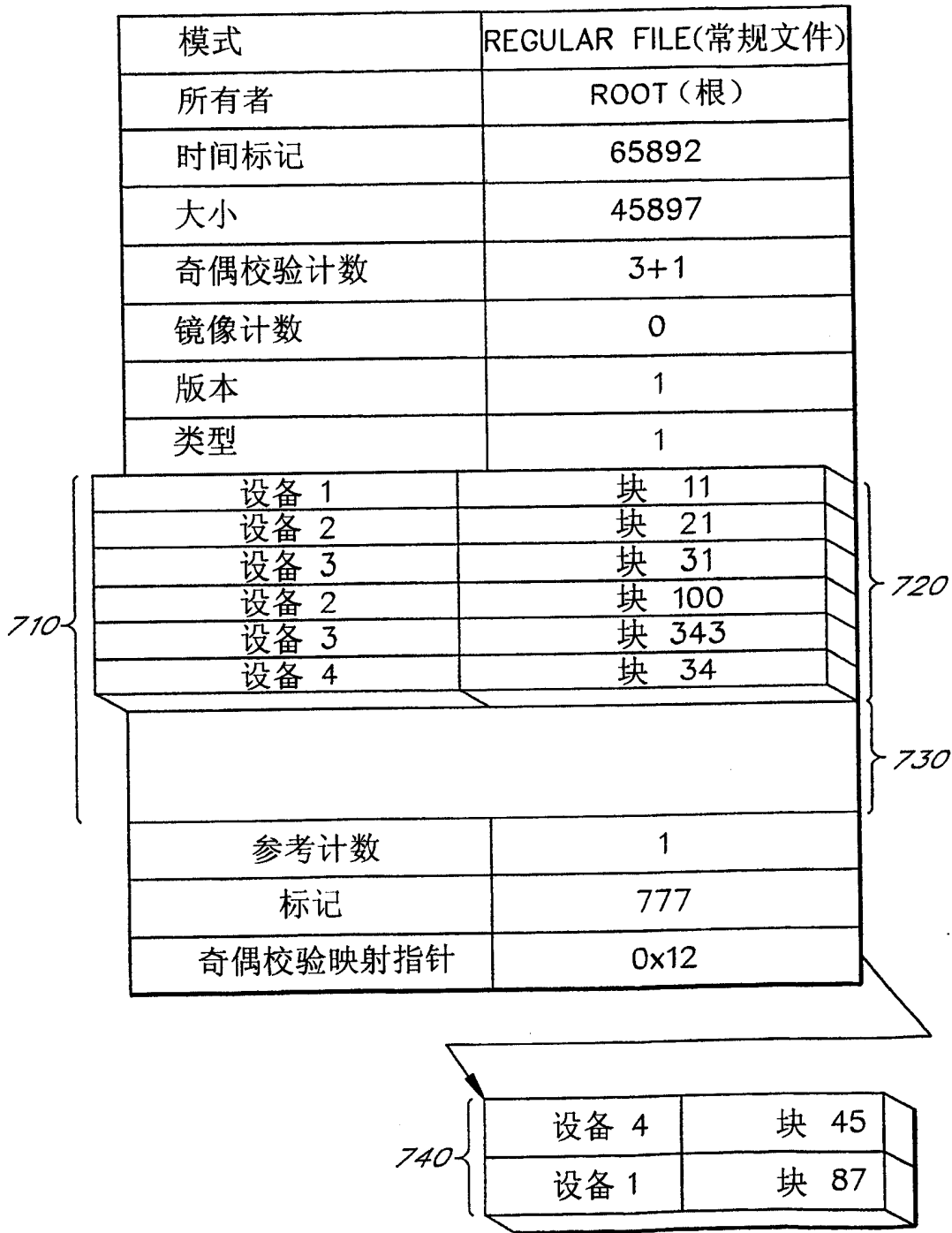


图 7B



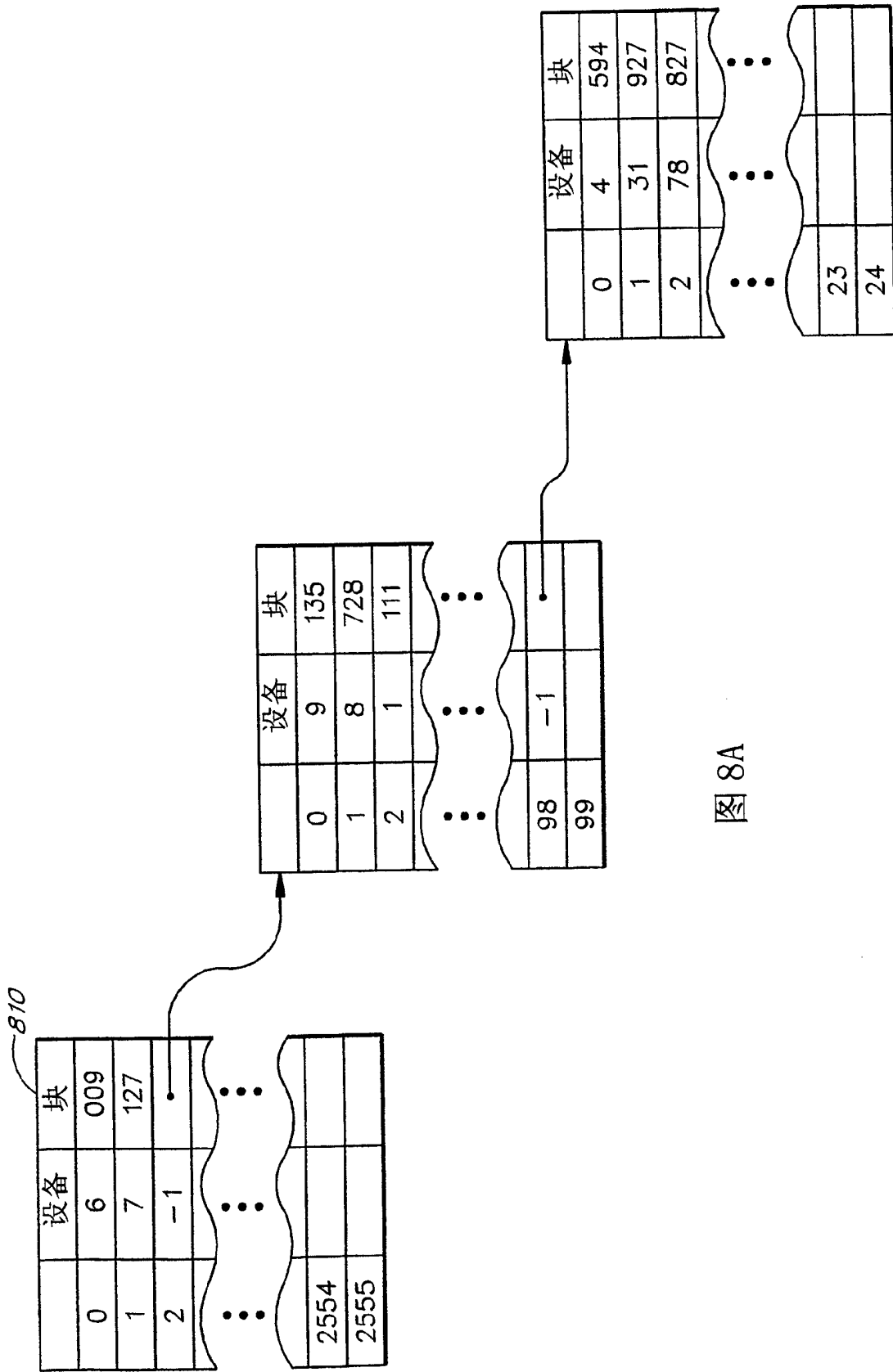


图 8A

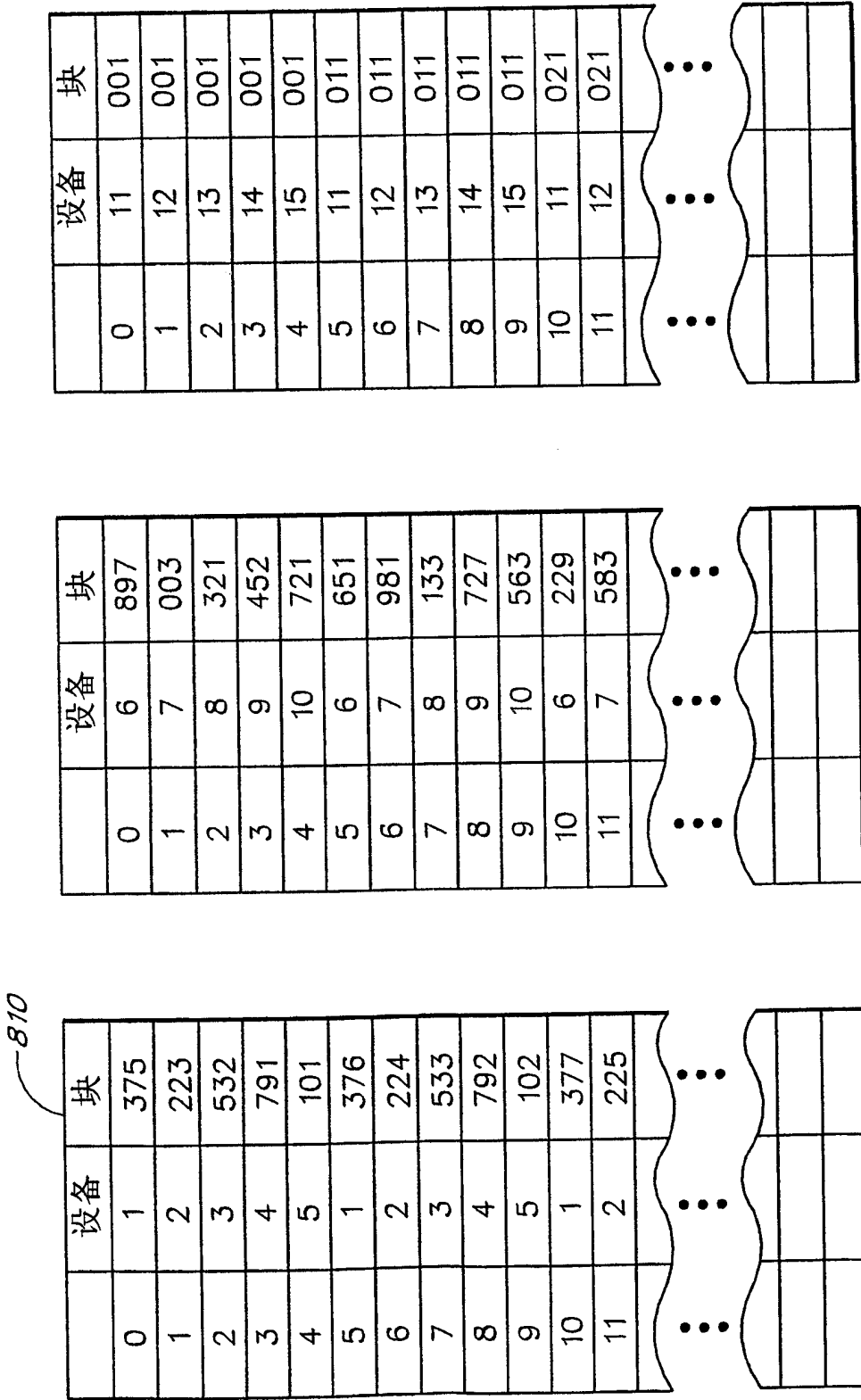


图 8B

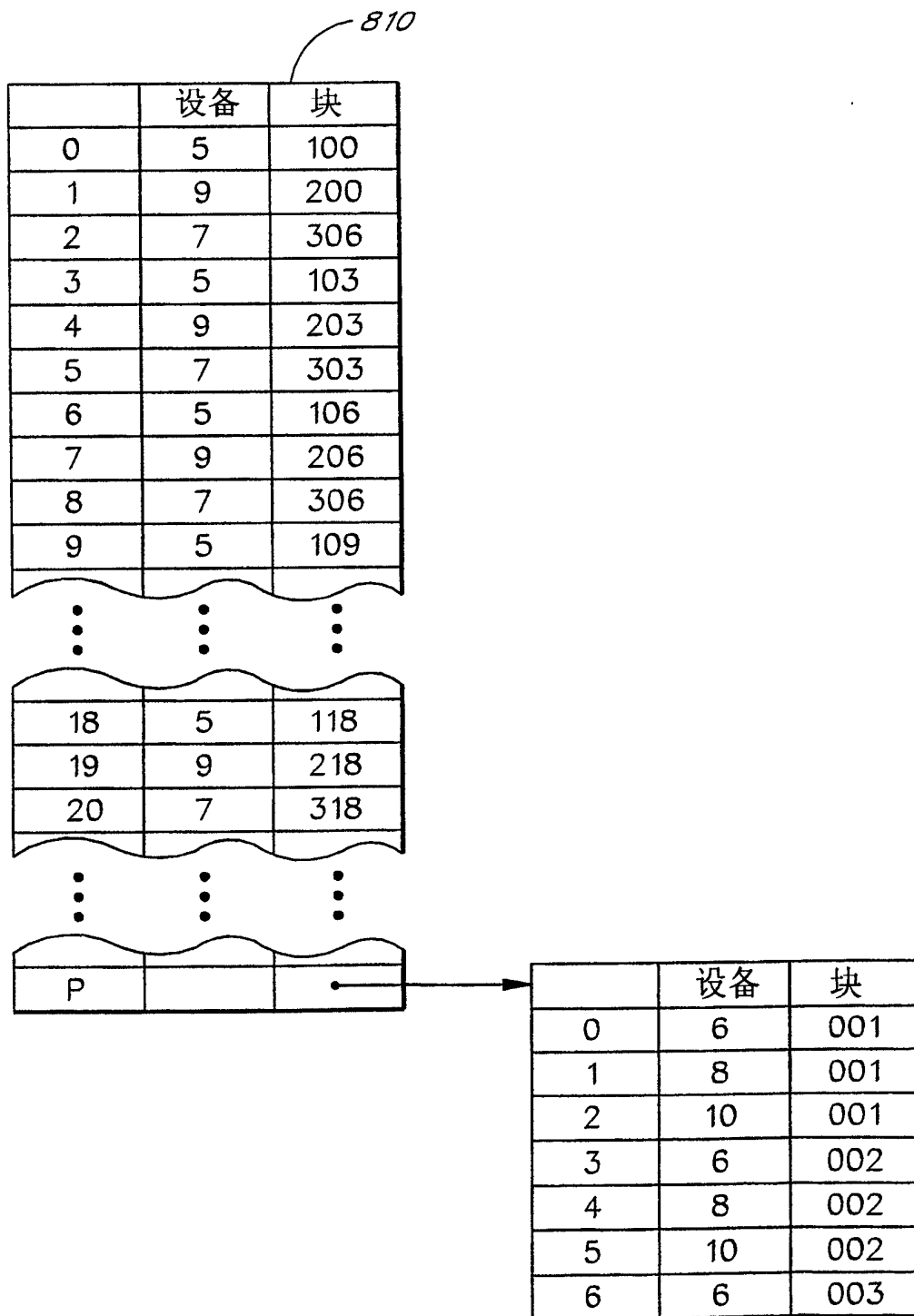


图 8C

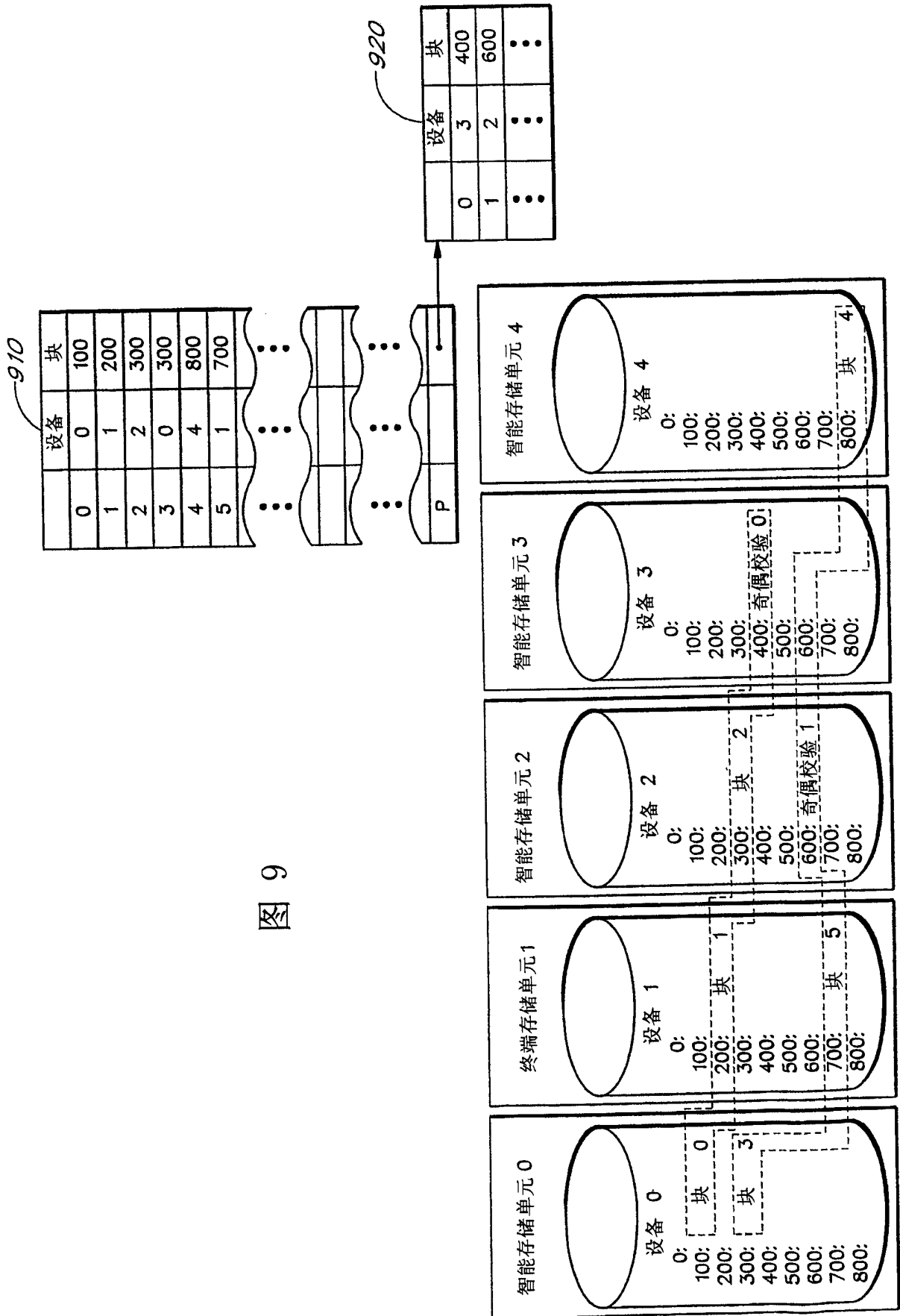


图 9

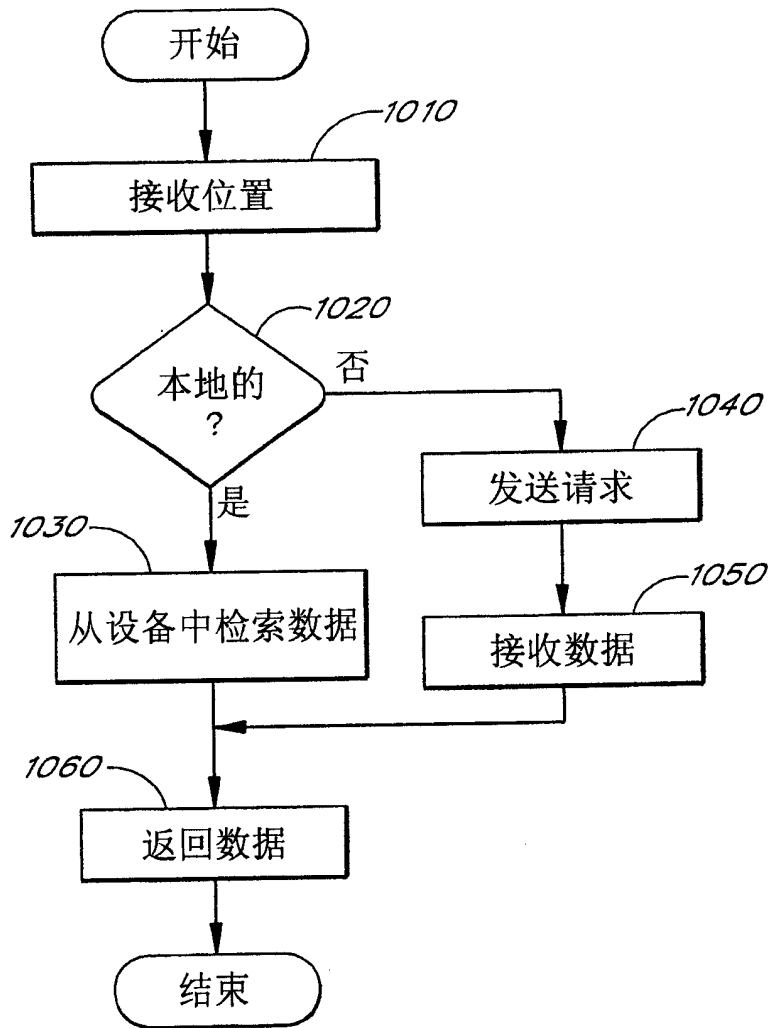


图 10

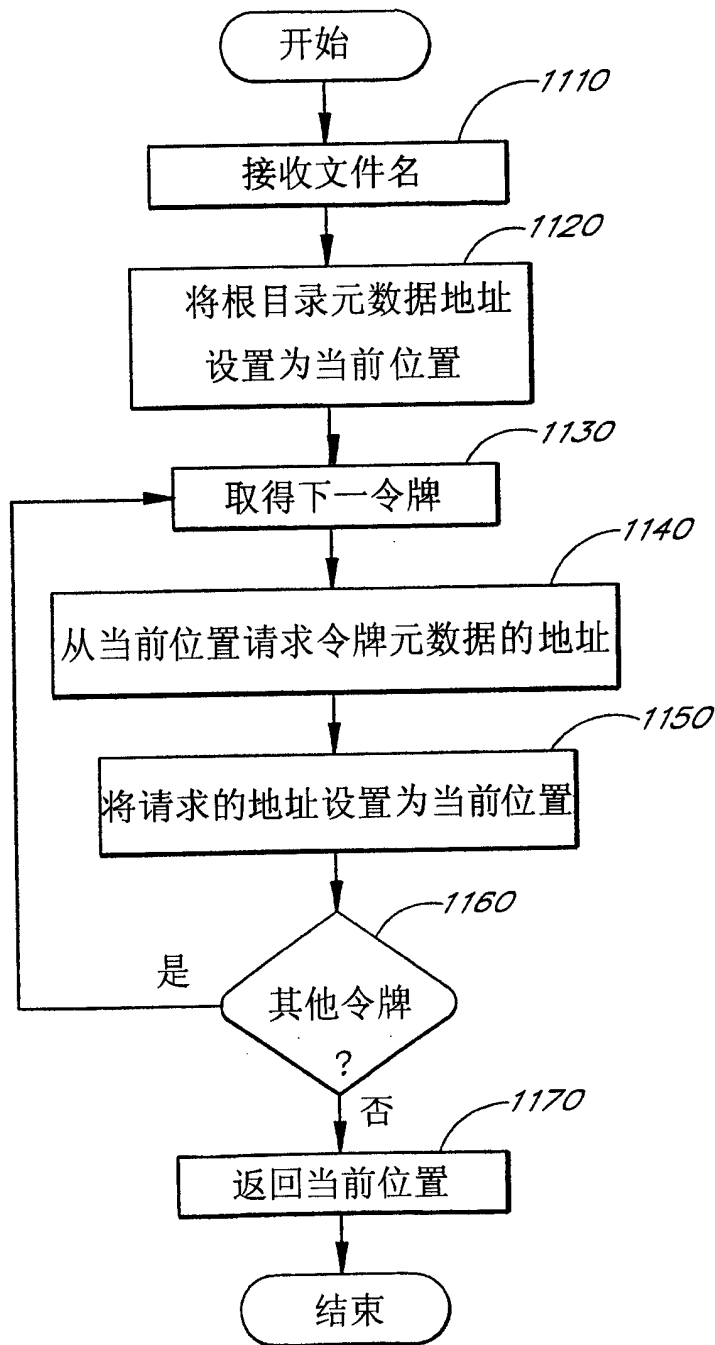


图 11

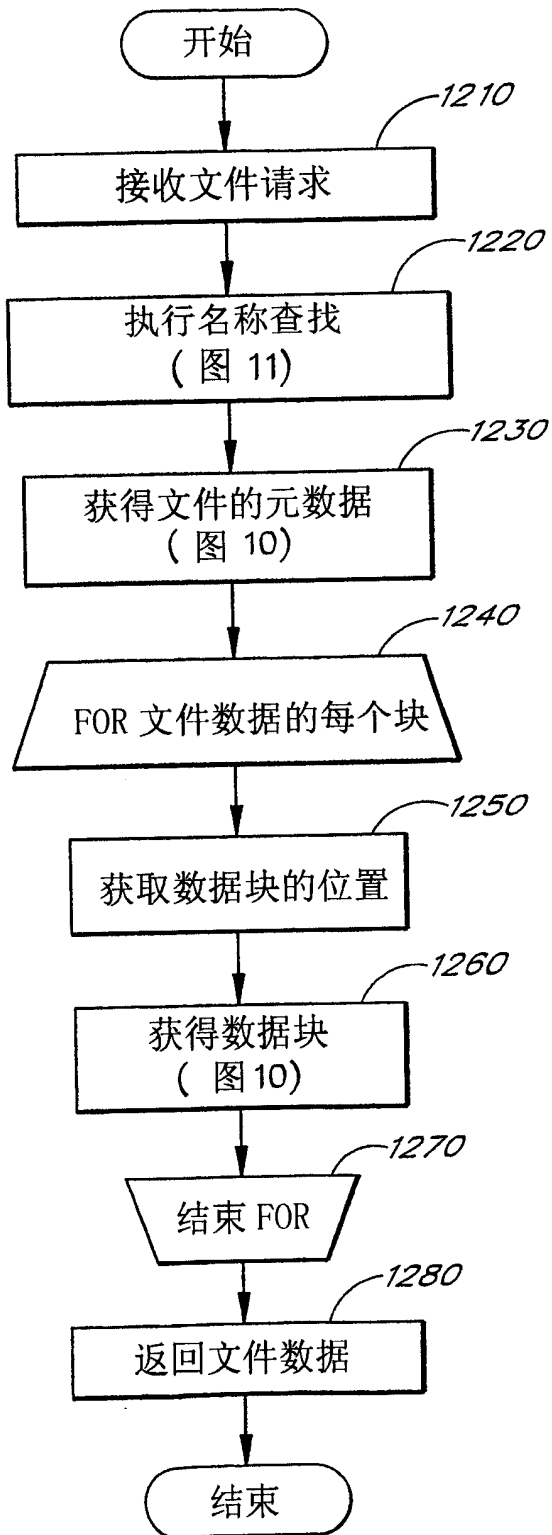


图 12

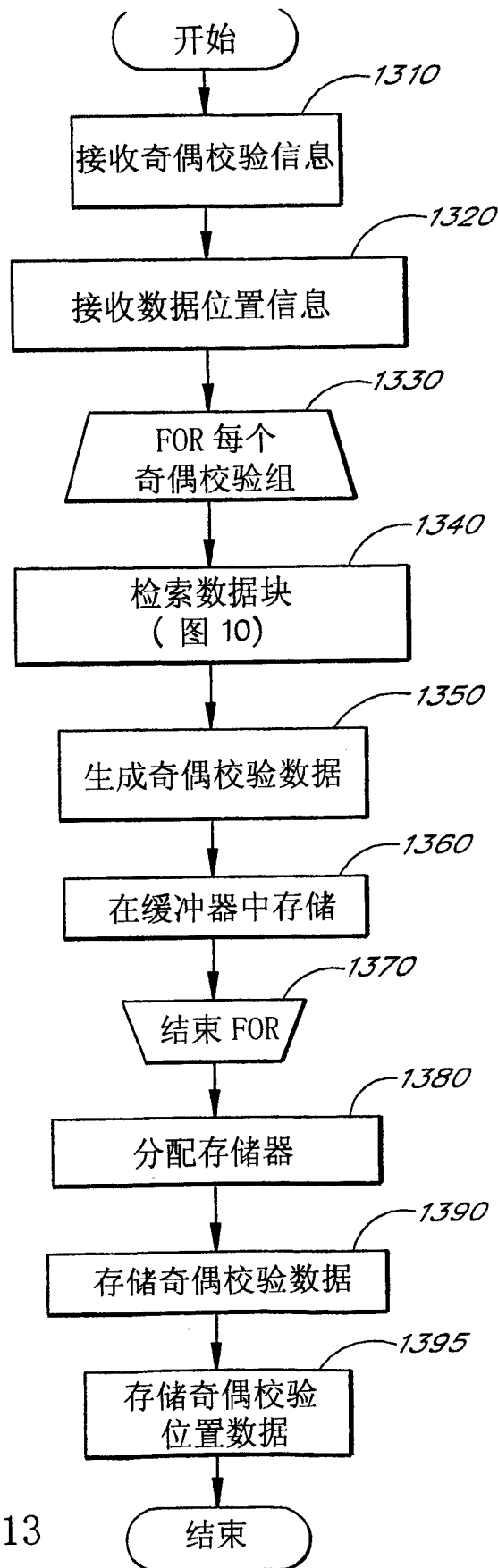


图 13



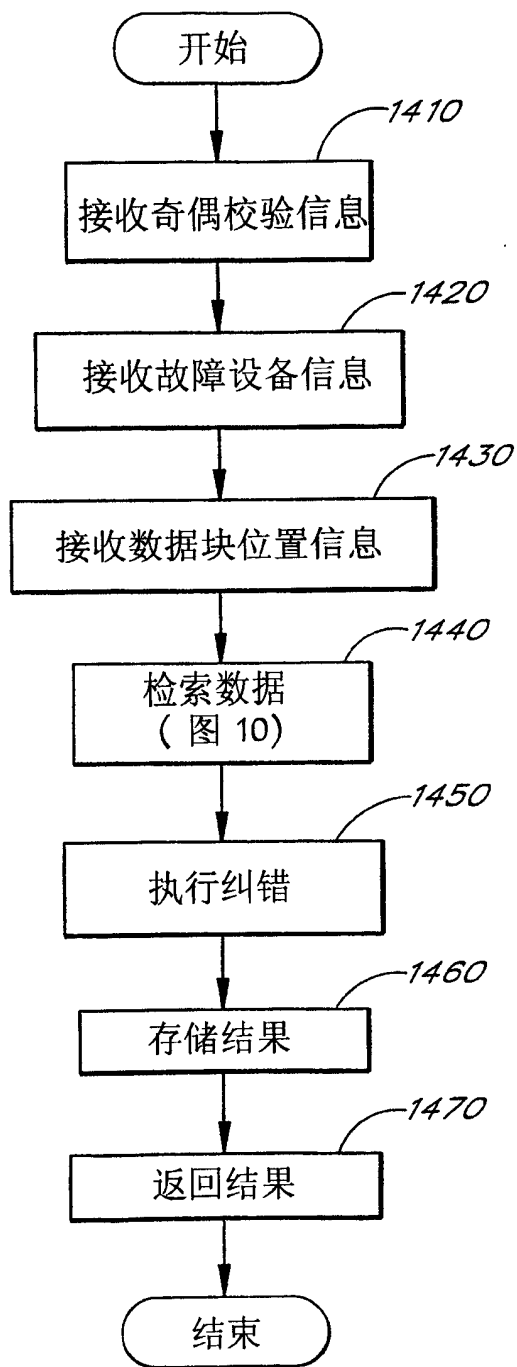


图 14