



(12)发明专利

(10)授权公告号 CN 104428752 B

(45)授权公告日 2018.06.26

(21)申请号 201380032691.7

(22)申请日 2013.06.12

(65)同一申请的已公布的文献号
申请公布号 CN 104428752 A

(43)申请公布日 2015.03.18

(30)优先权数据
13/529,747 2012.06.21 US

(85)PCT国际申请进入国家阶段日
2014.12.19

(86)PCT国际申请的申请数据
PCT/US2013/045290 2013.06.12

(87)PCT国际申请的公布数据
W02013/191972 EN 2013.12.27

(73)专利权人 微软技术许可有限责任公司
地址 美国华盛顿州

(72)发明人 S·坎杜拉 C·金 A·达巴格
D·班塞尔 D·A·马尔茨

(74)专利代理机构 上海专利商标事务所有限公
司 31100

代理人 罗婷婷

(51)Int.Cl.
G06F 9/50(2006.01)

(56)对比文件
CN 102132511 A,2011.07.20,
CN 1826768 A,2006.08.30,
US 2011/0320632 A1,2011.12.29,
Yan Luo,Eric Murray,Timothy
L.Ficarra.Accelerated Virtual Switching
with Programmable NICs for Scalable Data
Center Networking.《ACM》.2010,第65-72页.

审查员 刘朝兵

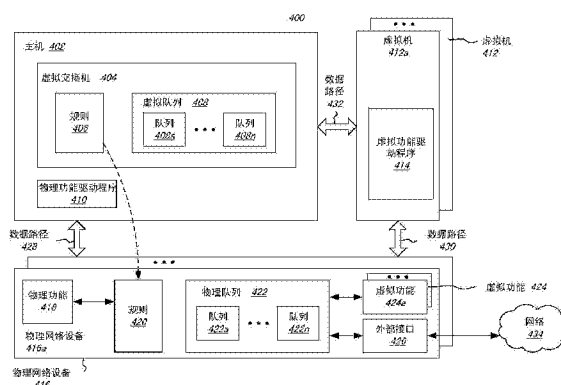
权利要求书3页 说明书8页 附图5页

(54)发明名称

将虚拟机流卸载至物理队列

(57)摘要

本发明涉及用于将虚拟机流卸载到物理队列的方法、系统和计算机程序产品。计算机系统执行一个或多个虚拟机,并用管理虚拟机的网络话务的一个或多个规则来对物理网络设备进行编程。该计算机系统还将网络设备编程为使用这些规则来管理网络话务。具体地,网络设备被编程为确定可用于处理虚拟机的网络流的一个或多个物理队列在该网络设备处的可用性。网络设备还被配置成标识虚拟机的网络流,包括标识每一网络流的特性。网络设备还被配置成基于网络流的特性并基于规则来将网络流中的一个或多个分配给物理队列中的至少一个。



1. 一种虚拟化主机计算机系统 (402), 该系统被配置成管理在所述虚拟化主机计算机系统处执行的一个或多个虚拟机 (412) 的网络话务, 所述虚拟化主机计算机系统包括:

一个或多个处理器;

一个或多个物理网络设备 (416); 以及

其上存储有计算机可执行指令的一个或多个计算机可读存储介质, 所述计算机可执行指令在被所述一个或多个处理器执行时, 使得所述虚拟化主机计算机系统执行虚拟交换机 (404), 所述虚拟交换机被配置成用一个或多个相应的规则 (420) 来对所述一个或多个物理网络设备中的每一个进行编程, 并执行以下操作:

确定一个或多个物理队列 (422) 在所述物理网络设备处的可用性, 所述一个或多个物理队列能用于处理所述一个或多个虚拟机的网络流;

标识所述一个或多个虚拟机的多个网络流, 包括标识所述多个网络流中的每一个的一个或多个特性; 以及

基于所述多个网络流中的每一个的一个或多个特性并基于所述一个或多个规则, 将所述多个网络流中的一个或多个分配给所述一个或多个物理队列中的至少一个, 其中将所述多个网络流中的一个或多个分配给所述一个或多个物理队列中的至少一个包括将所述多个网络流中具有相对较低话务等级的至少两个网络流一起分配给所述一个或多个物理队列中的单个物理队列, 而将所述多个网络流中具有相对较高话务等级的网络流中的每一者排他地分配给所述一个或多个物理队列中相应的物理队列。

2. 如权利要求1所述的虚拟机主机计算机系统, 其特征在于, 所述多个网络流在数目上大于所述一个或多个物理队列, 并且其中将所述多个网络流中的一个或多个分配给所述一个或多个物理队列中的至少一个进一步包括:

将所述多个网络流全部分配给所述一个或多个物理队列, 使得所述一个或多个物理队列中的至少一个被分配所述网络流中一个以上的网络流。

3. 如权利要求1所述的虚拟机主机计算机系统, 其特征在于, 所述多个网络流在数目上大于所述一个或多个物理队列, 并且其中将所述多个网络流中的一个或多个分配给所述一个或多个物理队列中的至少一个进一步包括:

将所述多个网络流的第一子集分配给所述一个或多个物理队列; 以及

将所述多个网络流的第二子集分配给所述虚拟交换机处的一个或多个基于软件的队列。

4. 如权利要求1所述的虚拟化主机计算机系统, 其特征在于, 标识所述多个网络流中的每一个的一个或多个特性包括:

标识所述多个网络流中的至少两个网络流具有相似的特性; 以及

基于所述多个网络流中具有相似特性的所述至少两个网络流, 标识所述多个网络流中的所述至少两个网络流能被分配给单个物理队列。

5. 如权利要求4所述的虚拟化主机计算机系统, 其特征在于, 所述多个网络流中具有相似特性的至少两个网络流包括所述多个网络流中具有相似速率限制特性的至少两个网络流。

6. 一种在计算机系统 (400) 处实现的用于管理网络话务的方法, 所述计算机系统包括一个或多个处理器和一个或多个物理网络设备 (416), 所述方法包括:

执行一个或多个虚拟机 (412) ;

用一个或多个规则 (420) 来对物理网络设备 (416a) 进行编程, 所述一个或多个规则被配置成管理所述一个或多个虚拟机的网络话务; 以及

将所述物理网络设备编程为管理网络话务, 包括:

确定一个或多个物理队列 (422) 在所述物理网络设备处的可用性, 所述一个或多个物理队列能用于处理所述一个或多个虚拟机的网络流;

标识所述一个或多个虚拟机的多个网络流, 包括标识所述多个网络流中的每一个的一个或多个特性; 以及

基于所述多个网络流中的每一个的一个或多个特性并基于所述一个或多个规则, 将所述多个网络流中的一个或多个分配给所述一个或多个物理队列中的至少一个, 其中将所述多个网络流中的一个或多个分配给所述一个或多个物理队列中的至少一个包括将所述多个网络流中具有相对较低等级的至少两个网络流一起分配在所述一个或多个物理队列中的单个物理队列上, 而将所述多个网络流中具有相对较高话务等级的网络流中的每一者排他地分配给所述一个或多个物理队列中相应的物理队列。

7. 如权利要求6所述的方法, 其特征在于, 所述多个网络流在数目上大于所述一个或多个物理队列, 并且其中将所述多个网络流中的一个或多个分配给所述一个或多个物理队列中的至少一个进一步包括:

将所述多个网络流全部分配给所述一个或多个物理队列, 使得所述一个或多个物理队列中的至少一个被分配所述网络流中一个以上的网络流。

8. 如权利要求6所述的方法, 其特征在于, 所述多个网络流在数目上大于所述一个或多个物理队列, 并且其中将所述多个网络流中的一个或多个分配给所述一个或多个物理队列中的至少一个进一步包括:

将所述多个网络流的第一子集分配给所述一个或多个物理队列; 以及

将所述多个网络流的第二子集分配给所述虚拟交换机处的一个或多个基于软件的队列。

9. 一种虚拟化主机计算机系统 (402), 该系统被配置成管理在所述虚拟化主机计算机系统上执行的一个或多个虚拟机 (412) 的网络话务, 所述虚拟化主机计算机系统包括:

一个或多个处理器;

一个或多个物理网络设备 (416); 以及

其上存储有计算机可执行指令的一个或多个计算机可读存储介质, 所述计算机可执行指令在被所述一个或多个处理器执行时, 使得所述虚拟化主机计算机系统执行虚拟交换机 (404), 所述虚拟交换机被配置成用一个或多个相应的规则 (420) 来对所述一个或多个物理网络设备中的每一个进行编程, 并执行以下操作:

确定一个或多个物理队列 (422) 在所述物理网络设备处的可用性, 所述一个或多个物理队列能用于处理所述一个或多个虚拟机的网络流;

标识所述一个或多个虚拟机的多个网络流, 包括标识所述多个网络流中的每一个的一个或多个特性, 其中所述多个网络流在数目上大于所述一个或多个物理队列; 以及

基于所述多个网络流中的每一个的一个或多个特性并基于所述一个或多个规则, 将所述多个网络流分配给所述一个或多个物理队列中的至少一个, 使得所述一个或多个物理队

列中的至少一个被分配所述多个网络流中一个以上的网络流,并且其中将所述多个网络流分配给所述一个或多个物理队列中的至少一个包括将所述多个网络流中具有相对较低等级的至少两个网络流一起分配在所述一个或多个物理队列中的单个物理队列上,而将所述多个网络流中具有相对较高话务等级的网络流中的每一者排他地分配给所述一个或多个物理队列中相应的物理队列。

将虚拟机流卸载至物理队列

[0001] 背景

[0002] 1.背景和相关技术

[0003] 计算机系统及及相关技术影响社会的许多方面。的确,计算机系统处理信息的能力已转变了人们生活和工作的方式。现在,计算机系统通常执行在计算机系统出现以前手动执行的许多任务(例如,文字处理、日程安排、帐目管理等)。最近,计算机系统彼此耦合并耦合到其他电子设备以形成计算机系统及其他电子设备可在其上传输电子数据的有线和无线计算机网络。因此,许多计算任务的执行分布在多个不同的计算机系统和/或多个不同的计算环境中。

[0004] 一些计算机系统被配置成提供用于托管(host)一个或多个虚拟机的虚拟化环境。例如,并行虚拟化执行环境包括管理程序(hypervisor)。管理程序提供父分区(有时被称为主机)及一个或多个子分区。父分区与物理硬件通信并管理物理硬件,并被配置成运行主操作系统并管理虚拟栈。每个子分区被配置成运行相应的客操作系统的“虚拟机”。

[0005] 虚拟化中的常见场景涉及管理在虚拟化主机计算机系统处执行的虚拟机间的网络分组,以及管理在这些虚拟机以及在主机计算机系统远程的计算机系统之间流动的网络分组。如此,主操作系统处的虚拟栈可包括联网虚拟栈,从而包括虚拟交换机。虚拟交换机被配置成截取、检查以及操纵在与各虚拟机的连接中通信的网络分组。然而,这样做可能是低效的,因为这导致主操作系统和客操作系统之间的频繁而昂贵(例如,在CPU使用方面)的上下文切换,并可在网络通信中引入等待时间。

[0006] 虚拟化的最近发展包括单根(single-root) I/O虚拟化(SRIOV)。SRIOV是外围组件互连快速(PCIe)总线体系结构的扩展,其使PCIe设备能够直接与子分区通信。如此,SRIOV使得PCIe设备能够通过管理程序将其自身暴露于子分区/虚拟机。例如,遵循SRIOV的物理网络接口卡(NIC)或交换机可向父分区呈现一物理功能并向相应的子分区呈现一个或多个虚拟功能。主操作系统随后可包括与物理功能通信的物理功能驱动程序,且每个客操作系统可执行与相应的虚拟功能通信的虚拟功能驱动程序。物理NIC随后可直接与客操作系统传递网络分组(绕过主操作系统),这可以大大改善网络性能。

[0007] 尽管SRIOV带来了优点,然而在虚拟化环境中的网络分组处理领域中仍存在某些低效。

[0008] 简要概述

[0009] 本发明涉及用于将虚拟机网络流卸载到网络硬件的物理队列的方法、系统和计算机程序产品。由此,本发明的各实施例可使得虚拟机网络话务能够在虚拟机和物理硬件之间直接传递,从而绕过父分区并避免与通过父分区来路由网络话务相关联的低效率。具体地,本发明的各实施例包括配置物理网络硬件以将来自虚拟机的网络流分配到物理网络硬件处的物理队列,并且潜在地将与物理网络硬件处存在的物理队列的数目相比更多的网络流分配到物理队列。

[0010] 在一些实施例中,用于管理网络话务的方法包括计算机系统执行一个或多个虚拟机。该方法还包括计算机系统用被物理网络设备用来管理虚拟机的网络话务的一个或多个

规则来对该物理网络设备进行编程。尤其地,物理网络设备被编程为确定一个或多个物理队列在物理网络设备处的可用性。物理队列可用于处理虚拟机的网络流。物理网络设备还被编程为标识虚拟机的多个网络流,包括标识这些网络流中的每一个的特性。物理网络设备还被编程为基于网络流的特性并基于这些规则将多个网络流中的一个或多个分配给物理队列中的至少一个。

[0011] 提供本概述是为了以简化的形式介绍将在以下详细描述中进一步描述的一些概念。本概述不旨在标识出所要求保护的的主题的关键特征或必要特征,也不旨在用于帮助确定所要求保护的的主题的范围。

[0012] 本发明的附加特征和优点将在以下描述中叙述,且其一部分根据本说明书将是显而易见的,或可通过对本发明的实践来获知。本发明的特征和优点可通过在所附权利要求书中特别指出的工具和组合来实现和获得。本发明的这些以及其它特征、优点和特征将根据以下描述和所附权利要求而变得更显而易见,或者可通过如此后阐述的对本发明的实践而获知。

[0013] 附图简述

[0014] 为了描述可获得本发明的上述和其它优点和特征的方式,将通过参考附图中示出的本发明的具体实施例来呈现以上简要描述的本发明的更具体描述。可以理解,这些附图仅描述本发明的典型实施例,从而不被认为是对其范围的限制,本发明将通过使用附图用附加特征和细节来描述和说明,在附图中:

[0015] 图1示出了可采用本文中所描述的原理的示例性计算系统。

[0016] 图2示出了可采用本文中所描述的原理的环境。

[0017] 图3示出了可采用本文中所描述的原理的主机。

[0018] 图4示出了促成将虚拟机流卸载到物理队列的示例计算机体系结构。

[0019] 图5示出了用于管理网络话务的示例方法的流程图。

[0020] 详细描述

[0021] 本发明涉及用于将虚拟机网络流卸载到网络硬件的物理队列的方法、系统和计算机程序产品。由此,本发明的各实施例可使得虚拟机网络话务能够在虚拟机和物理硬件之间直接传递,从而绕过父分区并避免与通过父分区来路由网络话务相关联的低效率。具体地,本发明的各实施例包括配置物理网络硬件以将来自虚拟机的网络流分配到物理网络硬件处的物理队列,并且潜在地将与物理网络硬件处存在的物理队列的数目相比更多的网络流分配到物理队列。

[0022] 首先,将参考图1-3来描述关于可采用本文中描述的原理的通用计算系统和计算环境的一些介绍性讨论。随后,将参考图4和图5来描述用于将虚拟机网络流卸载到网络硬件的物理队列的基本原理。

[0023] 计算系统现在越来越多地采取多种多样的形式。例如,计算系统可以是手持式设备、电器、膝上型计算机、台式计算机、大型机、分布式计算系统或甚至常规上不被认为是计算系统的设备。在本说明书以及权利要求书中,术语“计算系统”被广义地定义为包括任何设备或系统(或其组合),该设备或系统包含至少一个物理有形的处理器以及其上可能已存储了可由处理器执行的计算机可执行指令的物理有形的存储器。存储器可以采取任何形式,并可以取决于计算系统的性质和形式。计算系统可以分布在网络环境中,并可包括多个

组分计算系统。

[0024] 本文中描述的各实施例可包括或利用专用或通用计算机,该专用或通用计算机包括诸如例如一个或多个处理器和系统存储器等计算机硬件。例如,图1示出了示例性计算系统100。如图1所示,在其最基本的配置中,计算系统100通常包括至少一个处理单元102和存储器104。存储器104可以是物理系统存储器,该物理系统存储器可以是易失性、非易失性、或两者的某种组合。术语“存储器”在此也可用来指诸如物理存储介质等非易失性大容量存储。如果计算系统100是分布式的,则处理、存储器和/或存储能力也可以是分布式的。如此处所使用的那样,术语“模块”或“组件”可以指代在计算系统100上执行的软件对象或例程。此处所描述的不同组件、模块、引擎,以及服务可以实现为在计算系统100上执行的对象或进程(例如,作为分开的线程)。

[0025] 在随后的描述中,参考由一个或多个计算系统(诸如,计算系统100)执行的动作描述了各实施例。如果这样的动作是以软件实现的,则执行动作的相关联计算系统的一个或多个处理器响应于已经执行了计算机可执行指令来引导计算系统的操作。这样的操作的示例涉及对数据的操纵。在计算系统100的上下文中,计算机可执行指令(以及被操纵的数据)可被存储在存储器104中。计算系统100还可包含允许计算系统100例如通过网络110与其他消息处理器通信的通信信道108。

[0026] 这里描述的各实施例还包括用于承载或存储计算机可执行指令和/或数据结构的物理和其他计算机可读介质。这样的计算机可读介质可以是可由通用或专用计算机系统访问的任何可用介质。存储计算机可执行指令的计算机可读介质是物理存储介质。承载计算机可执行指令的计算机可读介质是传输介质。因此,作为示例而非限制,本发明的各实施例可包括至少两种明显不同种类的计算机可读介质:计算机存储介质和传输介质。

[0027] 计算机存储介质包括可记录型存储介质,诸如RAM、ROM、EEPROM、CD-ROM或其他光盘存储、磁盘存储或其他磁存储设备、或可用于存储计算机可执行指令或数据结构形式的所需程序代码装置且可由通用或专用计算机访问的任何其他介质。

[0028] “网络”被定义为使得电子数据能够在计算机系统和/或模块和/或其它电子设备之间传输的一个或多个数据链路。当信息通过网络或另一个通信连接(硬连线、无线、或者硬连线或无线的组合)传输或提供给计算机时,该计算机将该连接适当地视为传输介质。传输介质可包括可用于携带计算机可执行指令或数据结构形式的所需程序代码装置且可由通用或专用计算机访问的网络(例如,网络110)和/或数据链路。上述的组合也应被包括在计算机可读介质的范围内。

[0029] 此外,在到达各种计算机系统组件之后,计算机可执行指令或数据结构形式的程序代码装置可从传输介质自动转移到计算机存储介质(或者相反)。例如,通过网络或数据链路接收到的计算机可执行指令或数据结构可被缓存在网络接口模块(例如,“NIC”)内的RAM中,然后最终被传输到计算机系统RAM和/或计算机系统处的较不易失性的计算机存储介质。因而,应当理解,计算机存储介质可被包括在还利用(或甚至主要利用)传输介质的计算机系统组件中。

[0030] 计算机可执行指令例如包括,当在处理器处执行时使通用计算机、专用计算机、或专用处理设备执行某一功能或某组功能的指令和数据。计算机可执行指令可以是例如二进制代码、诸如汇编语言之类的中间格式指令、或甚至源代码。尽管已用对结构特征和/或方

法动作专用的语言描述了本主题,但可以理解,所附权利要求书中定义的主题不必限于此处所描述的所描述特征或动作。更具体而言,本文中描述的特征和动作是作为实现权利要求书的示例形式而公开的。

[0031] 本领域的技术人员将理解,本发明可以在具有许多类型的计算机系统配置的网络计算环境中实践,这些计算机系统配置包括个人计算机、台式计算机、膝上型计算机、消息处理器、手持式设备、多处理器系统、基于微处理器的或可编程消费电子设备、网络PC、小型计算机、大型计算机、移动电话、PDA、平板、寻呼机、路由器、交换机等等。本发明也可在其中通过网络链接(或者通过硬连线数据链路、无线数据链路,或者通过硬连线和无线数据链路的组合)的本地和远程计算机系统两者都执行任务的分布式系统环境中实施。在分布式系统环境中,程序模块可以位于本地和远程存储器存储设备二者中。

[0032] 图2抽象地示出了可采用本文中所描述的原理的环境200。在环境200中,多个客户机210使用接口202与系统210进行交互。环境200被示为具有三个客户机201A、201B和201C,但是省略号201D表示本文中描述的原理不受限于通过接口202与系统210对接的客户机的数目。系统210可按需向客户机201提供服务,并且由此从系统210接收服务的客户机201的数目可随时间改变。

[0033] 客户机201中的一个或多个例如可如上根据图1的计算系统100所述来结构化。替换地或另选地,客户机201中的一个或多个可以通过接口202与系统210对接的应用或其他软件模块。接口202可以是按以下方式来定义的应用程序接口(API):使得能够使用该API的任何计算系统或软件实体可与系统210进行通信。

[0034] 系统210可以是分布式系统,但这不是必须的。在一个实施例中,系统210是云计算环境。云计算环境可以是分布式的,但并不是必需的,并且可甚至分布在世界各地和/或具有跨多个组织拥有的组件。

[0035] 在该描述和下面的权利要求书中,“云计算”被定义为用于允许对可配置计算资源(例如,网络、服务器、存储、应用和服务)的共享池的按需网络访问的模型。“云计算”的定义不限于可从这样的模型(在被合适地部署时)中获得的任何其他多个优点。

[0036] 例如,云计算当前被用于市场,以便提供对可配置计算资源的共享池的普遍存在且方便的按需访问。此外,可配置计算资源的共享池可经由虚拟化而被快速地供应,并可利用低管理努力或服务提供商干预来发布,并随后因此被缩放。

[0037] 云计算模型可由各种特性组成,诸如按需自服务、广泛网络访问、资源池、快速灵活性、和所测定的服务等。云计算模型还可形成各种服务模型,诸如例如软件即服务(SaaS)、平台即服务(PaaS)以及基础结构即服务(IaaS)。云计算模型还可以使用不同的部署模型来部署,诸如私有云、社区云、公共云和混合云等。在该描述和权利要求书中,“云计算环境”是采用云计算的环境。

[0038] 如所描绘的,系统210包括多个主机211,这些主机各自能够运行虚拟机。虽然系统200可包括任何数目的主机211,但在图2中示出了三个主机211A、211B和211C,且省略号211D表示本文中描述的原理不受限于在系统210内的主机确切数目。可存在少至一个,但没有上限。此外,主机的数目可以是静态的,或者可由于向系统210添加新主机或从系统210中丢弃主机而随时间动态地改变。主机211中的每一个可以如上针对图1的计算系统100所述来结构化。

[0039] 每一主机能够运行一个或多个虚拟机,并潜在地运行许多虚拟机。例如,图3更详细地抽象示出主机300。作为示例,主机300可表示图2中的主机211中的任一个。在图3的情况下,主机300被示为操作三个虚拟机310,包括虚拟机310A、310B和310C。然而,省略号310D再次表示本文中描述的原理不受限于在主机300上运行的虚拟机的数目。可存在少至零个虚拟机在主机上运行,而仅上限由主机300的物理能力来定义。

[0040] 在操作期间,虚拟机模拟完全可操作的计算系统,包括至少一个操作系统并且也许还包括一个或多个其他应用。每一虚拟机均被分配给特定客户机,并负责支持那个客户机的桌面环境。

[0041] 虚拟机生成桌面映像或其他表示桌面的当前状态的呈现指令,并接着将该映像或指令传送到客户机以供呈现桌面。例如,参考图2和图3,假设图3的主机300表示图2的主机211A,并且虚拟机310A被分配给客户机201A(本文中被称为“主要示例”),则虚拟机310A可生成桌面映像或指令,并经由服务协调系统213和系统接口202将这样的指令从主机211A分派至相应的客户机201A。

[0042] 随着用户与客户机处的桌面进行交互,用户输入从客户机被传送到虚拟机。例如,在该主要示例中,并参考图2和图3,客户机201A的用户与桌面进行交互,并且用户输入是经由接口201、服务协调系统213和主机211A从客户机201传送到虚拟机310A的。

[0043] 虚拟机处理用户输入,并且如果合适,则改变桌面状态。如果桌面状态中这样的改变将导致在所呈现的桌面中的改变,则虚拟机更改映像或呈现指令(如果合适的话),并将经更改的映像或呈现的指令传送到客户机计算系统以供合适的呈现。从用户的角度而言,就好像客户机计算系统本身执行桌面处理。

[0044] 主机300包括使用从虚拟机310的示图中抽象出的物理资源321来模拟虚拟机310的虚拟资源的管理程序320。管理程序321还在虚拟机310之间提供合适的隔离。因此,从任何给定虚拟机的角度来看,管理程序320提供以下幻像:该虚拟机正与物理资源对接,即使该虚拟机仅与物理资源的外观(例如,虚拟机)对接而未与物理资源直接对接。在图3中,物理资源321被抽象地表示为包括资源321A到321F。物理资源321的示例包括处理能力、存储器、盘空间、网络带宽、和媒体驱动器等。

[0045] 主机300可操作监视主机的性能并执行管理该主机的其他操作的主机代理302。此外,主机300可包括其他主机303,诸如如稍后描述的虚拟交换机。

[0046] 参考回图2,系统200还包括服务212。在示出的示例中,主机200包括五个不同的服务212A、212B、212C、212D和212E,但是省略号211F表示本文中描述的原理不受限于在系统210中的服务的数目。服务协调系统213与主机211和服务212进行通信,以由此提供客户机201所请求的服务以及可作为所请求的服务的先决条件的其他服务(诸如认证、和计费等)。

[0047] 现在转向图4,图4示出了促成将虚拟机流卸载到物理队列的示例计算机体系结构400。如所描绘的,计算机体系结构400包括主机402、一个或多个虚拟机412(包括虚拟机412a)、和一个或多个物理网络设备416(包括物理网络设备416a)。

[0048] 主机402被配置成提供虚拟化环境。在一些实施例中,主机402可对应于图3中的主机300。例如,主机402可包括父分区(其执行主操作系统)和一个或多个子分区。每一子分区可被视为提供用于执行相应的虚拟机(诸如虚拟机412a)的虚拟化硬件环境。主机402可使用云计算环境的代表承租者来托管虚拟机的一部分。

[0049] 虚拟机412中的每一个(包括虚拟机412a)执行一个或多个虚拟化应用,诸如操作系统、应用软件等。虚拟机412中的每一个能够发送和接收网络分组。例如,虚拟机412中的每一个包括网络栈(例如,TCP/IP栈),并能够通过主机402在数据路径432上和/或通过物理网络设备416在数据路径430上发送和/或接收网络分组和其他信息。由此,虚拟机412可创建网络流。

[0050] 每一物理网络设备416使用一个或多个外部接口连接到其他计算机系统和/或网络。图4描绘了物理网络设备416a使用外部接口426连接到网络434。物理网络设备416可包括任何合适类型的物理联网设备,诸如NIC、交换机等。

[0051] 此外,每一物理网络设备416包括与虚拟化环境兼容的物理硬件。例如,图4描绘了物理网络设备416a向虚拟机412呈现虚拟功能424。具体地,物理网络设备416a可向虚拟机412中的每一个呈现一个或多个虚拟功能。例如,图4描绘了物理网络设备416a向虚拟机412a呈现虚拟功能424a。虚拟机412中的每一个进而包括相应的虚拟功能驱动程序。例如,图4描绘了虚拟机412a包括虚拟功能驱动程序414。由此,虚拟机412中的每一个可通过数据路径430来访问其相应的虚拟功能424,并可使用数据路径430来与物理网络设备416a通信网络分组,而无需通过主机402来路由网络分组。在与通过主机402来路由网络分组时相比,这样做可降低处理器使用和网络等待时间。

[0052] 此外,图4还描绘了物理网络设备416a将物理功能418呈现给主机402。图4还描绘了主机402包括相应的物理功能驱动程序410,并且数据路径428将物理网络设备416a处的物理功能418和主机402处的物理功能驱动程序410相连接。如此,物理功能418和物理功能驱动程序410可操作以在物理网络设备416a和主机402之间交换网络分组。

[0053] 如先前指示的,在一些实施例中,物理NIC 110包括遵循SRIOV的PCIe硬件。在此类实施例中,虚拟功能424或物理功能418中的一者或多者可包括PCIe功能。然而,将领会,本发明描述的原理可应用于各种硬件设备,而限于遵循SRIOV的设备或PCIe设备。

[0054] 物理网络设备416中的每一个可包括一个或多个物理队列,此一个或多个物理队列可由物理网络设备416在处理与虚拟机412相关联的网络流时使用。例如,图4描绘了物理网络设备416a包括物理队列422,包括队列422a和任何附加数目的(即,零个或更多个)附加物理队列,如水平省略号和队列422n所表示的。根据一个或多个实施例,主机402将物理网络设备416中的一个或多个配置为在处理虚拟机412的网络流时管理其物理队列的使用。如所描绘的,例如,主机402处的虚拟交换机404可包括规则406。使用规则406,虚拟交换机404可用规则420来编程物理网络设备416a,并可将物理网络设备416a编程为基于这些规则来管理到物理队列422的网络流分配。规则420可与规则406相同,可以某种方式改变,和/或可包括规则子集406。由此,物理网络设备416a可被配置成高效地处理来自虚拟机412的网络流,包括作出网络流到物理队列422的分配,而无需为了每一网络流而牵涉主机402。

[0055] 规则420可包括使得物理网络设备416a能够将多个网络流分配给各物理队列422的规则,网络流在数目上比物理队列422处存在的队列的数目要多。在简单示例中,来自虚拟机412的网络流可涉及八个活跃网络流,而物理网络设备416a可使用规则420将这八个流分配给物理队列422中的仅四个可用的队列。物理网络设备416a可被配置成使得网络流基于流的特性和/或基于流的分类来进行队列分配。在一些实施例中,物理网络设备416a基于网络流的特性并基于规则420将这些流置于不同的分类中。在一些附加或替换实施例中,

物理网络设备416a基于虚拟机412所作出的建议将网络流置于不同的分类中。例如,虚拟机412a可向流附上某属性,或可将所建议的分类传递给与流分开的物理功能418。

[0056] 规则420可允许各种类型的队列分配算法。例如,规则420可指定具有相对较低话务等级的多个网络流可一起被分配在单个物理队列上,而具有相对较高话务等级的流将各自被排他地分配给相应的物理队列。在另一实施例中,规则420可指定具有相似或相容要求的多个流可被组合在同一队列上。例如,如果多个流的网络分组将以类似速率行进(速率有限),则那些流可一起被分配在单个物理队列上。其他相似或相容要求可包括优先级(例如,将低优先级的流一起分组在单个队列上)、服务质量(QoS)(例如,将具有低QoS要求的流一起分组在单个队列上)等。规则420还可指定来自同一虚拟机的流将被分组在物理队列中的单个物理队列或组上。由此,本发明中的各实施例可促成在各虚拟机412中划分各硬件资源。

[0057] 在一些实施例中,物理网络设备416和虚拟交换机404可一起用于使其间的网络流的执行平衡。例如,图4描绘了虚拟交换机404可包括基于软件的虚拟队列408(包括队列408a和任何附加数目(即,零个或更多个)的附加队列,如水平省略号和队列408n所表示的)。由此,一些网络流可被分配给物理队列422,且一些流可被分配给虚拟队列408。将领会,与虚拟队列408相比,物理队列422可提供更快、更细粒度和/或更可靠的性能。由此,网络流可被分类成应该被分配给物理网络设备416a处的物理队列422以利用更快、更细粒度和/或更可靠性能的流,以及可被分配给主机402处的虚拟队列(因为快速、细粒度和/或可靠性能对这些流而言并不一样重要)的流。这样的分配可由虚拟机412建议,和/或可由物理网络设备416和/或虚拟交换机404作出。

[0058] 在一些实施例中,流可通过多个物理网络设备416(例如,NIC和交换机),并且主机402可将每一物理网络设备编程为独立地处理流。例如,一个物理网络设备可被编程为将该流分配给该设备处的单个物理队列,而另一物理网络设备可被编程为将这个流和其他流结合在一起分配给该设备处的单个物理队列。

[0059] 图5示出了用于管理网络话务的方法500的流程图。方法500将参考计算机架构400的组件和数据来描述。

[0060] 方法500包括执行一个或多个虚拟机的动作(动作502)。例如,主机402可执行虚拟机412,虚拟机412可包括虚拟机412a。在一些实施例中,动作502可包括以半虚拟化形式(包括使用一个或多个遵循SRIOV的物理网络设备)来执行虚拟机。由此,至少一个物理网络设备(例如,物理网络设备418)可向虚拟机412a呈现虚拟功能(例如,虚拟功能424a),并且虚拟机412a可包括用于直接与物理网络设备传递网络分组的相应虚拟功能驱动程序(例如,虚拟功能驱动程序414)。

[0061] 方法500还包括用一个或多个规则来对物理网络设备进行编程的动作,这一个或多个规则被配置成管理一个或多个虚拟机的网络话务(动作504)。例如,虚拟交换机404可用规则420来对物理网络设备416a进行编程。规则420可以是虚拟交换机404处的规则406的副本,或基于规则406。规则420可被配置成使得物理网络设备416a能够在与虚拟机412相关联的网络流和物理网络设备416a处的物理队列422之间作出分配。

[0062] 方法500还包括将物理网络设备编程为管理网络话务的动作(动作506)。例如,虚拟交换机404可将物理网络设备416a配置成基于规则420来作出流分配。在一些实施例中,

将物理网络设备416a编程为管理网络话务作为用规则420来对物理网络设备416a进行编程的结果而发生。在其他实施例中,将物理网络设备416a编程为管理网络话务包括用附加计算机可执行指令和/或附加配置设置来明确地对物理网络设备416a进行编程。

[0063] 动作506包括将物理网络设备编程为确定物理网络设备处可用于处理一个或多个虚拟机的网络流的一个或多个物理队列的可用性(动作508)。例如,物理网络设备416a可被配置成标识物理队列422,包括物理队列422的目前可用性。

[0064] 动作506包括将物理网络设备编程为标识一个或多个虚拟机的多个网络流,包括标识这多个网络流中的每一个的一个或多个特性(动作510)。例如,物理网络设备416a可被配置成标识与虚拟机412相关联的网络流。物理网络设备416a还可被配置成分析这些流的特性、来自虚拟机412的分类建议、或任何其他合适的信息,以对这些流进行分类或以其他方式对其进行归类。

[0065] 动作506包括将物理网络设备编程为基于多个网络流中的每一个的一个或多个特性并基于一个或多个规则将多个网络流中的一个或多个分配给一个或多个物理队列中的至少一个(动作512)。例如,基于规则420并基于在动作510标识出的特性和分类,物理网络设备416a可将这些流分配给物理队列422。在这样做时,物理网络设备416a可将超过物理队列的数目的多个流分配给物理队列422。例如,物理网络设备416a可将具有类相似特性、相容优先级或话务负载等的流分配给同一物理队列。另选地或替换地,物理网络设备416a可与物理交换机404一起用于将第一流子集分配给虚拟交换机404处的虚拟队列408,并将第二流子集分配给物理网络设备416a处的物理队列422。

[0066] 因此,本文中描述的各实施例可通过使物理网络设备能够在流和队列之间作出分配来改善物理硬件的网络性能和利用。当作出这样的分配时,本文中描述的各实施例可使得物理硬件能够通过用物理队列来处理比可用物理队列的数目更多数目的流。另选地或替换地,当作出这样的分配时,本文中描述的实施例可使得物理硬件可在物理队列和虚拟队列之间平衡流处理。

[0067] 本发明可具体化为其它具体形式而不背离其精神或本质特征。所描述的实施例在所有方面都应被认为仅是说明性而非限制性的。因此,本发明的范围由所附权利要求书而非前述描述指示。落入权利要求书的等效方案的含义和范围内的所有改变应被权利要求书的范围所涵盖。

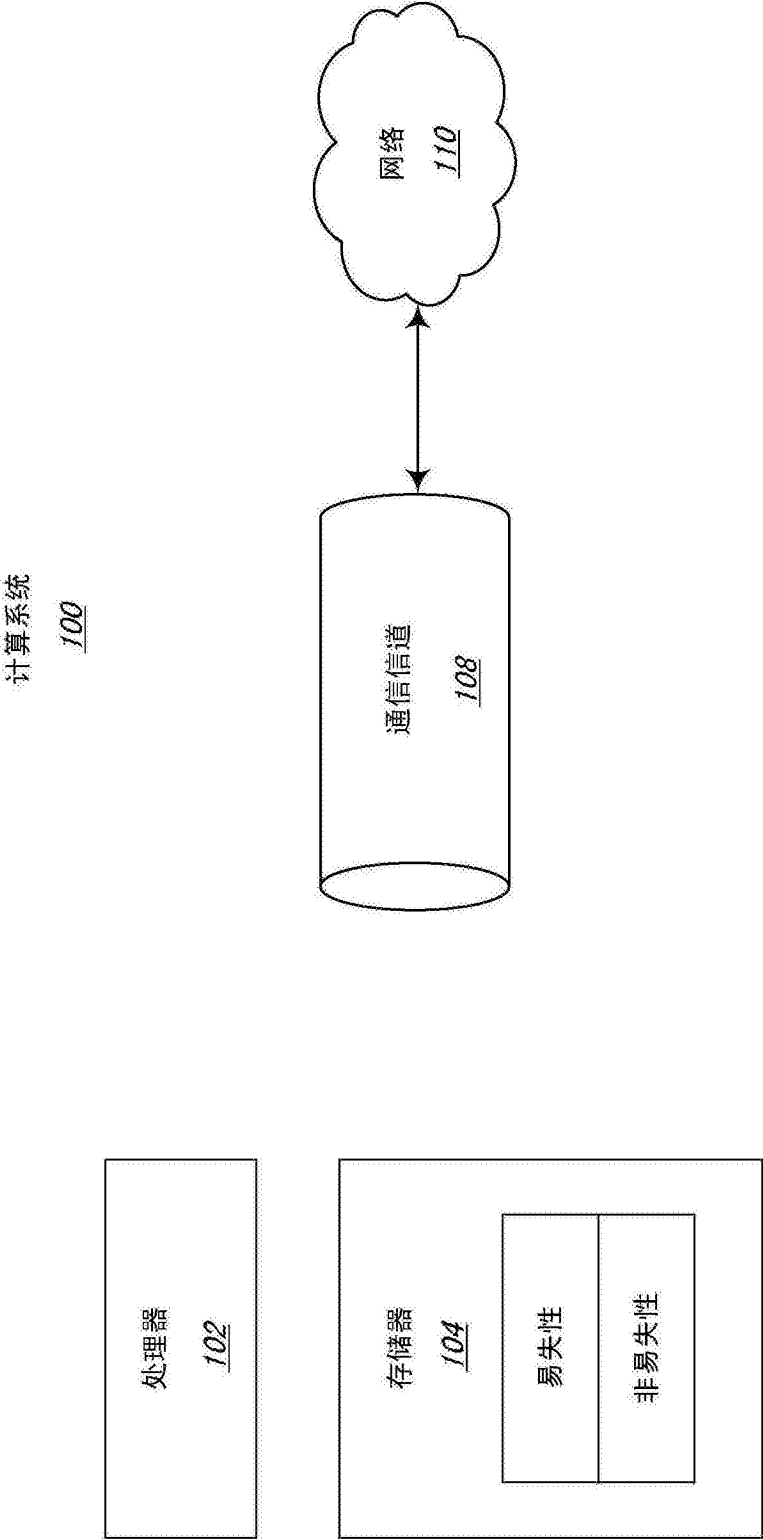


图1

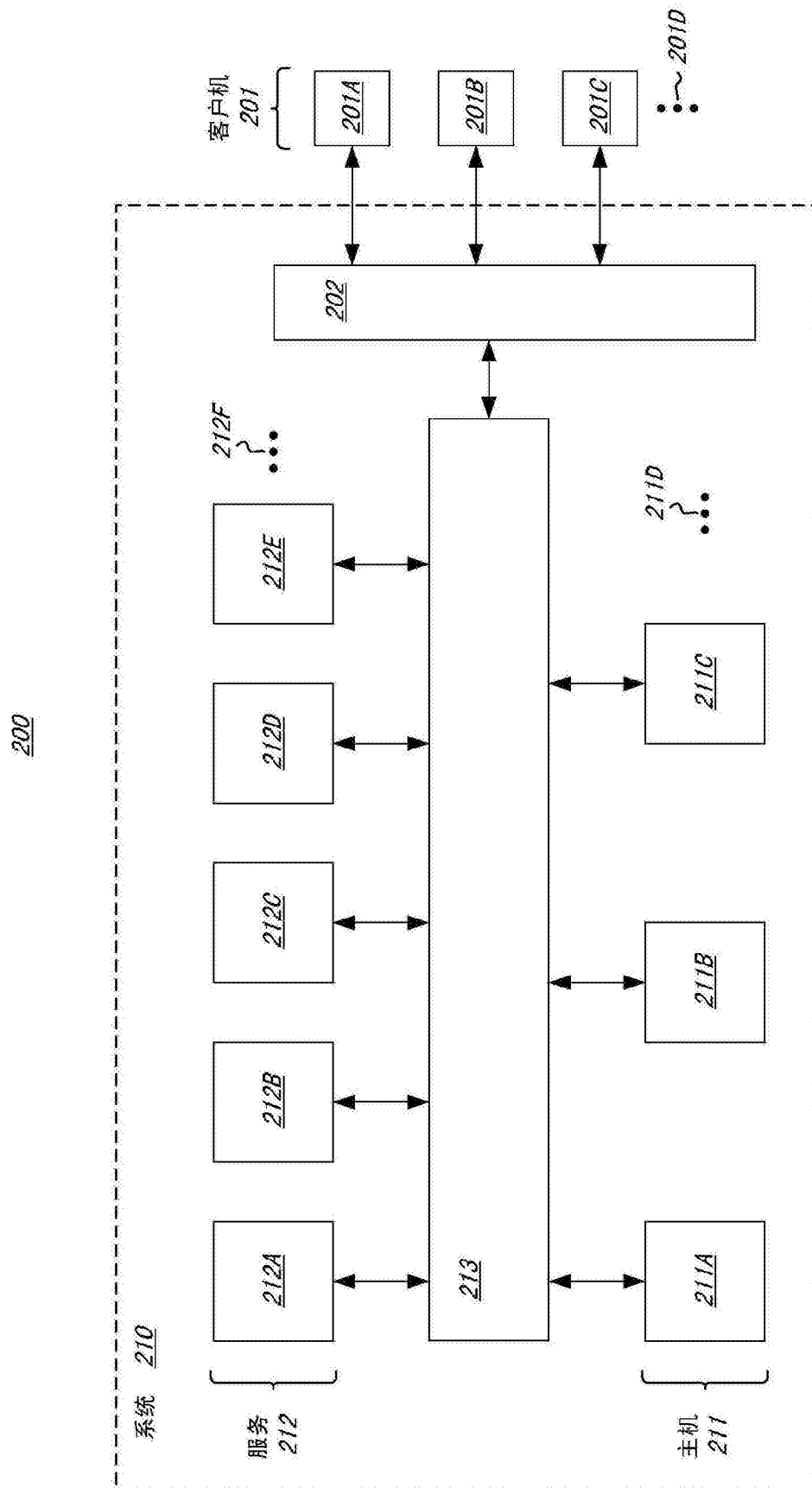


图2

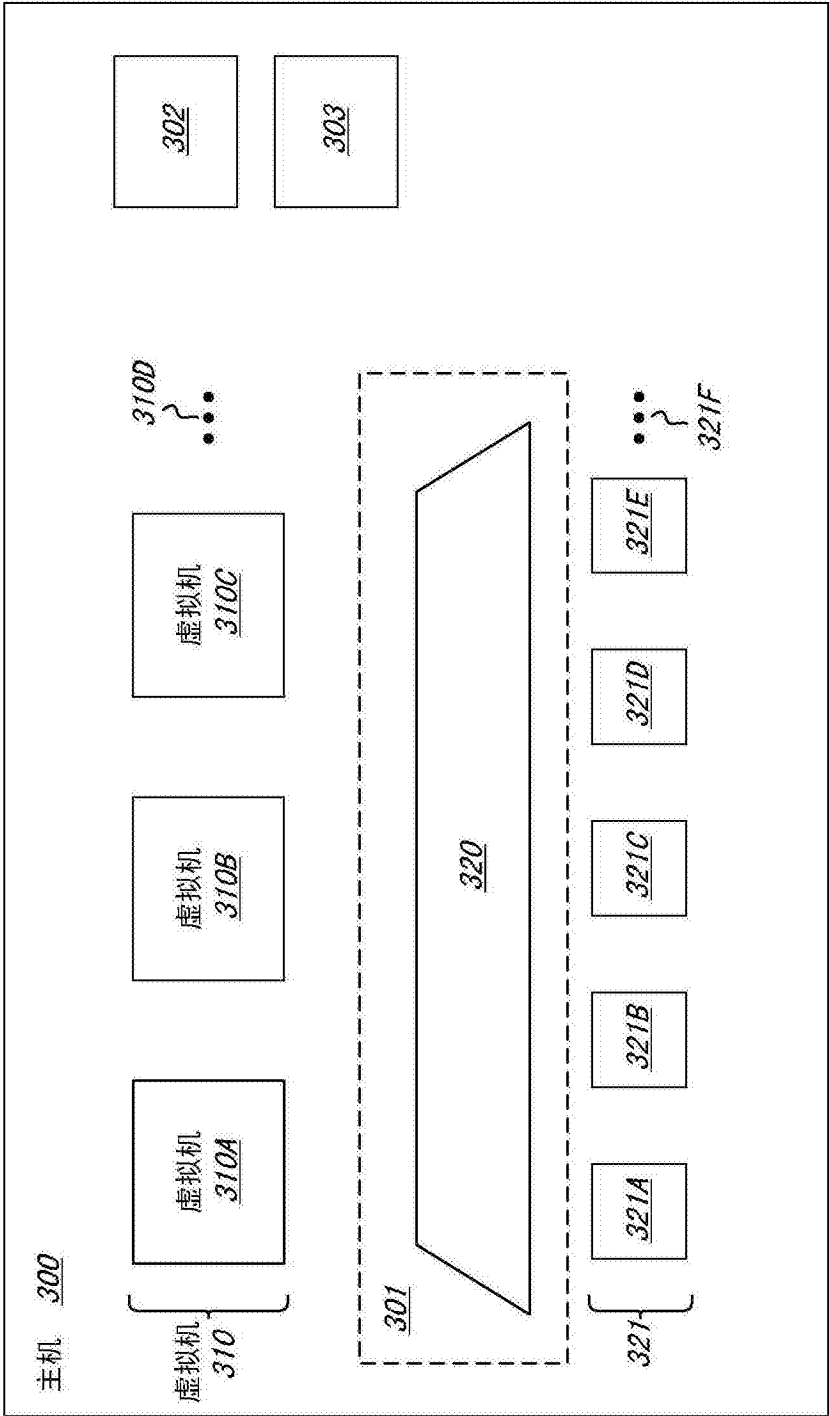


图3

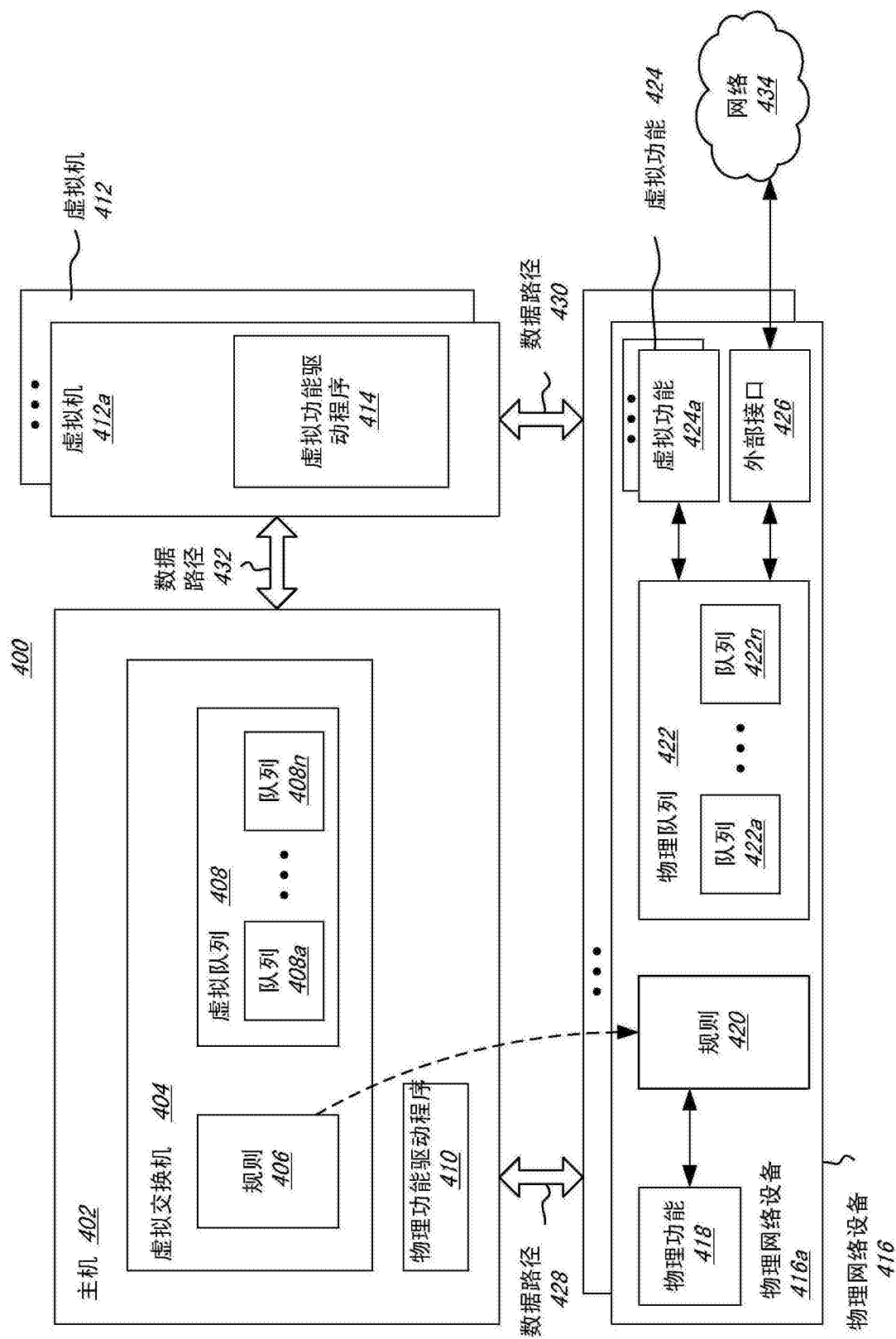


图4

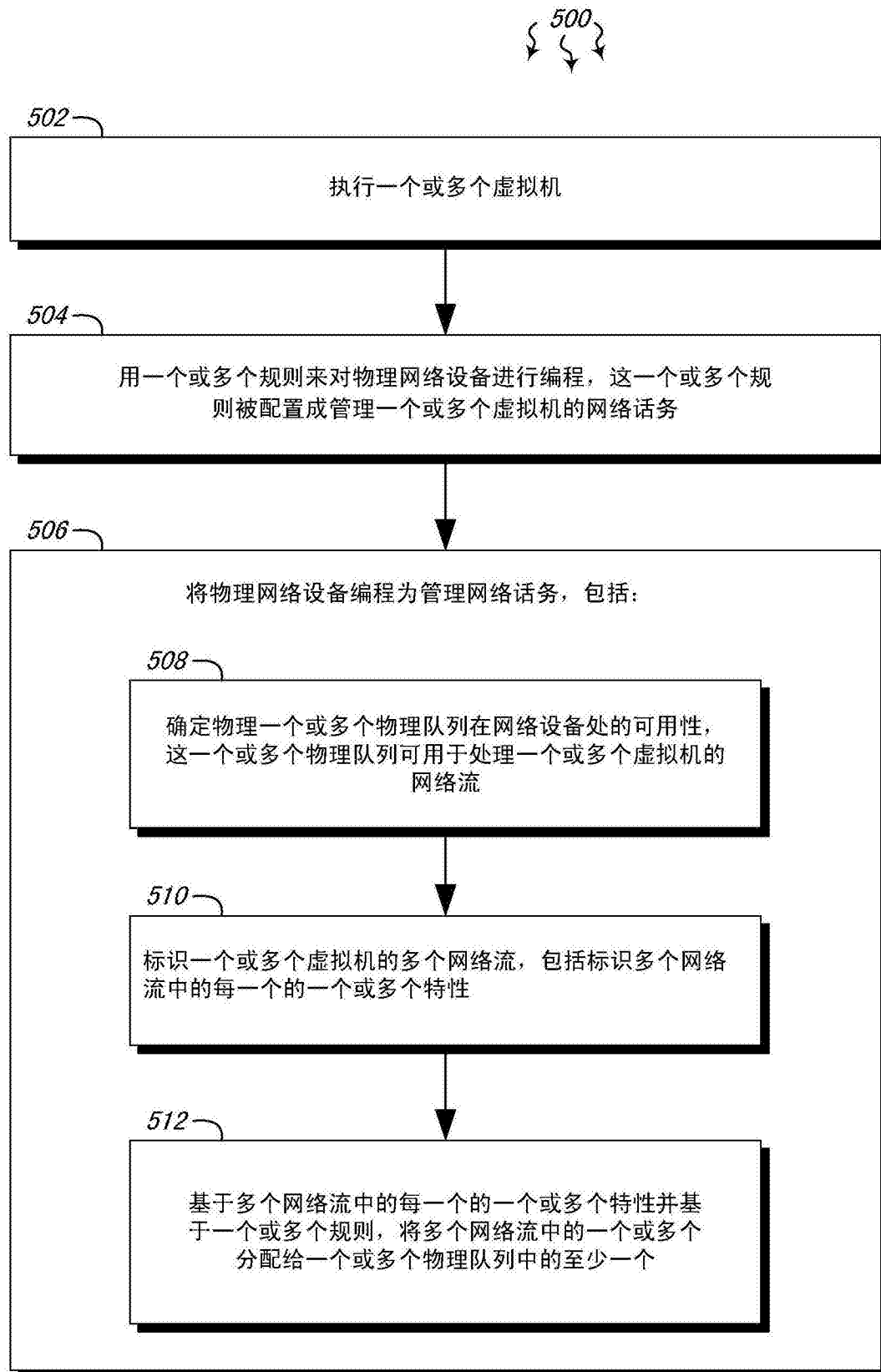


图5