



(12) 发明专利申请

(10) 申请公布号 CN 104717156 A

(43) 申请公布日 2015. 06. 17

(21) 申请号 201410690411. 8

(22) 申请日 2014. 11. 25

(30) 优先权数据

14/105, 442 2013. 12. 13 US

(71) 申请人 国际商业机器公司

地址 美国纽约阿芒克

(72) 发明人 C. M. 德库萨蒂斯 K. G. 坎布利

(74) 专利代理机构 北京市柳沈律师事务所

11105

代理人 邸万奎

(51) Int. Cl.

H04L 12/863(2013. 01)

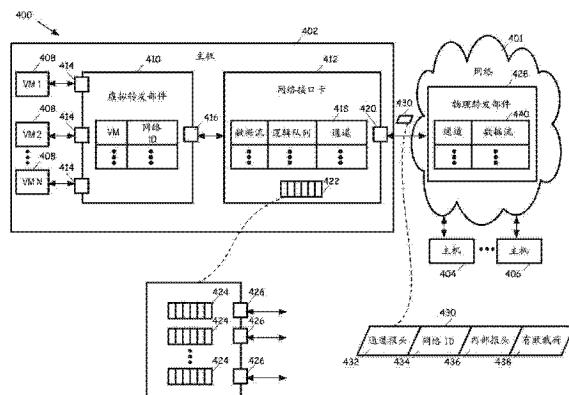
权利要求书2页 说明书11页 附图7页

(54) 发明名称

使用网络接口卡管理软件定义网络中的数据流的方法和系统

(57) 摘要

本发明提供一种用于在附接到主机的网络接口卡(NIC)处管理数据流的计算机实现的方法和系统以及用于管理数据流的系统。实施例涉及在NIC处管理数据流。NIC包括端口。所述方法从运行在所述主机中的虚拟转发部件接收属于数据流的分组。所述方法通过NIC基于在数据流、队列与多个通道之间的映射识别用于存储要通过所述端口发送到所述主机之外的数据流的若干队列之一。所述方法将所述分组放置于所识别的队列中，以便将所述分组通过所述端口发送到所述主机之外。



1. 一种用于在附接到主机的网络接口卡 NIC 处管理数据流的计算机实现的方法，所述 NIC 包括端口，所述方法包括：

从运行在所述主机中的虚拟转发部件接收属于数据流的分组；

通过所述 NIC 基于在数据流、队列与多个通道之间的映射识别用于存储要通过所述端口发送到所述主机之外的数据流的若干队列之一；以及

将所述分组放置于所识别的队列中，以便将所述分组通过所述端口发送到所述主机之外。

2. 如权利要求 1 所述的方法，其中所述多个队列是逻辑队列，所述方法还包括从与所述端口关联的物理缓冲器创建所述多个逻辑队列。

3. 如权利要求 2 所述的方法，还包括利用通道报头封装所述分组，所述通道报头定义所述主机与包括所述分组的目的地的另一主机之间的通道。

4. 如权利要求 3 所述的方法，其中所述通道报头包括所述主机的网络地址。

5. 如权利要求 1 所述的方法，其中所述虚拟转发部件在将所述分组发送到所述 NIC 之前向所述分组附接网络标识符。

6. 如权利要求 1 所述的方法，其中所述虚拟转发部件通过所述虚拟转发部件的不同虚拟端口从运行在所述主机上的不同虚拟机接收不同数据流，并将所述不同数据流转发到所述 NIC。

7. 如权利要求 1 所述的方法，还包括：

从位于所述 NIC 下游的转发部件接收拥塞消息，所述消息指示一个或多个特定数据流贡献网络拥塞；

针对所述特定数据流识别所述多个队列中的一个或多个特定队列；以及

通过使用所述特定队列调节所述特定数据流的数据率。

8. 如权利要求 7 所述的方法，还包括向所述虚拟转发部件通知所述拥塞消息。

9. 一种用于在附接到主机的网络接口卡 (NIC) 处管理数据流的系统，所述 NIC 包括端口，所述系统包括配置为执行权利要求 1 至 8 中的任一项的方法步骤的装置。

10. 一种用于管理数据流的系统，所述系统包括：

转发部件，其主机通信地连接，所述主机上运行多个虚拟机 (VM)，所述系统被配置为执行包括以下步骤的方法：

从所述主机接收分组，所述分组属于来源于所述主机的所述虚拟机之一的数据流，所述分组用通道报头和网络标识符封装；

通过检查所封装的分组的所述网络标识符以及报头来识别所述数据流；

将所述数据流映射到由所述通道报头定义的通道；以及

将所封装的分组转发到由所述通道报头指明的目的地。

11. 如权利要求 10 所述的系统，其中所述方法还包括：

接收第一拥塞消息，所述第一拥塞消息指示特定通道中的流量数据贡献网络拥塞；

识别映射到所述特定通道的一个或多个数据流；

生成第二拥塞消息，所述第二拥塞消息指示所识别的数据流贡献所述网络拥塞；

将所述第二拥塞消息发送到所述主机。

12. 如权利要求 11 所述的系统，其中所述主机包括网络接口卡，所述网络接口卡接收

所述第二拥塞消息并分别调节针对所识别的数据流的数据率。

13. 如权利要求 11 所述的系统，其中所述第一拥塞消息具有第一格式，所述第二拥塞消息具有不同所述第一格式的第二格式。

14. 如权利要求 10 所述的系统，其中由所述通道报头指明的目的地是另一主机。

15. 如权利要求 10 所述的系统，其中所述方法还包括创建在从所述主机接收的多个数据流与携带所述多个数据流的多个通道之间的映射。

# 使用网络接口卡管理软件定义网络中的数据流的方法和系统

## 技术领域

[0001] 本发明涉及计算机网络,更具体地涉及使用网络接口卡管理软件定义网络中的数据流。

## 背景技术

[0002] 在数据中心的环境中,典型的主机机器运行许多虚拟机(VM),这些虚拟机向其它虚拟或非虚拟机提供服务或从其接收服务。在提供或接收服务时,主机中的VM可以彼此通信或者可以与在其它主机上运行的其它VM通信。这些机器之间的通信以数据流的形式,数据流包括具有共同属性(例如共同报头)的数据分组。在某些情况下,主机中的VM共享附接到该主机中的一个或多个网络接口控制器,以发送或接收数据流。

## 发明内容

[0003] 实施例包括用于管理软件定义网络(SDN)中的数据流的方法、系统和计算机程序产品。根据本发明的一个实施例,提供了一种用于在主机的网络接口卡(NIC)处管理数据流的计算机实现的方法。所述NIC包括端口。所述方法从运行在所述主机中的虚拟转发部件接收属于数据流的分组。所述方法通过NIC基于在数据流、队列与多个通道之间的映射识别用于存储要通过所述端口发送到所述主机之外的数据流的若干队列之一。所述方法将所述分组放置于所识别的队列中,以便将所述分组通过所述端口发送到所述主机之外。

[0004] 根据本发明的另一实施例,提供了用于在主机的NIC处管理数据流的计算机程序产品。所述NIC包括端口。所述计算机程序产品包括有形存储介质,其可由处理电路读取并存储由处理电路执行来执行方法的指令。所述方法从运行在所述主机中的虚拟转发部件接收属于数据流的分组。所述方法通过NIC基于在数据流、队列与多个通道之间的映射识别用于存储要通过所述端口发送到所述主机之外的数据流的若干队列之一。所述方法将所述分组放置于所识别的队列中,以便将所述分组通过所述端口发送到所述主机之外。

[0005] 根据本发明的再一实施例,提供了一种用于管理数据流的系统。所述系统包括与主机通信地连接的转发部件,所述主机上运行多个虚拟机(VM)。所述系统被配置为执行方法。所述方法从所述主机接收分组。所述分组属于来源于所述主机的所述虚拟机之一的数据流。所述分组用通道报头和网络标识符封装。所述方法通过检查所封装的分组的所述网络标识符以及报头来识别所述数据流。所述方法将所述数据流映射到由所述通道报头定义的通道。所述方法将所封装的分组转发到由所述通道报头指明的目的地。

[0006] 通过本发明的技术实现附加特征和优点。本发明的其它实施例和方面在本文中进行了详细描述,并被认为是所要求权利的发明的一部分。为了更好地理解具有优点和特征的本发明,参考说明书和附图。

## 附图说明

[0007] 在说明书完结时的权利要求中特别地指出并清楚地主张了被认作本发明的主题。结合附图,从下面对的详细描述中,本发明的上述和 / 或其他特征和优点是显然的,其中:

- [0008] 图 1 描绘根据一实施例的云计算节点;
- [0009] 图 2 描绘根据一实施例的云计算环境;
- [0010] 图 3 描绘根据一实施例的抽象模型层;
- [0011] 图 4 描绘根据一实施例的用于管理数据流的系统的框图;
- [0012] 图 5 描绘根据一实施例的用于在 NIC 处管理数据流的流程图;
- [0013] 图 6 描绘根据一实施例的用于在物理转发部件处管理数据流的流程图;
- [0014] 图 7 描绘根据一实施例的用于处理拥塞消息的流程图;
- [0015] 图 8 描绘根据一实施例的用于配置虚拟转发部件的流程图。

## 具体实施方式

[0016] 示例实施例涉及通过主机的物理网络接口卡 (NIC) 以及通过与 NIC 通信地连接的物理转发部件 (例如交换机或网关),管理来源于运行在主机中的虚拟机的数据流。通常,对于 NIC 的物理输出端口有设定量的物理资源 (例如,一个或多个物理缓冲器或队列) 可用。通过 NIC 的输出端口发送到主机之外的所有数据流量共享所关联的物理资源。在一个实施例中,NIC 被配置为将物理缓冲器分区为若干逻辑队列,并将来源于主机内的每个数据流与一个逻辑队列关联。NIC 在将数据流发送到主机之外之前将数据流存储在关联的逻辑队列中。使用这些逻辑队列,NIC 能够分别地调节数据流的数据率。

[0017] 在一个实施例中,NIC 用通道报头 (也称为承载 (underlay) 网络报头或外部网络报头) 封装数据流的分组,该通道报头定义两个主机之间或者主机与覆盖网络网关交换机 (overlay gateway switch) 之间的通道。网络中除了覆盖网络网关交换机之外的连接两个主机的网络组件 (例如,交换机、路由器、网关等) 通常仅使用通道报头将分组从一个主机转发到另一个主机。即,网络组件不在意通道化数据流的内部封装分组。在一个实施例中,从 NIC 接收通道化数据流的网络组件被配置为通过检查内部分组的报头而辨识通道内不同的数据流。通过辨识通道中的数据流,网络组件可以请求 NIC 针对数据流分别调节数据率,而不是请求 NIC 针对所有通道化数据流整体调节数据率。

[0018] 首先应当理解,尽管本公开包括关于云计算的详细描述,但其中记载的技术方案的实现却不限于云计算环境,而是能够结合现在已知或以后开发的任何其它类型的计算环境而实现。

[0019] 云计算是一种服务交付模式,用于对共享的可配置计算资源池进行方便、按需的网络访问。可配置计算资源是能够以最小的管理成本或与服务提供者进行最少的交互就能快速部署和释放的资源,例如可以是网络、网络带宽、服务器、处理、内存、存储、应用、虚拟机和服务。这种云模式可以包括至少五个特征、至少三个服务模型和至少四个部署模型。

[0020] 特征包括:

[0021] 按需自助式服务:云的消费者在无需与服务提供者进行人为交互的情况下能够单方面自动地按需部署诸如服务器时间和网络存储等的计算能力。

[0022] 广泛的网络接入:计算能力可以通过标准机制在网络上获取,这种标准机制促进了通过不同种类的瘦客户机平台或厚客户机平台 (例如移动电话、膝上型电脑、个人数字

助理 PDA) 对云的使用。

[0023] 资源池 : 提供者的计算资源被归入资源池并通过多租户 (multi-tenant) 模式服务于多重消费者, 其中按需将不同的实体资源和虚拟资源动态地分配和再分配。一般情况下, 消费者不能控制或甚至并不知晓所提供的资源的确切位置, 但可以在较高抽象程度上指定位置 (例如国家、州或数据中心), 因此具有位置无关性。

[0024] 迅速弹性 : 能够迅速、有弹性地 (有时是自动地) 部署计算能力, 以实现快速扩展, 并且能迅速释放来快速缩小。在消费者看来, 用于部署的可用计算能力往往显得是无限的, 并能在任意时候都能获取任意数量的计算能力。

[0025] 可测量的服务 : 云系统通过利用适于服务类型 (例如存储、处理、带宽和活跃用户帐号) 的某种抽象程度的计量能力, 自动地控制和优化资源效用。可以监测、控制和报告资源使用情况, 为服务提供者和消费者双方提供透明度。

[0026] 服务模型如下 :

[0027] 软件即服务 (SaaS) : 向消费者提供的能力是使用提供者在云基础架构上运行的应用。可以通过诸如网络浏览器的瘦客户机接口 (例如基于网络的电子邮件) 从各种客户机设备访问应用。除了有限的特定于用户的应用配置设置外, 消费者既不管理也不控制包括网络、服务器、操作系统、存储、乃至单个应用能力等的底层云基础架构。

[0028] 平台即服务 (PaaS) : 向消费者提供的能力是在云基础架构上部署消费者创建或获得的应用, 这些应用利用提供者支持的程序设计语言和工具创建。消费者既不管理也不控制包括网络、服务器、操作系统或存储的底层云基础架构, 但对其部署的应用具有控制权, 对应用托管环境配置可能也具有控制权。

[0029] 基础架构即服务 (IaaS) : 向消费者提供的能力是消费者能够在其中部署并运行包括操作系统和应用的任意软件的处理、存储、网络和其他基础计算资源。消费者既不管理也不控制底层的云基础架构, 但是对操作系统、存储和其部署的应用具有控制权, 对选择的网络组件 (例如主机防火墙) 可能具有有限的控制权。

[0030] 部署模型如下 :

[0031] 私有云 : 云基础架构单独为某个组织运行。云基础架构可以由该组织或第三方管理并且可以存在于该组织内部或外部。

[0032] 共同体云 : 云基础架构被若干组织共享并支持有共同利害关系 (例如任务使命、安全要求、政策和合规考虑) 的特定共同体。共同体云可以由共同体内的多个组织或第三方管理并且可以存在于该共同体内部或外部。

[0033] 公共云 : 云基础架构向公众或大型产业群提供并由出售云服务的组织拥有。

[0034] 混合云 : 云基础架构由两个或更多部署模型的云 (私有云、共同体云或公共云) 组成, 这些云依然是独特的实体, 但是通过使数据和应用能够移植的标准化技术或私有技术 (例如用于云之间的负载平衡的云突发流量分担技术) 绑定在一起。

[0035] 云计算环境是面向服务的, 特点集中在无状态性、低耦合性、模块性和语意的互操作性。云计算的核心是包含互连节点网络的基础架构。

[0036] 现在参考图 1, 其中显示了云计算节点的框图。图 1 显示的云计算节点 10 仅仅是适合的云计算节点的一个示例, 不应对这里描述的实施例的功能和使用范围带来任何限制。总之, 云计算节点 10 能够被用来实现和 / 或执行以上所述的任何功能。

[0037] 云计算节点 10 具有计算机系统 / 服务器 / 主机 12, 其可与众多其它通用或专用计算系统环境或配置一起操作。众所周知, 适于与计算机系统 / 服务器 12 一起操作的计算系统、环境和 / 或配置的例子包括但不限于 : 个人计算机系统、服务器计算机系统、瘦客户机、厚客户机、手持或膝上设备、基于微处理器的系统、机顶盒、可编程消费电子产品、网络个人电脑、小型计算机系统、大型计算机系统和包括上述任意系统的分布式云计算技术环境, 等等。

[0038] 计算机系统 / 服务器 12 可以在由计算机系统执行的计算机系统可执行指令 ( 诸如程序模块 ) 的一般语境下描述。通常, 程序模块可以包括执行特定的任务或者实现特定的抽象数据类型的例程、程序、目标程序、组件、逻辑、数据结构等。计算机系统 / 服务器 12 可以在通过通信网络链接的远程处理设备执行任务的分布式云计算环境中实施。在分布式云计算环境中, 程序模块可以位于包括存储设备的本地或远程计算系统存储介质上。

[0039] 如图 1 所示, 云计算节点 10 中的计算机系统 / 服务器 12 以通用计算设备 ( 也称为处理设备 ) 的形式表现。计算机系统 / 服务器 12 的组件可以包括但不限于 : 一个或者多个处理器或者处理单元 16, 系统存储器 28, 连接不同系统组件 ( 包括系统存储器 28 和处理单元 16) 的总线 18。

[0040] 总线 18 表示几类总线结构中的一种或多种, 包括存储器总线或者存储器控制器, 外围总线, 图形加速端口, 处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说, 这些体系结构包括但不限于工业标准体系结构 (ISA) 总线, 微通道体系结构 (MAC) 总线, 增强型 ISA 总线、视频电子标准协会 (VESA) 局域总线以及外围组件互连 (PCI) 总线。

[0041] 计算机系统 / 服务器 12 可以包括多种计算机系统可读介质。这些介质可以是能够被计算机系统 / 服务器 12 访问的任意可获得的介质, 包括易失性和非易失性介质, 可移动的和不可移动的介质。

[0042] 系统存储器 28 可以包括易失性存储器形式的计算机系统可读介质, 例如随机存取存储器 (RAM) 30 和 / 或高速缓存存储器 32。计算机系统 / 服务器 12 可以进一步包括其它可移动 / 不可移动的、易失性 / 非易失性计算机系统存储介质。仅作为举例, 存储系统 34 可以用于读写不可移动的、非易失性磁介质 ( 图 1 未显示, 通常称为 “硬盘驱动器” )。尽管图 1 中未示出, 可以提供用于对可移动非易失性磁盘 ( 例如 “软盘” ) 读写的磁盘驱动器, 以及对可移动非易失性光盘 ( 例如 CD-ROM, DVD-ROM 或者其它光介质 ) 读写的光盘驱动器。在这些情况下, 每个驱动器可以通过一个或者多个数据介质接口与总线 18 相连。存储器 28 可以包括至少一个程序产品, 该程序产品具有一组 ( 例如至少一个 ) 程序模块, 这些程序模块被配置以执行各实施例的功能。

[0043] 具有一组 ( 至少一个 ) 程序模块 42 的程序 / 实用工具 40, 可以存储在存储器 28 中, 这样的程序模块 42 包括但不限于操作系统、一个或者多个应用程序、其它程序模块以及程序数据, 这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块 42 通常执行这里所描述的实施例中的功能和 / 或方法。

[0044] 计算机系统 / 服务器 12 也可以与一个或多个外部设备 14 ( 例如键盘、指向设备、显示器 24 等 ) 通信, 还可与一个或者多个使得用户能与该计算机系统 / 服务器 12 交互的设备通信, 和 / 或与使得该计算机系统 / 服务器 12 能与一个或多个其它计算设备进行通信的任何设备 ( 例如 NIC, 调制解调器等等 ) 通信。这种通信可以通过输入 / 输出 (I/O) 接

口 22 进行。并且,计算机系统 / 服务器 12 还可以通过网络适配器 20 与一个或者多个网络(例如局域网 (LAN),广域网 (WAN) 和 / 或公共网络,例如因特网)通信。如图所示,网络适配器 20 通过总线 18 与计算机系统 / 服务器 12 的其它模块通信。应当明白,尽管图中未示出,其它硬件和 / 或软件模块可以与计算机系统 / 服务器 12 一起操作,包括但不限于:微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID 系统、磁带驱动器以及数据备份存储系统等。

[0045] 现在参考图 2,其中显示了示例性的云计算环境 50。如图所示,云计算环境 50 包括云计算消费者使用的本地计算设备可以与其相通信的一个或者多个云计算节点 10,本地计算设备例如可以是个人数字助理 (PDA) 或移动电话 54A,台式电脑 54B、笔记本电脑 54C、数字摄像机 54D、数字音频记录设备 54E 和 / 或数字照相机 54N。云计算节点 10 之间可以相互通信。可以在包括但不限于如上所述的私有云、共同体云、公共云或混合云或者它们的组合的一个或者多个网络中将云计算节点 10 进行物理或虚拟分组(图中未显示)。这样,云的消费者无需在本地计算设备上维护资源就能请求云计算环境 50 提供的基础架构即服务 (IaaS)、平台即服务 (PaaS) 和 / 或软件即服务 (SaaS)。应当理解,图 2 显示的各类计算设备 54A-N 仅仅是示意性的,云计算节点 10 以及云计算环境 50 可以与任意类型网络上和 / 或网络可寻址连接的任意类型的计算设备(例如使用网络浏览器)通信。

[0046] 现在参考图 3,其中显示了云计算环境 50(图 2)提供的一组功能抽象层。首先应当理解,图 3 所示的组件、层以及功能都仅仅是示意性的,实施例不限于此。如图 3 所示,提供下列层和对应功能:

[0047] 硬件和软件层 60 包括硬件和软件组件。硬件组件的例子包括:主机,例如 IBM® zSeries® 系统;基于 RISC(精简指令集计算机)体系结构的服务器,例如 IBM pSeries® 系统;IBM xSeries® 系统;IBM BladeCenter® 系统;存储设备;网络和网络组件,诸如 NIC、集线器、交换机、路由器、网桥以及网关。软件组件的例子包括:网络应用服务器软件,例如 IBM WebSphere® 应用服务器软件;数据库软件,例如 IBM DB2® 数据库软件;软件定义联网软件,例如用于虚拟环境的 IBM SDN(SDN VE)。(IBM, zSeries, pSeries, xSeries, BladeCenter, WebSphere, DB2 以及 SDN VE 是国际商业机器公司在全世界各地的注册商标)。

[0048] 虚拟层 62 提供一个抽象层,该层可以提供下列虚拟实体的例子:虚拟服务器、虚拟存储、虚拟网络(包括虚拟覆盖网络、虚拟专用网络以及诸如虚拟交换机和路由器的虚拟网络部件)、虚拟应用和操作系统,以及虚拟客户端。

[0049] 在一个示例中,管理层 64 可以提供下述功能:资源供应功能:提供用于在云计算环境中执行任务的计算资源和其它资源的动态获取;计量和定价功能:在云计算环境内对资源的使用进行成本跟踪,并为此提供帐单和发票。在一个例子中,该资源可以包括应用软件许可。安全功能:为云的消费者和任务提供身份认证,为数据和其它资源提供保护。用户门户功能:为消费者和系统管理员提供对云计算环境的访问。服务水平管理功能:提供云计算资源的分配和管理,以满足必需的服务水平。服务水平协议 (SLA) 计划和履行功能:为根据 SLA 预测的对云计算资源未来需求提供预先安排和供应。

[0050] 工作负载层 66 提供云计算环境可能实现的功能的示例。在该层中,可提供的工作

负载或功能的示例包括：地图绘制与导航；软件开发及生命周期管理；虚拟教室的教学提供；数据分析处理；交易处理；以及移动桌面。

[0051] 现在转向图4，将更详细地说明系统400的示例。在图4所描绘的示例中，系统400是包括配置为通过网络401通信的若干主机402、404和406的数据中心环境。在一个实施例中，系统400中的主机托管若干虚拟机(VM)以及虚拟转发部件(例如虚拟交换机)。取决于其处理和存储资源，主机可以运行几百个VM以及一个或多个为VM转发数据的虚拟转发部件。系统400中的主机还包括附接到主机或集成在主机中的一个或多个NIC。

[0052] 在主机中运行的VM可能属于数据中心环境中的相同租户或不同租户。在系统100的主机中运行的虚拟转发部件可以隔离不同租户的VM，使得从一个租户的VM进出的数据流不能被另一租户的VM访问。在一个实施例中，虚拟转发部件通过用不同的网络标识符封装不同租户的数据流来实现数据流的隔离。在一个实施例中，网络标识符识别不同租户的VM通过其相连接的虚拟网络。虚拟转发部件不将用特定网络标识符封装的数据流转发到与该特定网络标识符不关联的VM。在某些情况下，一个租户可以具有超过一个虚拟网络，并且该租户的通过不同虚拟网络连接的VM彼此隔离。在本专利申请中，假设一个租户使用一个虚拟网络来连接该租户的VM。换言之，在本申请中，一个“租户”意味着一个“虚拟网络”，除非另有说明。

[0053] 网络标识符的示例包括虚拟可扩展LAN(VXLAN)标识符、SDN VE租户标识符、以及使用通用路由封装的网络虚拟化(NVGRE)租户标识符。在一个实施例中，网络标识符是24比特空间，从而可以唯一地识别超过1600万个虚拟覆盖网络。

[0054] 如图所示，在此示例中，系统400的主机402包括若干VM408、虚拟转发部件410以及NIC412。VM408由主机402的系统管理程序(hypervisor，未示出)提供。在一个实施例中，VM408是运行客户端和/或服务器应用的终端系统。VM408可以通过虚拟转发部件410彼此通信或与在系统400的其它主机中运行的其它终端系统进行通信。

[0055] 虚拟转发部件410包括若干端口414和416。端口414和416是虚拟转发部件410的虚拟端口或虚拟接口(VIF)。虚拟转发部件410通过这些虚拟端口接收或发送分组。在此示例中，端口414被配置为从VM408接收分组以及向VM408发送分组。端口416被配置为从NIC412接收分组以及向NIC412发送分组。特别地，通过端口416进出虚拟转发部件416的分组包括向运行在系统400的其它主机中的终端系统发送或从其接收的分组。

[0056] 如图所示，在此示例中，NIC412包括端口420、缓冲器422以及描绘为表418的映射。端口420是NIC412的物理端口。缓冲器422是I/O(输入/输出)缓冲器。换言之，缓冲器422是用于在分组通过端口420进入或离开主机402之前存储分组的物理存储空间。在一个实施例中，NIC被配置为将物理缓冲器422分区为多个逻辑或虚拟队列424(例如，8个逻辑队列)，如图4的底部所示。在将分组发送到端口420之外或发送到虚拟转发部件之前，NIC412将来自或进入虚拟转发部件410的数据流映射到逻辑队列424，并将数据流的分组存储在所关联的逻辑队列中。

[0057] NIC412使用逻辑队列424来针对数据流分别控制数据率。例如，NIC412可以为每个数据流实现不同的服务质量(QoS)策略。在一个实施例中，NIC412对逻辑队列424分配不同的优先等级，并因此将这些优先等级分配到不同的数据流。通过这些具有不同优先等级的逻辑队列，一个实施例的NIC412实现增强的传输选择(ETS)以及基于优先级的流

控制 (PFC)。换言之，NIC412 可以通过相对于具有较低优先等级的数据流照顾具有较高优先等级的数据流来发送出数据流。NIC 412 可以阻止或抑制与特定优先等级关联的特定数据流来促进特定数据流的分组的无损传输。

[0058] 通过将物理缓冲器分区为多个逻辑队列，NIC 在物理端口 420 中创建了相同数目的逻辑端口 426。在一个实施例中，逻辑端口 426 不具有它们自己的网络地址（例如因特网协议 (IP) 和介质访问控制 (MAC) 地址）而共享物理端口 420 的网络地址。在另一实施例中，每个逻辑端口 426 与主机 402 的 VM 408 的一个虚拟 NIC(未示出) 关联，并具有其自己的网络地址。

[0059] 在一个实施例中，NIC 412 利用通道报头封装数据流的分组（其已经由虚拟转发部件 410 用网络标识符进行了封装），通道报头定义主机 402 与数据流的目的地 VM 在其中运行的另一主机之间的通道。通道报头包括物理端口 420 的网络地址作为源地址，因为它们是主机 402 用于连接到另一主机的地址。

[0060] 一旦存储在逻辑队列中的数据流通过物理端口 420 离开主机 402，网络 401 中的网络组件基于通道报头中的地址将数据流转发到运行在另一主机中的目的地 VM。换言之，不在意通道化分组的内部报头和网络识别符的网络组件将分组转发到目的地 VM 的主机。这些网络组件因此不会辨识通道中的不同数据流。

[0061] 根据本发明的一个实施例，网络 401 中的网络组件被配置为通过检查数据流的分组的内部报头和网络标识符辨识通道中的不同数据流。例如，在一个实施例中，物理转发部件 428（例如，交换机）被制造为或配置为辨识通道中的不同数据流。当物理转发部件 428 从 NIC 412 接收通道化数据 430 时，物理转发部件 428 检查分组的通道报头 432、网络标识符 434 以及内部报头 436。内部报头 436 和有效载荷 438 构成 VM 408 之一发送到虚拟转发部件的分组的原始形式。识别虚拟覆盖网络的网络识别符 434 由虚拟转发部件 410 添加。通道报头由 NIC 412 添加。

[0062] 内部报头 436 包括源和目的地 VM 的网络地址。即，在一个实施例中，内部报头 436 包括内部以太网报头和内部 IP 报头。通道报头或外部报头 436 包括主机 402 和另一主机的网络地址，主机 402 和该另一主机是该通道的端点。主机 402 的网络地址是物理端口 420 的网络地址。即，在一个实施例中，通道报头 432 包括外部以太网报头、外部 IP 报头以及外部传输报头（例如，用户数据报协议 (UDP) 报头和传输控制协议 (TCP) 报头等）。

[0063] 在一个实施例中，物理转发部件 428 创建数据流与通道之间的映射。在该映射中，不同的通道由不同的通道报头定义（例如，具有在一个端点的主机 402 以及在另一端点的不同主机 404、406 的通道）。不同的数据流由不同的网络识别符和不同的内部报头定义。物理转发部件 428 创建的映射描绘为表 440。

[0064] 在一个实施例中，物理转发部件 428 使用数据流与通道之间的映射来帮助 NIC 412 实现 ETS 和 PFC。例如，物理转发部件 428 可以从网络 401 中的下游另一网络组件接收通道的拥塞消息。通过该映射，物理转发部件 428 可以识别与通道关联的数据流。换言之，因为物理转发部件 428 所接收的拥塞消息基于定义通道的网络地址，所以物理转发部件 428 使用该映射来识别针对该通道的数据流。物理转发部件 428 向 NIC 412 通知贡献拥塞（也称为网络拥塞）的数据流。NIC 412 接着可以使用与数据流关联的逻辑队列抑制数据流。在一个实施例中，NIC 412 向虚拟转发部分 410 通知拥塞，其继而可以通过抑制来源

于 VM 的数据流而控制数据流。

[0065] 在一个实施例中，物理转发部件 428 接收的拥塞消息具有不能够指明数据流的第一格式。在一个实施例中，物理转发部件 428 被配置为产生第二格式的新拥塞消息，该第二格式能够指明数据流。在一个实施例中，物理转发部件 428 通过向 NIC 412 发送第二格式的新消息而向 NIC 通知贡献拥塞的数据流。

[0066] 图 5 描绘用于在附接到主机或与主机集成的 NIC 处管理数据流的处理流。数据流来源于主机的一个或多个 VM。在一个实施例中，图 5 中示出的处理流由图 4 的 NIC 412 执行。在块 502，从与 NIC 的端口关联的物理缓冲器创建多个逻辑队列。物理缓冲器用于在通过 NIC 的物理端口将分组发送到主机之外之前存储分组。NIC 被配置为将物理缓冲器分区为多个（例如，八个）逻辑队列。在块 504，向块 502 中创建的逻辑队列分配不同优先等级。在一个实施例中，NIC 基于分配给逻辑队列的优先等级控制存储在逻辑队列中的分组的数据率。

[0067] 在块 506，从运行在主机中的虚拟转发部件接收分组。该分组来源于运行在主机中的 VM 并由虚拟转发部件转发到 NIC，因为分组的目的地是另一主机。该分组还被虚拟转发部件用网络标识符封装。网络标识符识别该 VM 连接到同一租户的其它终端系统所通过的虚拟覆盖网络。

[0068] 在块 508，分组所属的数据流被映射到通道。连接该主机与其它主机的通道可能在为数据中心环境设立主机时已经被先前创建。在一个实施例中，NIC 基于分组的目的地的位置（例如，托管分组的目的地的主机的网络地址）将数据流映射到通道。在一个实施例中，仅当分组是数据流的初始分组时才执行块 508。对于数据流的后续分组，数据流与通道之间的映射被用于识别通道。

[0069] 在块 510，利用通道报头封装分组。该通道报头定义块 508 中数据流被映射到的通道。在一个实施例中，通道报头包括主机的网络地址作为源和目的地地址。

[0070] 在块 512，分组所属的数据流被映射到逻辑队列。在一个实施例中，NIC412 基于数据流所分配的优先等级将数据流分配到逻辑队列。在一个实施例中，虚拟转发部件决定数据流应该得到的优先等级。优先等级分配也可以基于网络标识符，以便区别不同虚拟覆盖网络的流量。在一个实施例中，仅当分组是数据流的初始分组时才执行块 512。对于数据流的后续分组，数据流与逻辑队列之间的映射被用于识别逻辑队列。在块 514，分组被放置在逻辑队列中。

[0071] 通过在块 508 中将数据流映射到通道以及在块 512 中将数据流映射到逻辑队列，NIC 创建数据流、逻辑队列与通道之间的映射。使用此映射，NIC 可以通过将流控制方案（例如，数据率调节、QoS 等）应用到映射到通道的逻辑队列而将这些方案应用到通道。

[0072] 图 6 描绘用于在物理转发部件（例如，交换机、网关等）处管理数据流的处理流。物理转发部件连接到托管数据流所来源的若干 VM 的主机的 NIC。在一个实施例中，物理转发部件是在连接数据中心环境中的主机的网络的边缘处的网关。在一个实施例中，图 6 中所示的处理流由图 4 的物理转发部件 428 执行。在块 602，从 NIC 接收分组。利用网络标识符和通道报头封装该分组。通道报头的源和目的地地址包括该主机和另一主机或覆盖网络网关交换机的网络地址。

[0073] 在块 604，分组所属的数据流以及正在传输数据流的通道被识别。在一个实施例

中,物理转发部件被配置为检查分组的网络标识符以及内部报头以识别分组的数据流。具体的,物理转发部件被配置为检查内部报头中的源和目的地网络地址。物理转发部件可以基于这些网络地址和网络标识符唯一地识别数据流。物理转发部件通过检查分组的通道报头来识别通道。

[0074] 在块 606,将数据流映射到通道。在一个实施例中,物理转发部件将数据流(例如,内部分组的源和目的地地址以及网络标识符)映射到通道(例如,内部分组的源和目的地运行的主机的网络地址)。

[0075] 在块 608,从物理转发部件转发出分组。接着,网络的网络组件基于存储在通道报头中的信息转发分组。即,不在意封装分组的内部报头和网络标识符的网络组件将分组转发到通道的另一端点。

[0076] 图 7 描绘在物理转发部件处使用数据流与通道之间的映射处理拥塞消息的处理流。在一个实施例中,物理转发部件是执行图 6 中所示的处理流的同一物理转发部件。在一个实施例中,图 7 中所示的处理流由图 4 的物理转发部件 428 执行。在块 702,接收拥塞消息。该拥塞消息来源于网络中在物理转发部件下游的网络组件(例如,网关路由器)。拥塞消息指示通道化流量(例如,通道化 IP 流量)贡献拥塞并因此应当调节或阻止用于该流量的数据率。

[0077] 在块 704,为通道识别一个或多个数据流。物理转发部件识别接收拥塞消息所针对的通道。物理转发部件使用在图 6 的块 606 中创建的映射来识别与该通道关联的一个或多个数据流。在块 706,向物理转发部件发送数据流的 NIC 被通知块 704 中所识别的数据流。NIC 可以调节针对贡献拥塞的数据流的数据率。

[0078] 图 8 描绘在 NIC 处处理拥塞消息的处理流。在一个实施例中,NIC 是执行图 5 中所示的处理流的同一 NIC。在实施例中,图 8 中所示的处理流由图 4 的 NIC 412 执行。在块 802,从物理转发部件接收拥塞消息。物理转发部件所发送的拥塞消息指明哪个数据流正贡献拥塞。

[0079] 在块 804,基于拥塞消息识别贡献拥塞的一个或多个数据流。在一个实施例中,拥塞消息利用分组的源和目的地地址以及网络标识符指明数据流。NIC 使用在图 5 的块 512 中创建的映射来识别数据流的逻辑队列。

[0080] 在块 806,调节在块 806 中识别的数据流的数据率。NIC 使用与所识别的数据流关联的逻辑队列来调节数据率。在 808,向 NIC 发送数据流的虚拟转发部件被通知拥塞。在一个实施例中,NIC 将从物理转发部件接收的拥塞消息中继到虚拟转发部件。在一个实施例中,NIC 基于在块 802 中从物理转发部件接收的拥塞消息生产新拥塞消息。虚拟转发部件接收该新消息或所中继的消息,并可以使用其自己的流控制方案调节用于贡献数据流的数据率。

[0081] 在一个实施例中,即使 NIC 没有从物理转发部件接收到任何拥塞消息,NIC 也产生拥塞消息(例如,PFC 消息或电气和电子工程师协会(IEEE)802.3x 或等同消息)。在此实施例中,NIC 监视逻辑队列以判定是否有任何逻辑队列将溢出。当判定有任何逻辑队列将溢出时,NIC 产生针对与该逻辑队列关联的数据流的拥塞消息,并将该拥塞消息发送到虚拟转发部件。

[0082] 技术效果和益处包括能够在 NIC 处以每个流为基础控制数据流的数据率以及能

够在从 NIC 接收数据流的物理转发部件处辨识不同数据流。

[0083] 本领域的普通技术人员将理解,实施例的方面可以被实施为系统、方法或计算机程序产品。因此,实施例的方面可以采用完全硬件实施例的形式、完全软件实施例的形式(包括固件、常驻软件、微代码等)、或者软件和硬件方面组合的实施例形式,这里,它们全部可被一般地称为例如“电路”、“模块”或“系统”。此外,实施例的方面可以采用实现在一个或多个计算机可读存储设备中的计算机程序产品的形式,该计算机可读存储设备上实现由计算机可读程序代码。

[0084] 实施例的一个或多个能力可以被实现为软件、固件、硬件或它们的某些组合。此外,一个或多个所述能力可被仿真。

[0085] 一个实施例可以是用于使处理器电路能够执行本发明的要素的计算机程序产品,该计算机程序产品包括处理电路可读取的计算机可读存储介质,其存储由处理电路执行来执行方法的指令。

[0086] 计算机可读存储介质是有形的非暂时性存储介质,其上记录有指令用于使得处理器电路执行方法。“计算机可读存储介质”是非暂时性的至少因为一旦指令记录在介质上所记录的指令可以以后被处理器电路在独立于记录时间的时间读取一次或多次。非暂时性的“计算机可读存储介质”包括仅在上电时保持记录的信息的设备(易失性设备)以及无论是否上电都保持记录的信息的设备(非易失性设备)。例如,“非暂时性存储介质”的示例非穷举性列表包括但不限于:半导体存储设备,例如包括其上记录指令的诸如 RAM 的存储阵列或诸如锁存的存储电路;机械编码设备,诸如其上记录指令的打孔卡或凹槽中的突起结构;光学可读设备,诸如其上记录指令的 CD 或 DVD;以及磁编码设备,诸如其上记录指令的磁带或磁盘。

[0087] 计算机可读存储介质的示例的非穷举列表包括以下:便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM 或闪存)、便携式压缩盘只读存储器(CD-ROM)。程序代码可以通过网络、例如因特网、局域网、广域网和/或无线网从外部计算机或外部存储设备分布到相应计算/处理设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口卡从网络接收程序,并转发该程序,以供存储在各个计算/处理设备中的计算机可读存储设备中。

[0088] 用于执行实施例的方面的操作的计算机程序指令可以是例如汇编代码、机器代码、微代码、或者以一种或多种编程语言的任意组合编写的源代码或目标代码,所述编程语言包括面向对象的编程语言—诸如 Java、Smalltalk、C++ 等,以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0089] 这里参照根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述了实施例的各个方面。应当理解,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机程序指令实现。

[0090] 这些计算机程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器,从而生产出一种机器,使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时,产生了实现流程图和 / 或框图中的一个或多个方框中规定的功能 / 动作的装置。也可以把这些计算机程序指令存储在计算机可读存储介质中,这些指令使得计算机、其它可编程数据处理装置或其他设备以特定方式工作。

[0091] 也可以把计算机程序指令加载到计算机、其它可编程数据处理装置、或其它设备上,使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤,以产生计算机实现的过程,从而使得在计算机或其它可编程装置上执行的指令提供用于实现流程图和 / 或框图中的一个或多个方框中规定的功能 / 动作的处理。

[0092] 附图中的流程图和框图显示了根据多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分,所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个连续的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框图和 / 或流程图中的每个方框、以及框图和 / 或流程图中的方框的组合,可以用执行规定的功能或动作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

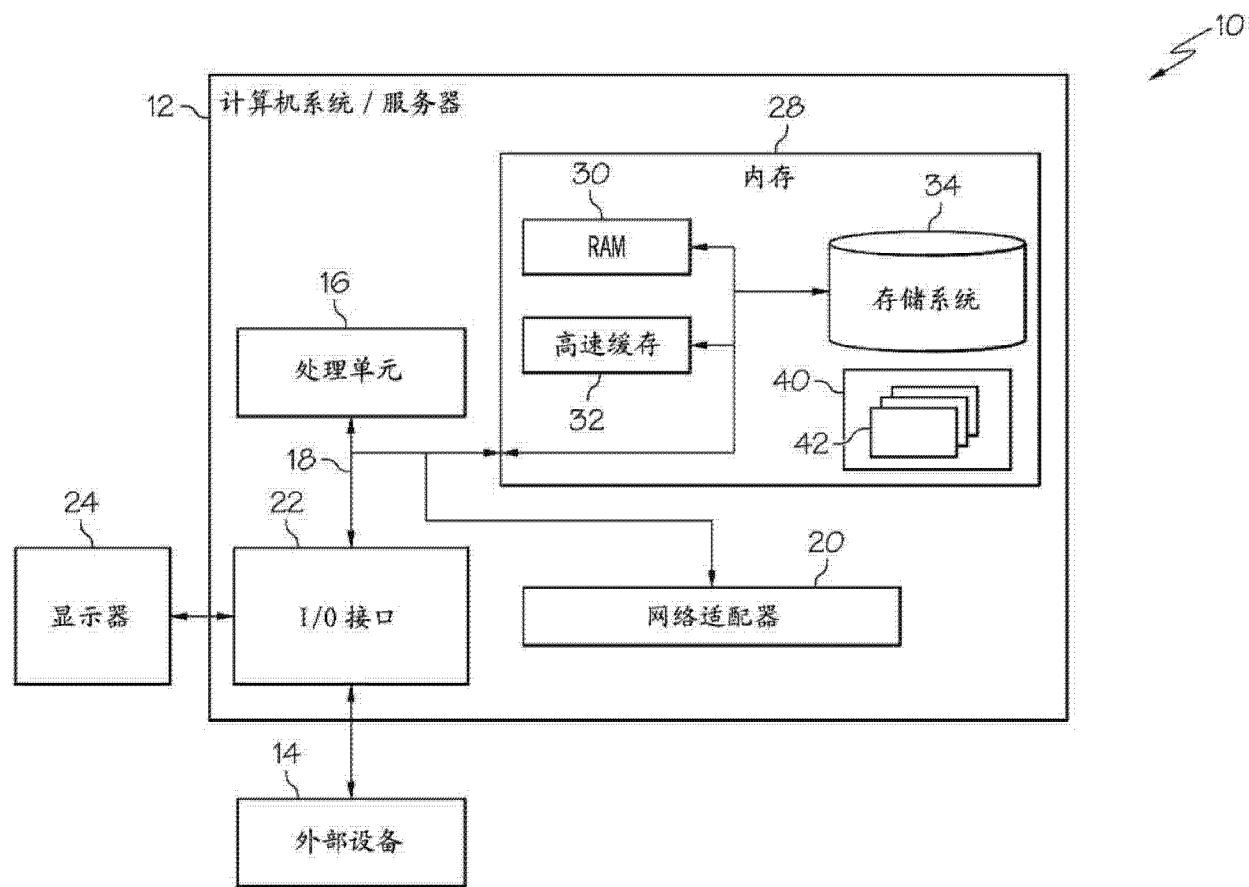


图 1

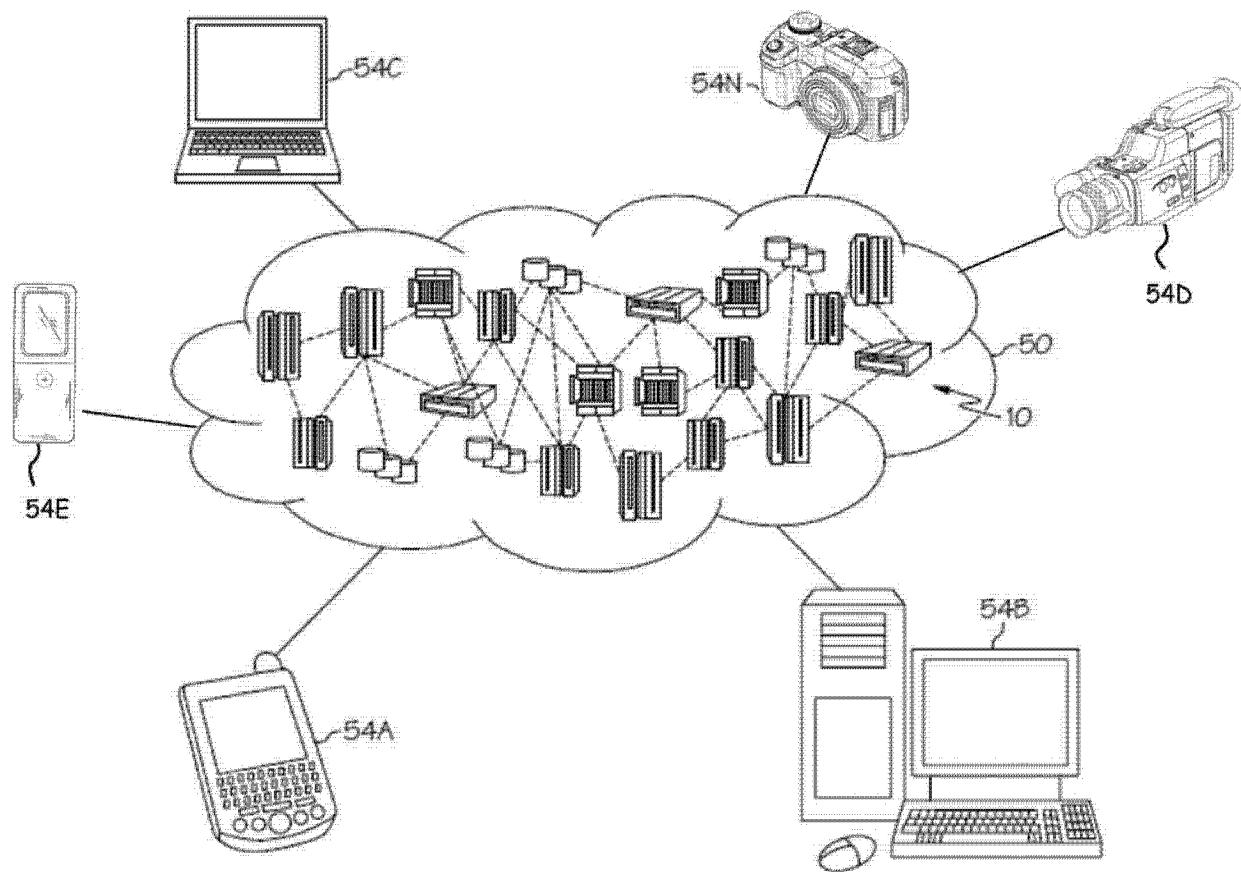


图 2

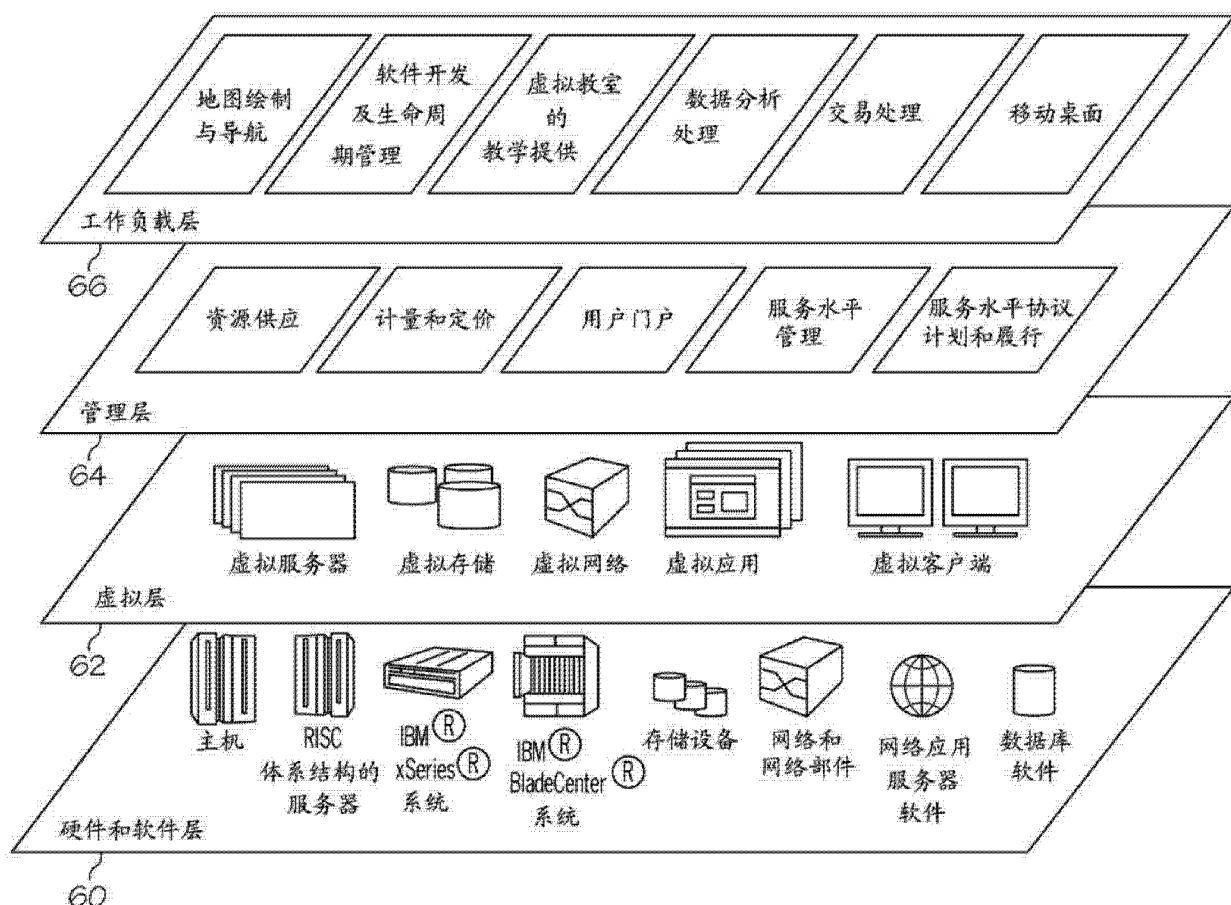


图 3

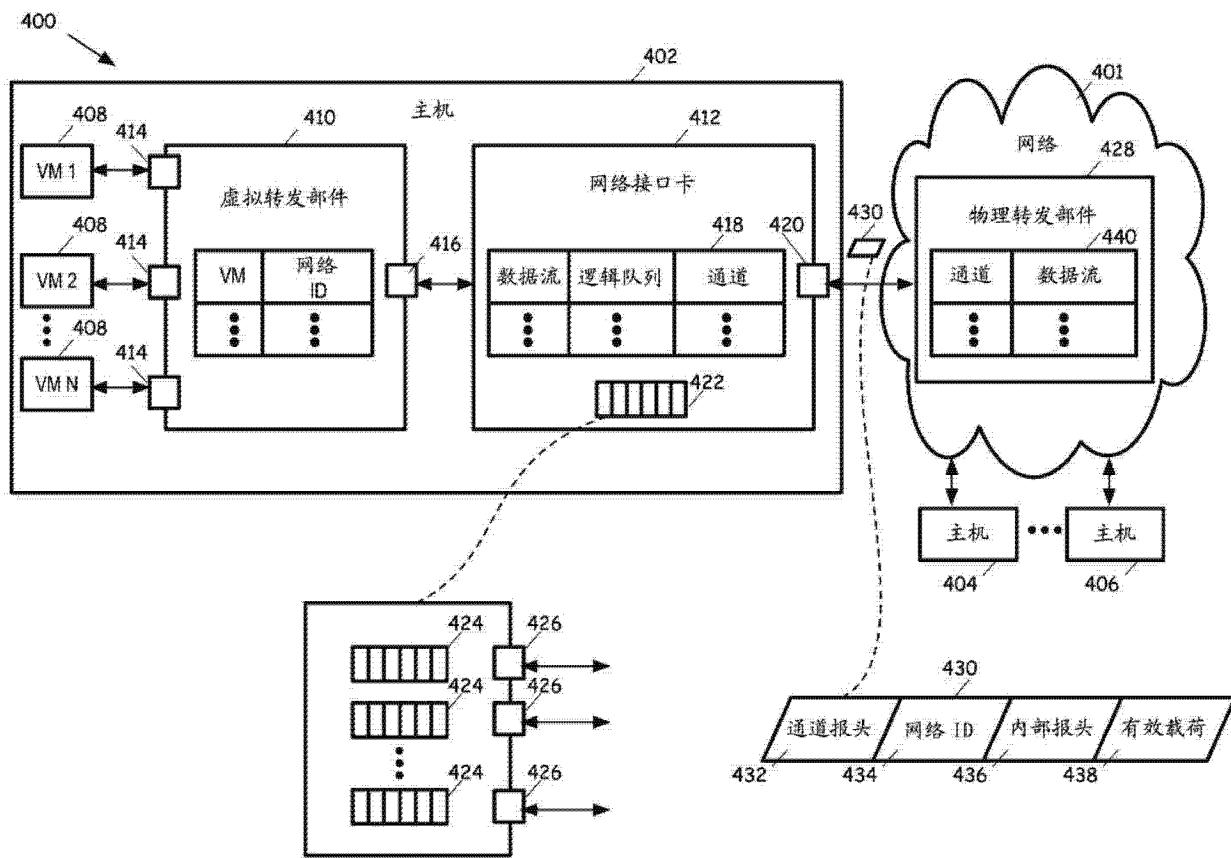


图 4

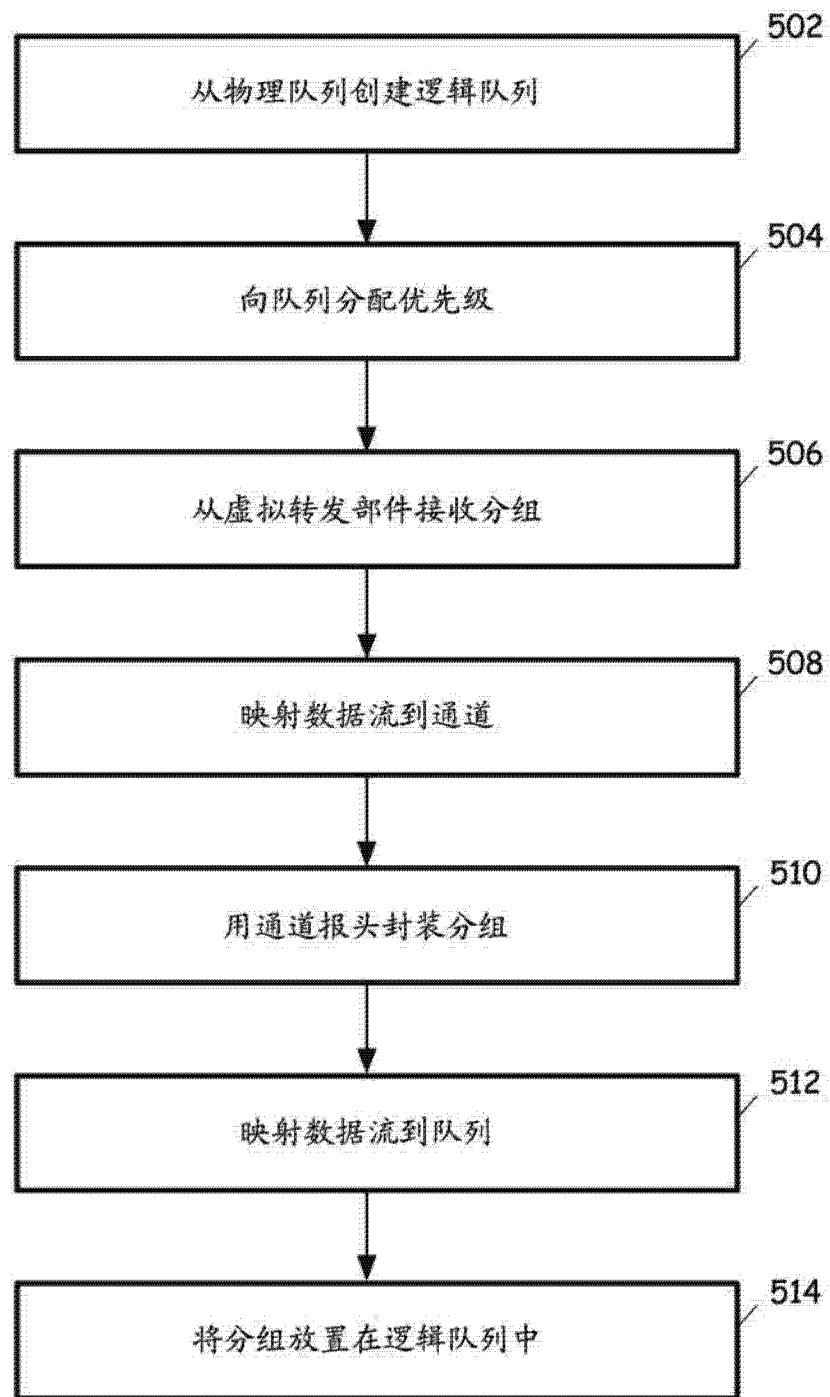


图 5

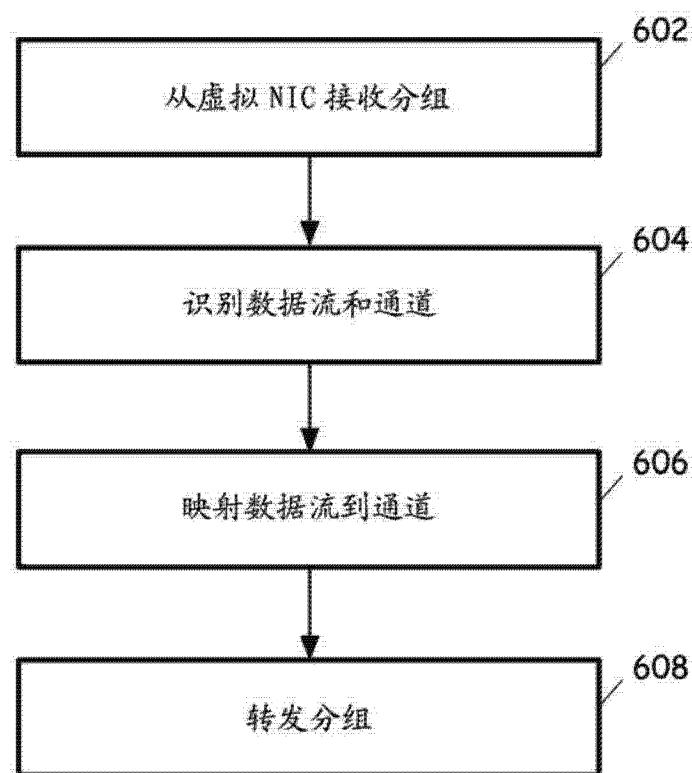


图 6

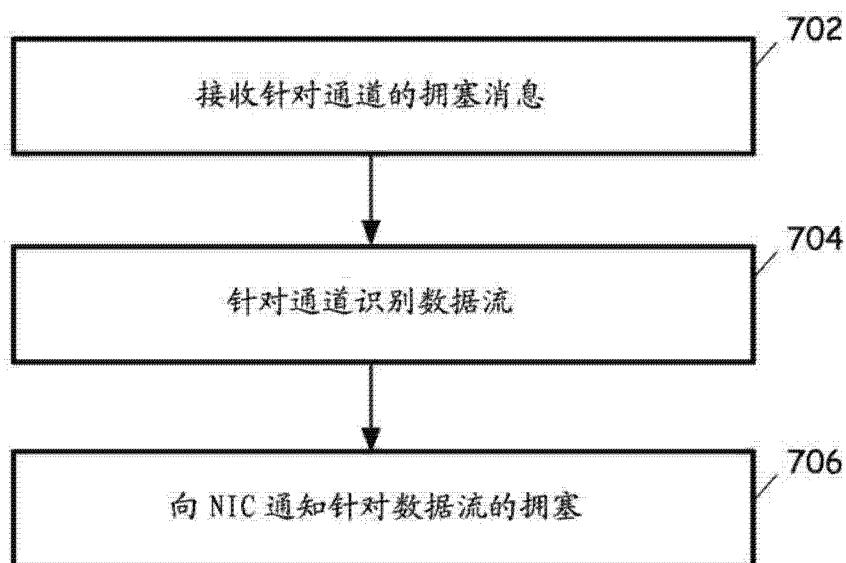


图 7

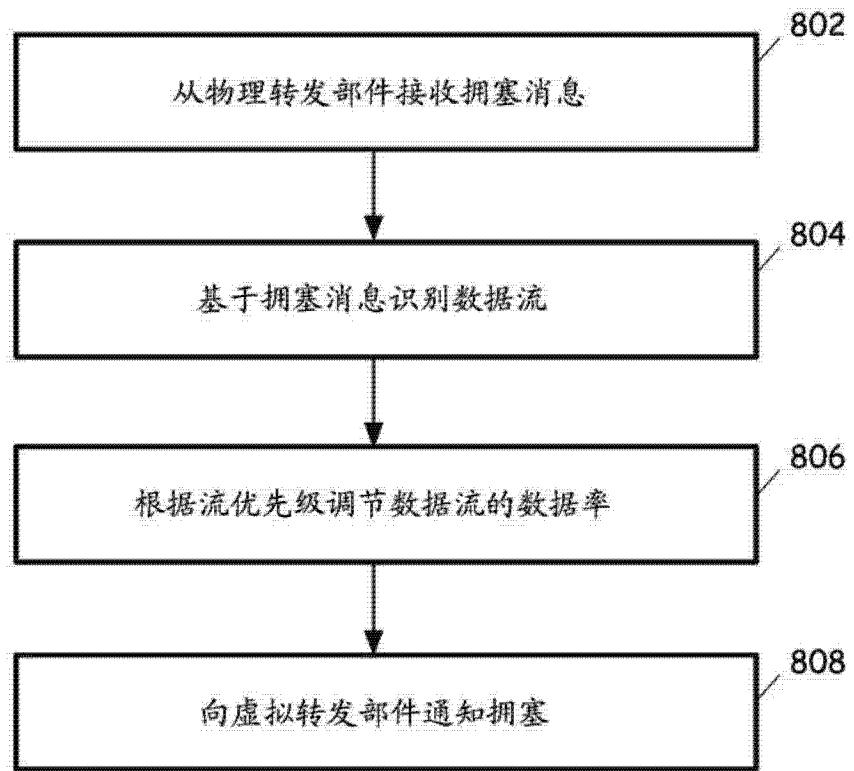


图 8