(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 102365625 A
(43)申请公布日 2012.02.29

(21)申请号 201080013779.0

(22)申请日 2010.03.19

(30)优先权数据
12/412272 2009.03.26 US

(85)PCT申请进入国家阶段日
2011.09.26

(86)PCT申请的申请数据
PCT/US2010/028034 2010.03.19

(87)PCT申请的公布数据
WO2010/111149 EN 2010.09.30

(71)申请人 微软公司
地址 美国华盛顿州

(72)发明人 J. 奥辛斯

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 李舒 刘鹏

(51)Int.Cl.
G06F 12/02(2006.01)
G06F 12/06(2006.01)
G06F 12/08(2006.01)
G06F 13/16(2006.01)

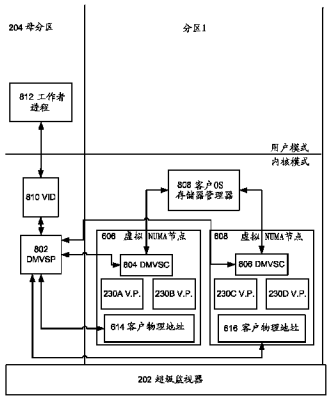
权利要求书 3 页 说明书 17 页 附图 17 页

(54)发明名称

用于虚拟机的虚拟非一致存储器体系结构

(57)摘要

这里描述了用于实现对于虚拟机的虚拟NUMA体系结构、以及调节虚拟NUMA节点中的存储器的技术。



CN 102365625 A

1. 一种方法，包括：

接收对于实例化虚拟机（240）的请求，所述请求包括用于虚拟机（240）的特征；

根据所述特征选择用于虚拟机（240）的虚拟 NUMA 节点拓扑，该虚拟 NUMA 节点拓扑包括多个虚拟 NUMA 节点（606-608）；

在计算机系统 (606) 上实例化该虚拟机 (240)，该虚拟机包括多个虚拟 NUMA 节点；以及

根据在该多个虚拟 NUMA 节点的特定虚拟 NUMA 节点 (606) 中的存储器压力，调节被指派给该特定虚拟 NUMA 节点的客户存储器的量。

2. 权利要求 1 的方法，还包括：

确定在该多个虚拟NUMA节点的第二虚拟NUMA节点（608）中的存储器压力大于预定值；以及

将第二虚拟 NUMA 节点 (608) 迁移到计算机系统 (700) 的第二 NUMA 节点 (704)。

3. 权利要求 1 的方法，其中调节客户存储器的量还包括：

从该特定虚拟 NUMA 节点（606）回收客户存储器的至少一个存储区；以及

将客户存储器的被回收的至少一个存储区调拨给第二虚拟 NUMA 节点（608）。

4. 权利要求 1 的方法，其中调节客户存储器的量还包括：

确定该特定虚拟 NUMA 节点（606）的客户存储器的至少一个存储区与系统存储器解除关联；以及

将客户存储器的该至少一个存储区映射到系统存储器的至少一个存储区上。

5. 权利要求 1 的方法，还包括：

将该特定虚拟 NUMA 节点（606）映射到该计算机系统的第一 NUMA 节点（702）上；以及

将该特定虚拟 NUMA 节点（606）迁移到该计算机系统的第二 NUMA 节点上（704）。

6. 权利要求 1 的方法，还包括：

将虚拟处理器（230B）添加到该特定虚拟 NUMA 节点（606）。

7. 权利要求 1 的方法，还包括：

接收对于执行虚拟机（240）的虚拟处理器（230A）的请求，该虚拟处理器（230A）被指派给逻辑处理器（212A），该逻辑处理器（212A）被指派给 NUMA 节点，以及该虚拟处理器被指派给虚拟 NUMA 节点（702）；

确定逻辑处理器（230A）不能执行该虚拟处理器（212A）；以及

选择第二逻辑处理器（230E）来执行虚拟处理器（230A），该第二逻辑处理器来自第二 NUMA 节点（704）。

8. 一种计算机系统，包括：

用于执行虚拟机（240）的电路，该虚拟机（240）具有包括多个虚拟NUMA节点（606-608）的拓扑，其中虚拟机（240）的拓扑是与计算机系统（700）的物理拓扑相独立地生成的；

用于确定在该多个虚拟 NUMA 节点的每个虚拟 NUMA 节点（606-608）中的存储器压力的电路；以及

用于根据在该多个虚拟NUMA节点的每个虚拟NUMA节点（606-608）中的存储器压力，调节被指派给该多个虚拟 NUMA 节点的至少一个虚拟 NUMA 节点（606）的客户存储器的电路。

9. 权利要求 8 的计算机系统，还包括：

用于发送该虚拟机(240)到第二计算机系统(510)的电路。

10. 权利要求 8 的计算机系统,还包括 :

用于将该多个虚拟 NUMA 节点的第一虚拟 NUMA 节点(606)映射到计算机系统(700)的第一 NUMA 节点(702)上的电路 ;以及

用于将该多个虚拟 NUMA 节点的第二虚拟 NUMA 节点(608)映射到计算机系统(700)的所述第一 NUMA 节点(702)上的电路。

11. 权利要求 8 的计算机系统,还包括 :

用于将该多个虚拟 NUMA 节点的第一虚拟 NUMA 节点(606)映射到计算机系统(700)的第一 NUMA 节点(702)上的电路 ;以及

用于将该多个虚拟 NUMA 节点的第二虚拟 NUMA 节点(608)映射到计算机系统(700)的第二 NUMA 节点(704)上的电路。

12. 权利要求 10 的计算机系统,还包括 :

用于确定在第二虚拟 NUMA 节点(608)中的存储器压力大于预定值的电路 ;以及

用于将第二虚拟 NUMA 节点 (608) 迁移到计算机系统的第二 NUMA 节点 (704) 的电路。

13. 权利要求 11 的计算机系统,还包括 :

用于确定第二虚拟机(242)的存储器压力大于预定值的电路 ;以及

用于将虚拟机(240)的第二虚拟 NUMA 节点 (608) 迁移到计算机系统(700)的第一 NUMA 节点 (704) 的电路。

14. 一种包括处理器可执行指令的计算机可读存储介质,所述计算机可读存储介质包括 :

用于执行第一虚拟机(240)的指令,该虚拟机(240)具有包括多个虚拟 NUMA 节点(606-608)的拓扑,该多个虚拟 NUMA 节点的每个虚拟 NUMA 节点包括虚拟处理器(230A-C)和客户物理地址,其中该虚拟机(240)的拓扑是与计算机系统(700)的物理拓扑相独立地生成的 ;以及

用于将附加的虚拟处理器(230D)添加到该多个虚拟 NUMA 节点中的虚拟 NUMA 节点(608)的指令。

15. 权利要求 14 的计算机可读存储介质,还包括 :

用于确定在该多个虚拟 NUMA 节点的每个虚拟 NUMA 节点(606-608)中的存储器压力的指令 ;以及

用于根据在该多个虚拟NUMA节点的每个虚拟NUMA节点(606-608)中的存储器压力,调节被指派给该多个虚拟 NUMA 节点的至少一个虚拟 NUMA 节点(606)的客户存储器的指令。

16. 权利要求 14 的计算机可读存储介质,还包括 :

用于从该多个虚拟 NUMA 节点中的虚拟 NUMA 节点(608)去除虚拟处理器(230D)的指令。

17. 权利要求 14 的计算机可读存储介质,还包括 :

用于将对于该多个虚拟 NUMA 节点 (606-608) 的 NUMA 比率报告给客户操作系统的指令。

18. 权利要求 14 的计算机可读存储介质,还包括 :

用于将虚拟机(240)从第一 NUMA 节点(606)迁移到多个 NUMA 节点(606-608)的指令。

19. 权利要求 15 的计算机可读存储介质，其中用于调节客户存储器的指令还包括：

用于根据虚拟 NUMA 节点（606）的当前存储器压力低于目标阈值的确定，从多个虚拟 NUMA 节点的第一虚拟 NUMA 节点（606）回收存储器的指令。

20. 权利要求 15 的计算机可读存储介质，其中用于调节客户存储器的指令还包括：

用于根据客户操作系统（220）的当前存储器压力大于目标阈值的确定，将存储器调拨给该多个虚拟 NUMA 节点的第一虚拟 NUMA 节点（606）的指令。

用于虚拟机的虚拟非一致存储器体系结构

背景技术

[0001] 虚拟化技术考虑了在多个分区之间共享硬件资源，每个分区可以宿有(host)客户操作系统。通常，虚拟机技术可被使用来整合服务器和增加它们的可移植性。随着虚拟机变得越来越大且随着它们的工作负荷增加，轻松地整合它们和/或将它们从一个计算机系统迁移到另一个计算机系统的能力变得更加困难。因此，用于提高整合和/或迁移更大虚拟机的能力的技术是所希望的。

发明内容

[0002] 本公开内容的示例性实施例描述了一种方法。在本例中，该方法包括，但不限于：接收对于实例化(instantiate)虚拟机的请求，所述请求包括用于虚拟机的特征；根据所述特征选择用于虚拟机的虚拟 NUMA 节点拓扑，该虚拟 NUMA 节点拓扑包括多个虚拟 NUMA 节点；在计算机系统上实例化该虚拟机，虚拟机包括多个虚拟 NUMA 节点；以及根据在该多个虚拟 NUMA 节点的特定虚拟 NUMA 节点中的存储器压力调节被指派给所述特定虚拟 NUMA 节点的客户存储器的量。除了上述的以外，在形成本公开内容的一部分的权利要求、附图和文本中描述了其它方面。

[0003] 本公开内容的示例性实施例描述了一种方法。在本例中，该方法包括，但不限于：执行虚拟机，该虚拟机具有包括多个虚拟 NUMA 节点的拓扑，其中虚拟机的拓扑是与计算机系统的物理拓扑相独立地生成的；确定在该多个虚拟 NUMA 节点的每个虚拟 NUMA 节点中的存储器压力；以及根据在该多个虚拟 NUMA 节点的每个虚拟 NUMA 节点中的存储器压力，调节被指派给该多个虚拟 NUMA 节点的至少一个虚拟 NUMA 节点的客户存储器。除了上述的以外，在形成本公开内容的一部分的权利要求、附图和文本中描述了其它方面。

[0004] 本公开内容的示例性实施例描述了一种方法。在本例中，该方法包括，但不限于：执行第一虚拟机，该虚拟机具有包括多个虚拟 NUMA 节点的拓扑，该多个虚拟 NUMA 节点的每个虚拟 NUMA 节点包括虚拟处理器和客户物理地址，其中虚拟机的拓扑是与计算机系统的物理拓扑相独立地生成的；以及将附加的虚拟处理器添加到该多个虚拟 NUMA 节点中的虚拟 NUMA 节点。除了上述的以外，在形成本公开内容的一部分的权利要求、附图和文本中描述了其它方面。

[0005] 本领域技术人员可以意识到，本公开内容的一个或多个各种各样的方面可包括，但不限于：用于实现本公开内容的这里提到的诸方面的电路和/或编程；所述电路和/或编程可以实际上是被配置来根据系统设计者的设计选择而实现这里提到的诸方面的硬件、软件和/或固件的任何组合。

[0006] 上述的内容是概要，因此必然包含细节的简化、一般化和省略。本领域技术人员应意识到，所述概要仅仅是说明性的，且无论如何不会被定为是限制性的。

附图说明

[0007] 图 1 描绘其中可以实施本公开内容的诸方面的示例性计算机系统。

[0008] 图 2 描绘用于实践本公开内容的诸方面的操作环境。

[0009] 图 3 描绘用于实践本公开内容的诸方面的操作环境。

[0010] 图 4 描绘在本公开内容的实施例中可以如何安排存储器。

[0011] 图 5 描绘实践本公开内容的诸方面的示例性操作环境。

[0012] 图 6 描绘实践本公开内容的诸方面的示例性操作环境。

[0013] 图 7 描绘实践本公开内容的诸方面的示例性操作环境。

[0014] 图 8 描绘图示本公开内容的诸方面的示例性框图。

[0015] 图 9 描绘用于实践本公开内容的诸方面的操作程序。

[0016] 图 10 描绘图 9 的操作程序的替换实施例。

[0017] 图 11 描绘用于实践本公开内容的诸方面的操作程序。

[0018] 图 12 描绘图 11 的操作程序的替换实施例。

[0019] 图 13 描绘图 12 的操作程序的替换实施例。

[0020] 图 14 描绘图 12 的操作程序的替换实施例。

[0021] 图 15 描绘用于实践本公开内容的诸方面的操作程序。

[0022] 图 16 描绘图 15 的操作程序的替换实施例。

[0023] 图 17 描绘图 16 的操作程序的替换实施例。

具体实施方式

[0024] 实施例可以在一个或多个计算机上执行。图 1 和以下的讨论用来提供对于在其中可以实施本公开内容的适当计算环境的概略的一般性描述。本领域技术人员可以意识到，图 1 的计算机系统在某些实施例中可以实现计算机系统 200、300、600 和 700。在这些示例性实施例中，计算机系统可包括图 1 上描述的某些或所有的部件以及被配置成实例化本公开内容的诸方面的电路。

[0025] 在本公开内容各处中使用的术语电路可包括硬件部件，诸如硬件中断控制器、硬驱动、网络适配器、图形处理器、基于硬件的视频 / 音频编解码器，以及被使用来操作这样的硬件的固件 / 软件。在相同的或其它的实施例中，术语电路可包括微处理器，其被配置成通过固件或以某种方式通过开关组（switch set）而执行功能。在相同的或其它的示例性实施例中，术语电路可包括一个或多个逻辑处理器，例如多核通用处理单元的一个或多个核。在本例中，逻辑处理器可以通过软件指令而被配置，所述软件指令具体化能够运行来执行从存储器 -- 例如 RAM、ROM、固件和 / 或虚拟存储器 -- 装载的功能的逻辑。在其中电路包括硬件和软件的组合的示例性实施例中，实施者可以编写具体化逻辑的源代码，其随后被汇编成可以由逻辑处理器处理的机器可读代码。由于本领域技术人员可以意识到现有技术水平已发展到在硬件、软件、或硬件 / 软件的组合之间几乎没有差别的程度，所以，对于实现功能来说，硬件相对（versus）软件的选择仅仅是一种设计选择。因此，由于本领域技术人员可以意识到，软件处理可以转换成等价的硬件结构、以及硬件结构本身也可以转换成等价的软件处理，所以硬件实现相对软件实现的选择是一种设计选择，并留给实施者处理。

[0026] 现在参照图 1，图上描绘示范性通用计算系统。通用计算系统可包括常规的计算机 20 等等，其包括逻辑处理器 21、系统存储器 22 和系统总线 23，系统总线 23 将包括系统存储器在内的各种系统部件耦合到逻辑处理器 21。系统总线 23 可以是几种类型的总线结构的

任何一种，所述总线结构包括存储器总线或存储器控制器、外围总线和使用各种总线体系结构中任何总线体系结构的本地总线。系统存储器可包括只读存储器 (ROM)24 和随机存取存储器 (RAM)25。基本输入输出系统 26（BIOS）被存储在 ROM 24 中，其包含诸如在启动期间帮助在计算机 20 内的单元之间传送信息的基本例行程序。计算机 20 还可包括 :硬盘驱动 27，用于从硬盘（未示出）读出或向其写入 ;磁盘驱动 28，用于从可拆卸磁盘 29 读出或向其写入 ;以及光盘驱动 30，用于从可拆卸光盘 31——诸如 CD ROM 或其它光学介质——读出或向其写入。硬盘驱动 27、磁盘驱动 28 和光盘驱动 30 被显示为分别通过硬盘驱动接口 32、磁盘驱动接口 33 和光盘驱动接口 34 而连接到系统总线 23。驱动和它们的相关联的计算机可读存储介质提供计算机可读指令、数据结构、程序模块和用于计算机 20 的其它数据的非易失性存储。虽然这里描述的示范性环境采用硬盘、可拆卸磁盘 29 和可拆卸光盘 31，但本领域技术人员应当意识到，可以存储计算机可访问数据的其它类型的计算机可读存储介质，诸如盒式磁带、快闪存储卡、数字视频盘、Bernoulli（伯努利）盒式磁带、随机存取存储器 (RAM)、只读存储器 (ROM) 等等，也可以在本示范性操作环境下被使用。通常，在某些实施例中，这样的计算机可读存储介质可被使用来存储体现本公开内容的诸方面的处理器可执行指令。

[0027] 在硬盘、磁盘 29、光盘 31、ROM 24 或 RAM 25 上可以存储若干程序模块，包括操作系统 35、一个或多个应用程序 36、其他程序模块 37 和程序数据 38。用户可以通过诸如键盘 40 和指示装置 42 之类的输入设备将命令和信息输入到计算机 20。其它输入设备（未示出）可包括话筒、操纵杆、游戏手柄（game pad）、碟型卫星天线、扫描器等等。这些和其它输入设备常常通过被耦合到系统总线的串行端口接口 46 而被连接到逻辑处理器 21，但也可以通过诸如并行端口、游戏端口或通用串行总线 (USB) 之类的其它接口被连接。显示器 47 或其它类型的显示设备也可以经由诸如视频适配器 48 之类的接口被连接到系统总线 23。除了显示器 47 以外，计算机典型地包括诸如扬声器和打印机之类的其它外围输出设备（未示出）。图 1 的示范性系统还包括主机适配器 55、小计算机系统接口 (SCSI) 总线 56 和被连接到 SCSI 总线 56 的外部存储装置 62。

[0028] 计算机 20 可以通过使用与一个或多个远程计算机——诸如远程计算机 49——的逻辑连接而在联网的环境下运行。远程计算机 49 可以是另一个计算机、服务器、路由器、网络 PC、对等设备或其它公共网络节点，且典型地可包括以上相对于计算机 20 描述的许多或所有的单元，虽然在图 1 上仅仅图示了存储器存储装置 50。图 1 所描绘的逻辑连接可包括局域网 (LAN)51 和广域网 (WAN)52。这样的联网环境在办公室、企业的广的计算机网络、内联网和互联网中是很平常的。

[0029] 当在 LAN 联网环境下被使用时，计算机 20 可通过网络接口或适配器 53 被连接到 LAN 51。当在 WAN 联网环境下被使用时，计算机 20 典型地可包括调制解调器 54 或用于建立在诸如互联网那样的广域网 52 上的通信的其它装置。调制解调器 54-- 可以是内部的或外部的 -- 可以经由串行端口接口 46 被连接到系统总线 23。在联网的环境下，相对于计算机 20 描绘的程序模块或部分程序模块可被存储在远程存储器存储装置中。应意识到，所显示的网络连接是示范性的，以及可以使用在计算机之间建立通信链路的其它手段。而且，虽然设想本公开内容的许多实施例特别适合于计算机化的系统，但本文档中没有什么要被确定为将公开内容限制于这样的实施例。

[0030] 现在参照图2和3，它们描绘计算机系统的高级别框图。如图所示，计算机系统200可包括物理的硬件设备，诸如举例而言硬驱动的存储装置208、网络接口控制器(NIC)210、图形卡234、至少一个逻辑处理器212和随机存取存储器(RAM)214。计算机系统200还可包括与图1的计算机20类似的部件。虽然只图示了一个逻辑处理器，但在其他实施例中，计算机系统200可以具有多个逻辑处理器，例如每个处理器多个执行核，和/或多个处理器，可以每个具有多个执行核。继续图2的说明，图上描绘了超级监视器(hypervisor)202，在本领域中它也可以被称为虚拟机监视器。在所描绘的实施例中，超级监视器202包括可执行指令，用于控制和仲裁对于计算机系统200的硬件的接入。广义地，超级监视器202可以生成被称为分区的执行环境，诸如子分区1到子分区N(其中N是大于1的整数)。在实施例中，子分区可被认为是由超级监视器202支持的隔离(isolation)的基本单元，也就是，每个子分区可被映射到一组硬件资源，例如存储器、设备、逻辑处理器循环等，其处在超级监视器202和/或母分区的控制下。在实施例中，超级监视器202可以是独立的软件产品、操作系统的一部分、被嵌入在母板的固件内、专用的集成电路或它们的组合。

[0031] 在描绘的例子中，计算机系统200包括母分区204，它也可以被认为是在开放源社区(open source community)中的域0。母分区204可被配置成通过使用虚拟化服务提供器228(VSP)--它们在开放源社区中也被称为后端驱动器--而提供资源给在子分区1-N中执行的客户操作系统。在这个示例性体系结构中，母分区204可以选通接入(gate access)到底层的(underlying)硬件。广义地，VSP 228可被使用来通过虚拟化服务客户机(VSC)--它们在开放源社区中也被称为前端驱动器--使接口对硬件资源复用。每个子分区可包括一个或多个虚拟处理器，诸如客户操作系统220到222可管理和调度线程来在其上执行的虚拟处理器230到232。通常，虚拟处理器230到232是可执行的指令和相关联的状态信息，其提供了具有特定体系结构的物理处理器的表示。例如，一个虚拟机可以具有虚拟处理器，其具有Intel x86处理器的特征，而另一个虚拟处理器可具有PowerPC处理器的特征。在本例中，虚拟处理器可被映射到计算机系统的逻辑处理器，使得实现虚拟处理器的指令将由逻辑处理器进行支持(back)。因此，在这些示例性实施例中，多个虚拟处理器可以同时在执行，而例如另一个逻辑处理器正在执行超级监视器指令。一般地说，且正如附图所例示的，在分区中的虚拟处理器、各种VSC和存储器的组合可被看作为虚拟机，诸如虚拟机240或242。

[0032] 通常，客户操作系统220到222可包括任何操作系统，诸如举例而言来自Microsoft®、Apple®、开放源社区等等的操作系统。客户操作系统可包括操作的用户/内核模式，且可以具有内核，该内核可包括调度器、存储器管理器等等。每个客户操作系统220到222可以具有相关联的文件系统和客户操作系统本身，文件系统可以具有存储在其上的应用，诸如电子商务服务器、电子邮件服务器等等。客户操作系统220-222可以调度线程来在虚拟处理器230-232上执行，且这样的应用的实例可以被实现。

[0033] 现在参照图3，图上图示可被使用的替换的体系结构。图3描绘与图2的那些部件类似的部件，然而，在本示例性实施例中，超级监视器202可包括虚拟化服务提供器228和设备驱动器224，以及母分区204可包含配置实用程序236。在这个体系结构中，超级监视器202可以执行与图2的超级监视器202相同的或类似的功能。图3的超级监视器202可以是独立的软件产品、操作系统的一部分、被嵌入在母板的固件内，或者超级监视器202的

一部分可以由专用集成电路实现。在本例中，母分区 204 可以具有可被使用来配置超级监视器 202 的指令，然而，硬件接入请求可以由超级监视器 202 处理，而不是被传递到母分区 204。

[0034] 现在参照图 4，图上图示在包括虚拟机的实施例中存储器可以如何安排。例如，诸如计算机系统 200 那样的计算机系统可以具有带有存储器地址的 RAM 214。代替将系统物理存储器地址报告给虚拟机，超级监视器 202 可以将对于该系统物理地址不同的地址——例如客户物理地址 (GPA)——呈递给客户操作系统的存储器管理器。客户操作系统然后可以操控客户物理地址，以及超级监视器 202 保持由 GPA 和 SPA 产生的关系。如图所示，在实施例中，GPA 和 SPA 可被安排在存储区中。广义地，存储区可以包括存储器的一页或多页。在 GPA 与 SPA 之间的关系可以通过影子页表（shadow page table）被保持，诸如，在标题为“Enhanced Shadow Page Table Algorithms”的共同转让的美国专利申请 No. 11/128,665 中描述的那些，该专利申请的内容通过引用的方式整体地合并到此处。在运行时，当客户操作系统将数据存储在块 1 的 GPA 中时，该数据实际上可被存储在系统上的不同 SPA 中，诸如块 6。

[0035] 概略地，图 5 描绘用于实践本公开内容的诸方面的操作环境。例如，若干计算机系统 504-510 可以在数据中心 500 处被耦合在一起（虽然只描绘了四个计算机系统，但本领域技术人员可以意识到，数据中心 500 可包括更多或更少的计算机系统）。所描绘的计算机系统可以具有不同的拓扑，而且，它们可以具有不同的特征，例如，不同的 RAM 的量、不同的 RAM 速度、不同的逻辑处理器的量和 / 或具有不同速度的逻辑处理器。

[0036] 管理系统 502 可以具有类似于图 1 的计算机系统 20 和 / 或计算机系统 200、300、600 或 700 的部件。也就是，在实施例中，管理系统 502 可以是包括下面相对于图 6 或图 7 描述的主题的计算机系统。

[0037] 继续进行附图的总体概述，图 6 描绘具有对称的多重处理（multiprocessing）拓扑 (SMP) 或‘平的（flat）’拓扑的计算机系统 600。通常，SMP 是包括被连接到单个共享存储器的多个处理器的计算机体系结构。在这种安排下，存储器控制器 602 可以管理去往和来自存储器的数据的流。存储器访问相对于每个逻辑处理器 212A-F 可以是一致的 (uniform)，且每个逻辑处理器可以访问全部范围的存储器，即，系统物理地址 622-632。这个拓扑对于具有相对较小数目的处理器的计算机系统适用，但当计算机系统包括许多处理器，全部都竞争对共享的存储器总线的访问时，系统的性能会降低。而且，计算机系统的复杂性大大地增加，这进而又驱使按每个处理器计的价格上升。

[0038] 概略地，计算机系统 600 可包括与计算机系统 200 或 300 相同的或类似的部件。如图所示，计算机系统 600 可以具有经由选通访问 RAM 214 的存储器控制器 602 而被耦合在一起的多个逻辑处理器 212A-212F（虽然描绘了六个逻辑处理器，但计算机系统可以具有更多或更少的逻辑处理器）。类似于以上描述的，每个逻辑处理器 212A-212F 可以具有不同的特征，例如，时钟速度、高速缓冲存储器尺寸等等。在这种安排下，存储器控制器 602 可以管理去往和来自 RAM 214 的数据的流。

[0039] 超级监视器 202 可以被实例化，并且它可以控制计算机系统 600 的硬件。超级监视器 202 可以管理一个或多个虚拟机 240 到 242，每个可具有虚拟 NUMA 节点，诸如虚拟 NUMA 节点 606-612。虚拟 NUMA 节点 606-612 可被使用来通过将虚拟拓扑报告到客户应用或客户

操作系统(诸如客户操作系统 220 和 222)而组织虚拟机的资源。如图所示,每个虚拟 NUMA 节点 606-612 可以具有一个或多个虚拟处理器 230A-D、232A-D 和客户物理地址 614-616 和 618-620。通常,超级监视器 202 可以用一个或多个逻辑处理器和来自 RAM 214 的系统物理地址对每个虚拟 NUMA 节点 606-612 进行支持。也就是,超级监视器 202 可以设置一个或多个逻辑处理器作为概念处理器(idea processor),它们可被使用来运行虚拟处理器线程。

[0040] 概略地,图 7 描绘具有包括 NUMA 节点 702-706 的拓扑的计算机系统 700。具有 NUMA 节点的计算机系统通常可被认为是由较小的计算机系统组成的计算机。在本例中,每个 NUMA 节点 606-612 可包括一个或多个逻辑处理器和本地存储器。在 NUMA 节点里面的存储器被认为是本地存储器,在其它 NUMA 节点中的存储器被认为是远程存储器,因为只有在该 NUMA 节点里面的存储器才被连接到相同的存储器总线。NUMA 节点通过高速缓冲存储器一致性(cache coherency)域互连而被相互连接,其允许在一个 NUMA 节点中的处理器以一致的方式访问在其它 NUMA 节点中的存储器。因此,系统物理地址 622-632 相对于每个处理器是一致的。或换一种说法,系统物理地址 20,000 对于在计算机系统中的每个处理器是相同的。差别在于,对于某些处理器,存储器地址 20,000 是本地存储器地址,例如在它们的 NUMA 节点里面,而对于其它处理器,存储器地址 20,000 是远程的,例如在它们的 NUMA 节点外面。通常,本地存储器可以比远程存储器更快地访问,以及在本地相对远程访问时间之间的关系被称为 NUMA 比率。1 比 2 的 NUMA 比率意味着,与访问本地系统物理地址相比,访问特定的远程系统物理地址要花费两倍的处理器循环。NUMA 通过限制在任一个存储器总线上的处理器数目而减轻由 SMP 系统造成的瓶颈,且通常没有具有相同量的逻辑处理器的 SMP 计算机系统那么昂贵。

[0041] 计算机系统 700 可包括与计算机 200 或 300 相同的或类似的部件。如图所示,在这个操作环境下,计算机系统 700 包括由互连 708 连接的三个 NUMA 节点 702-706(虽然计算机可以具有更多或更少的 NUMA 节点)。如图所例示的,在每个 NUMA 节点内的处理器的数目可以是可变的,以及每个节点可以具有它自己的 RAM。

[0042] 类似于图 7,超级监视器 202 可以控制计算机系统 700 的硬件。当客户操作系统或单片(monolithic)应用引导时,它们可以检测虚拟机 240 和 242 的拓扑,类似于以上描述的。每个虚拟 NUMA 节点 606-612 可被指派以来自可被用来运行虚拟处理器的线程的同一个 NUMA 节点的一个或多个理想处理器(ideal processor)和系统物理地址。

[0043] 虽然计算机系统 600 和 700 被描绘为包括两个虚拟机 240 和 242,但在其他实施例中,它们可以执行更多或更少的虚拟机。而且,虽然每个虚拟机被描绘为具有两个虚拟 NUMA 节点,但在其他实施例中,虚拟机可以具有更多或更少的虚拟 NUMA 节点。另外,虽然虚拟 NUMA 节点被描绘为具有两个虚拟处理器,但在其他实施例中,虚拟 NUM 节点可以具有更多或更少的虚拟处理器。此外每个虚拟 NUMA 节点可以具有与其它虚拟 NUMA 节点不同的拓扑,例如,一个虚拟 NUMA 节点可以具有 4 个虚拟处理器和 8G 字节的 RAM,而另一个虚拟 NUMA 节点可以具有 2 个虚拟处理器和 4G 字节的 RAM。

[0044] 图 8 描绘可被使用于本公开内容的诸方面的环境的框图。如图所示,图示了管理被指派给虚拟机的存储器的部件,该虚拟机可被称为动态存储器虚拟化服务提供器 802(DMVSP),且该部件可被使用来调节虚拟 NUMA 节点可访问的存储器的量。如图所示,DMVSP 820 可以与可被称为虚拟化服务客户机 -- 即,动态存储器虚拟化服务客户机 804 和 / 或

806（DMVSC）--的一个或多个气球膨胀驱动器（ballooning driver）相关联（虽然描绘了每个虚拟 NUMA 节点一个 DMVSC，但在其它实施例中，可以使用每个分区一个 DMVSC）。广义地，DMVSC 804 和 / 或 806 可以提供可被 DMVSP 802 使用来调节虚拟 NUMA 节点的存储器的信息，以及每个 DMVSC 也可以帮助调拨（commit）和回收（de-commit）来自与其相关联的虚拟 NUMA 节点的存储器。DMVSC 804、806 和 DMVSP 802 可以借助于在标题为“Partition Bus”的美国专利申请 No.11/128,647 中描述的虚拟化总线进行通信，该专利申请的内容通过引用的方式整体地合并到此处。另外，DMVSC 和 DMVSP 的另外的方面在标题为“Dynamic Virtual Machine Memory Management”的美国专利申请 No.12/345,469 中描述，该专利申请的内容通过引用的方式整体地合并到此处。

[0045] 继续进行图 8 的说明，描绘了可包括可以管理子分区的工作者进程（worker process）812。工作者进程 812 可以结合可分配存储器给子分区的虚拟化基础结构驱动器 810（VID）一起工作。例如，VID 810 可以建立和除去在客户物理地址与系统物理地址之间的关系。图 8 还描绘了可包括诸如客户操作系统 220 的客户操作系统的分区，所述客户操作系统可包括存储器管理器 808。通常，存储器管理器 808 可以应应用的请求来分配存储器给应用，以及当应用不再需要存储器时释放存储器。

[0046] 以下是描绘进程的实现的一系列流程图。为了易于理解起见，流程图被组织成使得初始流程图经由总体“大图片（big picture）”视点给出实现，且随后的流程图提供另外的补充和 / 或细节。而且，本领域技术人员将会意识到，用虚线画出的操作程序被认为是可选的。

[0047] 现在转到图 9，图上描绘了包括操作 900-910 的、用于实践本公开内容的诸方面的操作程序。操作程序从操作 900 开始，且操作 902 图示接收对于实例化虚拟机的请求，该请求包括用于该虚拟机的特征。例如，并参照图 6 或 7，超级监视器 202 可以接收对于创建虚拟机——诸如虚拟机 240——的请求。例如，所述请求可以从管理系统 502、图 2 或 3 的母分区 204 等等被接收。所述请求可以是对于新的虚拟机的请求，或者它可以是对于实例化以前保存的虚拟机的请求。当虚拟机 240 是新的虚拟机时，虚拟机的特征，例如被指派给虚拟机的 RAM 的量、虚拟处理器的数目、或虚拟机应当具有什么类型的 I/O 设备，可以由例如管理员设置。

[0048] 继续进行图 9 的描述，操作 904 显示 :根据特征选择用于该虚拟机的虚拟 NUMA 节点拓扑，虚拟 NUMA 节点拓扑包括多个虚拟 NUMA 节点。例如，在母分区 204（和 / 或超级监视器 202）中的进程可以根据所接收的特征来确定用于虚拟机 240 的拓扑。例如，母分区 204 可包括标识用于诸如虚拟 NUMA 节点 606 那样的虚拟 NUMA 节点的缺省尺寸的信息。在母分区 204 中的进程可以使用描述该缺省尺寸和想要的特征的信息来确定用于虚拟机 240 的虚拟 NUMA 的数目。在特定的例子中，想要的特征可以是具有 10G 字节的 RAM 的 6 处理器虚拟机。如果虚拟 NUMA 节点的缺省尺寸包括 2 个虚拟处理器和 4G 字节的 RAM，则管理系统 502 可以生成指示虚拟机 240 将包括 3 个虚拟 NUMA 节点的配置文件。

[0049] 在实施例中，缺省虚拟 NUMA 节点尺寸可以由管理员或由管理系统 502 设置。转到图 5，管理系统 502 可以执行一个或多个程序，其可以获得标识在数据中心 500 中的计算机系统 504-510 的物理拓扑的信息，例如，标识每个计算机系统 504-510 具有多少 NUMA 节点（如果有的话）、每个计算机系统 504-510 具有多少 RAM、RAM 的速度、RAM 被如何安排、处理

器速度、每个处理器具有多少核等等的信息。

[0050] 通常，虚拟 NUMA 节点的尺寸影响数据中心 500 中的虚拟机的操作。例如，随着虚拟 NUMA 节点例如在存储器和 / 或处理器方面的尺寸增加，虚拟 NUMA 节点的可移植性减小。换句话说，大的虚拟 NUMA 节点可能使得迁移虚拟机更困难。出现这种情况是因为虚拟 NUMA 节点需要被指派到具有足够的‘平的’资源来实现虚拟 NUMA 节点的计算机系统或 NUMA 节点。如果例如虚拟 NUMA 节点太大，例如，它具有太多的 RAM 或太多的虚拟处理器，则它不能适应于数据中心 500 的较小的 NUMA 节点，因此限制了迁移虚拟机的能力。另外，如果较大的虚拟 NUMA 节点被简单地指派到多个较小的 NUMA 节点，则由于在本地存储器与远程存储器访问时间之间存在的差别，虚拟机的性能将降低。

[0051] 另一方面，随着虚拟 NUMA 节点的尺寸减小，客户操作系统的性能可能受到不利的影响。这个低效率可能会出现，因为客户操作系统将试图分离（segregate）应用和它自己对单个虚拟 NUMA 节点的执行。客户操作系统在这种情形下将受到限制，并且性能将降低。

[0052] 因此，在实施例中，管理系统 502 可以通过确定对于数据中心 500 的最佳虚拟 NUMA 节点尺寸，而在可移植性与效率之间权衡利弊。例如，在实施例中，管理系统 502 的逻辑处理器可以执行程序且确定数据中心中的 NUMA 节点的平均尺寸，例如，逻辑处理器的平均数目、RAM 的平均量等等，并将虚拟 NUMA 节点的尺寸设置为与系统中的平均 NUMA 节点相同或小于系统中的平均 NUMA 节点。在另一个实施例中，程序可被配置成 :将虚拟 NUMA 节点的尺寸设置为稍微小于数据中心 500 中的最小的 NUMA 节点。在实施例中，虚拟 NUMA 节点的尺寸可被设置为稍微小于平均尺寸或最小的尺寸，以使得如果计算机系统变为大量地调拨，则一个以上的虚拟 NUMA 节点可被指派到单个 NUMA 节点。在特定的例子中，如果最小的 NUMA 节点具有 4 个逻辑处理器和 8G 字节的 RAM，则虚拟 NUMA 节点的尺寸可被设置为例如 2 个虚拟处理器和 4G 字节的 RAM。

[0053] 操作 906 显示在计算机系统上实例化该虚拟机，虚拟机包括多个虚拟 NUMA 节点。在实施例中，超级监视器 202 可以由逻辑处理器执行，以及具有多个虚拟 NUMA 节点的虚拟机可以被实例化。例如，并参照图 6 和 / 或图 7，具有虚拟 NUMA 节点 606-608 的虚拟机 240 可以由计算机系统 600 或 700 来实现。也就是，VID 810 可以用来自 RAM 的系统物理地址对虚拟机 240 的客户物理地址进行支持，以及用一个或多个逻辑处理器对虚拟处理器进行支持。例如，客户物理地址块 614 可以用系统物理地址块 622 进行支持，以及客户物理地址块 616 可以用系统物理地址块 624 进行支持。超级监视器线程然后可以在对虚拟处理器进行支持的逻辑处理器上被调度，以及可以执行指示虚拟处理器的指令。如图 6 和图 7 所示，每个虚拟机的拓扑可以与底层硬件的拓扑相独立地被创建。也就是，每个虚拟机的拓扑是与实现它的计算机系统的底层物理拓扑相分离的。

[0054] 在实施例中，虚拟机 BIOS 或引导固件可以向单片应用的客户操作系统描述虚拟机的拓扑，例如，它是否具有虚拟 NUMA 节点，任何虚拟 NUMA 节点的尺寸，以及对于虚拟 NUMA 节点的 NUMA 比率。数据结构可以被处理并且客户 OS 220 或应用以及它可以被所述 OS 或应用使用以利用虚拟 NUMA 节点的存在。例如，客户操作系统 220 可以试图将不是 NUMA 感知（NUMA aware）的应用的线程与虚拟 NUMA 节点建立亲缘关系（affinitize），以使得应用的执行保持为本地的。在另一个例子中，数据库管理程序，诸如 SQL 服务器，可以分配对于虚拟 NUMA 节点而言是本地的锁，并且数据库可以跨多个虚拟 NUMA 节点上来分解读 / 写请

求。在再一个例子中，客户操作系统 220 可以创建用于虚拟机中的每个虚拟 NUMA 节点的页池（page pool）。

[0055] 继续进行图 9 的说明，操作 908 显示：根据在多个虚拟 NUMA 节点的特定虚拟 NUMA 节点中的存储器压力来调节被指派给该特定虚拟 NUMA 节点的客户存储器的量。例如，逻辑处理器，举例而言图 6 或 7 的逻辑处理器 212A，可以执行指示 DMVSP 802 的指令，并且可以调节诸如虚拟 NUMA 节点 606 那样的虚拟 NUMA 节点可得到的客户物理地址的量。也就是，可以执行 DMVSP 802，并且可以根据虚拟 NUMA 节点正经受的压力而调拨或回收存储器。

[0056] 在实施例中，存储器压力可以标识客户操作系统 220 的性能如何受每个虚拟 NUMA 节点 606-608 可得到的存储器的量影响。这个信息可以在客户操作系统 220 的运行时期间由例如像 DMVSC 804 和 / 或 806 那样的 DMVSC 进行计算，并被发送到 DMVSP 802。例如，存储器压力可以由一系列值来代表，所述值可以标识在虚拟 NUMA 节点中不同的存储器压力水平。随着虚拟 NUMA 节点中的资源变得压力更大，即，随着对于有效执行在虚拟 NUMA 节点上的当前工作负荷所需要的存储器量增加，DMVSC 804 可以修订所述值，并把这个信息传送到 DMVSP 802。

[0057] 在实施例中，存储器压力信息可以由 DMVSC 804 从接收自客户操作系统 220 的信息进行计算。例如，DMVSC 804 可被配置成：从存储器管理器 808 接收对于虚拟 NUMA 节点 606 的操作系统分页信息。客户操作系统分页速率可以通过由存储器管理器 808 和高速缓冲存储器管理器暴露（expose）的两个计数器来监视，即，分页速率和高速缓冲存储器旋转速率。

[0058] 在同一个或另一个实施例中，DMVSC 804 可以接收来自存储器管理器 808 的、与虚拟 NUMA 节点 606 相关联的物理存储器通知，并使用这个信息来计算虚拟 NUMA 节点 606 的存储器压力。例如，存储器管理器 808 可以根据与虚拟 NUMA 节点 606 相关联的客户操作系统 220 中的活动，来输出高存储量通知和低存储量通知。存储器管理器 808 可以根据低存储量阈值（LMT）和高存储量阈值（HMT）来触发这些通知。在特定的示例性实施例中，以信号告知低存储量资源通知事件的可得到存储器的缺省水平可以是每 4GB 大约 32MB，到 64MB 的最大值。以信号告知高存储量资源通知事件的缺省水平，例如，可以是缺省的低存储量值的三倍。在这二者之间的中间存储器可用性水平可以通过在高存储量阈值和低存储量阈值水平之间划分出间隔而被确定。本领域技术人员可以意识到，这些值是示范性，并且可以作出改变而不背离本公开内容的精神。

[0059] 这些通知，连同其它的一起，可以被 DMVSC 804 使用来计算虚拟 NUMA 节点 606 的存储器压力。例如，每个水平可以与例如 0-4 的一个值相关联，以及如果考虑任何其它性能计数器，那么它们也可以与值相关联。用于每个性能计数器的值然后可被使用来计算虚拟 NUMA 节点 606 的当前的存储器压力。在特定的例子中，存储器压力可以通过取性能计数器值的较高或较低的值而进行计算。在另一个例子中，性能计数器的平均值可被用作为存储器压力。在再一个实施例中，可以使用更精巧的算法来计算存储器压力，它考虑了先前的性能计数器值，并给每个性能计数器指派一个标量以影响它在计算中的相对权重。

[0060] 当作出调拨存储器的决定时，DMVSP 802 可以使用各种各样的技术，其中的一个技术是热添加（hot-add）操作。例如，某些操作系统可以支持热添加，这允许成系列的物理存储器被添加到正运行的操作系统而不需要系统重新引导。也就是，存储器管理器 808 可被

配置成支持将存储器动态添加到正运行的系统。在热添加环境下，DMVSC 804 可被配置成：接入存储器管理器 808 的热添加接口，以及 DMVSC 804 可以将描述热添加的 GPA 和它们与哪个虚拟 NUMA 节点相关联的消息发送到客户操作系统 220。存储器管理器 808 然后可以使得新的存储器对于客户操作系统 220、驱动器、应用、或正在虚拟 NUMA 节点 606 上运行的任何其它进程是可得到的。例如，在 VID 810 生成在 GPA 与 SPA 之间的关系后，DMVSC 804 可以接收来自 DMVSP 802 的热添加的存储器地址。

[0061] 同样地，热去除（hot-remove）操作可被使用来从诸如虚拟 NUMA 节点 606 那样的虚拟 NUMA 节点中回收存储器地址。例如，DMVSC 804 可以将指示存储器已被热去除的消息发送到客户操作系统 220。DMVSC 804 可以请求：存储器管理器 808 提供来自虚拟 NUMA 节点 606 的 GPA 块以便去除。在本例中，DMVSC 804 然后可以调用存储器管理器 808 的去除 API，并从客户操作系统 220 中去除 GPA。在使用热去除的实施例中，被去除的存储器没有不利于（count against）客户当前的调拨，以及存储器管理器 808 可以使用由操作系统用来去除从母板上被物理地去除的存储器的类似技术，来调节它的内部计数器。

[0062] 在另一个实施例中，可以通过使用气球膨胀（Ballooning）技术而将存储器回收到虚拟 NUMA 节点。也就是，存储器可以通过使虚拟 NUMA 节点 606 中的客户物理地址与对它们进行支持的物理地址解除关联（dissociate）而被回收。例如，逻辑处理器 212B 可以执行指示 DMVSC 804 的指令，并且可以将请求存储器管理器 808 保留一定量的存储器 -- 例如一个或多个存储区 -- 供 DMVSC 804 使用的消息发送到存储器管理器 808。存储器管理器 808 可以锁定存储器来供在 DMVSC 804 内的专有使用，以及 DMVSC 804 可以发送存储器的 GPA 到 DMVSP 802。在本例中，DMVSP 802 可以发送 GPA 到 VID 810，以及 VID 810 可以在影子页表中去除用于这些 GPA 到 SPA 的条目。在本例中，存储器管理器 808 可包括标识 GPA 仍旧有效的信息，然而，实际上 GPA 不再由系统物理地址进行支持。在本例中，存储器管理器 808 将不使用锁定的 GPA，以及对它们进行支持的 SPA 可被重新分配。

[0063] 已经解除关联的客户物理地址可以与物理地址重新关联。在本例中，对于调拨存储页面的请求可以由 VID 810 接收，以及 VID 810 可以获得 SPA 以满足该请求，并把地址范围发送到 DMVSP 802。在实施例中，VID 810 可被配置成获得邻接范围的 SPA，以便提高系统效率。在本例中，VID 810 可以确定：客户操作系统 220 具有被锁定以供与虚拟 NUMA 节点 606 相关联的 DMVSC 804 专有使用的 GPA。VID 810 可以创建在被锁定的 GPA 与 SPA 之间的关系，并发送消息到 DMVSP 802。DMVSP 802 然后可以发送消息到 DMVSC 804，以及 DMVSC 804 可以发送消息给存储器管理器 808，指示 GPA 可被解锁并被归还到与虚拟 NUMA 节点 606 相关联的存储器管理器 808 的存储器池。

[0064] 在实施例中，VID 810 可以根据 GPA 是否气球般膨胀，而确定是使用热添加技术还是使用气球膨胀技术。例如，当 VID 810 接收 SPA 以调拨给虚拟 NUMA 节点 606 时，它可以确定是否有任何 GPA 被 DMVSC 804 锁定。在有 GPA 被锁定的情形下，VID 810 可以在它热添加存储器之前用 SPA 对它们进行支持。在存储器被调拨给虚拟 NUMA 节点 606 之前，它可以被置零（zero），以及它的相关联的高速缓冲存储器线为了安全原因可被清洗。通过使存储器置零，先前与一个分区相关联的存储器的内容不泄漏到另一个分区。

[0065] 现在转到图 10，图上描绘图 9 的操作程序的替换实施例，其包括附加的操作 1010-1020。操作 1010 显示：确定在多个虚拟 NUMA 节点的第二虚拟 NUMA 节点中的存储器

压力大于预定值；以及将第二虚拟 NUMA 节点迁移到计算机系统的第二 NUMA 节点。例如，且转到图 7，在实施例中，在第二虚拟 NUMA 节点 608 中的存储器压力可以增加。也就是，指示存储器压力的值可以由 DMVSP 802 接收，其指示虚拟 NUMA 节点 608 压力过大。在本例中，虚拟机 240 或个体虚拟 NUMA 节点 608 可以具有目标压力值，以及当前的压力值可能比管理员设置的目标值大。目标压力值可被存储在可由 DMVSP 802 访问的数据结构中。然后可以接收运行的虚拟机或虚拟 NUMA 节点的当前压力值。DMVSP 802 可以连续地步进通过（step through）运行的虚拟机或虚拟 NUMA 节点的列表，并调拨存储器以便把存储器压力值减小到目标值，以及回收存储器以便把压力增加到目标值。

[0066] 在例子中，DMVSP 802 可被配置成：确定当前宿有虚拟 NUMA 节点 606 和 608 的 NUMA 节点，例如 NUMA 节点 702，不能分配足够的存储器以便获得用于这两个虚拟 NUMA 节点的目标存储器压力值。在本例中，DMVSP 802 可被配置成：发送信号到超级监视器 202，以及超级监视器 202 可被配置成企图将虚拟 NUMA 节点之一移出 NUMA 节点 702。超级监视器 202 可以检查 NUMA 节点 702-706 的当前的工作负荷，并确定：例如 NUMA 节点 704 可以宿有该虚拟 NUMA 节点，并分配足够的资源给它以便将存储器压力减小到目标值。在本例中，超级监视器 202 可被配置成重新指派虚拟 NUMA 节点 608 给 NUMA 节点 704。也就是，超级监视器 202，与 VID 810 相结合地，可以将客户物理地址 616 重新映射到系统物理地址 714，以及把逻辑处理器 212E 和 F 设置为用于虚拟处理器 230C 和 D 的理想处理器。

[0067] 继续进行图 10 的说明，操作 1012 图示：从特定的虚拟 NUMA 节点回收客户存储器的至少一个存储区；以及将客户存储器的被回收的该至少一个存储区调拨给第二虚拟 NUMA 节点。例如，DMVSP 802 可被配置成：从例如虚拟 NUMA 节点 606 回收存储器，以及将该存储器调拨给虚拟 NUMA 节点 608。在本例中，虚拟 NUMA 节点 606 和 608 可以用单个 NUMA 节点或‘平的’体系结构进行支持。在本示例性实施例中，当例如没有可被调拨给虚拟 NUMA 节点 608 的可用存储器可得到时，DMVSP 802 可以尝试从虚拟 NUMA 节点 606 释放存储器。在另一个例子中，DMVSP 802 可被配置成：从例如虚拟 NUMA 节点 610 回收存储器，并把该存储器调拨给虚拟 NUMA 节点 608。也就是，可以从一个虚拟机取来存储器，并将它给予另一个虚拟机。

[0068] 在特定的例子中，并参照图 6，虚拟 NUMA 节点 606 和 608 可被映射到计算机系统 600 的资源。在本例中，DMVSP 802 可以检查其它虚拟 NUMA 节点，例如按照存储器优先权的次序，例如，从虚拟机 240 中的低优先权虚拟 NUMA 节点开始，或从最低优先权虚拟机开始。如果例如检测到一个虚拟 NUMA 节点，诸如虚拟 NUMA 节点 606，具有小于目标阈值的存储器压力值，则 DMVSP 802 可以发起存储器回收，并从虚拟 NUMA 节点 606 中去除存储器。在回收完成后，可以发起调拨操作，以及存储器可被热添加到虚拟 NUMA 节点 608，或者气球般膨胀的客户物理地址可以与系统物理地址重新关联。

[0069] 在特定的例子中，且参照图 7，DMVSP 802 可以例如以存储器优先权的次序检查由相同的 NUMA 节点 702 进行支持的其它虚拟 NUMA 节点。如果例如在与虚拟 NUMA 节点 608 相同的 NUMA 节点上的虚拟 NUMA 节点被检测到具有小于目标阈值的存储器压力值，则 DMVSP 802 可以发起存储器回收。在回收完成后，可以发起调拨操作，以及存储器可被热添加到虚拟 NUMA 节点 608，或者气球般膨胀的客户物理地址可以与系统物理地址重新关联。

[0070] 继续进行图 10 的说明，操作 1014 描绘：确定特定的虚拟 NUMA 节点的客户存储器

的至少一个存储区与系统存储器解除关联；以及将客户存储器的该至少一个存储区映射到系统存储器的至少一个存储区上。例如，在实施例中，DMVSP 802 可以由逻辑处理器执行，以及可以作出决定来用 SPA 624 对虚拟 NUMA 节点 606 中的 GPA 进行支持。例如，GPA 可以由 DMVSC 804 保留，以及 SPA 可以被重新分配给另一个虚拟 NUMA 节点或母分区 204。在本例中，对于调拨存储页面的请求可以被 VID 810 接收，以及 VID 810 可以获得 SPA 以满足该请求，并发送地址范围到 DMVSP 802。在实施例中，VID 810 可被配置成获得邻接范围的 SPA，以便提高系统效率。在 NUMA 实施例中，VID 810 可被配置成从正在运行虚拟 NUMA 节点 606 的相同的 NUMA 节点获得邻接范围的 SPA。VID 810 可以创建在锁定的 GPA 与 SPA 之间的关系，并发送消息到 DMVSP 802。DMVSP 802 然后可以发送消息到 DMVSC 804，以及 DMVSC 804 可以发送消息给存储器管理器 808，指示 GPA 可被解锁并被归还到与虚拟 NUMA 节点 606 相关联的存储器池。

[0071] 继续进行图 10 的说明，操作 1016 图示：将特定的虚拟 NUMA 节点映射到计算机系统的第一 NUMA 节点上；以及将该特定的虚拟 NUMA 节点迁移到计算机系统的第二 NUMA 节点上。例如，并参照图 7，客户操作系统 220 可被散布在至少两个 NUMA 节点上，诸如 NUMA 节点 702 和 704。例如，并参照图 7，超级监视器 202 可以调度虚拟 NUMA 节点 606 和 608 在 NUMA 节点 702 上运行。在本例中，超级监视器 202 可以接收指示 NUMA 节点 702 压力过大的信号。例如，客户操作系统 220 可以生成指示虚拟 NUMA 节点 606 和 608 在存储量方面较低的信号。在本例中，超级监视器 202 可被配置成：通过将虚拟 NUMA 节点 608 移出 NUMA 节点 702 ，而减小在压力过大的 NUMA 节点上的工作负荷。

[0072] 继续进行图 10 的说明，操作 1018 图示：将虚拟处理器添加到特定的虚拟 NUMA 节点。例如，在实施例中，可以在虚拟机 240 的运行时执行期间，通过使用例如处理器热添加操作而添加虚拟处理器，诸如虚拟处理器 230B。也就是，虚拟 NUMA 节点 606 可以在一个点只有单个虚拟处理器 230A，然后添加另一个。在实施例中，可以把新添加的处理器指派给对虚拟处理器 230A 进行支持的处理器，或者可以把另一个逻辑处理器分配来运行虚拟处理器 230B 线程。在 NUMA 实施例中，如果另一个逻辑处理器正在被使用来支持虚拟处理器 230B，则它可以从正在支持虚拟 NUMA 节点 606 中的另外虚拟处理器的同一个 NUMA 节点 702 被分配。

[0073] 继续进行图 10 的说明，操作 1020 图示：接收对于执行虚拟机的虚拟处理器的请求，虚拟处理器被指派给逻辑处理器，逻辑处理器被指派给 NUMA 节点，以及虚拟处理器被指派给虚拟 NUMA 节点；确定逻辑处理器不能执行该虚拟处理器；以及选择第二逻辑处理器来执行该虚拟处理器，第二逻辑处理器来自第二 NUMA 节点。例如，并参照图 7，在实施例中，超级监视器 202 可以从虚拟处理器 230A 接收对于执行虚拟处理器线程的请求，并尝试在概念处理器 212A 上，例如在支持虚拟处理器 230A 的处理器上调度该线程。在本例中，超级监视器 202 可以检测逻辑处理器 212A 被过量调拨，从而不能执行该虚拟处理器线程。在这种情形下，可以执行超级监视器 202，且它可以选择另一个逻辑处理器来执行该虚拟处理器线程。例如，超级监视器 202 可以尝试选择在同一个 NUMA 节点上的不同的逻辑处理器。如果例如 NUMA 节点被过量调拨，则超级监视器 202 可被配置成：选择远程处理器来执行虚拟处理器 230A，例如，逻辑处理器 212E。在本例中，关于是等待还是在远程节点上调度线程的决定，可以通过使用与 NUMA 节点 704 相关联的 NUMA 比率而做出。如果 NUMA 比率是低的，并

且预期的对理想处理器的等待较长，则可以做出在远程节点上调度线程的决定。另一方面，如果 NUMA 比率是高的并且预期的等待时间较低，则可以做出等待的决定。

[0074] 现在转到图 11，图上描绘包括操作 1100、1102、1104 和 1106 的用于实践本公开内容的诸方面的操作程序。操作 1100 开始该操作程序，以及操作 1102 显示执行虚拟机，虚拟机具有包括多个虚拟 NUMA 节点的拓扑，其中虚拟机的拓扑是与计算机系统的物理拓扑相独立地生成的。例如，超级监视器 202 可以执行具有多个虚拟 NUMA 节点的虚拟机。如图 6 所示，可以创建包括虚拟 NUMA 节点 606 和 608 的虚拟机 240。虚拟 NUMA 节点每个可以具有一个或多个虚拟处理器 230A-D 和客户物理地址 614 和 616。在本实施例中，虚拟 NUMA 节点 606 和 608 可以与底层硬件的拓扑相独立地被创建。也就是，虚拟机的拓扑是与诸如由图 6 和图 7 所描绘的底层硬件无关的。因此，在本实施例中，每个虚拟机的拓扑是与实现它的计算机系统的底层物理拓扑相分离的。

[0075] 继续进行图 11 的说明，操作 1104 图示：确定在多个虚拟 NUMA 节点的每个虚拟 NUMA 节点中的存储器压力。例如，并参照图 8，每个虚拟 NUMA 节点 606 和 608 的存储器压力可以被获得，例如被生成和 / 或被接收。存储器压力信息可以标识客户的性能如何受每个虚拟 NUMA 节点 606-608 可得到的存储器的量影响。这个信息可以在客户操作系统 220 的运行时期间例如由诸如 DMVSC 804 和 / 或 806 那样的一个或多个 DMVSC 进行计算，并被发送到 DMVSP 802。也就是，在特定的实施例中，逻辑处理器可以执行指示 DMVSC 804 或 806 的指令，并且为每个虚拟 NUMA 节点生成存储器压力信息。这个信息然后可被发送到例如 DMVSP 802。

[0076] 在示例性实施例中，存储器压力信息可包括范围从 0 到 4 的一系列值，且每个值可以标识由于虚拟 NUMA 节点 606-608 的资源引起的客户 OS 正在经受的不同的存储器压力水平。随着客户操作系统变得压力更大，即，随着对于有效执行当前工作负荷所需要的存储器量增加，DMVSC 804 和 806 可以修订它们的值，并把这个信息传送到 DMVSP 802。

[0077] 继续进行图 11 的说明，操作 1106 显示：根据在多个虚拟 NUMA 节点的每个虚拟 NUMA 节点中的存储器压力，调节被指派给多个虚拟 NUMA 节点中的至少一个虚拟 NUMA 节点的客户存储器。在包括操作 1206 的实施例中，逻辑处理器 212A 可以执行指示 DMVSP 802 的指令，以及可以调节例如在虚拟 NUMA 节点 606 中的客户物理地址的量。也就是，DMVSP 802 可以根据在虚拟 NUMA 节点 606 中的存储器压力而调拨或回收处理器，例如，如果被分配给虚拟 NUMA 节点 606 的进程压力过大，则可以调拨存储器。

[0078] 在实施例中，当由 DMVSP 802 配置的逻辑处理器 212 确定调拨或回收存储器时，它可以以每个存储区为基础做到这一点。例如，DMVSP 802 可以调拨 / 回收存储区，并检查存储器状态如何改变。如果存储器状态没有改变，则 DMVSP 802 可以调拨 / 回收另一个存储区。

[0079] 现在转到图 12，图上描绘图 11 的操作程序 1100 的替换实施例，其包括操作 1208-1216。如图所示，操作 1208 显示发送虚拟机到第二计算机系统。例如，并参照图 5，在实施例中，虚拟机的状态可被保存在一个或多个配置文件中，并被发送到另一个计算机系统，例如，从计算机 504 发送到 506。计算机系统 506 的超级监视器可以读取该文件或该多个文件，并实例化虚拟机。

[0080] 虚拟机拓扑影响迁移和复原虚拟机的能力。具体地，关于允许底层硬件的拓扑被

检测的决定和虚拟 NUMA 节点的尺寸影响虚拟机将执行得多好以及它是否可以轻松地迁移。举例而言，虚拟 NUMA 节点的尺寸影响迁移的能力，例如，随着虚拟 NUMA 节点的尺寸增加，虚拟 NUMA 节点的可移植性降低，而随着虚拟 NUMA 节点的尺寸减小，虚拟机的性能同样降低。另外，可以检测底层计算机的拓扑的虚拟机不能轻松地迁移，这是由于 NUMA 感知的（NUMA aware）操作系统和应用在引导时间根据它们检测到的第一个拓扑来最佳化其本身，且这些最佳化在虚拟机在未来可被迁移到的计算机上可能不适用。因此，通过在客户操作系统引导时将虚拟 NUMA 节点暴露给它，操作系统可被最佳化来使用 NUMA 节点。通过给虚拟 NUMA 节点正确地定尺寸，虚拟机可以对于数据中心 500 中的许多不同的计算机系统进行最佳化。

[0081] 例如，并参照图 6，虚拟机 240 可包括两个或更多个虚拟 NUMA 节点 606 和 608。超级监视器 202 可以用逻辑处理器 212A-D 对虚拟处理器 230A-D 进行支持。当客户操作系统 220 引导时，它可以检测虚拟 NUMA 节点 606 和 608，并且可被配置成使得使用虚拟 NUMA 节点 606 和 608 的进程的调度和执行最佳化。在一段时间后，虚拟机 240 可被迁移到具有类似于图 7 所描绘的物理拓扑的计算机系统。图 7 的超级监视器 202 可以用逻辑处理器 212A 和 B 对虚拟处理器 230A 和 B 进行支持，以及用逻辑处理器 212E 和 F 对虚拟处理器 230C 和 D 进行支持。即使底层计算机拓扑已经从 SMP 改变到 NUMA，客户操作系统 220 仍可以以与它在图 6 的计算机系统上运行时的相同方式继续操作。

[0082] 继续进行图 12 的说明，操作 1210 显示：将多个虚拟 NUMA 节点的第一虚拟 NUMA 节点映射到计算机系统的第一 NUMA 节点上；以及将多个虚拟 NUMA 节点的第二虚拟 NUMA 节点映射到计算机系统的第一 NUMA 节点上。例如，并参照图 7，诸如逻辑处理器 212A 的逻辑处理器可以执行超级监视器指令，并可以使诸如虚拟 NUMA 节点 606 和 608 的虚拟 NUMA 节点与诸如 NUMA 节点 702 的 NUMA 节点建立亲缘关系。更具体地，逻辑处理器可以用来自 NUMA 节点 702 的 RAM 214 的系统物理地址对客户物理地址 614 和 616 进行支持、以及可以用逻辑处理器 212A 到 D 对虚拟处理器 230A 到 D 进行支持。

[0083] 继续进行图 12 的说明，操作 1212 显示：将多个虚拟 NUMA 节点的第一虚拟 NUMA 节点映射到计算机系统的第一 NUMA 节点上；以及将多个虚拟 NUMA 节点的第二虚拟 NUMA 节点映射到计算机系统的第二 NUMA 节点上。例如，并参照图 7，诸如逻辑处理器 212A 的逻辑处理器可以执行超级监视器指令，并可以把虚拟 NUMA 节点 606 指派给 NUMA 节点 702 和把虚拟 NUMA 节点 608 指派给 NUMA 节点 704。在本例中，当超级监视器 202 被执行时，超级监视器调度器可以在逻辑处理器 212A-D 上调度来自虚拟处理器 230A-B 的线程、以及在逻辑处理器 212E 到 G 上调度来自虚拟处理器 230C 或 D 的线程。

[0084] 现在转到图 13，图上图示图 12 的操作程序的替换实施例，其包括操作 1314，该操作 1314 显示：确定在第二虚拟 NUMA 节点中的存储器压力大于预定值；和将第二虚拟 NUMA 节点迁移到计算机系统的第二 NUMA 节点。例如，并转到图 7，在实施例中，在第二虚拟 NUMA 节点 608 中的存储器压力可以增加。也就是，指示存储器压力的值可以由 DMVSP 802 接收，其指示虚拟 NUMA 节点 608 压力过大。在本例中，每个虚拟 NUMA 节点 606-612 和 / 或虚拟机 240-242 可以具有目标压力值，且虚拟 NUMA 节点 608 的当前的压力值可大于由管理员设置的目标值。目标压力值可被存储在可由 DMVSP 802 访问的数据结构中。然后可以接收正在运行的虚拟机或虚拟 NUMA 节点的当前的压力值。DMVSP 802 可以连续地步进通过运行的

虚拟机或虚拟 NUMA 节点的列表，并调拨存储器以便把存储器压力值减小到目标值，以及回收存储器以便把压力增加到目标值。

[0085] 现在转到图 14，图上图示图 12 的操作程序的替换实施例，其包括操作 1416，该操作 1416 显示：确定第二虚拟机的存储器压力大于预定值；和将该虚拟机的第二虚拟 NUMA 节点迁移到计算机系统的第一 NUMA 节点。在实施例中，至少两个虚拟机可以正在执行，例如，虚拟机 240 和 242。在本例中，虚拟机 240 的虚拟 NUMA 节点可被映射到例如图 7 的 NUMA 节点 702 和 704，以及虚拟机 242 的虚拟 NUMA 节点可被映射到例如 NUMA 节点 706。在本例中，每个虚拟机 240 和 242，和 / 或每个虚拟 NUMA 节点 606-612 可以具有可被存储在可由 DMVSP 802 访问的数据结构中的目标压力值。在本例中，在第二虚拟机 242 中的存储器压力，由于在虚拟机中的活动——例如虚拟机 242 接收到许多读 / 写请求——而可能增加，以及该值可被 DMVSP 802 接收。DMVSP 802 可以接收运行的虚拟机或虚拟 NUMA 节点的当前的压力值，并且连续地步进通过运行的虚拟机或虚拟 NUMA 节点的列表，以确定存储器是否可被调拨给虚拟机 242 以便缓和存储器压力。

[0086] 在其中压力不能通过调拨或回收存储器而被减轻的情形下，DMVSP 802 可被配置成：发送信号到超级监视器 202，然后超级监视器 202 可以尝试重新分配计算机系统的资源，以便减轻存储器压力。例如，超级监视器 202 可以检查 NUMA 节点 702-706 的当前的工作负荷，并确定：例如 NUMA 节点 702 可以宿有来自虚拟机 240 的虚拟 NUMA 节点，并将虚拟 NUMA 节点 608 重新指派给 NUMA 节点 702。也就是，超级监视器 202，与 VID 810 相结合，可以将客户物理地址 616 重新映射到系统物理地址 712，并将逻辑处理器 212A 和 D 设置为用于虚拟处理器 230C 和 D 的理想处理器。然后，超级监视器可以将虚拟 NUMA 节点 610 重新映射到 NUMA 节点 704，并调节在虚拟机 242 的每个虚拟 NUMA 节点 610-612 内的存储器，以便减小它的存储器压力。

[0087] 现在转到图 15，图上图示包括操作 1500、1502 和 1504 的、用于实践本公开内容的诸方面的操作程序。操作 1500 开始该操作程序，以及操作 1502 显示执行第一虚拟机，虚拟机具有包括多个虚拟 NUMA 节点的拓扑，该多个虚拟 NUMA 节点的每个虚拟 NUMA 节点包括虚拟处理器和客户物理地址，其中虚拟机的拓扑是与计算机系统的物理拓扑相独立地生成的。例如，图 7 的超级监视器 202 可以执行具有多个虚拟 NUMA 节点 608-610 的虚拟机 240。虚拟 NUMA 节点 606 和 608 每个可以具有一个或多个虚拟处理器 230A-D 和客户物理地址 614 和 616。在本实施例中，虚拟 NUMA 节点 606 和 608 可被映射到计算机系统 700 的资源。例如，逻辑处理器 212A 和 B 可被设置为用于虚拟处理器 230A 和 B 的理想处理器，以及客户物理地址 614 可以由系统物理地址 710 进行支持。同样地，逻辑处理器 212E 和 F 可被设置为用于虚拟处理器 230C 和 D 的理想处理器，以及客户物理地址 616 可以被映射到系统物理地址 714。在本实施例中，虚拟 NUMA 节点 606 和 608 是独立于底层硬件的拓扑的。也就是，虚拟机的拓扑是与诸如由图 6 和图 7 所描绘的底层硬件无关的。因此，在本实施例中，虚拟机的拓扑是与实现它的计算机系统的底层物理拓扑相分离的。

[0088] 继续进行图 15 的说明，操作 1504 显示：将附加虚拟处理器添加到多个虚拟 NUMA 节点的一个虚拟 NUMA 节点。例如，在实施例中，附加虚拟处理器可被加到虚拟 NUMA 节点，诸如举例而言虚拟 NUMA 节点 606。在本例中，在虚拟机 240 的运行时执行期间，可以通过使用例如处理器热添加操作而添加虚拟处理器，诸如虚拟处理器。在实施例中，新添加的处理

器可被指派给对虚拟处理器 230A 进行支持的处理器，或另一个逻辑处理器可被设置为理想处理器来运行新添加的虚拟处理器的线程。在 NUMA 实施例中，逻辑处理器可以从正在对虚拟 NUMA 节点 606 进行支持的同一个 NUMA 节点 702 分配。

[0089] 现在转到图 16，图上描绘由图 15 所描绘的操作程序的替换实施例，其包括附加操作 1606-1612。操作 1606 显示：确定在多个虚拟 NUMA 节点的每个虚拟 NUMA 节点中的存储器压力；和根据在多个虚拟 NUMA 节点的每个虚拟 NUMA 节点中的存储器压力，调节被指派给多个虚拟 NUMA 节点的至少一个虚拟 NUMA 节点的客户存储器。参照图 8，对于每个虚拟 NUMA 节点 606 和 608 的存储器压力可以被获得，例如被生成和 / 或被接收。存储器压力信息可以标识客户的性能如何受每个虚拟 NUMA 节点可得到的存储器的量影响。DMVSC 804 和 806 可被配置成：从存储器管理器 808 接收例如物理存储器通知和 / 或客户操作系统分页信息，并使用它来计算每个虚拟 NUMA 节点 606 和 608 的存储器压力。

[0090] 继续这个例子，逻辑处理器 212A 例如可以执行指示 DMVSP 802 的指令，并可以调节例如在虚拟 NUMA 节点 606 中的客户物理地址的量。也就是，DMVSP 802 根据由于在虚拟 NUMA 节点 606 中的资源引起的客户 OS 正经受的存储器压力，调拨或回收存储器。

[0091] 继续进行图 16 的说明，操作 1608 图示：从多个虚拟 NUMA 节点的一个虚拟 NUMA 节点去除虚拟处理器。在包括 1706 的实施例中，超级监视器 202 可以由逻辑处理器执行，且虚拟处理器可以从虚拟 NUMA 节点 606 被去除。例如，超级监视器 202 可以接入客户操作系统 220 的热去除 API，以及从虚拟 NUMA 节点 608 去除例如虚拟处理器 230B。

[0092] 现在转到操作 1610，它显示：把对于多个虚拟 NUMA 节点的 NUMA 比率报告给客户操作系统。例如，在实施例中，超级监视器 202 可以生成对于虚拟 NUMA 节点 606-608 的 NUMA 比率，并可以把这个信息报告给图 6 或图 7 的客户操作系统 220。在实施例中，超级监视器 202 可以在虚拟机的固件表中创建数据结构，其指示对于虚拟 NUMA 节点的 NUMA 比率，并且当客户操作系统 220 引导时客户可以读取该表并使用信息来做出调度决定。例如，客户操作系统，或者 NUMA 感知的应用，可以使用 NUMA 比率来确定是否使用来自远程 NUMA 节点的资源。例如，操作系统可以具有准备好要执行的待决线程。在本例中，操作系统也可以被配置成：等待一定量的时间以便让理想处理器变为空闲的，否则它将在具有小于预定的 NUMA 比率的远程处理器上调度该线程。在这种情形下，调度器愿意等待的时间量取决于 NUMA 比率。

[0093] 现在转到操作 1612，它显示：将虚拟机从第一 NUMA 节点迁移到多个 NUMA 节点。例如，在实施例中，超级监视器指令可以由逻辑处理器执行，且虚拟机 240 可被映射到多个 NUMA 节点 704 和 706。在本例中，计算机系统 700 可能处在繁重的使用下，例如，虚拟机 242 可能正在使用大部分的资源，所以，虚拟机 242 被迁移出计算机 700。在这种情形下，超级监视器 202 可以重新分配计算机系统 700 的资源，将虚拟机 240 重新映射到 NUMA 节点 704 和 706。

[0094] 现在转到图 17，图上描绘图 16 的操作程序的替换实施例，其包括操作 1714 和 1716。在实施例中，操作 1606 可包括操作 1714，其显示：根据虚拟 NUMA 节点的当前存储器压力低于目标阈值的确定，从多个虚拟 NUMA 节点的第一虚拟 NUMA 节点中回收存储器。例如，在实施例中，DMVSP 802 可被配置成：当虚拟 NUMA 节点 608 的存储器压力低于目标阈值时，从虚拟 NUMA 节点 606 回收存储器。例如，在实施例中，当例如在母分区中没有可得到的

存储器可被调拨且虚拟 NUMA 节点 608 正经受不可接受的压力时，DMVSP 802 可以尝试从虚拟 NUMA 节点 606 释放存储器。如果存储器从虚拟 NUMA 节点 606 被回收，则可以发送异步消息到 DMVSC 804，指引它回收存储器。当与客户 OS 220 相关联的 DMVSC 804 响应时，它可以指示在虚拟 NUMA 节点 606 内新的存储器压力。在一些情形下，响应于存储器去除操作，存储器压力可能增加。

[0095] 继续进行图 17 的说明，操作 1716 显示：根据客户操作系统的当前存储器压力大于目标阈值的确定，将存储器调拨给多个虚拟 NUMA 节点的第一虚拟 NUMA 节点。例如，在实施例中，DMVSP 802 可被配置成：当虚拟 NUMA 节点 606 的存储器压力大于目标阈值时，将存储器调拨给虚拟 NUMA 节点 606。在这种情形下，如果存储器是可得到的，则它可被分配给虚拟 NUMA 节点 606。也就是，DMVSP 802 可以获得标识虚拟 NUMA 节点 606 的性能如何受可得到的存储器的量影响的存储器压力信息，并把存储器添加到虚拟 NUMA 节点 606。在特定的例子中，存储器压力信息可以是一个值。在本例中，DMVSP 802 可以将当前存储器压力值与指示对于虚拟 NUMA 节点 606 的最小值的信息表进行比较，并调节存储器直至虚拟 NUMA 节点 606 的存储器压力等于最小为止。例如，管理员可以配置运行关键应用的客户操作系统，以便具有对于虚拟 NUMA 节点 606 和 608 的低的最小值。

[0096] 以上的详细说明经由例子和 / 或操作图阐述了系统和 / 或过程的各种实施例。就这样的框图和 / 或例子包含一个或多个功能和 / 或操作而论，本领域技术人员将懂得，在这样的框图或例子内的每个功能和 / 或操作可通过各种各样的硬件、软件、固件、或事实上它们的任何组合，而被单独地和 / 或共同地实施。

[0097] 虽然显示和描述了这里所描述的本主题的特定方面，但本领域技术人员将明白，根据这里的教导，可以做出改变和修改，而不背离这里描述的主题和它的更广义的方面，所以，所附权利要求是要将处在这里描述的主题的真实精神和范围内的所有的这样的改变和修改都包括在它们的范围内。

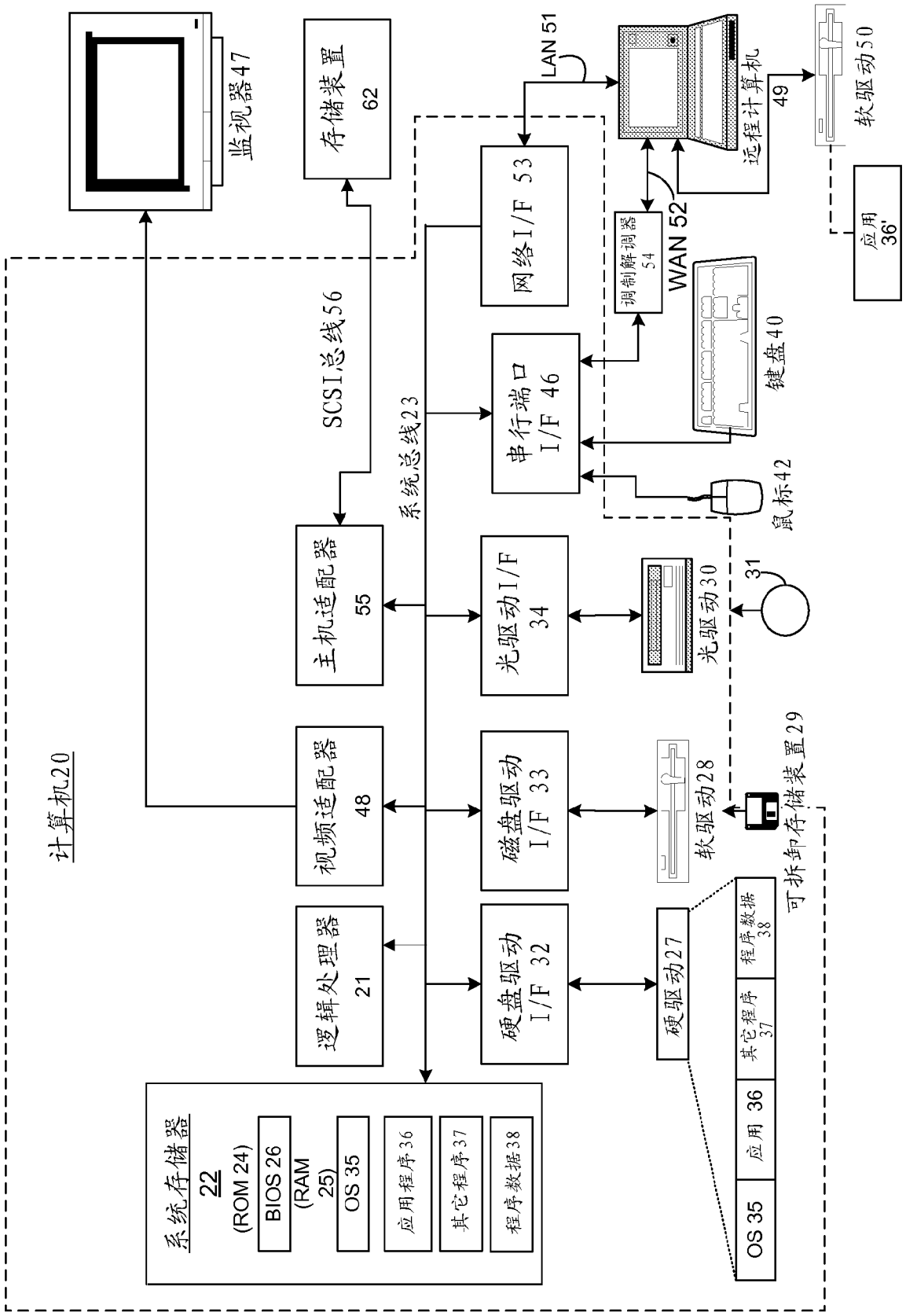

计算机20
监视器47
存储装置62
LAN 51
远程计算机49
软驱动50
应用36'
网络I/F 53
调制解调器54
WAN 52
SCSI总线56
系统总线23
串行端口I/F 46
键盘40
鼠标42
主机适配器55
光驱动I/F 34
光驱动30
31
视频适配器48
磁盘驱动I/F 33
软驱动28
可拆卸存储装置29
逻辑处理器21
硬盘驱动I/F 32
硬驱动27
系统存储器22
(ROM 24)
BIOS 26
(RAM 25)
OS 35
应用程序36
其它程序37
程序数据38
OS 35
应用 36
其它程序37
程序数据38

图 1

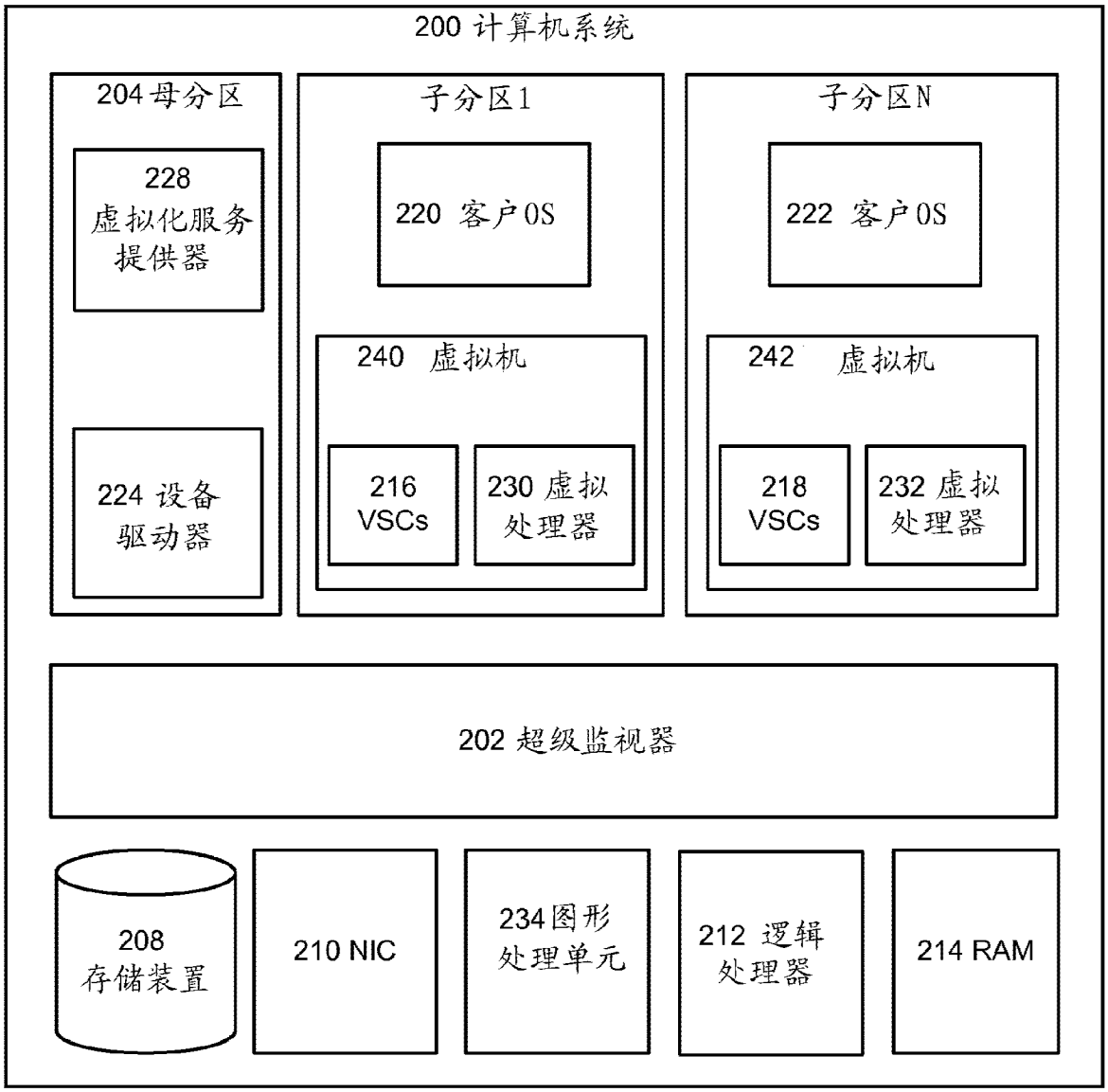

200 计算机系统
204 母分区
228 虚拟化服务提供器
224 设备驱动器
子分区1
220 客户OS
240 虚拟机
216 VSCs
230 虚拟处理器
子分区N
222 客户OS
242 虚拟机
218 VSCs
232 虚拟处理器
202 超级监视器
208 存储装置
210 NIC
234 图形处理单元
212 逻辑处理器
214 RAM

图 2

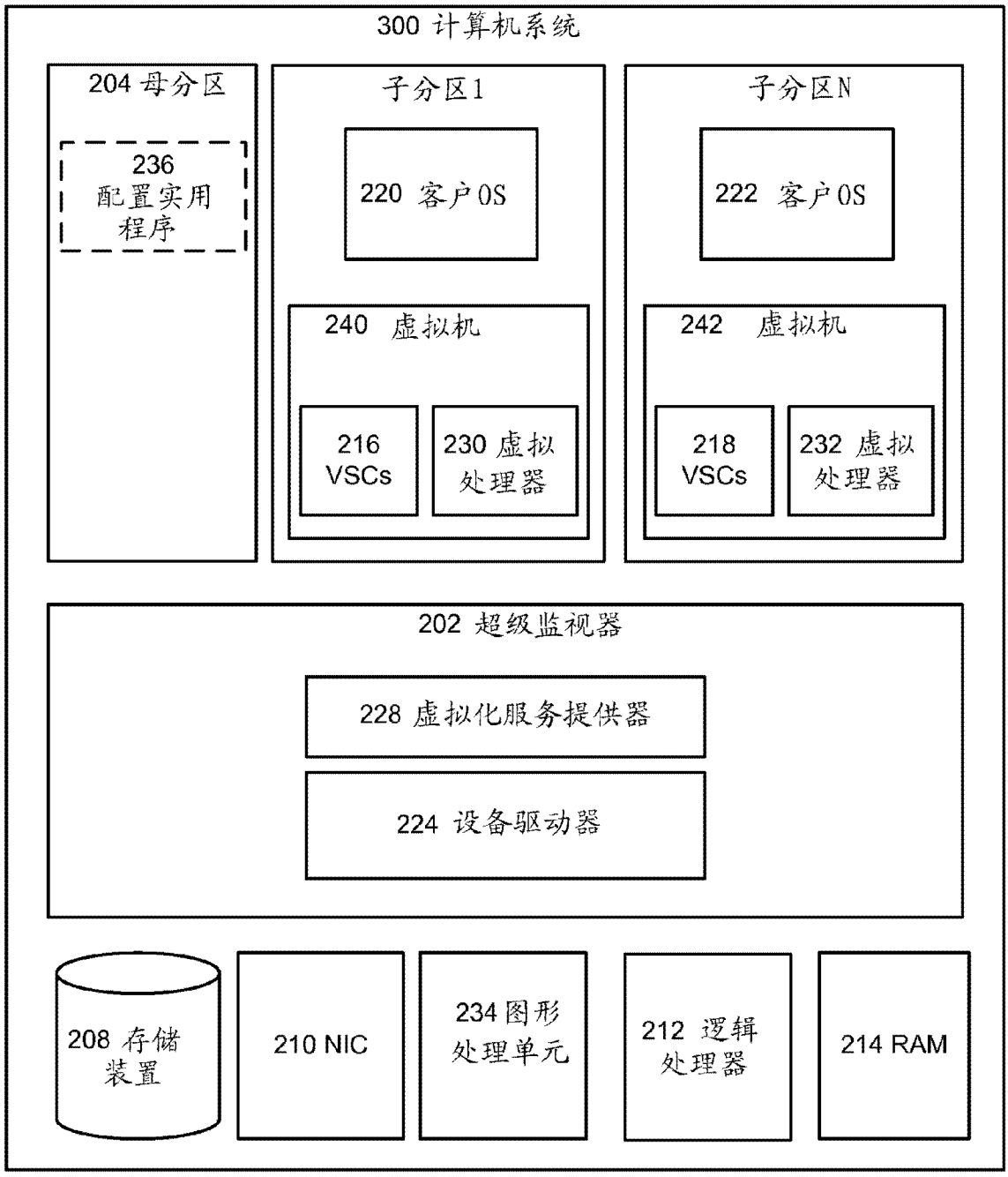

300 计算机系统
204 母分区
236 配置实用程序
子分区1
220 客户OS
240 虚拟机
216 VSCs
230 虚拟处理器
子分区N
222 客户OS
242 虚拟机
218 VSCs
232 虚拟处理器
202 超级监视器
228 虚拟化服务提供器
224 设备驱动器
208 存储装置
210 NIC
234 图形处理单元
212 逻辑处理器
214 RAM

图 3

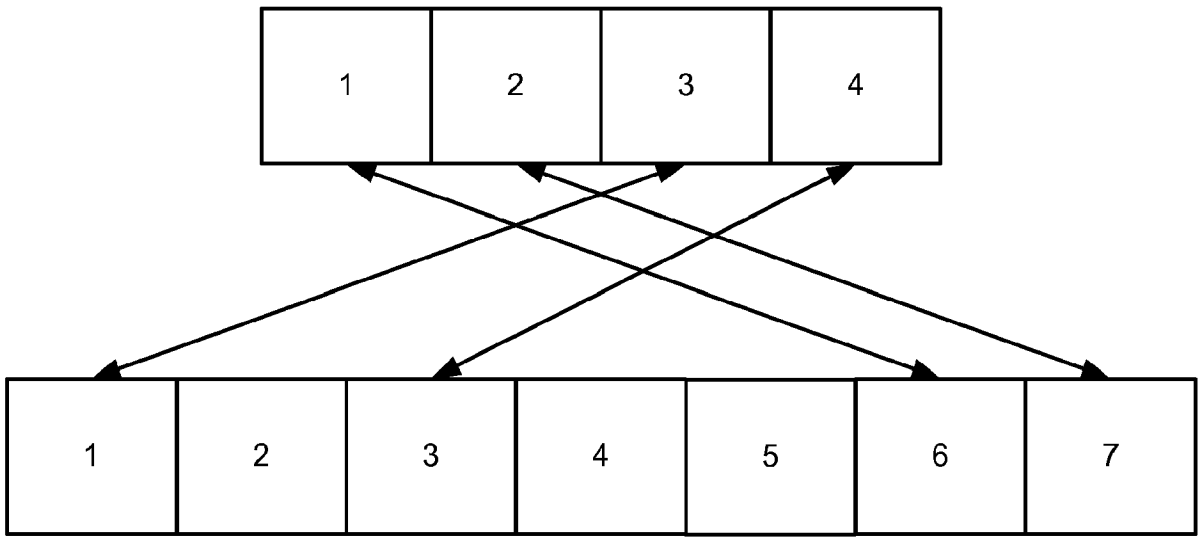


图 4

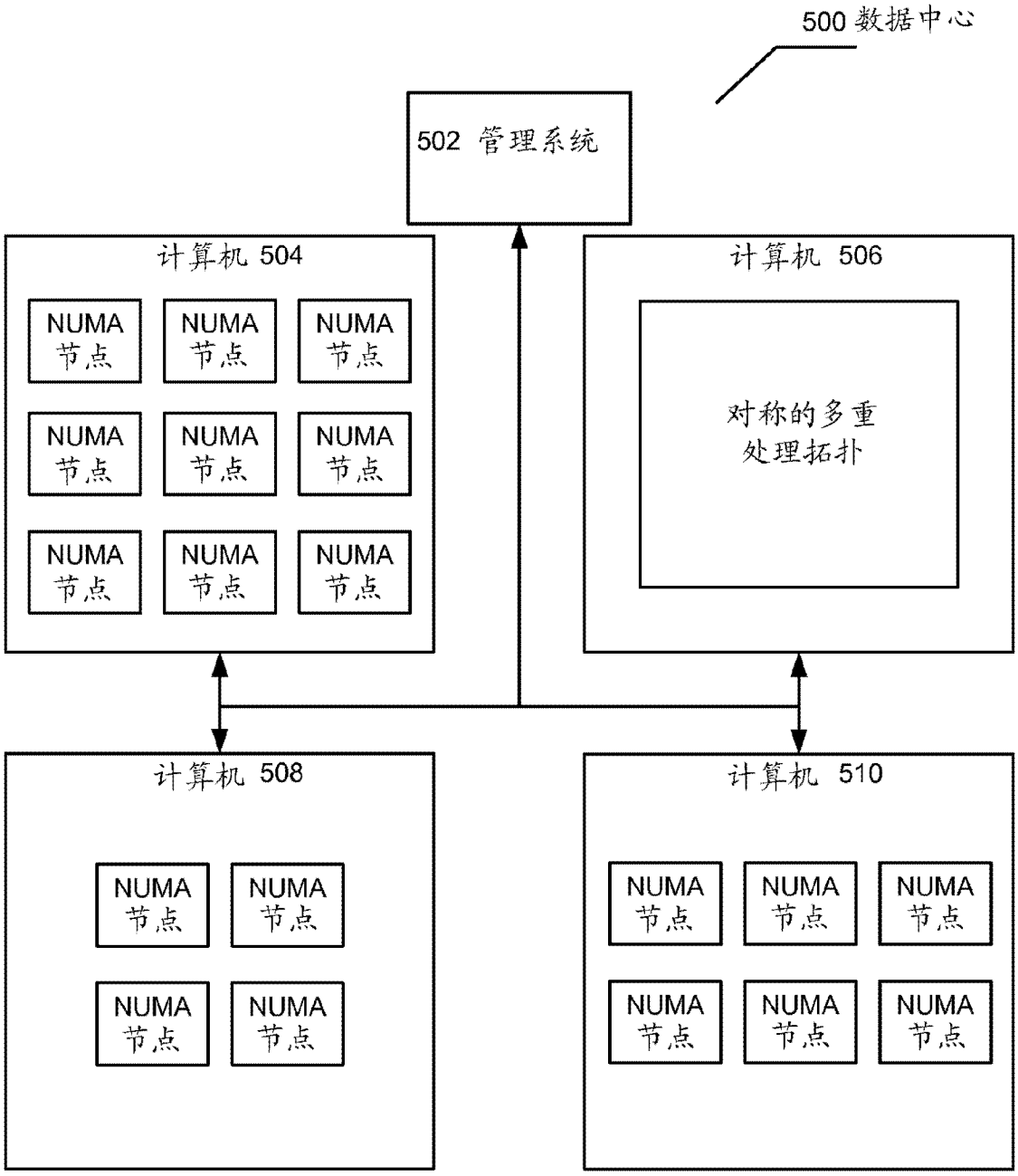

500 数据中心
502 管理系统
计算机 504
NUMA 节点
NUMA 节点
NUMA 节点
NUMA 节点
NUMA 节点
NUMA 节点
NUMA 节点
NUMA 节点
NUMA 节点
计算机 506
对称的多重处理拓扑
计算机 508
NUMA 节点
NUMA 节点
NUMA 节点
NUMA 节点
计算机 510
NUMA 节点
NUMA 节点
NUMA 节点
NUMA 节点
NUMA 节点
NUMA 节点

图 5

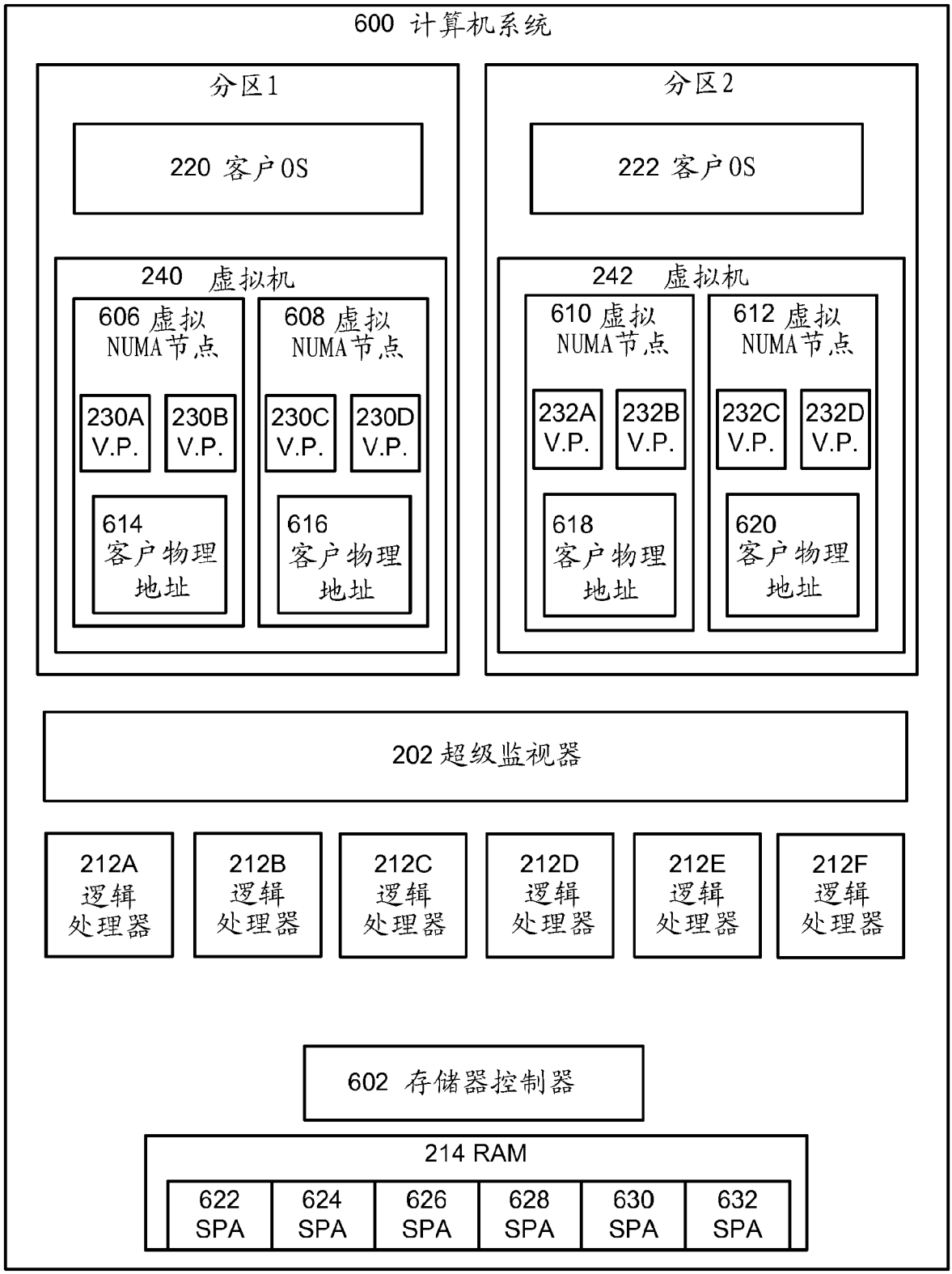

600 计算机系统
分区1
220 客户OS
240 虚拟机
606 虚拟NUMA节点
230A V.P.
230B V.P.
614 客户物理地址
608 虚拟NUMA节点
230C V.P.
230D V.P.
616 客户物理地址
分区2
222 客户OS
242 虚拟机
610 虚拟NUMA节点
232A V.P.
232B V.P.
618 客户物理地址
612 虚拟NUMA节点
232C V.P.
232D V.P.
620 客户物理地址
202 超级监视器
212A 逻辑处理器
212B 逻辑处理器
212C 逻辑处理器
212D 逻辑处理器
212E 逻辑处理器
212F 逻辑处理器
602 存储器控制器
214 RAM
622 SPA
624 SPA
626 SPA
628 SPA
630 SPA
632 SPA

图 6

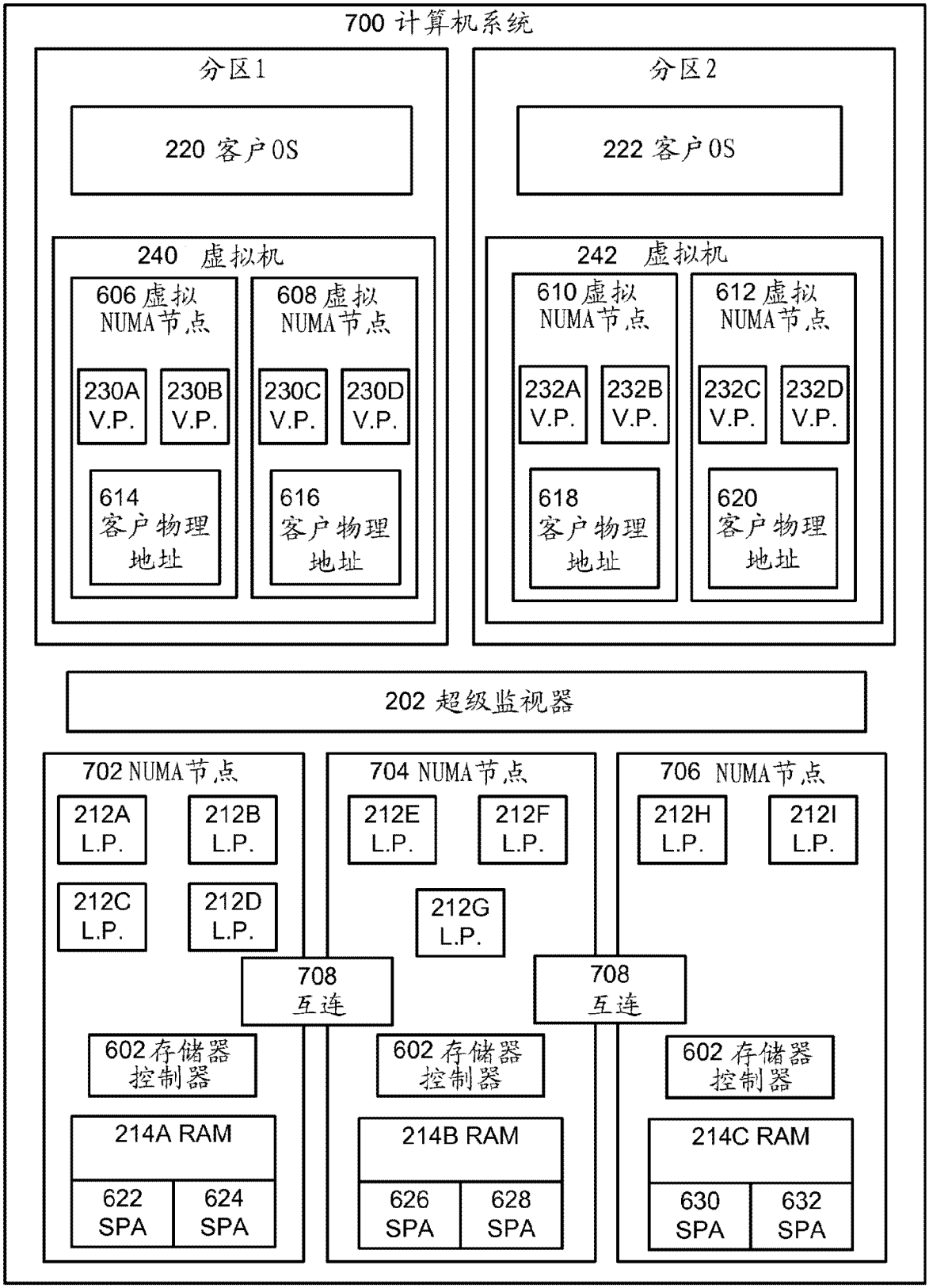

700 计算机系统
分区1
220 客户OS
240 虚拟机
606 虚拟NUMA节点
608 虚拟NUMA节点
230A V.P.
230B V.P.
230C V.P.
230D V.P.
614 客户物理地址
616 客户物理地址
分区2
222 客户OS
242 虚拟机
610 虚拟NUMA节点
612 虚拟NUMA节点
232A V.P.
232B V.P.
232C V.P.
232D V.P.
618 客户物理地址
620 客户物理地址
202 超级监视器
702 NUMA节点
704 NUMA节点
706 NUMA节点
212A L.P.
212B L.P.
212C L.P.
212D L.P.
212E L.P.
212F L.P.
212G L.P.
212H L.P.
212I L.P.
708 互连
708 互连
602 存储器控制器
602 存储器控制器
602 存储器控制器
214A RAM
622 SPA
624 SPA
214B RAM
626 SPA
628 SPA
214C RAM
630 SPA
632 SPA

图 7

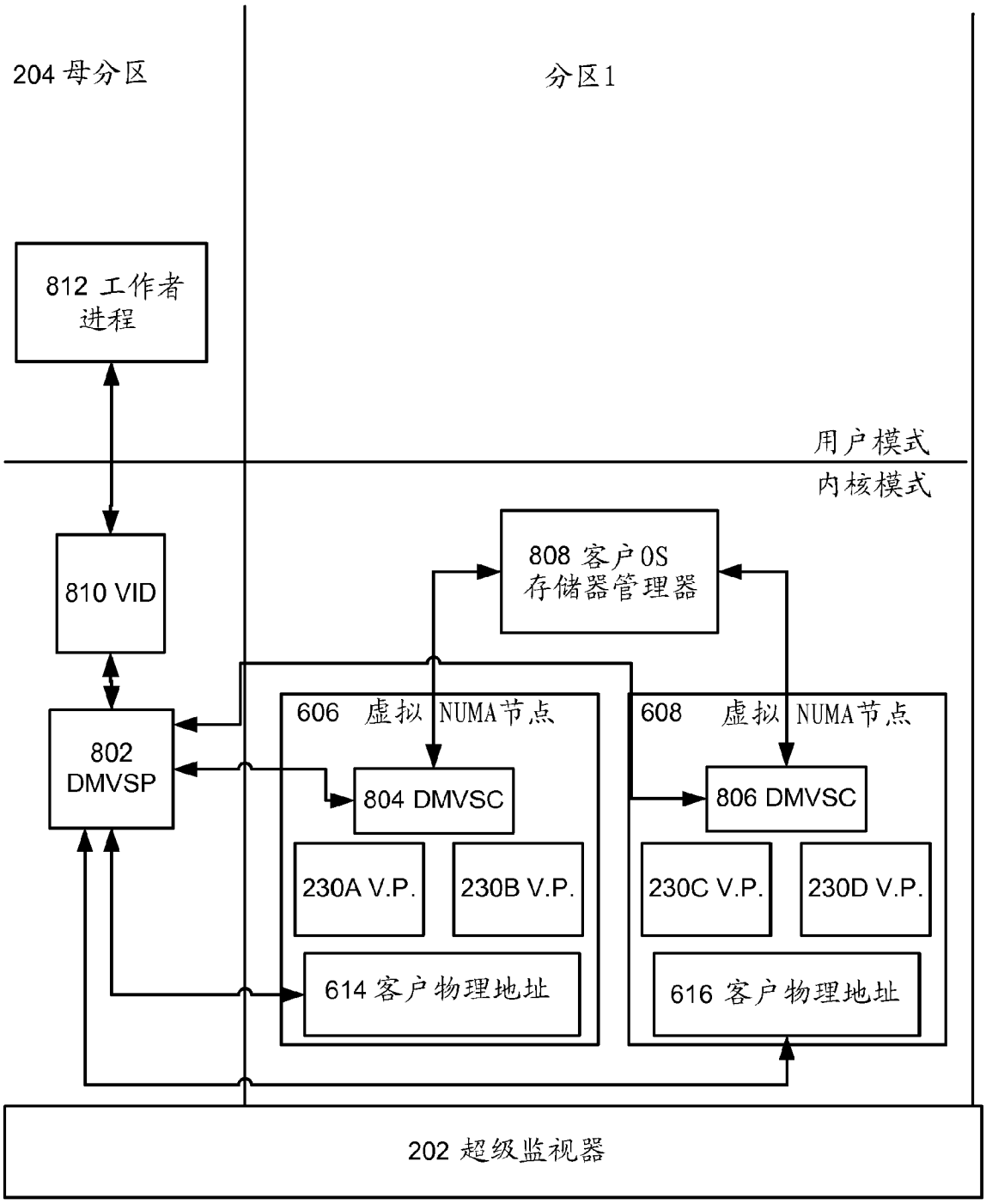

204 母分区
分区1
812 工作者进程
用户模式
内核模式
810 VID
808 客户OS存储器管理器
802 DMVSP
606 虚拟 NUMA节点
608 虚拟 NUMA节点
804 DMVSC
806 DMVSC
230A V.P.
230B V.P.
230C V.P.
230D V.P.
614 客户物理地址
616 客户物理地址
202 超级监视器

图 8

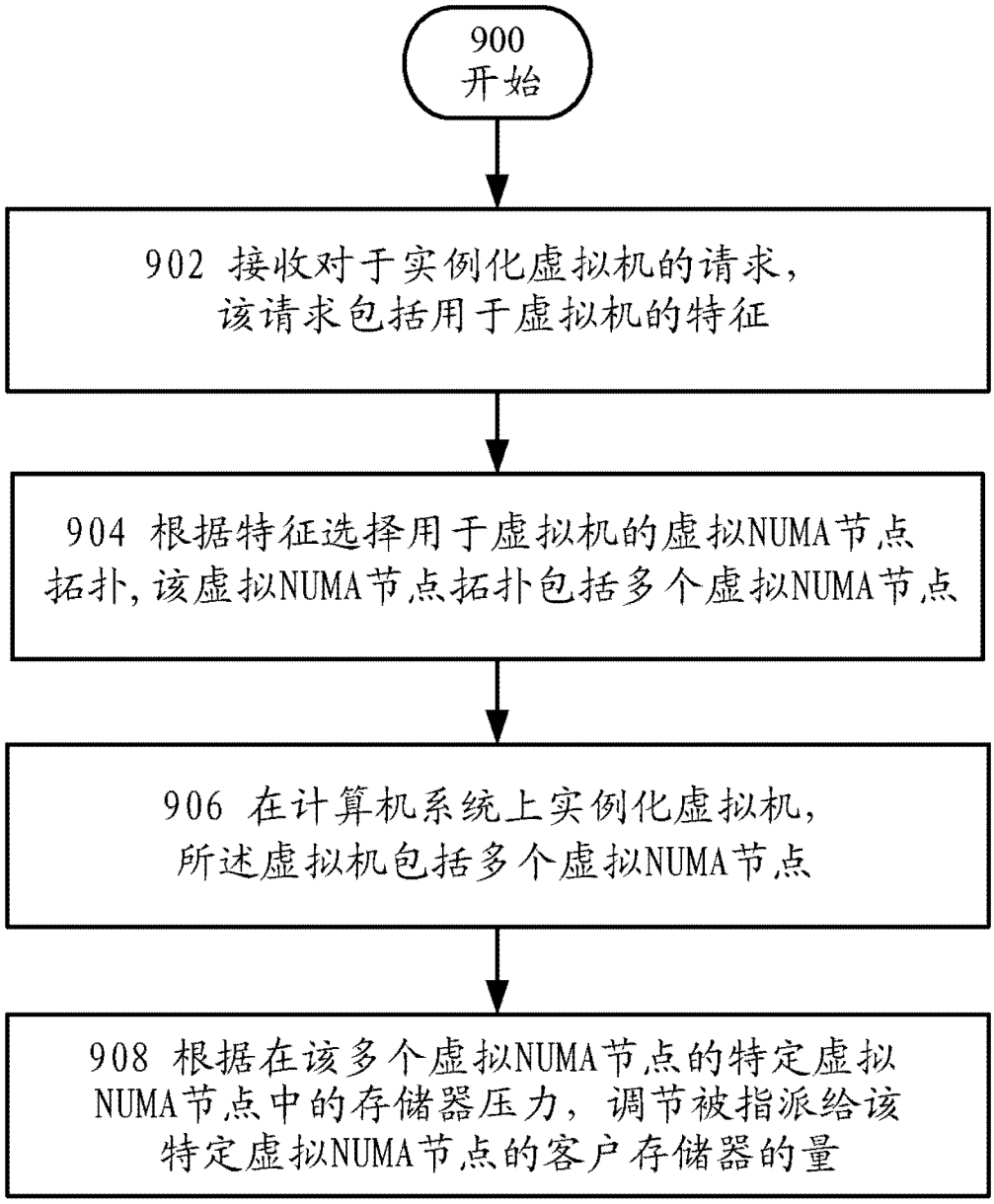

900
开始
902 接收对于实例化虚拟机的请求，
该请求包括用于虚拟机的特征
904 根据特征选择用于虚拟机的虚拟NUMA节点
拓扑，该虚拟NUMA节点拓扑包括多个虚拟NUMA节点
906 在计算机系统上实例化虚拟机，
所述虚拟机包括多个虚拟NUMA节点
908 根据在该多个虚拟NUMA节点的特定虚拟
NUMA节点中的存储器压力，调节被指派给该
特定虚拟NUMA节点的客户存储器的量

图 9

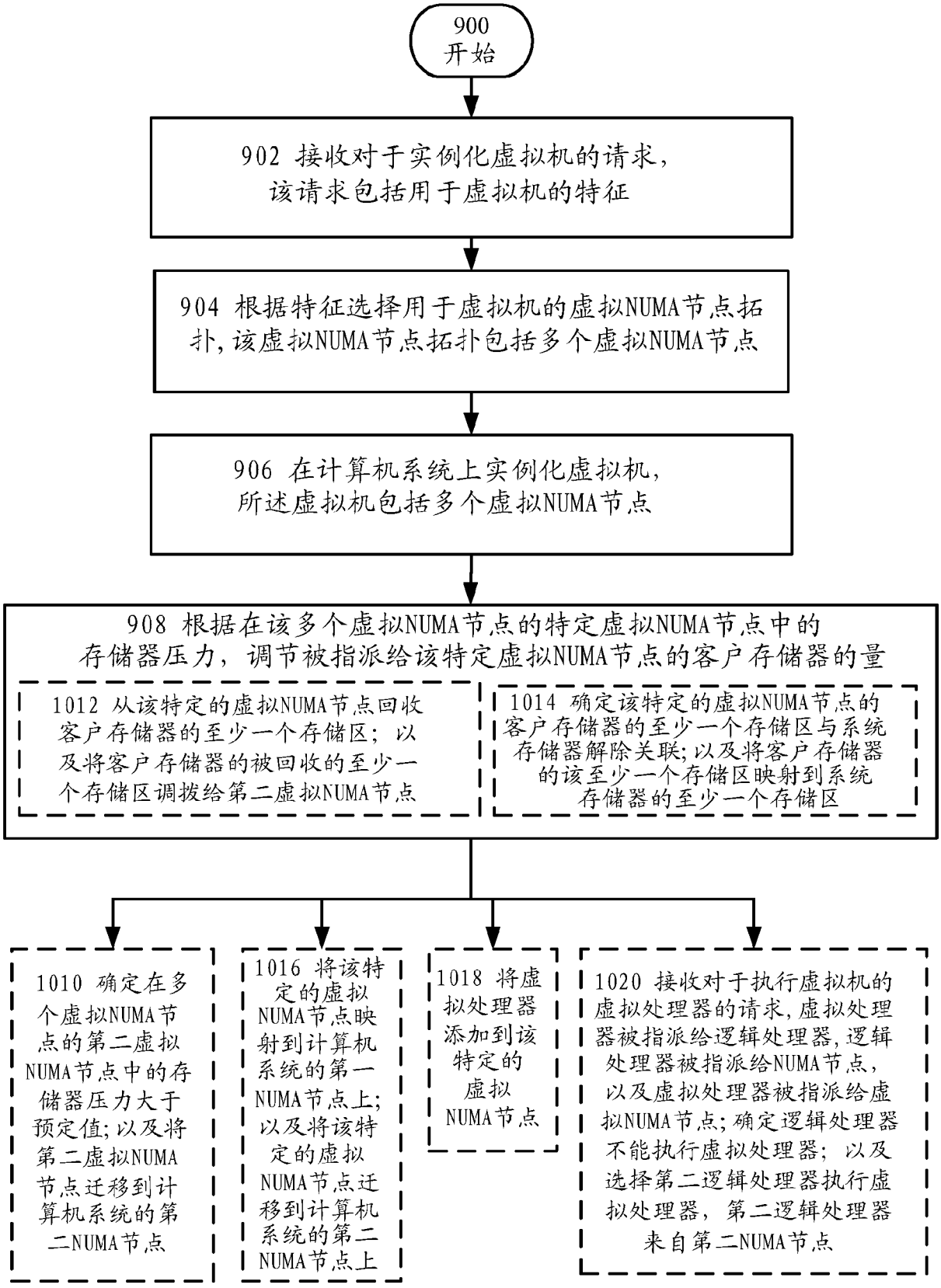

900
开始
902 接收对于实例化虚拟机的请求，
该请求包括用于虚拟机的特征
904 根据特征选择用于虚拟机的虚拟NUMA节点拓
扑，该虚拟NUMA节点拓扑包括多个虚拟NUMA节点
906 在计算机系统上实例化虚拟机，
所述虚拟机包括多个虚拟NUMA节点
908 根据在该多个虚拟NUMA节点的特定虚拟NUMA节点中的
存储器压力，调节被指派给该特定虚拟NUMA节点的客户存储器的量
1012 从该特定的虚拟NUMA节点回收
客户存储器的至少一个存储区；以
及将客户存储器的被回收的至少一
个存储区调拨给第二虚拟NUMA节点
1014 确定该特定的虚拟NUMA节点的
客户存储器的至少一个存储区与系统
存储器解除关联；以及将客户存储器
的该至少一个存储区映射到系统
存储器的至少一个存储区
1010 确定在多
个虚拟NUMA节
点的第二虚拟
NUMA节点中的存
储器压力大于
预定值；以及将
第二虚拟NUMA
节点迁移到计
算机系统的第
二NUMA节点
1016 将该特
定的虚拟
NUMA节点映
射到计算机
系统的第一
NUMA节点上；
以及将该特
定的虚拟
NUMA节点迁
移到计算机
系统的第二
NUMA节点上
1018 将虚
拟处理器
添加到该
特定的
虚拟
NUMA节点
1020 接收对于执行虚拟机的
虚拟处理器的请求，虚拟处理
器被指派给逻辑处理器，逻辑
处理器被指派给NUMA节点，
以及虚拟处理器被指派给虚
拟NUMA节点；确定逻辑处理器
不能执行虚拟处理器；以及
选择第二逻辑处理器执行虚
拟处理器，第二逻辑处理器
来自第二NUMA节点

图 10

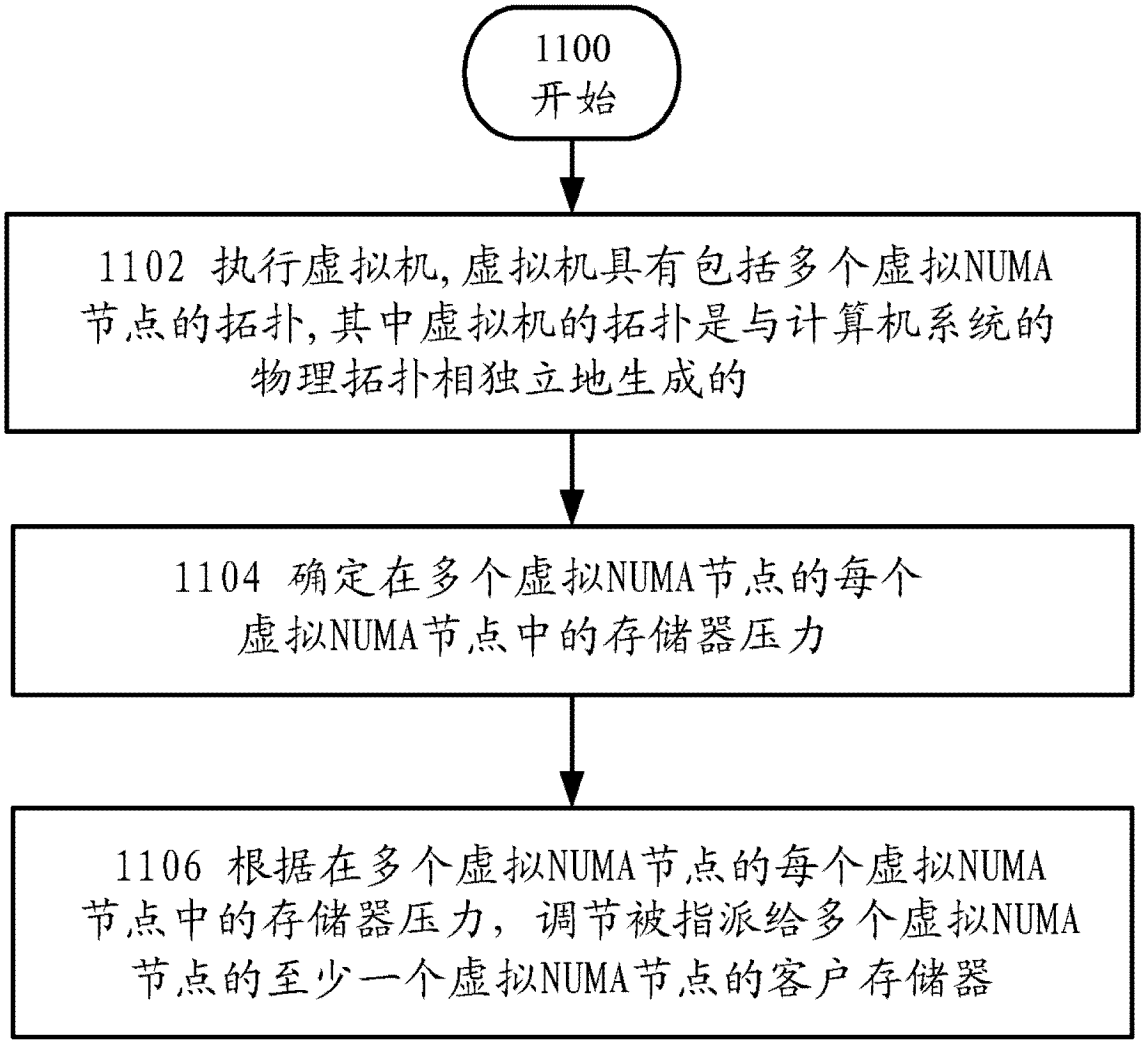

1100
开始
1102 执行虚拟机，虚拟机具有包括多个虚拟NUMA节点的拓扑，其中虚拟机的拓扑是与计算机系统的物理拓扑相独立地生成的
1104 确定在多个虚拟NUMA节点的每个虚拟NUMA节点中的存储器压力
1106 根据在多个虚拟NUMA节点的每个虚拟NUMA节点中的存储器压力，调节被指派给多个虚拟NUMA节点的至少一个虚拟NUMA节点的客户存储器

图 11

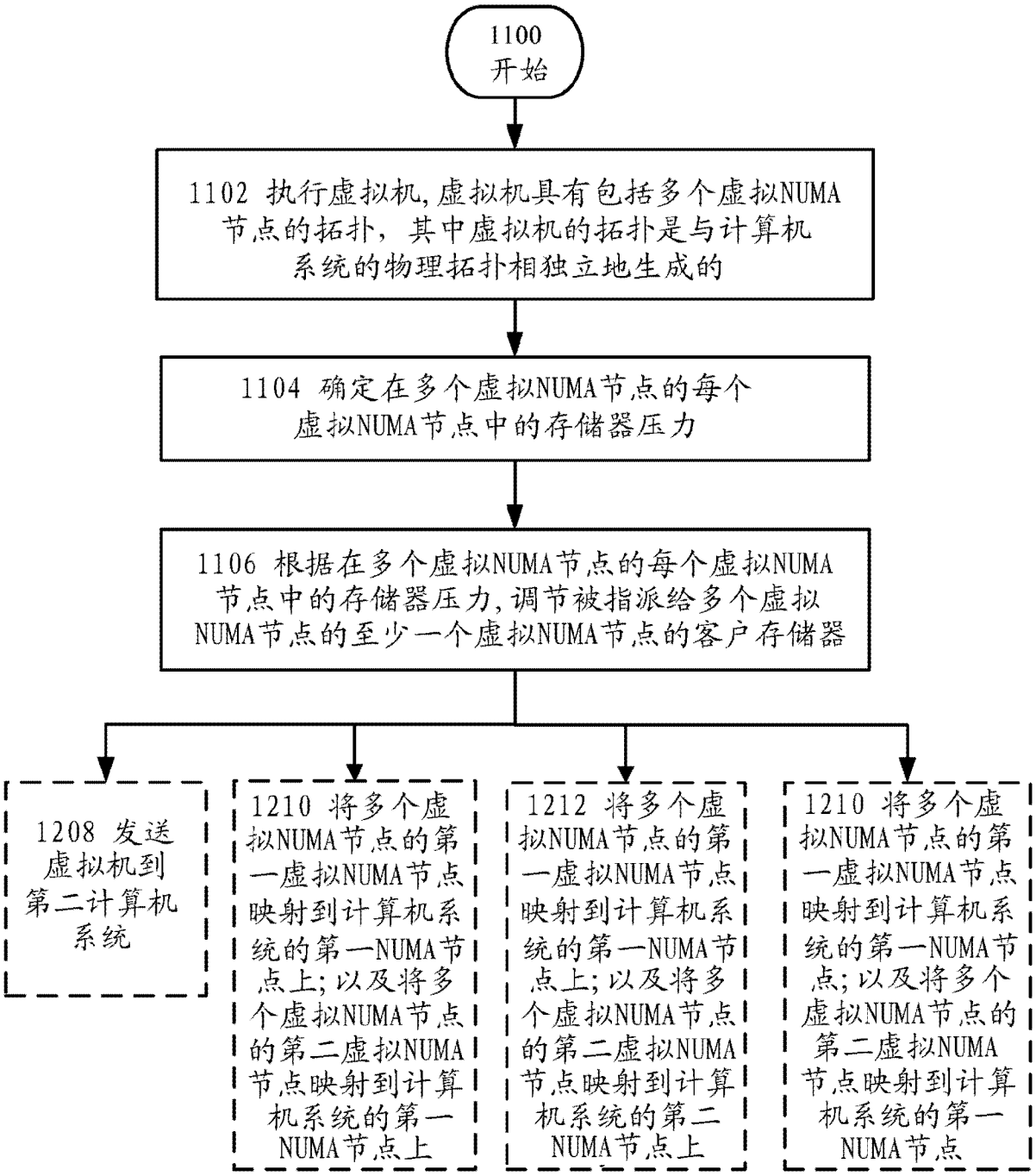

1100
开始
1102 执行虚拟机，虚拟机具有包括多个虚拟NUMA节点的拓扑，其中虚拟机的拓扑是与计算机系统的物理拓扑相独立地生成的
1104 确定在多个虚拟NUMA节点的每个虚拟NUMA节点中的存储器压力
1106 根据在多个虚拟NUMA节点的每个虚拟NUMA节点中的存储器压力，调节被指派给多个虚拟NUMA节点的至少一个虚拟NUMA节点的客户存储器
1208 发送虚拟机到第二计算机系统
1210 将多个虚拟NUMA节点的第一虚拟NUMA节点映射到计算机系统的第一NUMA节点上；以及将多个虚拟NUMA节点的第二虚拟NUMA节点映射到计算机系统的第一NUMA节点上
1212 将多个虚拟NUMA节点的第一虚拟NUMA节点映射到计算机系统的第一NUMA节点上；以及将多个虚拟NUMA节点的第二虚拟NUMA节点映射到计算机系统的第二NUMA节点上
1210 将多个虚拟NUMA节点的第一虚拟NUMA节点映射到计算机系统的第一NUMA节点；以及将多个虚拟NUMA节点的第二虚拟NUMA节点映射到计算机系统的第一NUMA节点

图 12

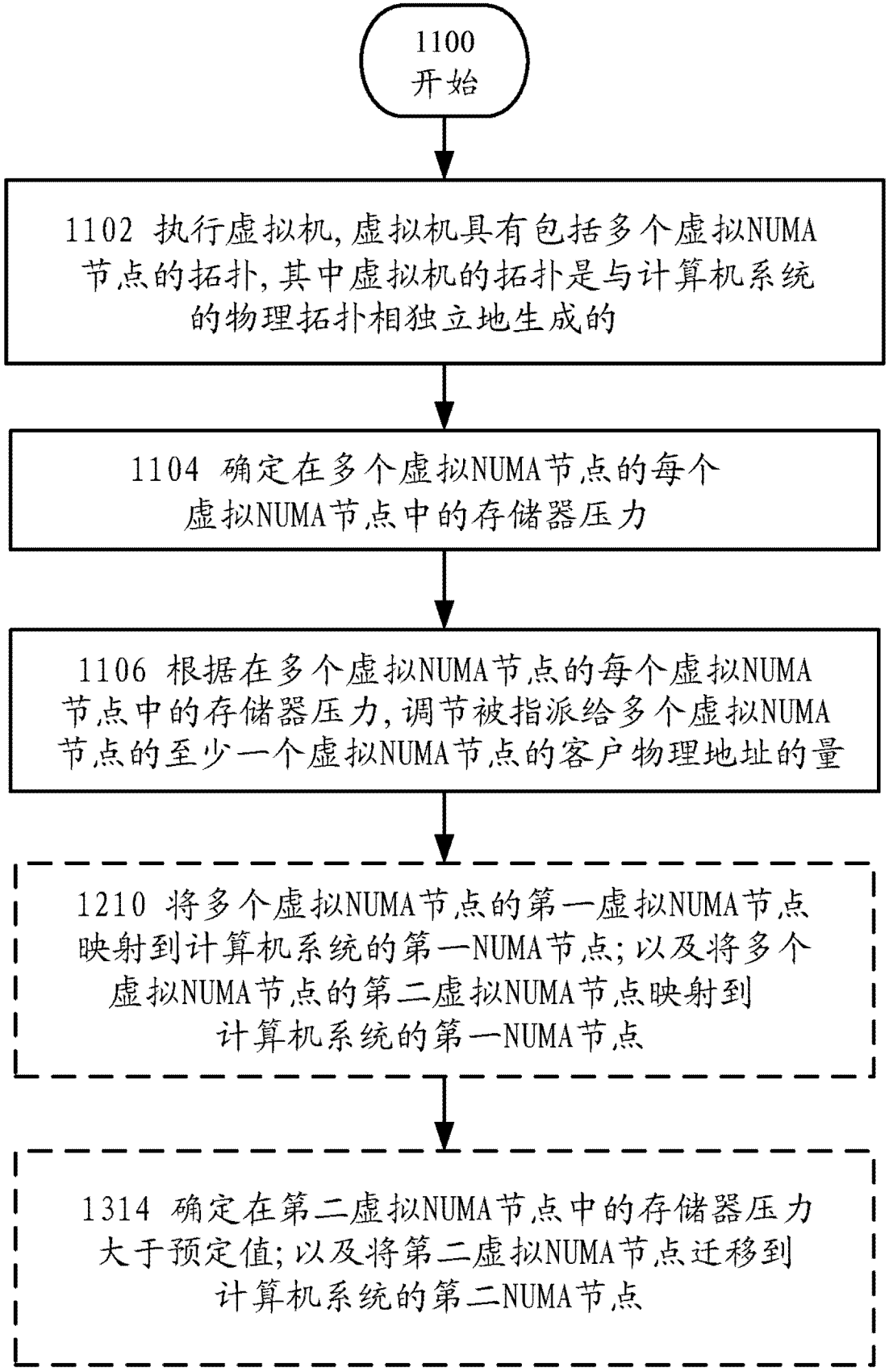

1100
开始
1102 执行虚拟机，虚拟机具有包括多个虚拟NUMA节点的拓扑，其中虚拟机的拓扑是与计算机系统的物理拓扑相独立地生成的
1104 确定在多个虚拟NUMA节点的每个虚拟NUMA节点中的存储器压力
1106 根据在多个虚拟NUMA节点的每个虚拟NUMA节点中的存储器压力，调节被指派给多个虚拟NUMA节点的至少一个虚拟NUMA节点的客户物理地址的量
1210 将多个虚拟NUMA节点的第一虚拟NUMA节点映射到计算机系统的第一NUMA节点；以及将多个虚拟NUMA节点的第二虚拟NUMA节点映射到计算机系统的第一NUMA节点
1314 确定在第二虚拟NUMA节点中的存储器压力大于预定值；以及将第二虚拟NUMA节点迁移到计算机系统的第二NUMA节点

图 13

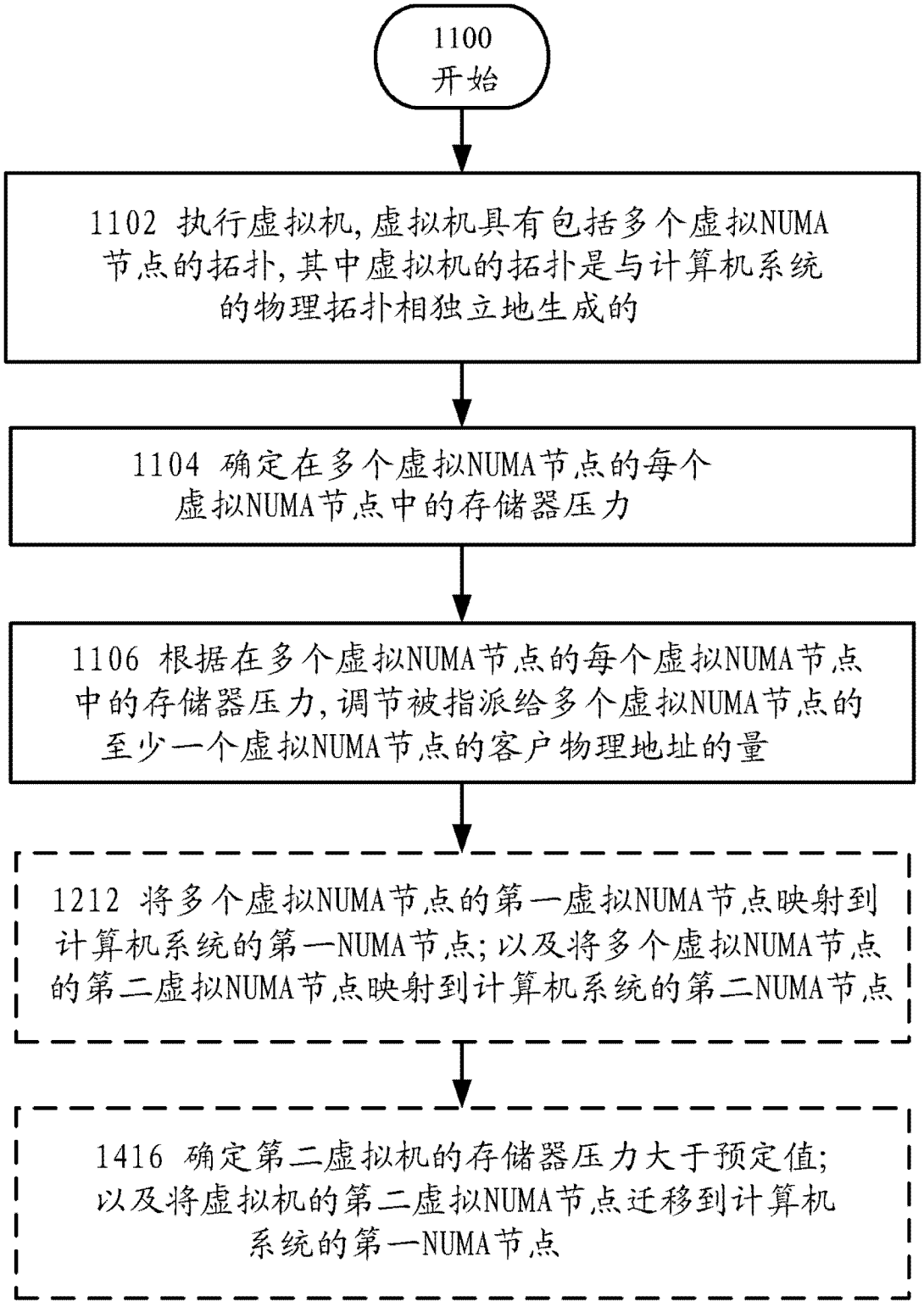

1100
开始
1102 执行虚拟机,虚拟机具有包括多个虚拟NUMA节点的拓扑,其中虚拟机的拓扑是与计算机系统的物理拓扑相独立地生成的
1104 确定在多个虚拟NUMA节点的每个虚拟NUMA节点中的存储器压力
1106 根据在多个虚拟NUMA节点的每个虚拟NUMA节点中的存储器压力,调节被指派给多个虚拟NUMA节点的至少一个虚拟NUMA节点的客户物理地址的量
1212 将多个虚拟NUMA节点的第一虚拟NUMA节点映射到计算机系统的第一NUMA节点;以及将多个虚拟NUMA节点的第二虚拟NUMA节点映射到计算机系统的第二NUMA节点
1416 确定第二虚拟机的存储器压力大于预定值;以及将虚拟机的第二虚拟NUMA节点迁移到计算机系统的第一NUMA节点

图 14

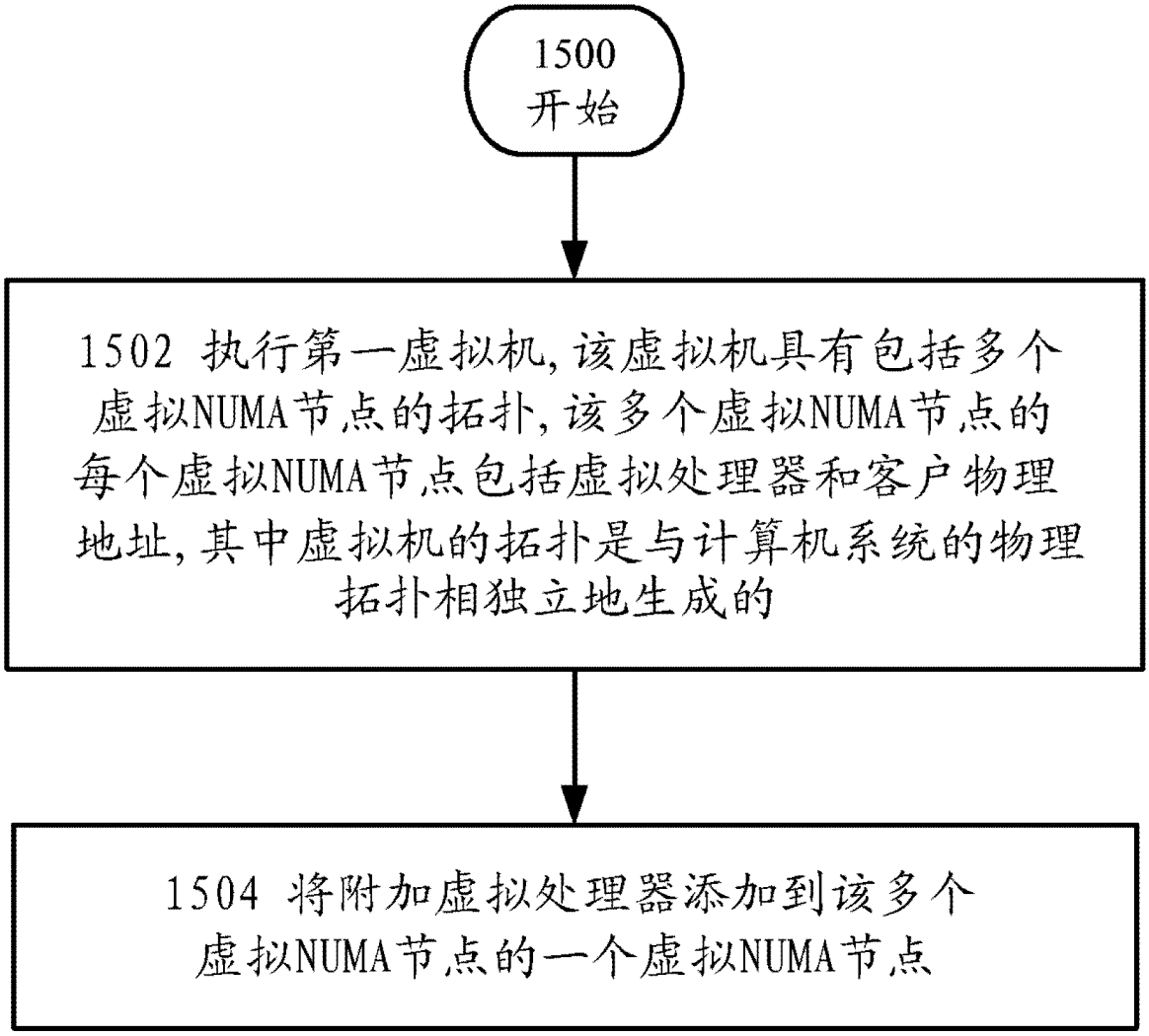

1500
开始
1502 执行第一虚拟机，该虚拟机具有包括多个虚拟NUMA节点的拓扑，该多个虚拟NUMA节点的每个虚拟NUMA节点包括虚拟处理器和客户物理地址，其中虚拟机的拓扑是与计算机系统的物理拓扑相独立地生成的
1504 将附加虚拟处理器添加到该多个虚拟NUMA节点的一个虚拟NUMA节点

图 15

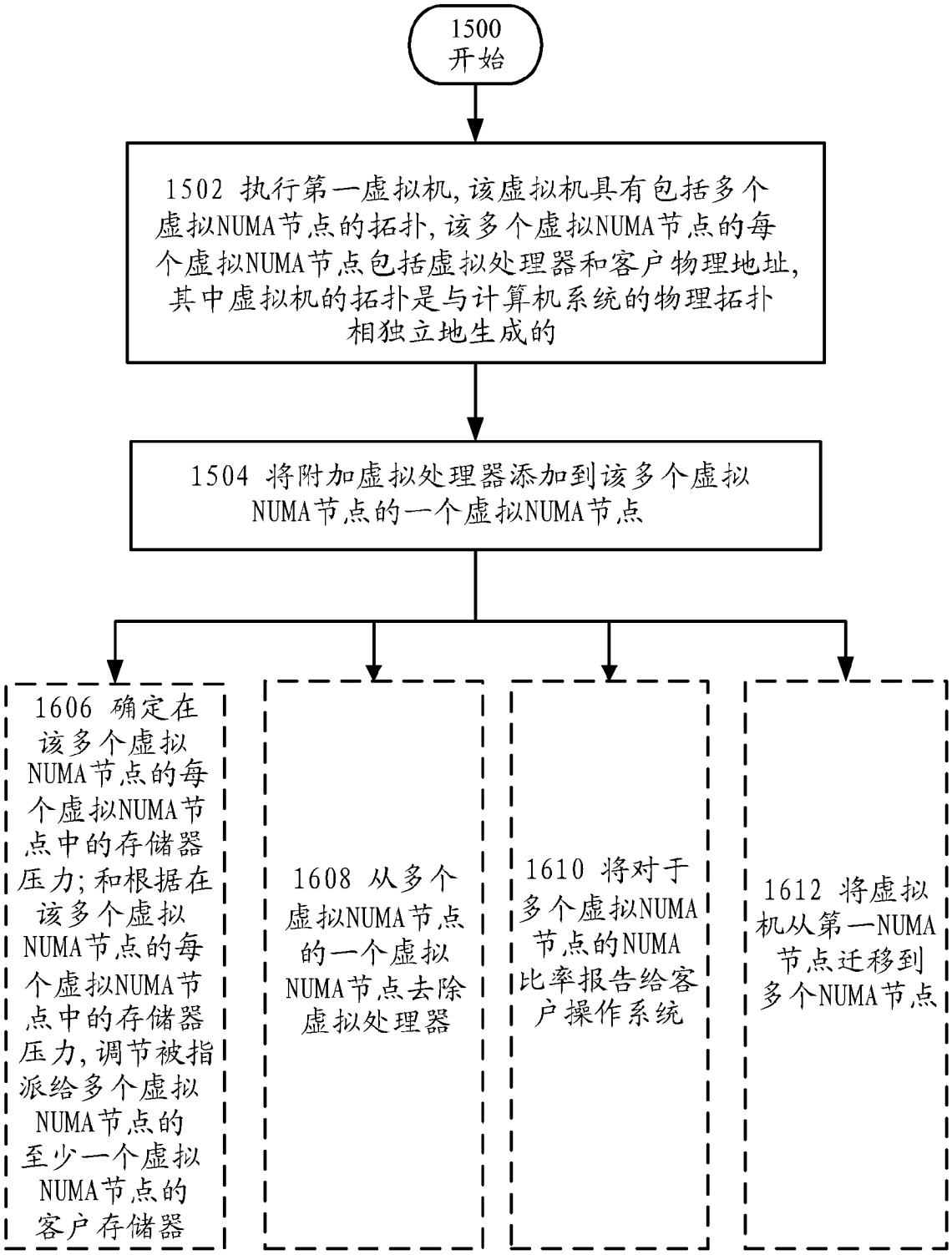

1500
开始
1502 执行第一虚拟机,该虚拟机具有包括多个虚拟NUMA节点的拓扑,该多个虚拟NUMA节点的每个虚拟NUMA节点包括虚拟处理器和客户物理地址,其中虚拟机的拓扑是与计算机系统的物理拓扑相独立地生成的
1504 将附加虚拟处理器添加到该多个虚拟NUMA节点的一个虚拟NUMA节点
1606 确定在该多个虚拟NUMA节点的每个虚拟NUMA节点中的存储器压力;和根据在该多个虚拟NUMA节点的每个虚拟NUMA节点中的存储器压力,调节被指派给多个虚拟NUMA节点的至少一个虚拟NUMA节点的客户存储器
1608 从多个虚拟NUMA节点的一个虚拟NUMA节点去除虚拟处理器
1610 将对于多个虚拟NUMA节点的NUMA比率报告给客户操作系统
1612 将虚拟机从第一NUMA节点迁移到多个NUMA节点

图 16

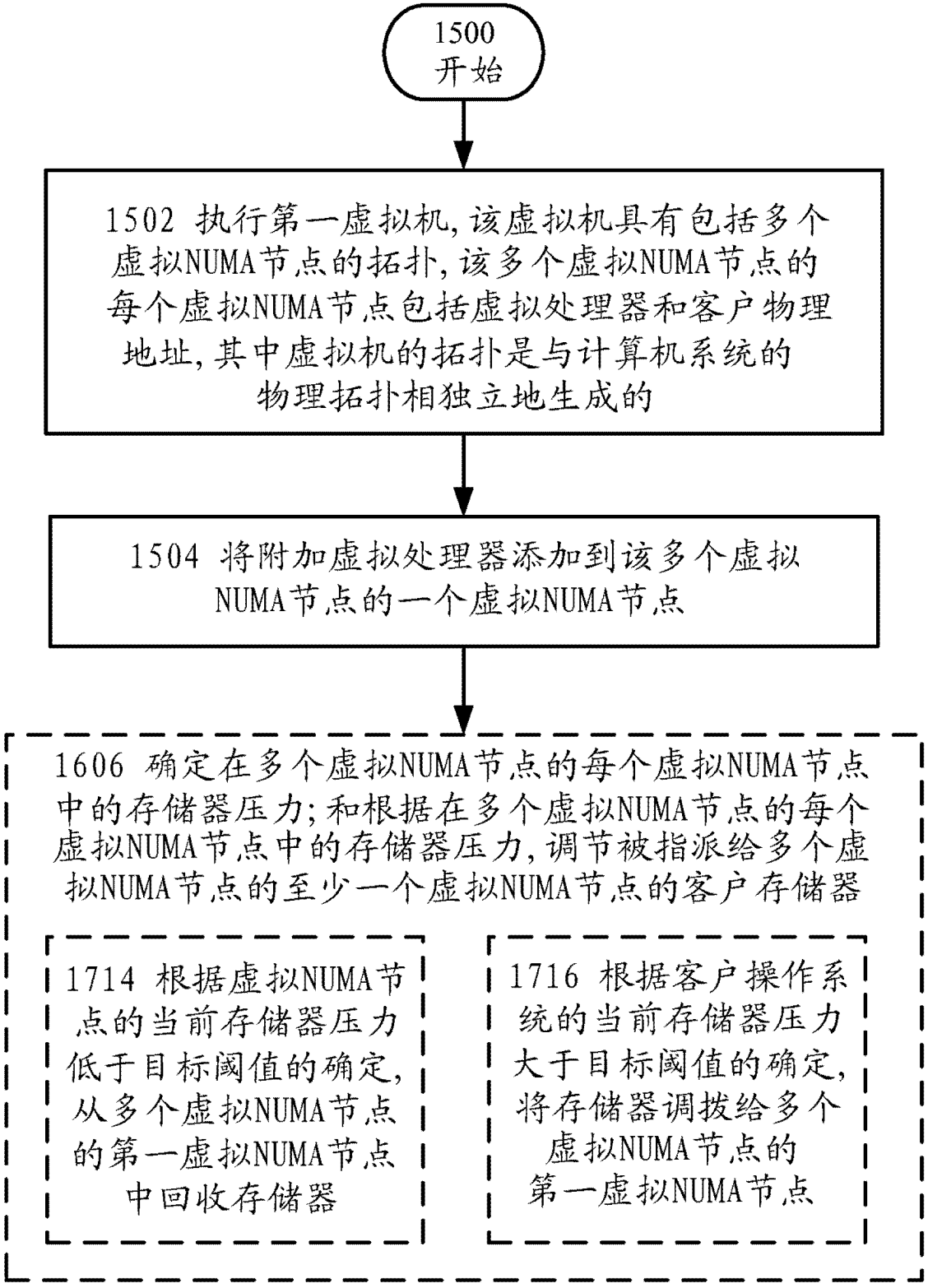

1500
开始
1502 执行第一虚拟机，该虚拟机具有包括多个虚拟NUMA节点的拓扑，该多个虚拟NUMA节点的每个虚拟NUMA节点包括虚拟处理器和客户物理地址，其中虚拟机的拓扑是与计算机系统的物理拓扑相独立地生成的
1504 将附加虚拟处理器添加到该多个虚拟NUMA节点的一个虚拟NUMA节点
1606 确定在多个虚拟NUMA节点的每个虚拟NUMA节点中的存储器压力；和根据在多个虚拟NUMA节点的每个虚拟NUMA节点中的存储器压力，调节被指派给多个虚拟NUMA节点的至少一个虚拟NUMA节点的客户存储器
1714 根据虚拟NUMA节点的当前存储器压力低于目标阈值的确定，从多个虚拟NUMA节点的第一虚拟NUMA节点中回收存储器
1716 根据客户操作系统的当前存储器压力大于目标阈值的确定，将存储器调拨给多个虚拟NUMA节点的第一虚拟NUMA节点

图 17