(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 102693230 B
(45)授权公告日 2015.12.09

(21)申请号 201110069907.X

(22)申请日 2011.03.23

(73)专利权人 伊姆西公司
地址 美国马萨诸塞州
专利权人 威睿公司

(72)发明人 常雷 杨子夜 毛文波 何英 堵俊平

(74)专利代理机构 北京市金杜律师事务所 11256
代理人 王茂华

(51)Int.Cl.
G06F 17/30(2006.01)
H04L 29/08(2006.01)

(56)对比文件
CN 101883042 A,2010.11.10, 全文.
CN 101911508 A,2010.12.08, 全文.
US 2009300302 A1,2009.12.03, 参见图1、2,段0030-0033,0036,0039-0040.
WO 2009145764 A1,2009.12.03, 参见图1,段0010-0012,0019-0020,0023-0024.
Sanjay Ghemawat etc..《The Google file system》.《SOSP '03 Proceedings of the nineteenth ACM symposium on Operating systems principles》.2003,

审查员 张伯

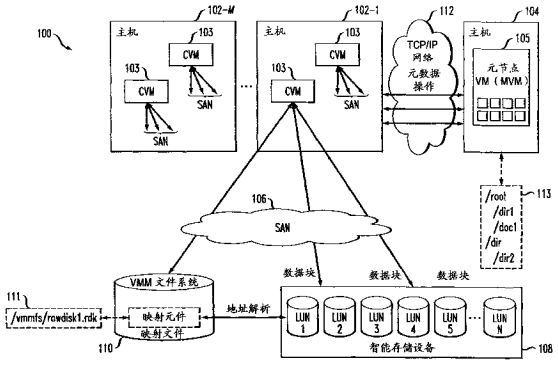
权利要求书3页 说明书9页 附图5页

(54)发明名称

用于存储区域网络的文件系统

(57)摘要

本发明涉及用于存储区域网络的文件系统。具体地,公开了用于管理对数据存储系统中的数据的访问的方法和装置。例如,一种装置,包括与分布式虚拟基础架构相关联的至少一个处理平台。该处理平台包括至少一个处理设备,其具有耦合至存储器的处理器。该处理平台可操作用于对元数据管理过程进行实例化,该元数据管理过程被配置为用于向至少一个客户端过程提供信息,以允许客户端过程通过存储区域网络、根据一个或多个数据存储设备而执行一个或多个操作。向客户端过程提供的信息可包括一个或多个数据块描述符。该一个或多个数据块描述符的每一个可包括:针对数据存储设备中至少一个数据存储设备的路径信息,以及在该至少一个数据存储设备中的偏移地址。



CN 102693230 B

1. 一种用于管理存储区域网络中的文件系统的装置，包括：

与分布式虚拟基础架构相关联的至少一个处理平台，所述处理平台包括至少一个处理设备，其具有耦合至存储器的处理器，

其中所述处理平台可操作用于对元数据管理过程进行实例化，所述元数据管理过程在所述至少一个处理设备所托管的虚拟机上执行，所述元数据管理过程被配置为用于向所述至少一个处理设备所托管的多个虚拟机上执行的多个客户端过程中的至少一个客户端过程提供信息，以允许所述客户端过程通过存储区域网络、根据一个或多个数据存储设备而执行一个或多个数据存储操作，

其中由所述元数据管理过程向所述至少一个客户端过程提供的所述信息包括元数据，所述客户端过程在数据存储操作期间向所述存储区域网络的文件系统中的映射元件提供所述元数据，使得所述映射元件执行地址解析操作，以根据所提供的所述元数据确定在所述一个或多个数据存储设备上执行的哪一个或多个逻辑存储单元对应于所提供的所述元数据。

2. 根据权利要求 1 所述的装置，其中所述客户端过程向所述元数据管理过程发送请求，以获取允许所述客户端过程执行所述一个或多个数据存储操作中至少一个数据存储操作的信息。

3. 根据权利要求 2 所述的装置，其中所述元数据管理过程向所述客户端过程返回信息，以允许所述客户端过程执行所述一个或多个操作中的至少一个操作。

4. 根据权利要求 1 所述的装置，其中向所述客户端过程提供的所述元数据包括一个或多个数据块描述符。

5. 根据权利要求 4 所述的装置，其中所述一个或多个数据块描述符的每一个包括：针对所述数据存储设备中至少一个数据存储设备的路径信息，以及针对所述至少一个数据存储设备的偏移地址。

6. 根据权利要求 5 所述的装置，其中所述一个或多个数据块描述符对应于跨所述一个或多个数据存储设备而存储或者可存储的给定数据文件的一个或多个数据块，并且其中所述给定数据文件的所述一个或多个数据块包括一个或多个主数据块和一个尾数据块。

7. 根据权利要求 1 所述的装置，其中执行所述元数据管理过程的所述虚拟机是元数据管理虚拟机，并且执行所述客户端过程的所述虚拟机是客户端虚拟机，并且其中所述元数据管理虚拟机和所述客户端虚拟机是所述分布式虚拟基础架构的部分。

8. 根据权利要求 7 所述的装置，其中由所述客户端虚拟机执行的所述一个或多个数据存储操作中的数据存储操作包括数据读取操作或者数据写入操作。

9. 根据权利要求 8 所述的装置，其中所述操作包括：所述客户端虚拟机请求所述元数据管理虚拟机发送由所述元数据管理虚拟机存储的一个或多个数据块描述符，所述数据块描述符对应于与所述一个或多个数据存储设备相关联的一个或多个地址。

10. 根据权利要求 9 所述的装置，其中所述操作还包括：所述元数据管理虚拟机向所述客户端虚拟机返回一个或多个所请求的数据块描述符的至少一部分。

11. 根据权利要求 10 所述的装置，其中所述操作还包括：所述客户端虚拟机本地存储由所述元数据管理虚拟机返回的一个或多个数据块描述符。

12. 根据权利要求 11 所述的装置，其中所述操作还包括：当一个或多个附加块描述符

未由所述客户端虚拟机本地存储时，所述客户端虚拟机从所述元数据管理虚拟机请求所述一个或多个附加块描述符。

13. 根据权利要求 12 所述的装置，其中所述操作还包括：所述元数据管理虚拟机向所述客户端虚拟机返回所请求的所述一个或多个附加块描述符。

14. 根据权利要求 10 所述的装置，其中当所述操作包括数据读取操作时，所述客户端虚拟机使用从所述元数据管理虚拟机获取的所述一个或多个数据块描述符的至少一部分，通过所述存储区域网络从所述一个或多个数据存储设备读取一个或多个相应的数据决。

15. 根据权利要求 10 所述的装置，其中当所述操作包括数据写入操作时，所述客户端虚拟机使用从所述元数据管理虚拟机获取的所述一个或多个数据块描述符的至少一部分，通过所述存储区域网络向所述一个或多个数据存储设备写入一个或多个相应的数据块。

16. 根据权利要求 1 所述的装置，其中所述处理平台包括云基础架构。

17. 一种用于管理存储区域网络中的文件系统的方法，包括：

在与分布式虚拟基础架构相关联的至少一个处理平台上对元数据管理过程进行实例化，

其中所述元数据管理过程在所述处理平台的至少一个处理设备所托管的虚拟机上执行，被配置为用于向所述至少一个处理设备所托管的多个虚拟机上执行的多个客户端过程中的至少一个客户端过程提供信息，以允许所述客户端过程通过存储区域网络、根据一个或多个数据存储设备而执行一个或多个数据存储操作，

其中由所述元数据管理过程向所述至少一个客户端过程提供的所述信息包括元数据，所述客户端过程在数据存储操作期间向所述存储区域网络的文件系统中的映射元件提供所述元数据，使得所述映射元件执行地址解析操作，以根据所提供的所述元数据确定在所述一个或多个数据存储设备上执行的哪一个或多个逻辑存储单元对应于所提供的所述元数据。

18. 根据权利要求 17 所述的方法，其中所述客户端过程向所述元数据管理过程发送请求，以获取允许所述客户端过程执行所述一个或多个数据存储操作中至少一个数据存储操作的信息；并且所述元数据管理过程向所述客户端过程返回信息，以允许所述客户端过程执行所述一个或多个数据存储操作中的至少一个数据存储操作。

19. 一种用于管理存储区域网络中的文件系统的装置，包括：

用于在与分布式虚拟基础架构相关联的至少一个处理平台上对元数据管理过程进行实例化的模块，

其中所述元数据管理过程在所述处理平台的至少一个处理设备所托管的虚拟机上执行，被配置为用于向所述至少一个处理设备所托管的多个虚拟机上执行的多个客户端过程中的至少一个客户端过程提供信息，以允许所述客户端过程通过存储区域网络、根据一个或多个数据存储设备而执行一个或多个数据存储操作，

其中由所述元数据管理过程向所述至少一个客户端过程提供的所述信息包括元数据，所述客户端过程在数据存储操作期间向所述存储区域网络的文件系统中的映射元件提供所述元数据，使得所述映射元件执行地址解析操作，以根据所提供的所述元数据确定在所述一个或多个数据存储设备上执行的哪一个或多个逻辑存储单元对应于所提供的所述元数据。

20. 一种存储区域网络文件系统，包括：

元数据管理虚拟机，其由至少一个处理平台实例化，被配置为用于向多个客户端虚拟机提供数据块描述符，以允许所述多个客户端虚拟机对给定数据文件执行数据读取操作和数据写入操作中的一个或多个，所述给定数据文件作为对应的数据块而被存储或者可存储在通过存储区域网络可访问的一个或多个数据存储设备上；以及

映射元件，其通过所述存储区域网络而与所述一个或多个数据存储设备以及与所述多个客户端虚拟机相耦合，其中所述映射元件解析由所述多个客户端虚拟机向其提供所述数据块描述符，以允许所述多个客户端虚拟机访问所述一个或多个数据存储设备上与所述数据块描述符对应的一个或多个逻辑存储单元上的所述给定数据文件的所述对应的数据块；

其中所述元数据管理虚拟机和所述映射元件由处理设备实现。

用于存储区域网络的文件系统

技术领域

[0001] 本发明总体上涉及数据存储系统领域，并且更具体地，涉及用于管理对此类数据存储系统中的数据的访问的技术。

背景技术

[0002] 数据密集型可扩展计算 (DISC) 系统是一种分布在计算机集群或者网格上的计算系统，这些计算机被设计用于处理可能在各种应用和环境中生成的大量数据。一般地，生成这种大量数据的应用和环境的示例包括但不限于：科学（例如，图像数据）、商业（例如，在线事务记录）和社会（例如，医疗或者其他个人记录、web 页面）。

[0003] 已经引入了支持对 DISC 系统中的大规模数据集进行处理的各种软件框架。一种这样的软件框架称为 MapReduce™，它由 Google™（美国加州山景城）开发，并且例如在美国专利号 7,650,331 中被描述，在此通过引用而并入其全部公开内容。MapReduce™是一种软件框架，它将涉及大规模数据集的计算分布到 DISC 计算机系统的计算机（节点）上。一般而言，MapReduce™使用“映射器工作者”节点和“还原器工作者”节点来获得给定任务并且将其分解为子任务，子任务被分布至 DISC 系统的一个或多个节点以进行处理。子任务被处理，并且结果被组合为针对给定任务的综合结果。通常在“映射”阶段中将给定任务分解为子任务，而通常在“还原”阶段中生成综合结果。

[0004] 另外，对 DISC 系统中大规模数据集的访问通常由存储文件系统来管理。对于 MapReduce™环境而言，可以使用诸如 Google 文件系统 (GFS) 的文件系统，例如参见以下文献：S.Ghemawat 等人，“The Google File System，”19th ACM Symposium on Operating SystemsPrinciples，Lake George，NY，October 2003，在此通过引用而并入其全部公开内容。在应用于 DISC 系统的 GFS 中，服务器将“数据区块”作为文件存储在本地文件系统中。由此，在采用 GFS 的 DISC 系统中，计算和数据是紧密耦合的。例如，利用 GFS，映射器工作者节点的中间结果被写入本地盘，并且该中间结果继而被传送给多个其他还原器工作者节点。遗憾的是，如果映射器工作者节点故障，则必须重新进行对其执行的任务。

发明内容

[0005] 本发明的原理提供一种用于管理对数据存储系统中的数据的访问的技术。

[0006] 在本发明的一个方面中，一种装置，包括与分布式虚拟架构相关联的至少一个处理平台。该处理平台包括至少一个处理设备，其具有耦合至存储器的处理器。处理平台可操作用于对元数据管理过程进行实例化，该元数据管理过程被配置为用于为至少一个客户端过程提供信息，以允许该客户端过程通过存储区域网络按照一个或多个数据存储设备执行一个或多个操作。

[0007] 例如，在说明性实施方式中，客户端过程向元数据管理过程发送请求，以获得允许该客户端过程执行一个或多个操作中至少一个操作的信息。元数据管理过程向客户端过程返回信息，以允许客户端过程执行一个或多个操作中的至少一个操作。向客户端过程提供

的信息可以包括一个或多个数据块描述符。该一个或多个数据块描述符的每一个可以包括针对至少一个数据存储设备的路径信息，以及针对该至少一个数据存储设备的偏移地址。一个或多个数据块描述符可以对应于跨一个或多个数据存储设备而存储或可存储的给定数据文件的一个或多个数据块。该给定数据文件的一个或多个数据块可以包括一个或多个主数据块以及一个尾数据块。

[0008] 此外，在说明性实施方式中，元数据管理过程由元数据管理虚拟机实现，并且客户端过程由客户端虚拟机实现。元数据管理虚拟机和客户端虚拟机是分布式虚拟基础架构的部分。由客户端虚拟机执行的一个或多个操作中的操作包括数据读取操作或者数据写入操作，它们是输入 / 输出 (I/O) 操作的示例。

[0009] 在本发明的第二方面，一种存储区域网络文件系统，包括元数据管理虚拟机和映射元件。元数据管理虚拟机由至少一个处理平台实例化，其被配置为用于向多个客户端虚拟机提供数据块描述符，以允许该多个客户端虚拟机对给定数据文件执行一个或多个数据读取操作和数据写入操作，该给定数据文件作为相应的数据块而被存储或者可存储在通过存储区域网络可访问的一个或多个数据存储设备上。映射元件通过存储区域网络而与一个或多个数据存储设备以及与多个客户端虚拟机耦合。映射元件解析数据块描述符，以允许多个客户端虚拟机访问一个或多个数据存储设备上的给定数据文件的相应数据块。

[0010] 有益地，本发明的技术提供了诸如 DISC 系统的计算机系统中的计算与数据的解耦。这至少是通过以下方式实现的，即，元数据管理过程（虚拟机）为客户端过程（虚拟机）提供元数据（以数据块描述符的形式），其允许客户端过程通过存储区域网络而直接访问（并发地或者说并行地）数据存储设备。本发明的原理所提供的这种计算与数据的解耦改善了资源利用，并且获得了更为能源有效的 DISC 解决方案，在此将对其做出进一步阐释。还可以在相同大小的集群上部署不同的工作负载，同时适应每个工作负载中的动态改变。此外，本发明的技术有益地提供了改进的系统性能。将会显而易见的是，利用根据本发明示范实施方式的文件系统，用于映射和还原功能的 I/O 路径被缩短，由此改进了系统性能。

[0011] 通过附图和下文详细描述，本发明的这些以及其他特征和优点将变得更为清楚。

附图说明

[0012] 图 1 示出了根据本发明一个实施方式的文件系统和存储区域网络。

[0013] 图 2 示出了根据本发明一个实施方式的用于实现图 1 中的文件系统和存储区域网络的处理平台。

[0014] 图 3 示出了根据本发明一个实施方式的数据文件的存储示例。

[0015] 图 4 示出了根据本发明一个实施方式的图 1 的文件系统和存储区域网络中的数据文件读取过程。

[0016] 图 5 示出了根据本发明一个实施方式的图 1 的文件系统和存储区域网络中的数据文件写入过程。

具体实施方式

[0017] 在此将参考示例性计算系统和数据存储系统以及关联的服务器、计算机、存储设

备和其他处理设备来描述本发明。然而，应理解，本发明不限于与所示的特定说明性系统和设备配置结合使用。而且，在此使用的词语“计算系统”和“数据存储系统”意在做广义地理解，从而涵盖例如专用或公共的云计算或存储系统，以及包括分布式虚拟基础架构的其他类型的系统。然而，给定的实施方式可以更为一般地包括一个或多个处理设备的任意布置。

[0018] 此外，在此使用的词语“文件系统”一般地表示提供针对数据存储系统上存储的数据的访问管理功能（例如，辅助数据读取和数据写入操作，以及任何其他 I/O 操作）的系统。可以理解，当数据块或者数据文件可由不止一个访问实体并发访问时，将期望文件系统确保数据一致性。

[0019] 如下文详述的，在本发明的说明性实施方式中，在虚拟平台上提供一种并行存储区域网络 (SAN) 文件系统，以供在数据密集型可扩展计算 (DISC) 系统中使用。换言之，在此实施方式中，与文件系统结合操作的计算系统是 DISC 系统，并且数据存储系统是 SAN。实现包括具有分布式虚拟基础架构的一个或多个处理平台。这样，在此说明性实施方式中，文件系统能够适应 DISC 系统的文件访问模式，并且为具有 SAN 和智能存储设备的虚拟平台提供优化的性能。然而，应当理解，本发明的原理不限于任何特定的 DISC 系统或者任何特定 SAN。

[0020] 虚拟平台（分布式虚拟基础架构）实现称为“虚拟化”的计算概念。一般而言，虚拟化允许一个或多个“虚拟机”(VM) 在单个物理机器上运行，每个虚拟机共享该一个物理机器的资源。由此，虚拟机是可以在一个或多个物理处理元件（例如，服务器、计算机、处理设备）上被实例化的逻辑处理元件。换言之，一般而言，“虚拟机”表示类似于物理机器而执行程序的机器（即，计算机）的软件实现。由此，不同的虚拟机可以在相同的物理计算机上运行不同的操作系统和多个应用。虚拟化可以这样来实现：在计算机硬件上直接插入软件层，以便提供虚拟机监控器或称“管理器 (hypervisor)”，其动态地并且透明地分配物理计算机的硬件资源。管理器为多个操作系统提供在单个物理计算机上并发运行并且彼此共享硬件资源的能力。

[0021] 诸如VMware® vSphere™之类商业上可获得的虚拟化软件可以用来构建跨数百个互连的物理计算机和存储设备而分布的复杂虚拟基础架构，包括专用和公共云计算以及存储系统。由此，“分布式虚拟基础架构”一般地表示通过一个或多个虚拟机的实例化（生成或者创建）而实现的计算和存储元件。这样的布置有益地避免了持久性地向每个应用指派服务器、存储设备或者网络带宽的需要。相反，可用的硬件资源在需要的时候和情况下被动态地分配。因此，高优先级应用可被分配以需要的资源，而无需以仅在高峰时间才被使用的专用硬件为代价。

[0022] 图 1 示出了根据本发明一个说明性实施方式的并行 SAN 和文件系统 100。如所示，系统 100 包括多个主机 102-1、...、102-M，其中每个主机实现有代表一个或多个客户端过程的一个或多个客户端虚拟机 (CVM) 103。系统 100 还包括主机 104，其实现有代表元数据管理过程的元节点虚拟机 (MVM) 105。每个 CVM 103 通过存储区域网络 (SAN) 106 耦合至数据存储系统 108。由此，因为每个 CVM 都通过 SAN 106 耦合至数据存储系统 108，并且多个 CVM 可以并行或者说并发访问相同的数据文件，因此这一布置被称为并行 SAN。还可以理解，当阐释主机或者其他计算元件实现虚拟机时，这一般地表示虚拟机被生成或者创建（即，实例化），以执行该特定虚拟机所需或者特定的任何功能（一个或多个过程）。

[0023] 如已知的，SAN 106 包括多个 SAN 交换机或者其他网元，其允许主机 102-1、...、102-M 直接与数据存储网络 108 连接。主机，或者更具体地说是 CVM 103，可以访问数据存储系统 108，例如用于执行数据读取或写入请求或者其他 I/O（输入 / 输出）操作。在一个实施方式中，SAN 106 的通信介质采用光纤通道 (FC) 通信协议。

[0024] 然而，将主机与数据存储系统连接的通信介质不限于 FC SAN 布置，而可以是各种网络或其他类型通信连接中的任意一个或多个，诸如网络连接、总线或者其他类型的数据链路，这是本领域技术人员已知的。例如，通信介质可以是因特网、内联网或者任何其他有线或无线连接，主机 102 可以借助于该通信介质访问数据存储系统 108 并与之通信，并且还可以与系统 100 中包括的其他组件通信。这样，介质 106 上的通信可以备选地遵循诸如小型计算机系统接口 (SCSI)、因特网 SCSI(iSCSI) 等已知协议。

[0025] 主机 102 和数据存储系统 108 可以位于相同的物理站点，或者可以位于不同的物理站点。每个主机 102 可以根据不同类型的任务而执行不同类型的数据操作。例如，任何一个主机 102 都可以向数据存储系统 108 发出数据请求以执行数据操作。更具体地，在主机 102 之一上执行的应用可以执行读取或者写入操作，这导致针对数据存储系统 108 的一个或多个数据请求。

[0026] 还应理解，数据存储系统 108 可以包括单个数据存储系统，诸如单个数据存储阵列；或者在使用本技术的实施方式中，也可以表示例如多个数据存储阵列，这些数据存储阵列可以是单独的，也可以与诸如 SAN 中具有适当连接性的其他数据存储设备、系统、装置或者其他组件相结合。还应注意，实施方式可以包括来自一个或多个供应商的数据存储阵列或者其他组件。例如，数据存储系统 108 可以实现为传统的Symmetrix® DMX™数据存储阵列或者可从美国马萨诸塞州霍普金顿市的 EMC 公司购得的CLARiiON®数据存储阵列。然而，本领域技术人员将会理解，在此公开的技术适于与其他供应商的其他数据存储阵列以及与这里出于示例目的而描述的那些组件之外的其他组件结合使用。

[0027] 数据存储系统 108 包括多个数据存储设备，其可以与一个或多个存储阵列相关联。这些物理数据存储设备（未在图 1 中单独地示出，但被总体上示为“智能存储设备”）可以包括一个或多个不同类型的数据存储设备，诸如一个或多个盘驱动器、一个或多个固态驱动器 (SSD) 等。由此，存储设备可以包括采用一个或多个不同闪存技术的闪存器件。在这样的实现中，数据存储设备可以包括盘设备和闪存器件的结合，其中闪存器件可以充当与数据存储阵列结合使用的各种软件工具的标准 FC 盘驱动器。盘设备可以是任何一个或多个不同类型的盘设备，诸如高级技术附件 (ATA) 盘驱动器、FC 盘驱动器，等等。闪存器件可以使用不同类型的存储器技术来构造，诸如形成一个或多个 SLC（单级单元）器件或者 MLC（多级单元）器件的非易失性半导体 NAND 闪存。闪存器件和盘设备是与在此描述的技术结合使用的数据存储系统中可以包括的存储设备的两种示例性类型。

[0028] 在主机 102 通过 SAN 106 而直接访问数据存储系统 108 的同时，主机访问被主机视作多个逻辑单元 (LU) 的已存储数据。LU 可以对应于也可以不对应于实际物理存储设备。例如，一个或多个 LU 可以驻留在单个物理驱动器或者多个驱动器上，或者驻留在多个驱动器的各种子集上。例如，如图 1 所示，数据存储系统 108 包括 LU 1、2、...、N(LUN 表示逻辑单元编号)。诸如单个数据存储阵列的单个数据存储系统中的数据可被多个主机访问，从而允许主机共享驻留于其中的数据。

[0029] 回想，主机 104 实现 MVM 105。MVM 105 存储和管理系统 100 的元数据。如已知的，一般而言，“元数据”是关于数据的数据。MVM 105 所处理的元数据的类型的示例包括但不限于：文件系统目录信息（例如，如 113 中所示）；关于数据存储系统 108 的原始（物理）存储设备的信息；以及关于数据存储系统 108 中存储的数据块的信息，此类信息的形式是数据块描述符，这将在下文进一步详述。由此，MVM 105 是元数据管理虚拟机。

[0030] MVM 105 向主机 102 上的每个 CVM 103 提供文件访问接口，这允许图 1 中的元件在其上具体化的处理平台中的计算处理和数据的解耦。如所示，借以执行元数据操作的、CVM 103 与 MVM 105 之间的连接是通过 TCP/IP（传输控制协议 / 网际协议）网络 112 的；然而，其他连接 / 通信协议也是可能的。

[0031] 进一步如图 1 中所示，系统 100 包括虚拟机监控器 (VMM) 文件系统映射元件 110。VMM 文件系统映射元件 110 维护映射文件（映射），其将逻辑单元地址与物理设备地址进行关联。换言之，VMM 文件系统映射元件 110 维护映射文件，其是指向由一个或多个后端存储系统（即，作为数据存储系统 108 部分的后端存储系统）暴露的物理存储（原始）设备的符号化链接。由此，当 CVM 103 使用从 MVM 105 获得的一个或多个数据块描述符来向数据存储系统 108 发送数据读取或者写入请求时，VMM 文件系统映射元件 110 使用映射文件来解析与 CVM 希望访问的一个或多个存储位置相对应的正确逻辑单元地址。注意，数据存储系统 108 中存储的或者可以存储的（可存储的）数据的每个块具有块描述符，其包括该数据块所驻留的原始（物理）设备的全局路径，以及在该原始设备中的偏移（开始地址）。作为示例，图 1 中的参考标号 111 示出了可以存储在映射元件 110 中的映射文件的一个示例，例如，/vmmfs/rawdisk1.rdk。在接收到块描述符之后，即接收到目标数据块所驻留的（或者一旦被写入则将驻留的）原始（物理）设备的全局路径以及在该原始设备中的偏移（开始地址）之后，映射元件 110 访问相应的映射文件（例如，/vmmfs/rawdisk1.rdk），并且从其获得关于哪一个或多个逻辑单元编号 (LUN) 对应于这些数据块的指示。由此，该数据块位置中的数据可被请求 CVM 读取（如果是读取操作的话），或者数据可由请求 CVM 写入该数据块位置（如果是写入操作的话）。

[0032] 应当理解，图 1 中所示的元件（例如，主机 102、主机 104、SAN106、数据存储系统 108、VMM 文件系统映射元件 110、TCP/IP 网络 112）中的一个或多个可以实现为专用或者公共云计算或存储系统的部分。此外，至少元件的子集可以共同实现在公共处理平台上，或者每个此类元件可以实现在包括一个或多个服务器、计算机或者其他处理设备的独立处理平台上。

[0033] 此类处理平台的一个示例是图 2 中所示的处理平台 200。此实施方式中的处理平台 200 至少包括系统 100 的一部分，并且包括由 202-1、202-2、202-3、...、202-P 表示的多个服务器，其通过网络 204 彼此通信。因此，系统 100 的一个或多个元件每一个都可以在服务器、计算机或者其他处理平台元件（其可以视作在此更为一般地称作“处理设备”的示例）上运行。如图 2 中所示，此类设备一般地包括至少一个处理器以及相关联的存储器，并且实现用于控制系统 100 的特定特征的一个或多个功能模块。再次，在给定的实施方式中，多个元件可由单个处理设备实现。

[0034] 处理平台 200 中的服务器 202-1 包括与存储器 212 耦合的处理器 210。处理器 210 可以包括微处理器、微控制器、专用集成电路 (ASIC)、现场可编程门阵列 (FPGA)、或者其他

类型的处理电路以及此类电路元件的部分或者组合。存储器 212 可以视作在此更为一般地称作“计算机程序产品”的示例，其具有包含于其中的可执行计算机程序代码。此类存储器可以包括电存储器，诸如随机访问存储器 (RAM)、只读存储器 (ROM) 或其他类型的存储器，或者其任意组合。当计算机程序代码由诸如服务器 202-1 的处理设备执行时，其使得设备执行与系统 100 的一个或多个元件相关联的功能。给出在此提供的教导，本领域技术人员将能够容易地实现此类软件。具体化本发明的各方面的计算机程序产品的其他示例可以包括例如光盘或者磁盘。

[0035] 服务器 202-1 中还包括网络接口电路 214，其用于将服务器与网络 204 和其他系统组件对接。此类电路可以包括本领域已知类型的传统收发机。

[0036] 处理平台 200 的其他服务器 202 被假设为按照与图中针对服务器 202-1 所示的类似方式进行配置。

[0037] 图 2 中所示的处理平台 200 可以包括其他已知的组件，诸如批处理系统、并行处理系统、物理机器、虚拟机、虚拟交换机、存储卷，等等。再次，图中所示的特定处理平台仅仅作为示例给出，并且系统 100 可以包括附加的或备选的处理平台，以及多个不同处理平台的任意组合。

[0038] 而且，在系统 100 中，服务器、计算机、存储设备或者其他组件的多种其他布置也是可能的。此类组件可以通过任何类型的网络与系统 100 的其他元件通信，诸如通过广域网 (WAN)、局域网 (LAN)、卫星网络、电话或者电缆网络或者这些和其他类型网络的各个部分或者组合。

[0039] 现在将参考图 3 到图 5 更为详细地描述系统 100 的元件的操作。

[0040] 在并行 SAN 文件系统 100 中，文件存储在数据存储系统 108 中的多个数据块中。每个数据块具有数据块描述符，其包括该块所驻留的原始设备（物理存储设备）的全局路径，以及在该原始设备中的偏移（开始地址）。存在两种类型的块：主块和尾块。主块是大小为 S 的大数据块，其中 S 通常大于或等于约 64MB（兆字节），并被用于提供快速顺序访问。尾块是可变大小（小于 S）的小数据块，其被用于避免为文件尾部只是较小的数据区块分配大块。DISC 文件的常见情况是：它包含多个主块，并且只包含一个尾块。然而，一个文件可以包含一个主块和一个尾块。

[0041] 图 3 示出了根据本发明一个实施方式的 DISC 文件的存储的示例。如所示，文件 A 包括三个主数据块 (B1、B2 和 B3) 以及一个尾数据块 (B4)，文件 B 包括四个主数据块 (B1、B2、B3 和 B4) 和一个尾数据块 (B5)。为便于说明，假设文件 A 和文件 B 存储于其中的数据存储系统包括四个物理存储设备（原始设备）1、2、3 和 4，每个具有实现于其上的两个逻辑单元 (LUN 1 和 2)。当然，这仅仅是出于说明目的，并且可以理解的是，数据存储系统（例如，图 1 中的 108) 可以具有更少或者更多数目的原始设备，每个原始设备可以具有实现于其上的更少或者更多数目的 LU。而且，并不要求如图 3 中说明性绘出的那样将特定文件的数据块存储在连续的原始设备或者连续的 LU 上。

[0042] 还假设：数据存储系统实现 2 的倍数的复制。例如，与文件 A 和文件 B 关联的应用要求：维护每个文件的两个拷贝，并且文件的相同数据块的拷贝不存储在相同的原始设备上。这些要求通过图 3 中所示的文件 A 和文件 B 的主数据块和尾数据块的分布而得以满足。

[0043] 存储在图 3 中的存储设备上的每个数据块由唯一的数据块描述符标识。存储在数据存储系统中的数据块的数据块描述符是图 1 的 MVM 105 所存储和维护的元数据的部分。应当理解，系统 100 的锁定粒度因而处于文件级或块级。对于块，在本发明的一个说明性实施方式中，系统 100 可以具有多个并发的读取方（执行读取操作的 CVM）而没有并发的写入方（执行写入操作的 CVM），或者只有一个写入方而没有并发读取方。利用此锁定模型，文件可被并发地读取，并且不同的块可被不同的写入方并发写入，例外是只允许一个追加 (appending) 写入方。

[0044] 图 4 示出了根据本发明一个实施方式的图 1 的文件系统和存储区域网络中的文件读取过程 400。文件读取过程 400 是这样的过程，CVM 103 在 MVM 105 的辅助下借助于该过程来读取与数据存储系统 108 中存储的文件相对应的一个或多个数据块。为了简化对文件读取过程 400 的描述，假设锁定步骤和查错被省略。给出这里的详细描述，本领域普通技术人员将会理解可以如何实现此类锁定步骤和查错。

[0045] 在步骤 1，CVM 103 请求 MVM 105 发送由 MVM 105 存储的一个或多个数据块描述符。如上所述，一个或多个数据块描述符对应于与一个或多个数据存储设备相关联的一个或多个地址。如图 4 中特别示出的，步骤 1 这样来实现：由 CVM 103 针对 MVM 105 调用开放远程过程调用 (RPC)。RPC 包括以下参数：（要读取的目标文件的）文件名，以及读取标志。在步骤 2，MVM 105 向 CVM 103 返回目标文件的块描述符的列表。如果文件的块数目较大，则只返回前 N 个块描述符 (N 是可选的系统参数)。块描述符的列表被缓存（本地存储）在客户端侧 (CVM 103 处)。

[0046] 在步骤 3，CVM 103 利用以下参数调用搜寻：文件描述符 (fd) 和偏移。在此步骤中，CVM 103 首先检查已缓存的块。如果所请求块的块描述符在本地缓存中，则过程去往步骤 5。否则，CVM 103 从 MVM 105 请求（附加的）相应块描述符。在步骤 4，MVM 5 返回所请求的块描述符，并且相邻的块描述符可被预取，以适应 DISC 文件的顺序访问模式。在步骤 5，使用所获取的块描述符，CVM 103 直接通过 SAN 106（图 3 中未示出）从数据存储系统 108 中的物理（原始）设备的适当存储位置读取目标文件的数据块。CVM 103 还可以预取其他数据块。注意，文件读取功能可以实现在并行文件系统库 (PFS Lib) 中。这些功能可被用于通过 SAN 106 直接从数据存储系统 108 读取文件。

[0047] 图 5 示出了根据本发明一个实施方式的图 1 的文件系统和存储区域网络中的文件写入过程 500。文件写入过程 500 是这样的过程，CVM 103 在 MVM 105 的辅助下借助于该过程来写入与数据存储系统 108 中要存储的（或者要更新的）文件相对应的一个或多个数据块。为了简化对文件写入过程 500 的描述，假设锁定步骤和查错被省略。给出这里的详细描述，本领域普通技术人员将会理解可以如何实现此类锁定步骤和查错。

[0048] 在步骤 1，CVM 103 请求 MVM 105 发送由 MVM 105 存储的一个或多个数据块描述符。如上所述，一个或多个数据块描述符对应于与一个或多个数据存储设备相关联的一个或多个地址。如图 5 中特别示出的，步骤 1 这样来实现：由 CVM 103 针对 MVM 105 调用开放远程过程调用 (RPC)。RPC 包括以下参数：（要写入或者更新的目标文件的）文件名，以及写入标志。在步骤 2，MVM 105 向 CVM 103 返回目标文件的块描述符的列表。如果文件的块数目较大，则只返回前 N 个块描述符 (N 是可选的系统参数)。块描述符的列表被缓存（本地存储）在客户端侧 (CVM 103 处)。

[0049] 在步骤 3，CVM 103 利用以下参数调用搜寻 :文件描述符 (fd) 和偏移。在此步骤中，CVM 103 首先检查已缓存的块。如果所请求块的块描述符在本地缓存中，则过程去往步骤 5。否则，CVM 103 从 MVM 105 请求（附加的）相应块描述符。在步骤 4，MVM 105 返回所请求的块描述符。如果处于文件的末端（偏移超过文件的末端），则由 MVM 105 分配新块。新块被写入客户端侧 (CVM) 缓存。在步骤 5，使用所获取的块描述符，CVM 103 直接通过 SAN 106（图 3 中未示出）向数据存储系统 108 中的物理（原始）设备的适当存储位置写入目标文件的数据块。注意，如果添加了新块，则当已缓存块变满或写入结束时，将块提交给 MVM 105。MVM 105 为已缓存的块分配主块 / 尾块。还要注意，文件写入功能可以实现在并行文件系统库 (PFS Lib) 中。这些功能可被用于通过 SAN 106 直接向数据存储系统 108 写入文件。

[0050] 应当理解，结合图 3 到图 5 的框图而描述的特定处理操作和其他系统功能仅仅是以说明性示例的方式给出的，并且不应当被理解为以任何方式限制本发明的范围。备选实施方式可以使用其他类型的处理操作来执行数据文件读取和写入。

[0051] 而且，如上文指出的，诸如结合图 3 到图 5 的框图而描述的功能至少可以部分地实现为以下形式，即，存储在存储器中并且由诸如计算机或者服务器的处理设备的处理器执行的一个或多个软件程序。具有包含于其中的此类程序代码的存储器是在此更为一般地称作“计算机程序产品”的示例。

[0052] 说明性实施方式相对于已有技术而言提供了多种优点。

[0053] 如上所述，在基于 GFS 的系统中，计算和数据是紧密耦合的。例如，在 MapReduce™ 计算模型中，映射器工作者的中间结果被写入本地盘中。中间结果将被传输给多个其他的还原器工作者。如果映射器故障，则其上的工作必须重做。对于根据本发明实施方式而提供的架构和方法，数据和计算被解耦，由此计算节点几乎是无状态的。本发明的技术允许自由的重启故障节点和任何节点的迁移。自由的迁移将获得很多优点。虚拟机可以容易和自由地移动，并且资源可得到有效利用。空闲的物理机器可被关闭以便节能，并且还可向空闲资源指派除 DISC 任务之外的其他类型的任务。而在已有方法中，即使当工作负载不重并且存在很多空闲资源时，整个计算机集群通常也被打开，这导致了明显的能源浪费。

[0054] 而且，已有的 DISC 文件系统没有考虑虚拟化的环境。它们受制于不佳的 I/O 性能。以 MapReduce™ 计算模型为例。在映射阶段，块访问路径是客户端→ GFS → Linux 文件系统 → SAN 文件系统→ SAN 存储设备。而且，对于还原阶段，块访问路径是客户端→ TCP/IP 网络→ Linux 文件系统→ SAN 文件系统→ SAN 存储设备。由此，长路径和路径的某些缓慢部分降低了此类已有系统中的 I/O 性能。

[0055] 根据本发明的说明性实施方式，计算和数据的解耦改进了资源利用，并且获得了更为能源有效的 DISC 解决方案。不同的工作负载甚至可以部署在相同的大型集群上，以适应工作负载的动态改变。

[0056] 本发明的另一优点是改进的性能。对于本发明的说明性实施方式中的文件系统而言，映射和还原的 I/O 路径都是简单的客户端→本发明中的文件系统→ SAN 存储设备。路径很短，并且路径中的缓慢部分被消除。由此，增强了 I/O 性能。在 DISC 系统中，最耗时的部分与 I/O 操作相关联。根据在此描述的本发明的原理，I/O 性能的改进获得了 DISC 系统的总体性能改进。

[0057] 应当再次强调的是，上文描述的本发明实施方式仅仅是出于说明目的而给出的。可以在示出的特定布置中进行很多变化。例如，尽管在特定的系统和设备配置的上下文中被描述，但是技术也适用于各种其他类型的信息处理系统、处理设备和分布式虚拟基础架构布置。而且，在描述说明性实施方式的过程中所做出的任何简化假设都应当被视为是示例性的，而不是对本发明的要求或限制。对于本领域的技术人员而言，所附权利要求范围内的其他备选实施方式将是显而易见的。

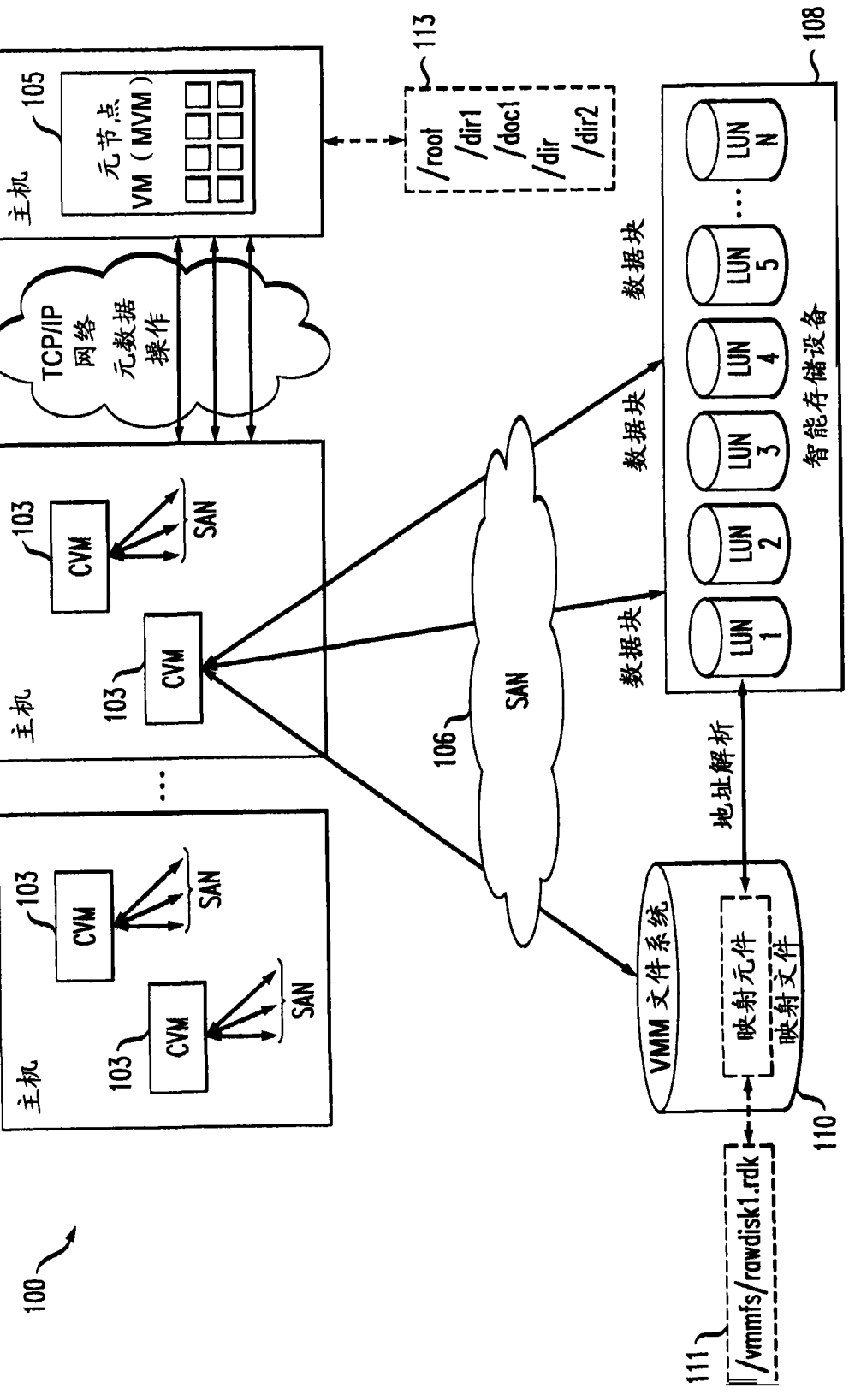

100
102-1
102-M
103
CVM
SAN
主机
104
105
元节点
VM（MVM）
112
TCP/IP
网络
元数据
操作
113
/root
/dir1
/doc1
/dir
/dir2
106
SAN
108
数据块
LUN
1
2
3
4
5
N
智能存储设备
地址解析
VMM 文件系统
映射元件
映射文件
110
111
/vmmfs/rawdisk1.rdk

图 1

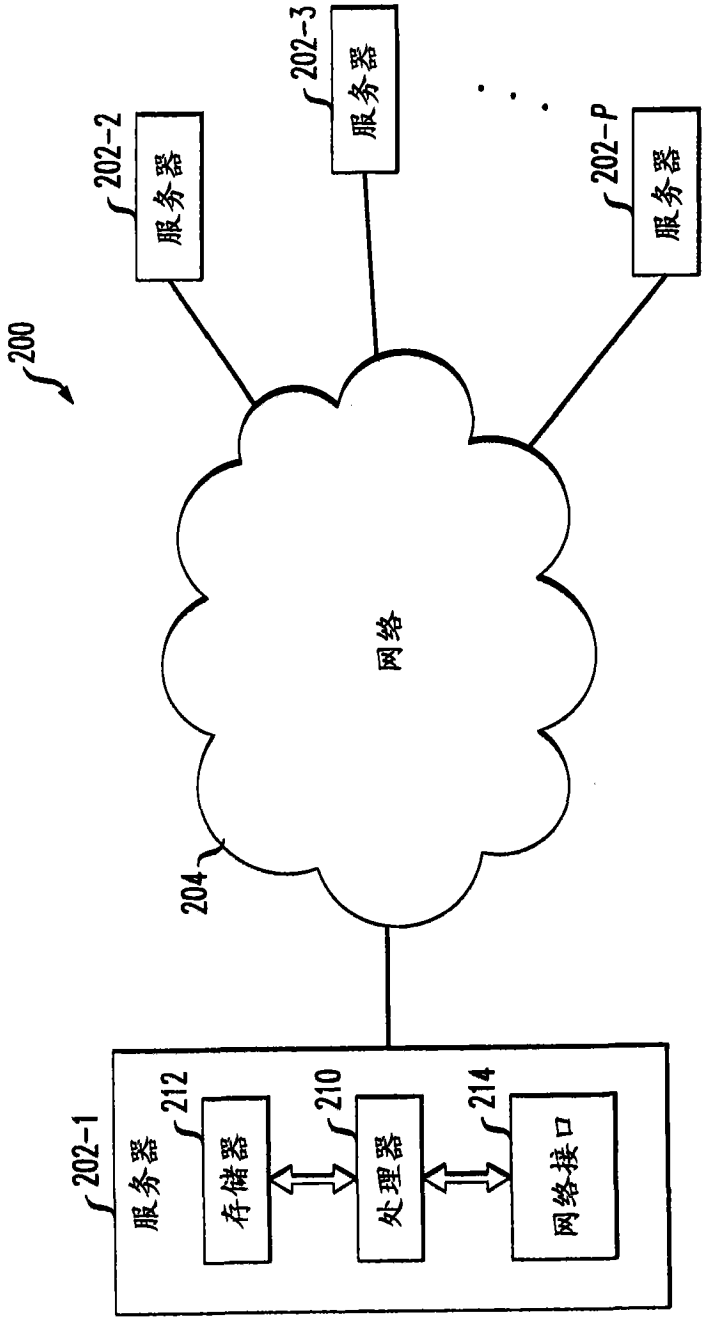

200
202-1
服务器
212
存储器
210
处理器
214
网络接口
204
网络
202-2
服务器
202-3
服务器
202-P
服务器

图 2

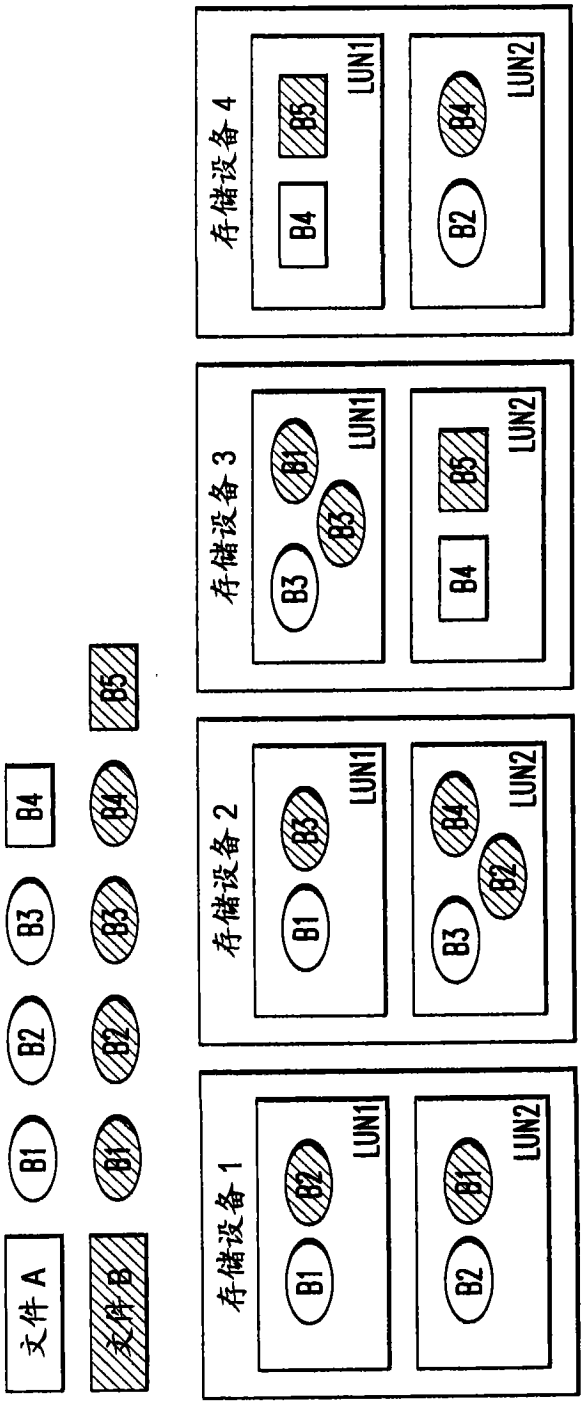

文件A
B1
B2
B3
B4
文件B
B1
B2
B3
B4
B5
存储设备1
LUN1
LUN2
存储设备2
LUN1
LUN2
存储设备3
LUN1
LUN2
存储设备4
LUN1
LUN2

图 3

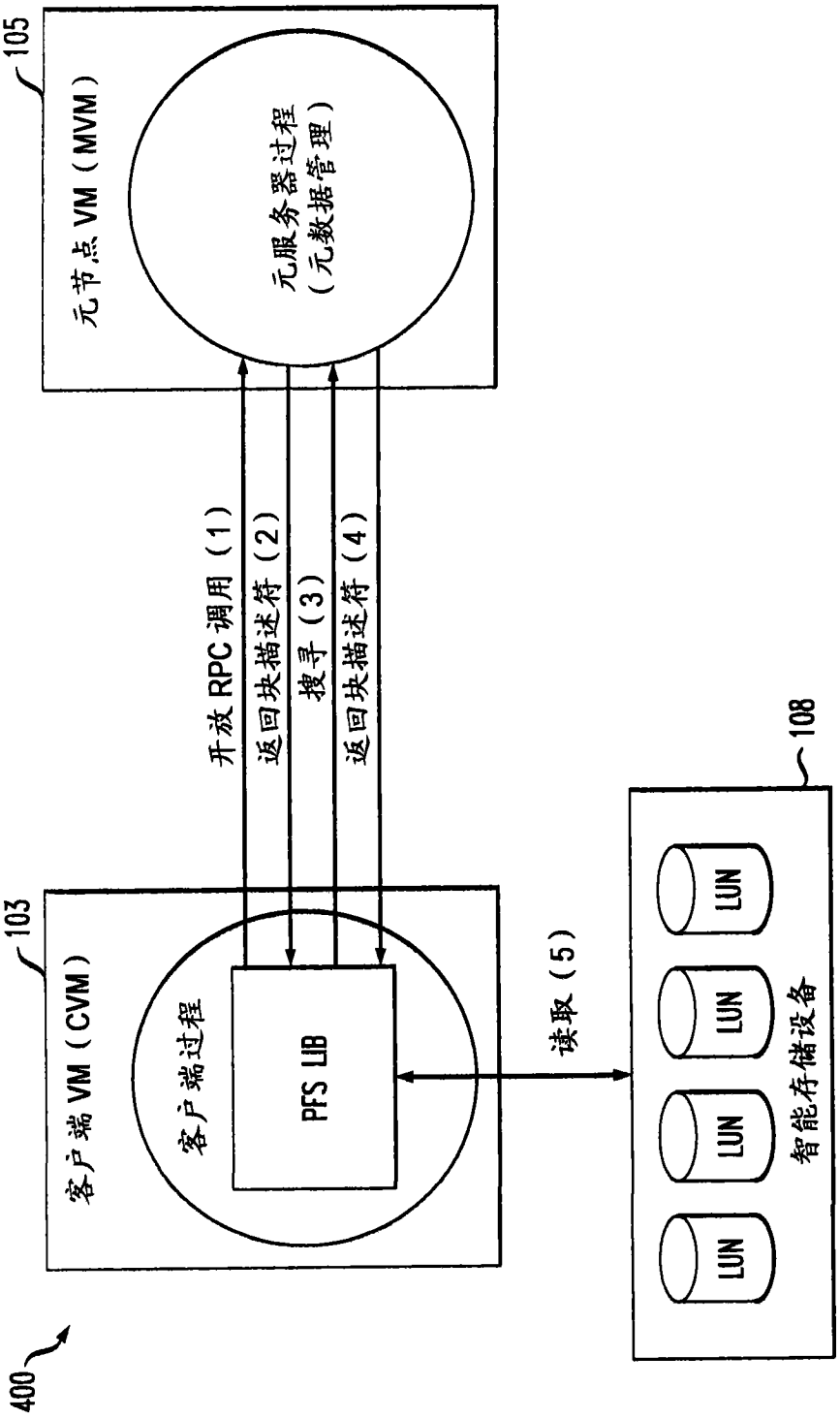

105
元节点 VM（MVM）
元服务器过程
（元数据管理）
开放 RPC 调用（1）
返回块描述符（2）
搜寻（3）
返回块描述符（4）
103
客户端 VM（CVM）
客户端过程
PFS LIB
读取（5）
108
LUN
LUN
LUN
LUN
智能存储设备
400

图 4

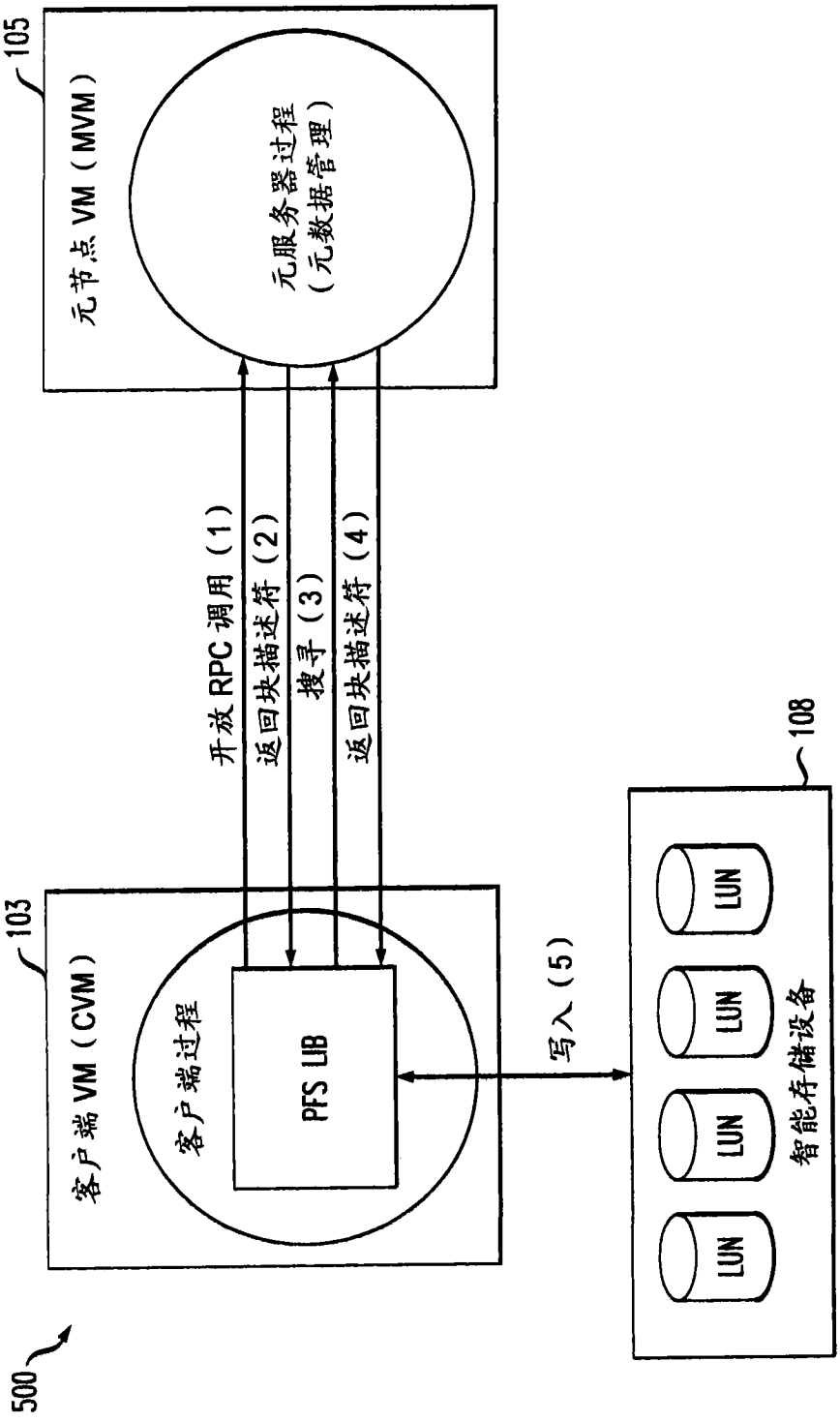

105
元节点 VM（MVM）
元服务器过程
（元数据管理）
开放 RPC 调用（1）
返回块描述符（2）
搜寻（3）
返回块描述符（4）
103
客户端 VM（CVM）
客户端过程
PFS LIB
写入（5）
108
LUN
LUN
LUN
LUN
智能存储设备
500

图 5