(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 102176198 A
(43)申请公布日 2011.09.07

(21)申请号 201110076346.6

(22)申请日 2009.06.25

(30)优先权数据
61/133,534 2008.06.30 US

(62)分案原申请数据
200980125013.9 2009.06.25

(71)申请人 枢轴3公司
地址 美国得克萨斯

(72)发明人 W·C·加罗威 R·A·卡里森

(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)

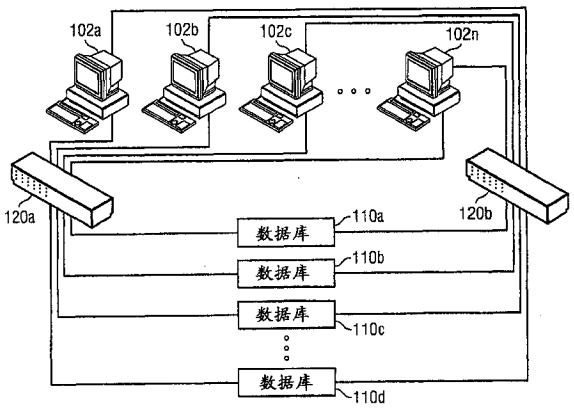
权利要求书 2 页 说明书 8 页 附图 4 页

(54)发明名称

用于结合RAID执行应用的方法和系统

(57)摘要

公开了允许在实施RAID系统的同一组计算设备上执行可利用该RAID系统的各种不同的应用（或其它类型的应用）的系统和方法。更具体地，在某些实施例中，可以在数据库上执行虚拟化层。通过使用这个虚拟化层，可以执行一组期望的应用程序，其中对于在虚拟化层上执行的应用的每个实例的上下文可被存储在卷中，被保持来利用RAID系统。



CN 102176198 A

1. 一种用于实施 RAID 的系统，所述系统包括：

耦合到一组主机的计算装置，该计算装置包括：

处理器；

数据存储器；

包括可执行指令的计算机可读的介质，所述可执行指令用于：

在计算装置上执行应用程序，其中应用程序是虚拟机；以及

实施与计算装置相关联的 RAID 应用，所述 RAID 应用用于：

接收对应于卷段的命令，其中由应用程序发布该命令，并且卷和对应于结合卷实施的 RAID 级别的冗余性数据被与计算装置相关联地存储；以及

针对与计算装置相关联地存储的第一段，执行第一命令。

2. 根据权利要求 1 的系统，其中执行应用程序包括使用计算装置上的虚拟化层，执行虚拟机。

3. 根据权利要求 2 的系统，其中虚拟机对应于第二卷，所述第二卷与计算装置相关联地存储。

4. 根据权利要求 3 的系统，其中计算机可读的介质在 RAID 控制器上实施。

5. 根据权利要求 4 的系统，其中 RAID 控制器用于管理虚拟机。

6. 一种用于在具有处理器的计算装置上结合各种应用程序实施 RAID 的方法，所述方法包括：

在计算装置上建立第一卷，其中第一卷包括段组，所述段组与计算装置相关联地存储；

建立与计算装置相关联的第二卷，其中第二卷对应于存储为虚拟机的应用程序；

在计算装置上执行应用程序；

结合卷实施 RAID 级别，其中实施 RAID 级别包括存储冗余性段组；

接收对应于卷的段组的第一段的命令，其中从在计算装置上执行的应用程序接收命令；

针对第一段执行第一命令，其中第一命令在计算装置上执行。

7. 根据权利要求 6 的系统，其中执行应用程序包括利用在计算装置上的虚拟化层，执行虚拟机。

8. 根据权利要求 7 的系统，其中虚拟机对应于第二卷，所述第二卷与计算装置相关联地存储。

9. 根据权利要求 8 的系统，其中在 RAID 控制器上实施所述方法。

10. 根据权利要求 9 的系统，其中 RAID 控制器用于管理虚拟机。

11. 一种包含计算机可执行指令的计算机可读的介质，所述计算机可执行指令用于结合各种应用程序实施 RAID，计算机指令可执行用于：

在计算装置上建立第一卷，其中第一卷包括段组，所述段组与计算装置相关联地存储；

在计算装置上建立第二卷，其中第二卷对应于存储为虚拟机的应用程序；

在计算装置上执行应用程序；

结合卷实施 RAID 级别，其中实施 RAID 级别包括存储冗余性段组；

接收对应于卷的段组的第一段的命令，其中从在计算装置上执行的应用程序接收命令；

相对于与计算装置相关联地存储的第一段执行第一命令。

12. 根据权利要求 11 的计算机可读的介质，其中执行应用程序包括利用在至少一个计算装置上的虚拟化层，执行虚拟机。

13. 根据权利要求 12 的计算机可读的介质，其中第二卷被存储在计算装置上。

14. 根据权利要求 13 的计算机可读的介质，其中在 RAID 控制器实施所述方法。

15. 根据权利要求 14 的计算机可读的介质，其中 RAID 控制器用于管理虚拟机。

用于结合 RAID 执行应用的方法和系统

[0001] 本申请是 2010 年 12 月 30 日提交的，申请号为 200980125013.9，题目为“用于结合分布式 RAID 执行应用的方法和系统”的分案申请。

技术领域

[0002] 本发明总的涉及存储设备的使用。更具体地，本发明的实施例涉及在存储设备上实施 RAID 和可以利用这个 RAID 功能的应用。甚至更具体地，本发明的某些实施例涉及在同一组计算设备上实施分布式 RAID 和一个或多个应用。

背景技术

[0003] 数据代表许多实体的重要的资产。因此，数据丢失，不管是偶然的还是由于恶意的活动造成的，都会在人力浪费、来自客户的信誉的丢失、时间的损失和潜在的法律义务方面代价昂贵。为了确保用于商业、法律的或其它目的的数据的适当的保护，许多实体可能希望通过使用各种各样的技术，包括数据存储、冗余性、保密性等等，来保护它们的数据。然而，这些技术可能与由被使用来处理或存储这个数据的计算设备的状态或配置所施加的其它竞争约束条件或要求冲突。

[0004] 用于处理这些紧张状态的一个方法是实施冗余磁盘阵列 (RAID)。通常，RAID 系统划分和复制在多个硬盘驱动器（或其它类型的存储介质）上的数据，统称为阵列，以增加可靠性，以及在某些情形下通过使用这些用于存储的 RAID 系统而提高计算设备（被称为主机）的吞吐量。然后对于主机，RAID 阵列可以呈现为一个或多个单片存储区域。当主机希望与 RAID 系统通信（读出、写入等等）时，主机就好像 RAID 阵列是单个盘那样通信。RAID 系统又处理这些通信，以结合这样的通信实施某个 RAID 级别。这些 RAID 级别可被设计成达到在各种各样的折衷，诸如可靠度、容量、速度等等之间的某个期望的平衡。例如，RAID（级别）0 把数据分布在几个盘上，以使得它给出提高的速度和几乎利用盘的全部容量，但如果盘发生故障，则在盘上的所有的数据将丢失；RAID（级别）1 使用两个盘（或更多的盘），每个盘存储相同的数据，以使得只要一个盘不出问题数据就不丢失。阵列的全部容量基本上是单个盘的容量，以及 RAID（级别）5 组合三个或更多的盘，以使得它保护数据免得遭受何一个盘的丢失；阵列的存储容量被减小一个盘。

[0005] 在许多情形下，在给定现代计算设备的计算功率后，其在实施利用 RAID 系统的主机的计算设备与实施 RAID 系统本身的计算设备之间可能存在一定的冗余量。除了物理部件的冗余性以外，主机和 RAID 系统在运行期间也会消耗许多相同的资源。因为二者都可能需要功率、冷却、机架空间等等。而且，因为主机和 RAID 系统沿单独的路径通信以便实施 RAID 系统，可能需要利用某些网络部件和路径。这种情形会引起许多不希望的问题，包括增加的硬件花费、通信瓶颈、需要大量物理空间来包含主机和 RAID 系统等等。

[0006] 因此，希望大大地改善这些问题。

发明内容

[0007] 给出了允许在实施 RAID 系统的同一组计算设备上执行各种不同的应用的系统和方法的实施例。具体地，在一个实施例中，为了允许在同一组计算设备上结合其它应用执行 RAID 应用，可以在数据库上执行虚拟化层。通过使用这个虚拟化层，可以执行一组期望的应用程序，其中对于在虚拟化层上执行的应用的每个实例的上下文可被存储在卷 (volume) 中，被保持来利用 RAID 系统。这些虚拟机（例如，被存储在卷中的应用和任何可应用的上下文信息）然后可以结合在数据库上的虚拟化层被执行。这样，可以利用一组计算设备来实施 RAID 系统和执行利用这样的分布式 RAID 系统的应用（这只是其中之一）。这些类型的应用，例如，可包括视频监视应用、游戏、零售或银行应用、视频流应用、内容操控应用等等。

[0008] 通过结合实施 RAID 系统的计算设备执行这些应用，可以得到许多优点。首先和最重要的，可以达到资源的联合，减小与物理资源和安装、配置、利用和保持这样的资源所需要的资源相关联的花费，如可以需要较少的空间、功率、冷却、备用部件等等。而且，在实施应用和分布式 RAID 系统时可以达到更大的速度，因为和这些应用以及分布式 RAID 系统的使用结合发生的通信可以较少或可以更快地进行。

[0009] 另外，故障容忍度水平可被引入到配置中，其中应用的实例作为虚拟机被保存，它们在实施分布式 RAID 系统的数据库的虚拟化层上被执行，正如在更详细地考察本申请的其余部分后将会看到的。概略地，这个故障容忍度水平可能起源于这样的事实，每个虚拟机（例如，执行可以在虚拟化层上被执行的应用的实例）被存储在分布式 RAID 系统的卷中，可能意味着，在数据库之一出现故障的情形下虚拟机可被恢复。另外，故障容忍度可被引入，因为虚拟机可以在任何数据库的虚拟化层上被执行，因此即使在单个数据库出现故障时，每个虚拟机仍旧可被执行。

[0010] 本发明的这些和其它方面，在结合以下的描述和附图被考虑时将更好地被意识到和理解。以下的描述，虽然指示本发明的各种实施例和本发明的许多具体的细节，但是作为说明而不是限制被给出的。在本发明的范围内可以作出许多替换、修改、添加、或重新安排，以及本发明包括所有的这样的替换、修改、添加、或重新安排。

附图说明

[0011] 伴随本技术说明书并作为技术说明书的一部分的附图被包括来描绘本发明的某些方面。通过参考在附图上显示的示例性和非限制性的实施例，本发明和本发明所提供的系统的部件和运行的更清晰的印象变得更为明显，其中相同的附图标记是指相同的部件。应当指出，在附图上显示的特性不一定按比例画出。

[0012] 图 1 是采用分布式 RAID 系统的结构体系的一个实施例的框图。

[0013] 图 2A 是数据库的一个实施例的框图。

[0014] 图 2B 是用于数据库的结构体系的一个实施例的框图。

[0015] 图 3 是数据库的一个实施例的流程图。

[0016] 图 4A 是用于数据库的结构体系的一个实施例的框图。

[0017] 图 4B 是用于数据库的结构体系的一个实施例的框图。

[0018] 图 4C 是用于数据库的结构体系的一个实施例的框图。

[0019] 图 5 是利用在其上可以执行其它应用的分布式 RAID 系统的结构体系的一个实施例的框图。

[0020] 图 6 是用于在分布式 RAID 系统上执行应用的一个实施例的流程图。

[0021] 图 7 是描绘结合分布式 RAID 系统的虚拟机的存储的一个实施例的框图。

具体实施方式

[0022] 本发明及其各种特性和有利的细节，将参照在附图上显示的和在以下的说明中详细阐述的非限制性实施例更全面地进行说明。熟知的开始资料、处理技术、部件和设备的描述被省略，以免用细节不必要地遮蔽本发明。然而，应当看到，详细说明和具体的例子，虽然指示本发明的优选实施例，但是作为说明而不是限制被给出的。在本发明概念的精神和 / 或范围内的各种替换、修改、添加、或重新安排，对于阅读本公开内容的本领域技术人员，将是显而易见。这里所讨论的实施例可以以适当的计算机可执行的指令被实施，这些指令可以放置在计算机可读的介质（例如，硬盘驱动器）、硬件电路等等，或它们的任何组合。

[0023] 在讨论具体的实施例之前，这里先描述用于实施某些实施例的硬件结构体系的实施例。一个实施例可包括与网络通信地耦合的一个或多个计算机。正如本领域技术人员所熟知的，计算机可包括中央处理单元（“CPU”）、至少一个只读存储器（“ROM”）、至少一个随机存取存储器（“RAM”）、至少一个硬盘驱动器（“HD”）、和一个或多个输入 / 输出（“I/O”）设备。I/O 设备可包括键盘、监视器、打印机、电子定向设备（诸如鼠标、跟踪球、指示笔等）等等。在各种实施例中，计算机通过网络接入到至少一个数据库。

[0024] ROM、RAM 和 HD 是用于存储由 CPU 可执行的、计算机可执行的指令的计算机存储器。在本公开内容内，术语“计算机可读的介质”并不限于 ROM、RAM、和 HD，而是可包括任何类型的、可以由处理器读出的数据存储介质。在某些实施例中，计算机可读的介质可以是指数据盒式磁带、数据备份磁带、软盘、快闪存储器驱动器、和光学数据存储驱动器、CD-ROM、ROM、RAM、HD 等等。

[0025] 这里描述的功能实体或处理过程的至少一部分可以以适当的计算机可执行的指令被实施。计算机可执行的指令可以作为软件代码部件或模块被存储在一个或多个计算机可读的介质上（诸如，非易失性存储器、易失性存储器、DASD 阵列、磁带、软盘、硬盘驱动器、光学存储设备等等、或任何其它适当的计算机可读的介质或存储设备）。在一个实施例中，计算机可执行的指令可包括成行的、汇编的 C++、Java、HTML、或任何其它编程或脚本代码。

[0026] 另外，所公开的实施例的功能可以在一个计算机上被实施，或在网络上在两个或更多个计算机之间被共享 / 分布。在实施本实施例的计算机之间的通信可以通过使用任何的电子的、光学的、射频的信号、或其它适当的方法和遵从已知的网络协议的通信工具而被完成。

[0027] 正如这里使用的，术语“包括”、“具有”、或它们的其它变例，意在覆盖非排他的包括。例如，包括一系列单元的过程、处理、物体、设备不一定仅仅限于那些单元，而是可包括没有明显列出的，或对于这样的过程、处理、物体、设备来说是固有的其它单元。而且，除非明示的相反表示，“或”是指“包括的或”，而不是“排他的或”。例如，条件 A 或 B 由以下的任一项被满足 :A 是正确（或存在）和 B 是错误（或不存在），A 是错误（或不存在）和 B 是正确（或存在），以及 A 和 B 都是正确（或存在）。

[0028] 另外，这里给出的例子或说明无论如何不看作为对于它们被利用的任何术语的约束、限制、极限或表达它的定义。而是，这些例子或说明被看作为是对于一个特定的实施例

描述的，以及仅仅作为说明性的。本领域技术人员将意识到，这些例子或说明利用的任何术语将包括在本说明书的此处或他处可能给出或没有给出的其他实施例，并且所有的这样的实施例意在被包括在该术语的范围内。指定这样的非限制性例子和说明的语言包括，但不限于：“例如”、“举例”、“在一个实施例中”。

[0029] 本申请涉及到由 Galloway 等在 2009 年 6 月 5 日提交的、题目为“Method and System for Distributed RAID Implementation”的美国专利申请 No.12/479,319；由 Galloway 等在 2009 年 6 月 5 日提交的、题目为“Method and System for Data Migration in a Distributed RAIDImplementation”的美国专利申请 No.12/479,360；由 Galloway 等在 2009 年 6 月 5 日提交的、题目为“Method and System for DistributingCommands to Targets”的美国专利申请 No.12/479,403；由 Galloway 等在 2009 年 6 月 5 日提交的、题目为“Method and System forInitializing Storage in a Storage System”的美国专利申请 No.12/479,377；由 Galloway 等在 2009 年 6 月 5 日提交的、题目为“Method and System for Rebuilding Data in a Distributed RAIDSystem”的美国专利申请 No.12/479,434；和由 Galloway 等在 2009 年 6 月 5 日提交的、题目为“Method and System for Placement of Data on aStorage Device”的美国专利申请 No.12/479,394；和由 Galloway 等提交的、题目为“Method and System for Protecting Against MultipleFailures in a RAID System”的专利申请 No.12/490,916；所有这些专利申请在此通过引用并入本文。

[0030] 现在，具体地对于数据存储的进行上下文的概略的讨论是有帮助的。正如以上讨论的，RAID 系统在多个硬盘驱动器（或其它类型的存储介质），统称为阵列上划分和复制数据，以便通过使用这些 RAID 系统用于存储而增加可靠度，和在某些情形下，提高计算设备（称为主机）的吞吐量。然而，RAID 的当前的实施方案可能有各种各样的问题。

[0031] 具体地，这些问题中的某些问题是来自于由这些 RAID 系统的结构体系所施加的限制，诸如以下的事实：在许多实例中与 RAID 系统的所有的通信必须寻址到控制和管理 RAID 系统的单个服务器。这个结构体系可以导致在包括利用 RAID 系统的主机和被使用来实施 RAID 系统的计算设备的物理部件中的冗余性。除了物理部件的冗余性以外，主机和 RAID 系统在运行时还会消耗许多相同的资源。因为主机和 RAID 系统都可能需要功率、冷却、机架空间等等。这种情形会引起许多不希望的问题，包括增加的硬件花费、通信瓶颈、需要大量物理空间来包含主机和 RAID 系统等等。

[0032] 针对特定的 RAID 系统描述本发明的某些实施例将是有帮助的，然而，应当指出，通过其描述某些实施例的特定的 RAID 系统没有对于本发明的其它实施例的可应用性或使用施加限制，以及这样的实施例可以在包括其它类型的 RAID 系统或一起包括其它的存储系统的任何的各种各样的上下文中被有用地利用。

[0033] 通过以上所述的，某些实施例可以相对于分布式 RAID 系统被有用地描述，其中具有相关的 RAID 级别的卷可以通过使用分布式 RAID 系统被创建。然后每个分布式 RAID 应用可以协调与该卷的数据相关联的操作，以使得与该卷相关联的数据或结合该卷的期望的 RAID 级别的实施可被存储在分布式 RAID 系统的多个数据库。通过利用在多个数据库的每个上执行的类似的分布式 RAID 应用，将卷的数据以及和 RAID 的实施相关联的数据存储在该多个数据库中，结合卷来协调 RAID 级别的实施，来实现多个优点。即，不同的存储卷可以被分配，一个或多个卷结合不同的 RAID 级别被实施。而且，由于在数据库上存储的协调和

RAID 的实施是通过使用基本上相同的分布式 RAID 应用而完成的，在许多情形下，可以利用标准的或现成的硬件，诸如基于标准 x86 的服务器和存储介质。通过利用这里给出的实施例或其它实施例，也可以实现许多其它优点，以及在阅读本公开内容后，将认识到这样的优点，这些优点可以或可以没有以具体的细节被指出。

[0034] 现在转到图 1，图上显示利用分布式 RAID 系统的一个实施例的系统的结构体系的框图。分布式 RAID 系统 100 包括一组数据库 110，每个数据库 110 通信地耦合到两个交换机 120。每个交换机 120 还通信地耦合到每个主机 102，以使得主机 102 可以通过对应于特定的数据库 110 的一组路径与每个数据库 110 通信，每条路径包括一个交换机 120。

[0035] 在数据库 110、交换机 120 和主机 102 之间的通信耦合可以通过使用几乎任何期望的输送介质（有线或无线），包括以太网、SCSI、iSCSI、光纤信道、串行附属 SCSI（“SAS”）、先进技术附件（“ATA”）、串行 ATA（“SATA”）、或在技术上已知的其它协议，而被完成。而且，通信耦合可以结合诸如互联网、LAN、WAN、无线网络或在技术上已知的任何其它通信网络那样的通信网络被实施。

[0036] 在一个实施例中，然后，通过使用诸如 iSCSI、SCSI 等命令协议，主机 102 可以与数据库 110 通信，以便操控数据。更具体地，每个数据库 110 包括存储介质（正如将在后面更详细地说明的）。总起来说，在数据库 110 中的存储介质可被虚拟化，并呈现给主机 102 作为一个或多个相邻的存储块，存储设备等等。例如，当利用 iSCSI 协议时，在数据库 110 中的存储介质可被呈现给主机 102 作为 SCSI 目标，在一个实施例中，其具有多个端口。

[0037] 因此，在运行期间，在一个实施例中，主机 102（或在主机 102 处或与数据库 110 接口的用户）可以请求创建卷，并规定结合该卷被实施的 RAID 的级别。与该卷相关联的数据和与该卷相关联的期望的级别 RAID 的实施被存储在数据库 110。主机 102 然后可以通过使用对应于该卷或该卷的一部分的逻辑地址而存取这个卷。这样，主机 102 可以利用所创建的存储的卷，以及可以结合这些卷达到对于主机 102 来说基本上觉察不到的故障容忍度。

[0038] 通过参考图 2A 可以更好地理解存储的虚拟化和利用数据库 110 的 RAID 的实施，在图 2A 上显示用来实施分布式 RAID 的数据库 110 的计算机的一个实施例的框图。这里，数据库 110 包括数据存储器 (store)250 和用来执行被存储在计算机可读的介质中的指令的处理器 202，其中指令用来实施分布式 RAID 应用 210。分布式 RAID 应用 210 可以周期地发布心跳通信到其它的数据库 110 上的分布式 RAID 应用 210，以确定对于该数据库 110 是否有故障。如果分布式 RAID 应用 210 确定另一个数据库 110 有故障，则它可以设置对应于该数据库 110 的一个或多个故障标记。通过使用这些故障标记，对于每个数据库 110 上的每个分布式 RAID 应用 210，特定的分布式 RAID 应用 210 可以确定某个数据库 110 是否有故障。

[0039] 分布式 RAID 应用 210 还可以存取（例如，读出、写入、发布命令等等）包括一个或多个存储介质的数据存储器 250，该存储介质例如可以是按照几乎任何已知的协议运行的盘 252，诸如 SATA、PATA、FC 等等。其中每个盘 252 可以，或可能不一定，具有相等的大小。在每个数据库 110 上执行的分布式 RAID 应用 210 可以允许通过使用在数据库 110 上的数据存储器 250 而分配和使用卷，以及通过利用在数据库 110 之间共享的一组全局表格 240、一组本地表格 245 和写高速缓存器 260 而结合这些卷实施 RAID，所有这些表格可被存储在存储器 230（它可以是数据存储器 250 或在一起的另外的存储器）。

[0040] 图 2B 显示可被使用来实施用来实施分布式 RAID 的数据库 110 计算机的硬件结构体系的一个实施例的框图。在这个结构体系例子中，数据库 110 包括一个或多个处理器 202，其可以一起附着到 Intel x86 结构体系或某个其它结构体系，以及存储器 230，通过总线被耦合到 I/O 控制器中心 212，其在一个实施例中可以是南桥芯片等等。I/O 控制器中心 212 又可以被耦合到和控制诸如 PCI-X 总线、串行总线等等那样的总线 272。被耦合到这个总线 272 的是一个或多个盘控制器 262，诸如，例如 LSI 1068 SATA/SAS 控制器。每个这些盘控制器 262 被耦合到一个或多个盘 252，其中总起来说，这些盘 252 可包括数据存储器 250。另外，一个或多个网络接口 282 也可以被耦合到总线 272。这些网络接口 282 可以是被包括在主板上的网络接口（诸如，以太网等等），它可包括被配置成经由诸如以太网、光纤信道等等那样的一个或多个协议进行接口的一个或多个网络接口卡，或这些网络接口 282 可以是某些其它类型的网络接口，以使得数据库 110 可以通过这些网络接口 282 与交换机 120 通信。

[0041] 正好，在某些实施例中，被使用来实施数据库 110 的计算设备的部件可具有（或可被构建成具有）比用于执行分布式 RAID 应用 210 所需要的计算功率更多的计算功率。在许多情形下，然后，被使用来实施数据库 110 的计算设备（例如，处理器、高速缓存器、存储器、电路板等等）当被使用来仅仅执行分布式 RAID 应用 210 时可能具有未利用的计算功率。然后，所期望的是，利用这个过量的计算功率来执行将在主机 102 上执行并利用分布式 RAID 应用 210 的应用。然而，这可能是成问题的，因为在许多情形下，可能有各种各样的这样的应用在不同的主机 102 上执行，其中每个主机可以执行不同的操作系统，应用可被配置成仅仅在一种类型的操作系统上执行，等等。

[0042] 所以，现在注意点被引导到允许可利用分布式 RAID 系统的各种应用（或其它类型的应用）在实施该分布式 RAID 系统的同一组计算设备上执行的本发明的系统和方法。为了允许在同一组计算设备上结合其它应用执行分布式 RAID 应用，可以在数据库上执行虚拟化层，通过使用这个虚拟化层可以执行一组期望的应用程序，其中对于在虚拟化层上执行的应用的每个实例的上下文可被存储在卷中，被保持来利用分布式 RAID 系统。这些虚拟机（例如，被存储在卷中的应用和任何可应用的上下文信息）然后可以结合在任一个数据库上的虚拟化层执行。这样，可以利用一组计算设备来实施分布式 RAID 系统和执行利用这样的分布式 RAID 系统（这只是其中之一）的应用。这些类型的应用，例如，可包括视频监视应用、游戏、零售业或银行应用、视频流应用、内容操控应用等等。

[0043] 通过结合实施分布式 RAID 系统的计算设备执行这些应用，可以得到许多优点。首先和最重要的，可以实现物理资源的联合，减小与物理资源和安装、配置、利用和保持这样的资源所需要的资源相关联的花费，因为可能需要较少的空间、功率、冷却、备用部件等等。而且，在实施应用和分布式 RAID 系统时可以达到更大的速度，因为结合应用和分布式 RAID 系统的使用发生的通信，可以较少或可以更快地进行。

[0044] 另外，故障容忍度水平可被引入到配置中，其中应用的实例作为虚拟机被保存，它们在实施分布式 RAID 系统的数据库的虚拟化层上被执行，正如在更详细地考察本申请的其余部分后将会看到的。概略地，这个故障容忍度水平可以来源于这样的事实，每个虚拟机（例如，执行可以在虚拟化层上被执行的应用的实例）被存储在分布式 RAID 系统的卷中，意味着，在数据库之一出现故障的情形下虚拟机可被恢复。另外，故障容忍度可被引入，因为

虚拟机可以在任何数据库的虚拟化层上执行，因此即使在单个数据库出现故障时，每个虚拟机仍旧可被执行。

[0045] 现在参照图 3，图上显示可以实施分布式 RAID 系统和允许执行各种不同的应用的数据库的一个实施例。这里，数据库 1110 包括数据存储器 350 和用来执行存储在计算机可读的介质上的指令的处理器（未示出）或其它硬件。这个硬件，例如，可以是 x86 平台等等。

[0046] 被存储在计算机可读的介质上的指令可以用来实施虚拟化层 312 和分布式 RAID 应用 310。虚拟化层 312 可以是，例如由 SunMicrosystems 的容器、Linux KVM、Linux Vserver、Oracle VM、Virtual PC、Microsoft 的虚拟服务器、IBM 的 PowerVM、SunMicrosystems 的逻辑域、VMware 服务器等等，或任何其它类型的虚拟化或模仿机应用，正如在技术上已知的。分布式 RAID 应用 310 可以具有类似于上述功能的功能。为了帮助在存储器 330 中一个或多个这些表格 340 的实施，可以跟踪诸如由分布式 RAID 系统所存储的哪些卷是虚拟机和哪个数据库 110 被分配给特定的虚拟机那样的信息。

[0047] 然后概略地参考图 4A、4B 和 4C，图上显示具有虚拟化层的数据库 1010 的结构体系的三个实施例。将会指出，这样的结构体系的其它实施例也是可能的，它们可以被利用。在图 4A 上，虚拟化层 312 可以存在于硬件层 402 上，这样，分布式 RAID 应用 310 和虚拟机 430（应用和它们的对应的上下文）可以在虚拟化层 312 上执行。图 4B 显示一个实施例，其中操作系统 420（例如，Windows、Solaris、MacOS 等）可以在具有在操作系统 420 上执行的虚拟化层 312 的硬件层 402 上执行，这样，分布式 RAID 应用 310 和虚拟机 430（应用和它们的对应的上下文）可以在虚拟化层 312 上执行。图 4C 显示一个实施例，其中操作系统 420 可以在硬件层 402 上执行。分布式 RAID 应用 310 和虚拟化层 312 可以在操作系统 420 上执行。虚拟机 430 然后可以在虚拟化层 312 上执行。

[0048] 因此，在图 5 中描述了可以用于通过使用一组数据库 1110 来实施分布式 RAID 系统并执行可以利用在一个或多个相同的数据库 1110 上的该分布式 RAID 系统的一个或多个应用的系统的一个实施例。

[0049] 移到图 6，图上显示用于在也正在执行分布式 RAID 应用 310 的数据库 1110 上执行一个或多个应用的方法的一个实施例。在步骤 610，可以在数据库 1110 的虚拟化层 312 上执行应用。这个安装过程可以在步骤 620 创建虚拟机，用来当被虚拟化层 312 执行时执行与特定的上下文相关联的应用。更具体地，在一个实施例中，可以通过使用分布式 RAID 应用 310 来创建卷，其中应用可以连同从在虚拟化层 312 上在一定的时间量内执行该应用所得到的上下文一起被存储在该卷中。

[0050] 这被更清晰地显示于图 7，在图上显示每个虚拟机 430 可以如何成为被存储在每个数据库 1110 的数据存储器 350 中的卷。每个虚拟机可以对应于应用和相关联的上下文的一个安装的实例，这样，虚拟机 430 可以在虚拟化层 312 上执行，以便运行具有对应的上下文的应用。

[0051] 回到图 6，一旦应用被安装和虚拟机被创建，在一个实施例中，在步骤 630，包括应用的实例的虚拟机就可被指定在特定的数据库 1110 上执行，因此虚拟机可以在步骤 640 在所指定的数据库 1110 的虚拟化层 312 上执行。

[0052] 在任何时候，如果因为任何原因期望停止这个虚拟机的执行，则虚拟机可以被存储，以使得虚拟机的执行可以在以后由该虚拟机被分配到的那个特定的数据库 1110 从相

同的点继续进行。这个存储可能使得把上下文存储在虚拟机（卷对应于虚拟机），如果需要的话。

[0053] 将会指出，因为每个虚拟机 430 被存储在包括分布式 RAID 系统的数据库 1110，该虚拟机被分配到的那个数据库 1110 可以存取和执行每个虚拟机。另外，注意，RAID 的级别可以结合该虚拟机的存储被实施，因为该虚拟机可被存储在与分布式 RAID 应用 310 相关联的卷中。因此，正如先前讨论的，通过使用作为被分布式 RAID 应用 310 管理的卷被存储的虚拟机而在数据库 1110 上执行应用，在分布式 RAID 系统的结构体系中可以固有地实现故障容忍度。

[0054] 更具体地，在一个实施例中，因为每个数据库 1110 可以存取任何虚拟机以及每个虚拟机 430 可以结合 RAID 级别被存储，即使在特定的数据库 1110 出现故障时，应用仍旧能够被执行，因为关于任何虚拟机的所有的数据仍旧可以由剩余的工作的数据库 1110 存取。事实上，在一个实施例中，如果特定的数据库 1110 有故障，则有故障的数据库 1110 可以被检测，以及被分配给有故障的数据库 110 的每个 VM 可以基本上自动地被分配给另一个工作的数据库 1110，这样，在有故障的数据库 1110 上执行的所有的 VM 现在将在另一个数据库 1100 上执行。这个分配可以随机地或基于一个或多个准则完成，诸如，被分配给每个数据库 1110 的 VM 的数目等等。通过自动地重新分配有故障的数据库 1110 的 VM，除了给予应用以某个水平的故障容忍度以外，停机时间可被最小化。

[0055] 在以上的技术说明书中，本发明是参照具体的实施例描述的。然而，本领域技术人员将会看到，可以作出各种不同的修改和改变，而不背离如在下文所附权利要求中阐述的本发明的范围。因此，技术说明书和附图应当为说明的意义而不是限制的意义，以及所有的这样的修改意在被包括在本发明的范围内。

[0056] 在上面对于具体的实施例描述了好处、其它优点和对于问题的解决方案。然而，好处、其它优点和对于问题的解决方案，以及可以使得任何好处、优点或解决方案发生或变为更明确的任何部件不被看作为任何或所有的权利要求的关键的、需要的、或本质的特性或部件。

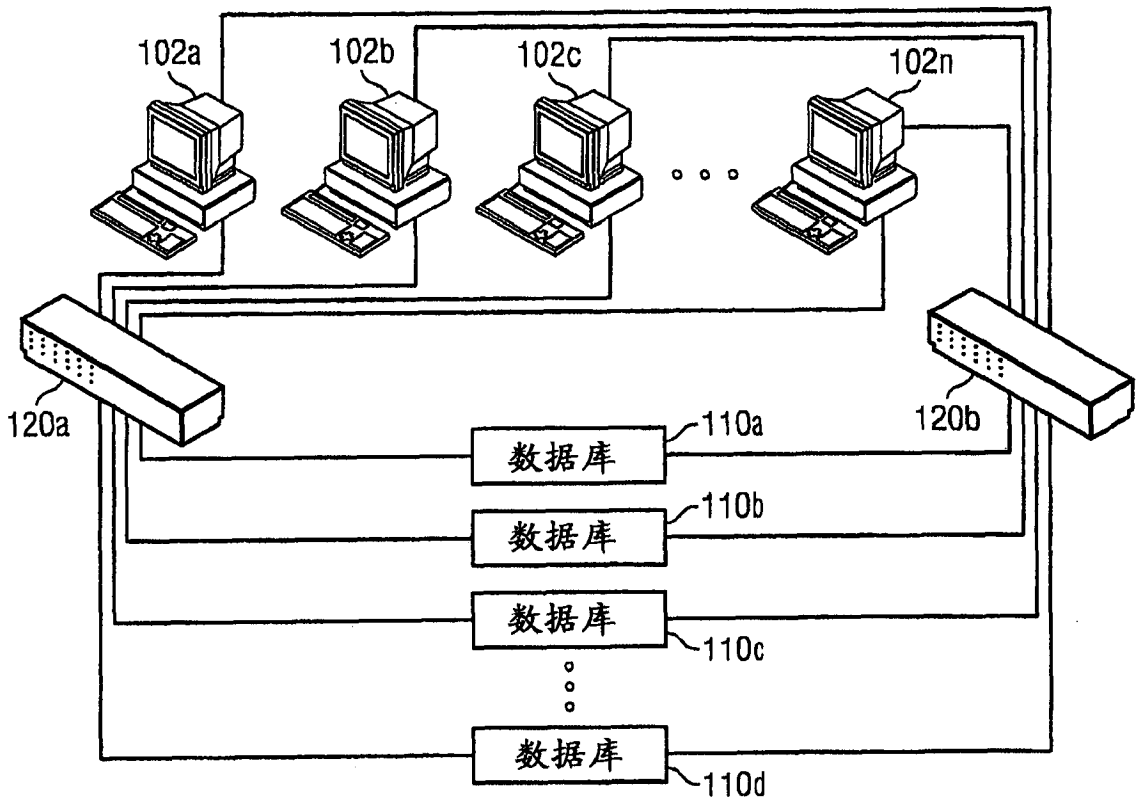

102a
102b
102c
102n
120a
120b
数据库
110a
数据库
110b
数据库
110c
数据库
110d

图 1

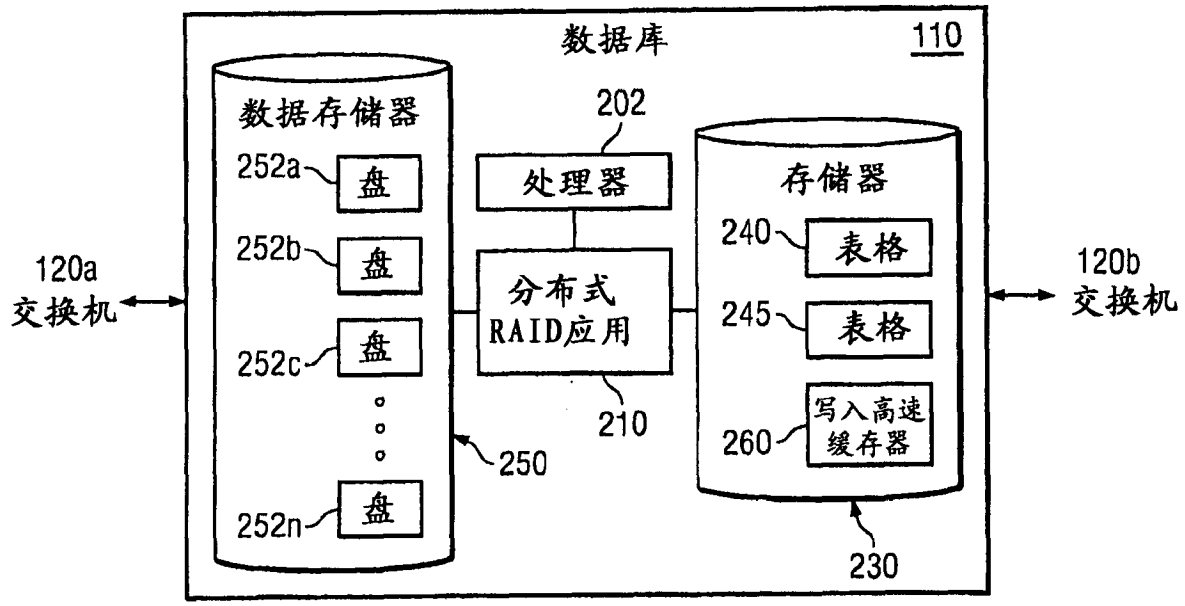

数据库
110
数据存储器
252a
盘
252b
盘
252c
盘
252n
盘
250
202
处理器
分布式
RAID应用
210
存储器
240
表格
245
表格
260
写入高速
缓存器
230
120a
交换机
120b
交换机

图 2A

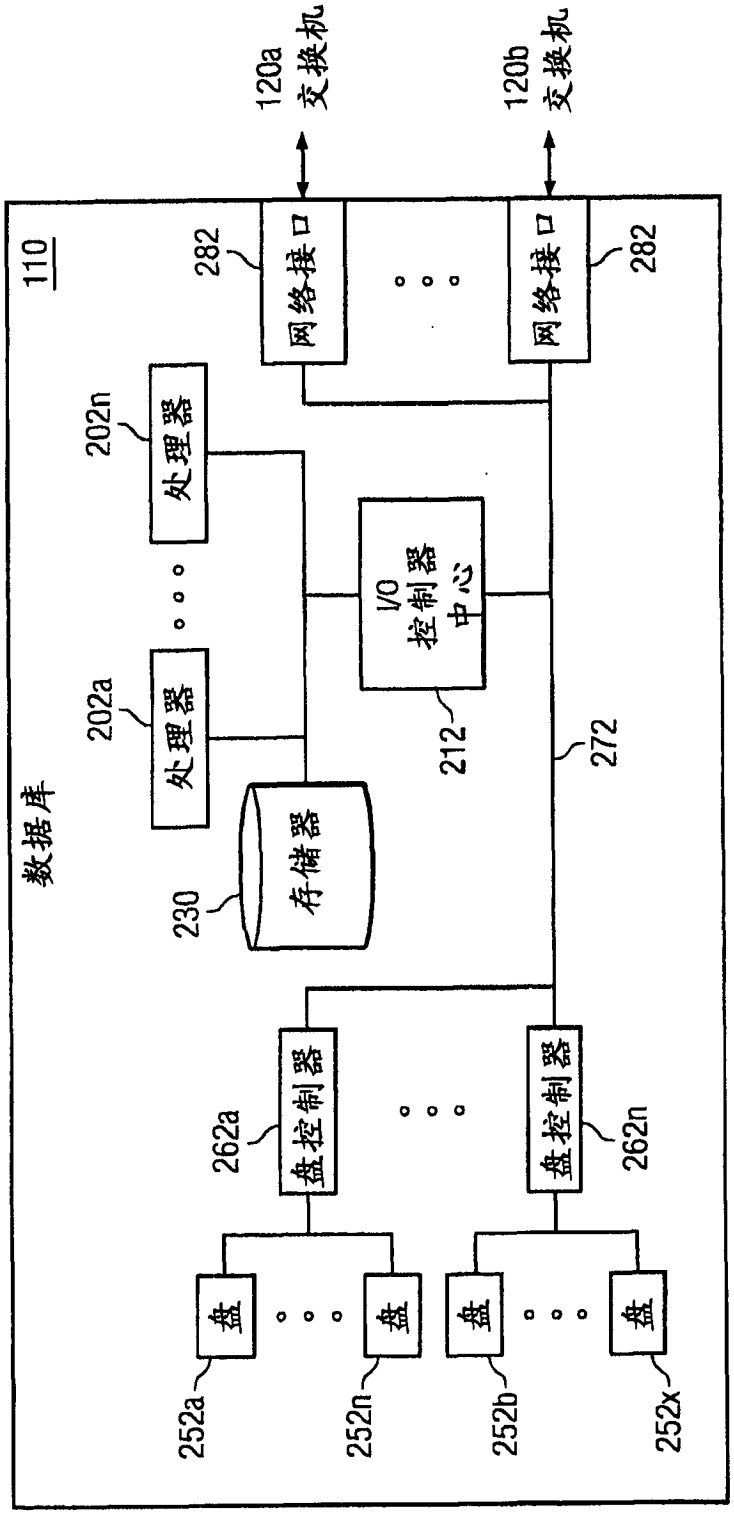

120a
交换机
120b
交换机
110
282
网络接口
282
网络接口
202n
处理器
202a
处理器
I/O
控制器
中心
212
272
数据库
230
存储器
262a
盘控制器
262n
盘控制器
盘
盘
盘
盘
252a
252n
252b
252x

图 2B

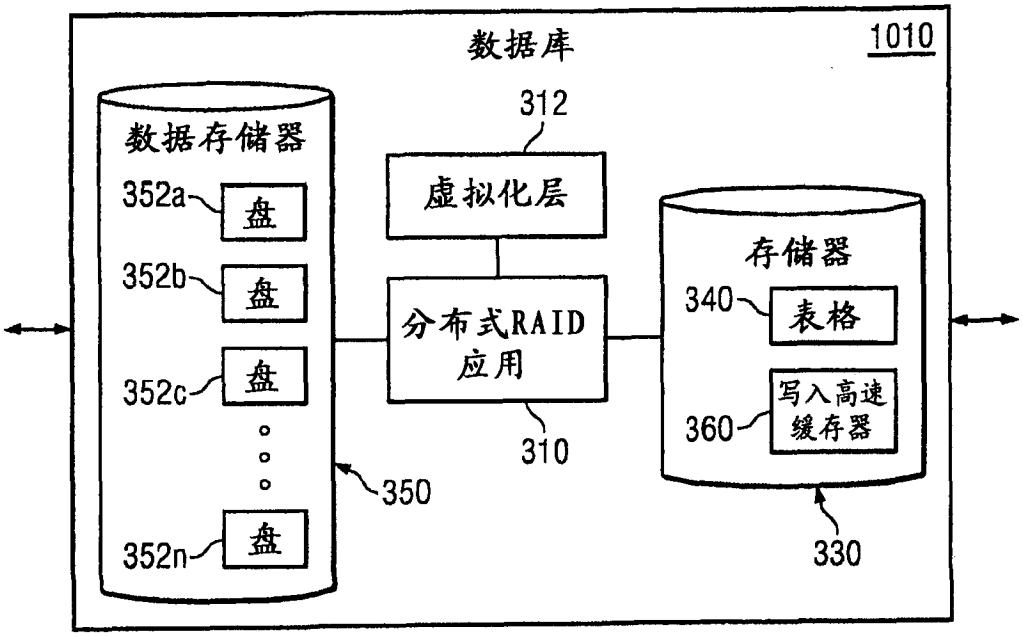


图 3

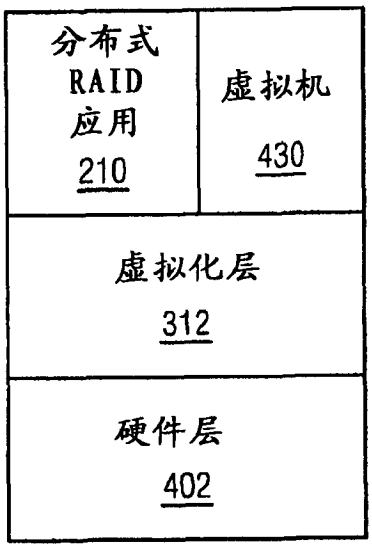


图 4A

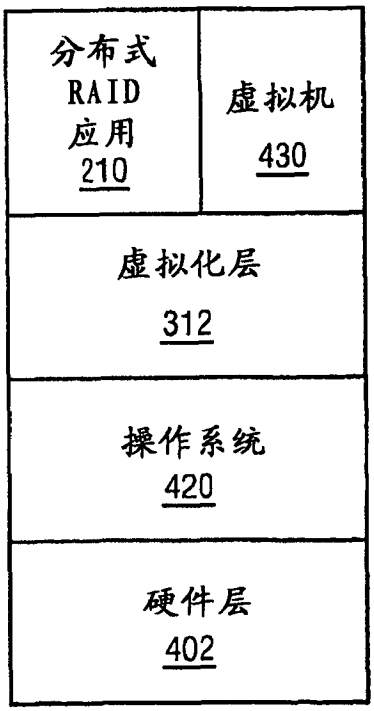


图 4B

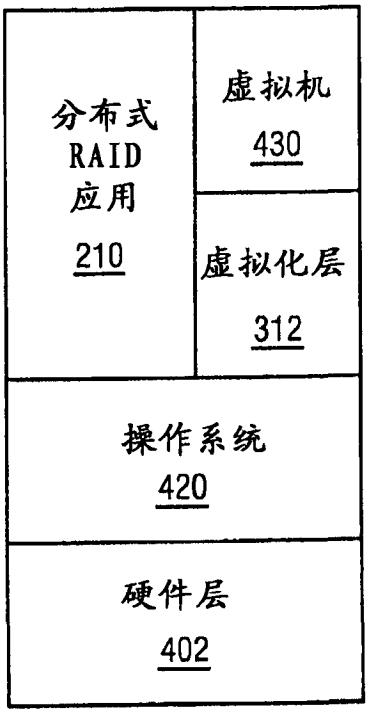


图 4C

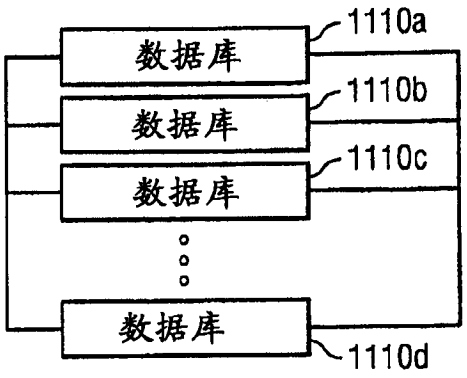

数据库
1110a
数据库
1110b
数据库
1110c
数据库
1110d

图 5

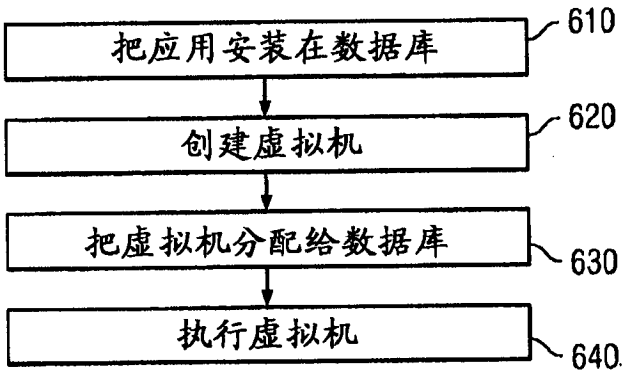

把应用安装在数据库
610
创建虚拟机
620
把虚拟机分配给数据库
630
执行虚拟机
640

图 6

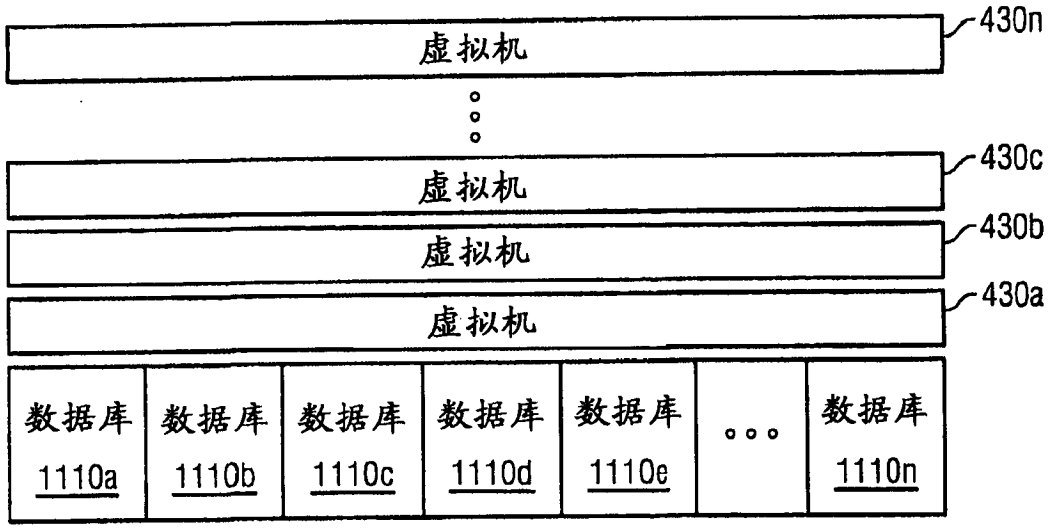

虚拟机
430n
虚拟机
430c
虚拟机
430b
虚拟机
430a
数据库 1110a
数据库 1110b
数据库 1110c
数据库 1110d
数据库 1110e
...
数据库 1110n

图 7