(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 102648455 B
(45)授权公告日 2015.11.25

(21)申请号 201080055153.6
(22)申请日 2010.11.30
(30)优先权数据
2009-276679 2009.12.04 JP
(85)PCT国际申请进入国家阶段日
2012.06.04
(86)PCT国际申请的申请数据
PCT/JP2010/071316 2010.11.30
(87)PCT国际申请的公布数据
WO2011/068091 JA 2011.06.09
(73)专利权人 日本电气株式会社
地址 日本东京都
(72)发明人 狩野秀一
(74)专利代理机构 北京东方亿思知识产权代理有限责任公司 11258
代理人 宋鹤
(51)Int.Cl.
H04L 12/721(2013.01)
H04L 12/713(2013.01)

(56)对比文件
CN 101305561 A,2008.11.12, 说明书第10页第2段, 第11页第3段,第16页第2-5段,第17页第1-5段,第18页第6-8段,第19页第1-2段,说明书附图1-5.
CN 1943179 A,2007.04.04, 全文.
EP 1359724 A1,2003.11.05, 全文.
US 20060208718 A,2006.09.21, 全文.

审查员 谢永坚

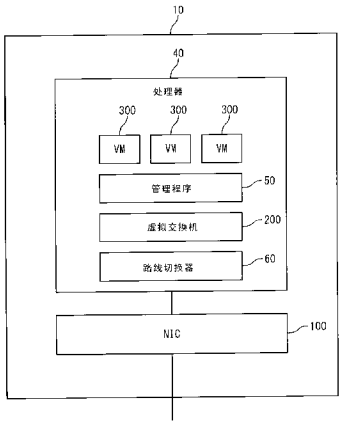
权利要求书4页 说明书14页 附图24页

(54)发明名称

服务器和流控制程序

(57)摘要

公开了一服务器,其具有处理器、连接到处理器的网络适配器以及路线切换器。处理器具有虚拟机和中继在虚拟机与外界之间交换的分组的虚拟交换机。网络适配器具有在不通过虚拟交换机的情况下向虚拟机发送分组和从虚拟机接收分组的传送功能。路线切换器在第一路线模式流和第二路线模式流之间动态地切换由虚拟机发送和接收的分组的流。并且,路线切换器指令传送功能处理第一路线模式流并且指令虚拟交换机处理第二路线模式流。



CN 102648455 B

1. 一种服务器，包括：

处理器；

配置为硬件并连接到所述处理器和设置在所述服务器外侧的外部数据链路的物理网络适配器；以及

路线切换器，

其中所述处理器包括：

多个虚拟机；以及

虚拟交换机，该虚拟交换机中继在所述虚拟机与外界之间交换的分组；

其中，所述物理网络适配器具有在不通过所述虚拟交换机的情况下向所述虚拟机发送分组和从所述虚拟机接收分组的传送功能；

其中，所述路线切换器在第一路线模式流和第二路线模式流之间动态地切换由所述虚拟机中的一者发送和接收的分组的特定流，并且

其中，所述第一路线模式流包括其中分组在不通过所述虚拟交换机的情况下使用所述传送功能在所述物理网络适配器和所述虚拟机中的所述一者之间直接发送和接收的流，

其中，所述第二路线模式流包括其中分组在不通过所述物理网络适配器和所述虚拟机之间的直接连接的情况下通过所述虚拟交换机在所述物理网络适配器和所述虚拟机中的所述一者之间发送和接收的流，其中，所述物理网络适配器包括：

接收过滤器，该接收过滤器接收接收分组；

存储单元，该存储单元存储指示流与接收动作之间的关系的接收过滤器表格，

第一接收队列，每个所述第一接收队列存储要直接发送到所述虚拟机中的相关一者的分组；以及

第二接收队列，所述第二接收队列存储要发送到所述虚拟交换机的分组，

其中，当将所述特定流切换到所述第一路线模式流时，所述路线切换器设定所述接收过滤器表格，以使得由所述虚拟机中的所述一者发送和接收的所述分组的所述特定流与第一接收动作相关，

其中，当将所述特定流切换到所述第二路线模式流时，所述路线切换器设定所述接收过滤器表格，以使得由所述虚拟机中的所述一者发送和接收的所述分组的所述特定流与第二接收动作相关，

其中，所述第一接收动作包括将所述接收分组存储到与所述第一接收队列的所述虚拟机中的所述一者相关的接收队列，并利用所述传送功能将所述接收分组从所述相关的接收队列发送到所述虚拟机中的所述一者，

其中，所述第二接收动作包括将所述接收分组存储到所述第二接收队列，并将所述接收分组从所述第二接收队列发送到所述虚拟交换机，并且

其中，所述接收过滤器参考所述接收过滤器表格以对所述接收分组执行与所述特定流相关的接收动作。

2. 根据权利要求 1 所述的服务器，其中，所述物理网络适配器还包括：

发送队列，每个所述发送队列存储使用所述传送功能从所述虚拟机中的相关一者直接接收的分组；以及

发送过滤器，其从所述虚拟机接收发送分组，

其中，所述存储单元还存储发送过滤器表格，该发送过滤器表格指示流与发送动作之间的关系，

其中，当将所述特定流切换到所述第一路线模式时，所述路线切换器设定所述发送过滤器表格，以使得由所述虚拟机中的所述一者发送和接收的所述分组的所述特定流与第一发送动作相关，

其中，当将所述特定流切换到所述第二路线模式时，所述路线切换器设定所述发送过滤器表格，以使得由所述虚拟机中的所述一者发送和接收的所述分组的所述特定流与第二发送动作相关，

其中，所述第一发送动作包括将从所述虚拟机的所述一者接收并存储在与所述发送队列的所述虚拟机中的所述一者相关的发送队列中的发送分组发送到外界，

其中，所述第二发送动作包括将从所述虚拟机的所述一者接收并存储在与所述虚拟机中的所述一者相关的所述发送队列中的发送分组作为所述接收分组环回到所述接收过滤器，并且

其中，所述发送过滤器参考所述发送过滤器表格以对所述发送分组执行与所述特定流相关的发送动作。

3. 根据权利要求 1 所述的服务器，其中，每个所述虚拟机包括 ：

第一发送 / 接收功能，该第一发送 / 接收功能在不通过所述虚拟交换机的情况下向所述物理网络适配器发送分组并从所述物理网络适配器接收分组 ；

第二发送 / 接收功能，该第二发送 / 接收功能向所述虚拟交换机发送分组并从所述虚拟交换机接收分组 ；以及

分支功能，该分支功能参考指示出流与分组传送目的地之间的关系的流表格并且将来自每个所述虚拟机的发送分组转发到从所述分组传送目的地中选择的与所述发送分组的流相关的分组传送目的地，

其中，所述路线切换器设定所述流表格，以使得所述第一路线模式流与第一分组传送目的地相关并且所述第二路线模式流与第二分组传送目的地相关，

其中，所述第一分组传送目的地是所述第一发送 / 接收功能，并且

其中，所述第二分组传送目的地是所述第二发送 / 接收功能。

4. 根据权利要求 1 至 3 中任何一项所述的服务器，其中，所述路线切换器基于由所述虚拟机中的所述一者发送和接收的所述分组来测量所述特定流的负荷，

其中，当所述特定流的所述负荷超过预定阈值时，所述路线切换器将所述特定流切换成所述第一路线模式流。

5. 根据权利要求 1 至 3 中任何一项所述的服务器，其中，在接收到属于所述特定流的预定分组时，所述路线切换器将所述特定流切换成所述第一路线模式流。

6. 根据权利要求 1 至 3 中任何一项所述的服务器，其中，所述路线切换器被结合在所述虚拟交换机中。

7. 根据权利要求 1 至 3 中任何一项所述的服务器，其中，所述路线切换器被结合在所述物理网络适配器中。

8. 一种要连接到服务器的处理器和设置在所述服务器外侧的外部数据链路的物理网络适配器，所述处理器包括多个虚拟机和中继在所述虚拟机与外界之间交换的分组的虚拟

交换机；

其中，所述物理网络适配器具有向所述虚拟机发送分组和从所述虚拟机接收分组的传送功能，

其中，所述物理网络适配器包括：

路线切换器，

接收过滤器，该接收过滤器接收接收分组，

存储单元，该存储单元存储指示流与接收动作之间的关系的接收过滤器表格，

第一接收队列，每个所述第一接收队列存储要直接发送到所述虚拟机中的相关一者的分组；以及

第二接收队列，所述第二接收队列存储要发送到所述虚拟交换机的分组，

其中，所述路线切换器在第一路线模式流和第二路线模式流之间动态地切换由所述虚拟机中的一者发送和接收的分组的特定流，并且

其中，当将所述特定流切换到所述第一路线模式流时，所述路线切换器设定所述接收过滤器表格，以使得由所述虚拟机中的所述一者发送和接收的所述分组的所述特定流与第一接收动作相关，

其中，当将所述特定流切换到所述第二路线模式流时，所述路线切换器设定所述接收过滤器表格，以使得由所述虚拟机中的所述一者发送和接收的所述分组的所述特定流与第二接收动作相关，

其中，所述第一接收动作包括将所述接收分组存储到与所述第一接收队列的所述虚拟机中的所述一者相关的接收队列，并利用所述传送功能将所述接收分组从所述相关的接收队列发送到所述虚拟机中的所述一者，

其中，所述第二接收动作包括将所述接收分组存储到所述第二接收队列，并将所述接收分组从所述第二接收队列发送到所述虚拟交换机，并且

其中，所述接收过滤器参考所述接收过滤器表格以对所述接收分组执行与所述特定流相关的接收动作。

9. 一种用于包括处理器和物理网络适配器的服务器的流控制方法，所述物理网络适配器连接到所述处理器和设置在所述服务器外侧的外部数据链路，所述处理器包括多个虚拟机和中继在所述虚拟机与外界之间交换的分组的虚拟交换机，并且所述物理网络适配器配置成硬件并具有在不通过所述虚拟交换机的情况下向所述虚拟机发送分组和从所述虚拟机接收分组的传送功能，其中，所述物理网络适配器包括：存储单元，该存储单元存储指示流与接收动作之间的关系的接收过滤器表格；第一接收队列，每个所述第一接收队列存储要直接发送到所述虚拟机中相关一者的分组；以及第二接收队列，所述第二接收队列存储要发送到所述虚拟交换机的分组，

所述方法包括：

在第一路线模式流和第二路线模式流之间动态地切换由所述虚拟机中的一者发送和接收的分组的特定流；

其中，所述第一路线模式流包括其中分组在不通过所述虚拟交换机的情况下使用所述传送功能在所述物理网络适配器和所述虚拟机中所述一者之间直接发送和接收的流，

其中，所述第二路线模式流包括其中分组在不通过所述物理网络适配器和所述虚拟机

之间的直接连接的情况下通过所述虚拟机在所述物理网络适配器和所述虚拟机中的所述一者之间发送和接收的流，

其中，当将所述特定流切换到所述第一路线模式流时，设定所述接收过滤器表格，以使得由所述虚拟机中的所述一者发送和接收的所述分组的所述特定流与第一接收动作相关，

其中，当将所述特定流切换到所述第二路线模式流时，设定所述接收过滤器表格，以使得由所述虚拟机中的所述一者发送和接收的所述分组的所述特定流与第二接收动作相关，

其中，所述第一接收动作包括将所述接收分组存储到与所述第一接收队列的所述虚拟机中的所述一者相关的接收队列，并利用所述传送功能将所述接收分组从所述相关的接收队列发送到所述虚拟机中的所述一者，

其中，所述第二接收动作包括将所述接收分组存储到所述第二接收队列，并将所述接收分组从所述第二接收队列发送到所述虚拟交换机，并且

其中，在所述接收过滤器表格中对所述接收分组执行与所述特定流相关的接收动作。

服务器和流控制程序

技术领域

[0001] 本发明涉及基于虚拟化技术的服务器和这样的服务器执行的流控制程序。

背景技术

[0002] 虚拟化技术在服务器的领域中是重要的。具体而言，使用诸如 VMware（注册商标）和 Xen（注册商标）之类的虚拟化软件的虚拟化技术使得一台物理机器能够作为多个虚拟机 (VM) 工作。这实现了高效的服务器操作。

[0003] 虚拟化技术还在物理服务器内连同虚拟机建立虚拟交换机。作为基于软件的分组交换机，虚拟交换机在虚拟机之间以及虚拟机与外界之间中继通信，如图 1A 和 1B 中所示。由于虚拟交换机位于虚拟机附近，所以流量控制是容易的。另外，由于虚拟交换机是基于软件的，所以虚拟交换机在灵活性和可扩展性上是优良的。

[0004] 另外，已知一种 I/O（输入 / 输出）虚拟化技术，例如 VT-d/VT-c（注册商标）。I/O 虚拟化技术使得能够在虚拟机与网络接口卡 (NIC) 之间直接交换数据，而无需使用虚拟交换机。具体而言，如图 2 中所示，为每个虚拟机建立虚拟 NIC。然后，对虚拟 NIC 的使用允许了完全绕开虚拟交换机。以下，这种处理被称为“NIC 卸荷”(NIC offload)。

[0005] 作为与虚拟化有关的技术，已知以下技术。

[0006] 在日本早期公开专利申请 No.P2007-522583A 中，公开了一种装置，其包括至少一个路由器和一数据结构。该数据结构用于通过利用该路由器组织一个或多个虚拟网络接口卡 (VNIC) 之间的连接来创建虚拟网络。

[0007] 日本早期公开专利申请 No.P2008-102929A 公开了一种技术，其使用队列数据结构来与网络适配器通信。设备驱动器调用设备驱动器服务以便最初针对该队列数据结构设定路由复合体内部的地址转化和保护表 (ATPT) 的项目。该设备驱动器服务将未转换地址返回到设备驱动器，并且该未转换地址随后被提供给网络适配器。响应于通过搜索队列数据结构而获得队列元素，网络适配器请求对指定到该队列元素的未转换地址的转换，这使得能够在接收到针对与该队列元素有关的缓冲器的数据分组之前将经转换的地址保存在网络适配器中。

[0008] 日本早期公开专利申请 No.P2009-151745A 公开了一种虚拟机监视器，其在多处理器系统上运行虚拟服务器。该虚拟机监视器包括物理硬件信息获取部、接收器部和指派处理器部。物理硬件信息获取部获取硬件的配置信息，该配置信息包含包括该多处理器系统中的处理器、存储器和 I/O 设备在内的硬件的物理位置信息。接收器部接收生成请求，该生成请求包括要生成的虚拟服务器中的处理器的数目、存储器量和 I/O 设备和资源的指派策略。指派处理器部根据接收到的生成请求向虚拟服务器指派 I/O 设备，然后向虚拟服务器指派处理器和存储器，以满足指派策略。

发明内容

[0009] 在图 1A 和图 1B 的情况下，虚拟交换机中继虚拟机与外界之间的所有流量。换言

之，流量集中在虚拟交换机上。另外，虚拟交换机是基于软件的，并且交换处理可在单线程中进行。在该情况下，集中的流量无法被处理。鉴于这种情况，虚拟交换机很有可能充当网络处理中的瓶颈。

[0010] 另一方面，使用图 2 中所示的 NIC 卸荷使得能够完全绕开虚拟交换机。然而，在此情况下，分组通信路径是固定的，这消除了基于虚拟交换机的灵活流量控制的优点。

[0011] 本发明的一个目的是抑制虚拟交换机上流量的集中，同时实现基于虚拟交换机的灵活流量控制。

[0012] 在本发明的一个方面中，提供了一种服务器。该服务器包括处理器、连接到处理器的网络适配器以及路线切换器。处理器包括虚拟机和中继在虚拟机与外界之间交换的分组的虚拟交换机。网络适配器具有在不通过虚拟交换机的情况下向虚拟机发送分组和从虚拟机接收分组的传送功能。路线切换器在第一路线模式流和第二路线模式流之间动态地切换由虚拟机发送和接收的分组的流。并且，路线切换器指令传送功能处理第一路线模式流并且指令虚拟交换机处理第二路线模式流。

[0013] 在本发明的另一方面中，提供了一种要由服务器执行的流控制程序。服务器包括处理器和连接到处理器的网络适配器。处理器包括虚拟机和中继在虚拟机与外界之间交换的分组的虚拟交换机。网络适配器具有在不通过虚拟交换机的情况下向虚拟机发送分组和从虚拟机接收分组的传送功能。流控制程序允许服务器实现路线切换功能。路线切换功能在第一路线模式流和第二路线模式流之间动态地切换由虚拟机发送和接收的分组的流。并且，路线切换功能指令传送功能处理第一路线模式流并且指令虚拟交换机处理第二路线模式流。

[0014] 在本发明的另一方面中，提供了一种适于连接到服务器的处理器的网络适配器。处理器包括虚拟机和中继在虚拟机与外界之间交换的分组的虚拟交换机。网络适配器具有向虚拟机发送分组和从虚拟机接收分组的传送功能。网络适配器包括路线切换器。路线切换器在第一路线模式流和第二路线模式流之间动态地切换由虚拟机发送和接收的分组的流。并且，路线切换器指令传送功能处理第一路线模式流并且指令虚拟交换机处理第二路线模式流。

[0015] 本发明使得能够抑制虚拟交换机上流量的集中，同时实现基于虚拟交换机的灵活流量控制。

附图说明

[0016] 从以下结合附图理解的描述中将更清楚本发明的以上和其他优点和特征，附图中：

[0017] 图 1A 是示出虚拟交换机的一个示例的概念图；

[0018] 图 1B 是示出虚拟交换机的另一示例的概念图；

[0019] 图 2 是示出 NIC 卸荷功能的概念图；

[0020] 图 3 是示意性示出根据本发明实施例的网络系统的示例性配置的框图；

[0021] 图 4 是示出根据本发明实施例的服务器的硬件配置的框图；

[0022] 图 5 是概念性示出根据本发明一个实施例的服务器的配置的框图；

[0023] 图 6 是示出根据本发明一个实施例的网络适配器的整体配置的框图；

[0024] 图 7 是示出本发明一个实施例中的接收过滤器表格的一个示例的概念图；

[0025] 图 8 是示出根据本发明一个实施例的路线切换器的功能的示意图；

[0026] 图 9 是示出根据本发明一个实施例的路线切换处理的一个示例的概念图；

[0027] 图 10 是示出本发明一个实施例中的发送过滤器表格的一个示例的概念图；

[0028] 图 11 是示出本发明一个实施例中的两个路线模式的概念图；

[0029] 图 12 是示出本发明一个实施例中的发送 / 接收过滤器表格的一个示例的概念图；

[0030] 图 13 是示出根据本发明一个实施例的虚拟交换机的配置示例的框图；

[0031] 图 14 是示出本发明一个实施例中的缓存控制的概念图；

[0032] 图 15 是示出根据本发明第一实施例的虚拟交换机的配置的框图；

[0033] 图 16 是示出第一实施例中的处理的流程图；

[0034] 图 17 是示出本发明一个实施例中的流表格的一个示例的概念图；

[0035] 图 18 是示出本发明实施例中的端口 -VM 对应表格的一个示例的概念图；

[0036] 图 19 是示出第一实施例中的处理的概念图；

[0037] 图 20 是示出第一实施例中的处理的概念图；

[0038] 图 21 是示出根据本发明第二实施例的配置示例的框图；

[0039] 图 22 是示出第二实施例中的处理的流程图；

[0040] 图 23 是示出根据本发明第三实施例的配置示例的框图；

[0041] 图 24 是描述根据本发明一个实施例的路线切换处理的另一示例的框图；

[0042] 图 25 是示出图 24 中所示的虚拟机的分支功能所参考的流表格的一个示例的概念图；并且

[0043] 图 26 是示出图 24 的情况中的虚拟交换机的配置的框图。

具体实施方式

[0044] 以下将参考附图描述本发明的实施例。

[0045] 1. 整体配置

[0046] 图 3 是示意性示出根据一个实施例的网络系统 1 的配置示例的框图。网络系统 1 包括连接到网络（未示出）的多个服务器 10。多个交换机被部署在服务器 10 之间。网络系统 1 通过诸如防火墙和负荷均衡器之类的网络用具连接到外部网络。网络系统 1 例如是设在数据中心内的网络系统。

[0047] 图 4 是示出根据此实施例的每个服务器（物理服务器）10 的硬件配置的框图。服务器 10 包含 CPU（中央处理单元）20、主存储器 30 和网络适配器（网络接口装置）100。网络适配器 100 也可被称为网络卡或 NIC（网络接口卡）。CPU 20、主存储器 30 和网络适配器 100 相互连接。

[0048] 主存储器 30 存储虚拟化软件和流控制程序 PROG。虚拟化软件包括由 CPU 20 执行的计算机程序，并且虚拟机 (VM) 和虚拟交换机被建立在服务器 10 上。流控制程序 PROG 是由 CPU 20 执行并用于在服务器 10 中实现稍后将描述的“路线切换功能”的计算机程序。虚拟化软件和流控制程序 PROG 可被记录在非暂态计算机可读记录介质上。流控制程序 PROG 可被结合在虚拟化软件中。

[0049] 图 5 是概念性示出根据此实施例的服务器 10 的配置的框图。服务器 10 包括处理器 40 和连接到处理器 40 的网络适配器 100。处理器 40 是由上述 CPU 20、主存储器 30、虚拟化软件和流控制程序 PROG 协作实现的，并且设有基于虚拟环境的各种功能。具体而言，处理器 40 包括管理程序 (hypervisor)50、虚拟交换机 200 和一个或多个虚拟机（虚拟服务器）300。管理程序 50 管理各个主存储器 30 的操作并且还提供虚拟机 300 之间的通信路径。管理程序 50 也可被称为虚拟机监视器 (VMM)。虚拟交换机 200 向外界中继从虚拟机 300 发送的分组或者从外界中继由虚拟机 300 接收的分组。虚拟交换机 200 可在控制虚拟机（控制 VM）上操作（参考图 1A）或者可在管理程序 50 上操作（参考图 1B）。各个应用在各个虚拟机 300（客 VM）上运行。控制虚拟机（控制 VM）也可被称为输入 / 输出虚拟机 (IOVM)。

[0050] 在此实施例中，“NIC 卸荷”是由网络适配器 100 实现的。即，可直接在网络适配器 100 与虚拟机 300 之间交换数据，而不通过虚拟交换机 200。

[0051] 图 6 是示出根据此实施例的网络适配器 100 的整体配置的框图。网络适配器 100 包括虚拟 NIC（在图 6 中以虚线框指示）、接收过滤器 110、发送过滤器 120、存储单元 130 以及直接数据传送功能 140。直接数据传送功能 140 是不通过虚拟交换机 200、直接向或自虚拟机 300 发送或接收分组的功能。具体地，直接数据传送功能 140 在网络适配器 100 的发送 / 接收队列和虚拟机 300 所使用的地址空间之间直接传送数据。

[0052] 虚拟 NIC 是分别为虚拟机 300(VM1，VM2，---) 准备的。每个虚拟 NIC 包括接收队列 101 和发送队列 102。接收队列 101 存储由网络适配器 100 从数据链路接收的接收分组。接收队列 101 中存储的接收分组被直接数据传送功能 140 直接发送到相应的虚拟机 300。另外，由网络适配器 100 利用直接数据传送功能 140 从虚拟机直接接收的发送分组被存储在与该虚拟机相对应的发送队列 102 中。

[0053] 另外，为虚拟交换机 200 准备了另一虚拟 NIC。连接到虚拟交换机 200 的虚拟 NIC 中的接收队列 101 和发送队列 102 以下分别被称为接收队列 101-S 和发送队列 102-S。接收队列 101-S 存储由网络适配器 100 从外部数据链路接收的接收分组。接收队列 101-S 中存储的接收分组被发送到虚拟交换机 200。另外，由网络适配器 100 从虚拟交换机 200 接收的发送分组被存储在发送队列 102-S 中。

[0054] 发送过滤器 120 以预定的顺序或在预定的定时选择发送队列 102 和 102-S。发送过滤器 120 随后从所选的发送队列 102 或 102-S 中提取发送分组并将发送分组发送到数据链路。应当注意，发送队列 102 可以只存储分组的元数据，例如存储分组的虚拟机 300 的地址，而不存在分组的原始数据。在此情况下，发送过滤器 120 在选择了接下来从其提取分组的发送队列 102 时，通过使用存储在相应队列中的分组的元数据来指令直接数据传送功能 140 从虚拟机 300 传送分组。

[0055] 接收过滤器 110 接收来自数据链路的接收分组。接收过滤器 110 选择要将接收分组存储在其中的接收队列 101 或 101-S。接收过滤器表格 FILT1 用于此选择。接收过滤器表格 FILT1 被存储在存储单元 130 中。存储单元 130 的示例包括 DRAM、SRAM、内容可寻址存储器 (CAM) 等等。

[0056] 接收过滤器表格 FILT1 是示出流与接收动作之间的关系的表格。接收过滤器 110 参考接收过滤器表格 FILT1 来对接收分组执行与接收分组流相关的接收动作。有两个接收

动作可用。第一接收动作是利用直接数据传送功能 140 将接收分组直接发送到指定的虚拟机 300。在此情况下，接收过滤器 110 将接收分组存储在指定的接收队列 101 中。第二接收动作是将接收分组发送到虚拟交换机 200。在此情况下，接收过滤器 110 将接收分组存储在与虚拟交换机 200 相关联的接收队列 101-S 中。

[0057] 图 7 示出了接收过滤器表格 FILT1 的一个示例。接收过滤器表格 FILT1 具有多个过滤器条目。每个过滤器条目指示用于标识流的键 (key) 和要对相应的流的接收分组执行的接收动作。键是流标识信息并且由接收分组的头部信息中的预定协议头部字段的组合构成。此键与例如 OpenFlowSwitch（参考 http://www.openflowswitch.org/）的流表格中的键类似。接收动作指示要在其中存储接收分组的接收队列。例如，意味着与虚拟机 VM1 相关联的接收队列 101 的“接收动作 :VM1”相当于上述的第一接收动作。另外，意味着与虚拟交换机 200 相关联的接收队列 101-S 的“接收动作 :vswitch”相当于上述的第二接收动作。

[0058] 在接收到接收分组后，接收过滤器 110 通过使用接收分组的头部信息来检索接收过滤器表格 FILT1 中的精确匹配条目 (exact match entry)。如果有与接收分组的流匹配的精确匹配条目，则接收队列 101 对接收分组执行由精确匹配条目指定的第一接收动作。在图 7 的示例中，例如，接收过滤器 110 将属于流“流 1”的接收分组存储在与虚拟机 VM1 相关联的接收队列 101 中。另一方面，如果没有与接收分组的流匹配的精确匹配条目，则接收过滤器 110 对接收分组执行第二接收动作。即，接收分组被存储在与虚拟交换机 200 相关联的接收队列 101-S 中。这种操作提供 NIC 卸荷。

[0059] 此实施例的服务器 10 还包括路线切换器 60。图 8 是示出根据此实施例的路线切换器 60 的功能的示意图。在此实施例中，路线切换器 60 “动态”切换由虚拟机 300 发送或接收的分组的传送路线。

[0060] 具体地，作为由虚拟机 300 发送或接收的分组的传送路线，有两个模式可用。在第一路线模式中，通过使用如上所述的网络适配器 100 中的直接数据传送功能 140，而不通过虚拟交换机 200，在网络适配器 100 与虚拟机 300 之间直接交换分组 (NIC 卸荷）。另一方面，在第二路线模式中，至少通过虚拟交换机 200 向或自虚拟机 300 发送或接收分组。第一和第二路线模式的流以下分别被称为“第一路线模式流”和“第二路线模式流”。

[0061] 路线切换器 60 把由虚拟机 300 发送和接收的分组的流路线设定成第一和第二路线模式之一。另外，路线切换器 60 基于预定的条件动态地切换路线设定。即，路线切换器 60 把由虚拟机 300 发送或接收的分组的流动态地切换（或配置）到第一路线模式流或第二路线模式流。路线切换器 60 随后指令网络适配器 100 中的直接数据传送功能 140 提供第一路线模式流，并且指令虚拟交换机 200 提供处理第二路线模式流。

[0062] 如这样所述，在此实施例中不是所有的流都始终固定地绕开虚拟交换机 200。仅对期望的流（第一路线模式流）执行 NIC 卸荷以绕开虚拟交换机 200。其余的流（第二路线模式流）像通常操作中那样经过虚拟交换机 200。这有效地抑制了流量在虚拟交换机 200 上的集中，同时提供了基于虚拟交换机 200 的灵活流量控制。

[0063] 应当注意，路线切换器 60 是通过在服务器 10(CPU 20) 上执行流控制程序 PROG 实现的。路线切换器 60 可如图 5 中所示被结合在处理器 40 中。取而代之，路线切换器 60 可被结合在网络适配器 100 中（稍后将在第 3-3 节中描述）。通常，路线切换器 60 被结合在处理器 40 中的虚拟交换机 200 或管理程序 50 中；应当注意本发明并不限于这种配置。

[0064] 在下文中，将对根据此实施例的路线切换处理给出详细描述。

[0065] 2. 路线切换处理的示例

[0066] 图 9 是示出根据此实施例的路线切换处理的一个示例的概念图。在此处理示例中，网络适配器 100 设有发送过滤器表格 FILT2 以及接收过滤器表格 FILT1。与接收过滤器表格 FILT1 类似，发送过滤器表格 FILT2 也被存储在存储单元 130 中。应当注意，接收过滤器表格 FILT1 和发送过滤器表格 FILT2 可被统称为“过滤器表格 FILT”。

[0067] 2-1. 发送过滤器表格

[0068] 发送过滤器表格 FILT2 指示流与发送动作之间的关系。发送过滤器 120 参考发送过滤器表格 FILT2 来对发送分组执行与流相关的发送动作。作为发送动作，有两种模式可用。第一发送动作是将发送分组发送到外部数据链路。在此情况下，发送过滤器 120 将发送分组发送到数据链路。第二发送动作是将发送分组作为接收分组环回到接收过滤器 110（即接收路线）。在此情况下，发送过滤器 120 将发送分组作为接收分组环回到接收过滤器 110。

[0069] 图 10 示出了发送过滤器表格 FILT2 的一个示例。发送过滤器表格 FILT2 具有我个过滤器条目。每个过滤器条目指示用于标识流的键和要对相应的流的发送分组执行的发送动作。键是流标识信息并且由发送分组的头部信息中的预定协议头部字段的组合构成。此键与例如 OpenFlowSwitch（参考 http://www.openflowswitch.org/）的流表格中的键类似。发送动作指示第一发送动作“向外”或第二发送动作“环回”。

[0070] 在从所选的发送队列 102 提取了发送分组后，发送过滤器 120 通过使用发送分组的头部信息来检索发送过滤器表格 FILT2 中的精确匹配条目。如果有与发送分组的流匹配的精确匹配条目，则发送过滤器 120 对发送分组执行由精确匹配条目指定的第一发送动作（向外）。即，发送分组被发送到数据链路。另一方面，如果没有与发送分组匹配的精确匹配条目，则发送过滤器 120 对发送分组执行第二发送动作（环回）。即，发送分组被作为接收分组环回到接收过滤器 110（即接收路线）。

[0071] 以下将参考图 9 和图 11 来描述这两个路线模式。在图 9 和 11 的示例中，在接收过滤器表格 FILT1 中只有流“流 1”和“流 2”与第一接收动作相关，而其他流与第二接收动作相关。另外，在发送过滤器表格 FILT2 中只有流“流 1”和“流 2”与第一发送动作相关，而其他流与第二发送动作相关。

[0072] 在此处理示例中，从虚拟机 300 发送的发送分组首先被输入到网络适配器 100。此时，利用网络适配器 100 中的直接数据传送功能 140，不通过虚拟交换机 200，将发送分组直接输入到网络适配器 100。发送过滤器 120 从所选的发送队列 102 提取发送分组。

[0073] 如果发送分组属于流“流 1”或“流 2”，则在发送过滤器表格 FILT2 中命中到精确匹配条目。从而，发送过滤器 120 将发送分组发送到数据链路。即，发送分组被从虚拟机 300 通过网络适配器 100 发送到外界，而不经过虚拟交换机 200。这相当于第一路线模式。

[0074] 另一方面，如果发送分组属于另一不同的流，则在发送过滤器表格 FILT2 中没有命中到精确匹配条目。从而，发送过滤器 120 将发送分组作为接收分组环回到接收过滤器 110。在接收过滤器表格 FILT1 中也没有命中到键。因此，接收过滤器 110 通过接收队列 101-S 将接收分组发送到虚拟交换机 200。换言之，分组一度被输入到网络适配器 100，然后被虚拟交换机 200 处理。这相当于第二路线模式。

[0075] 对从数据链路接收的接收分组的处理如下。如果接收分组属于流“流 1”或“流 2”，则在接收过滤器表格 FILT1 中命中到精确匹配条目。从而，接收过滤器 110 将接收分组存储在与相应的虚拟机 300 相关联的接收队列 101 中。利用直接数据传送功能 140 将接收分组直接发送到虚拟机 300，而不通过虚拟交换机 200。这相当于第一路线模式。

[0076] 另一方面，如果接收分组另一不同的流，则在接收过滤器表格 FILT1 中没有命中到精确匹配条目。从而，接收过滤器 110 将接收分组存储在与虚拟交换机 200 相关联的接收队列 101-S 中。因此，接收分组被虚拟交换机 200 处理。这相当于第二路线模式。

[0077] 应当注意，接收过滤器表格 FILT1 和发送过滤器表格 FILT2 可被组合并作为单个发送 / 接收过滤器表格提供，如图 12 中所示。在图 12 的示例中，第二接收动作和第二发送动作共同包括将分组存储在与虚拟交换机 200 相关联的接收队列 101-S 中，如符号“vswitch”所指示。这也实现了发送分组到接收路线的环回。

[0078] 2-2. 路线切换器 60

[0079] 如上所述，根据接收过滤器表格 FILT1 和发送过滤器表格 FILT2 中的条目设定，由虚拟机 300 发送或接收的分组的流路线被设定为第一路线模式或第二路线模式。此外，通过修改接收过滤器表格 FILT1 和发送过滤器表格 FILT2 中的条目设定，可以“动态地”切换流路线。路线切换器 60 执行这种条目设定和设定的修改。

[0080] 具体而言，路线切换器 60 根据预定的标准把由虚拟机 300 发送或接收的分组的流指派到第一路线模式流或第二路线模式流。该指派可被动态修改。路线切换器 60 设定接收过滤器表格 FILT1，以使得第一路线模式流与第一接收动作相关，并且第二路线模式流与第二接收动作相关。另外，路线切换器 60 设定发送过滤器表格 FILT2，以使得第一路线模式流与第一发送动作相关，并且第二路线模式流与第二发送动作相关。结果，不通过虚拟交换机 200 处理第一路线模式流，即对第一路线模式流执行 NIC 卸荷。另一方面，由虚拟交换机 200 处理第二路线模式流。

[0081] 应当注意，可以只在接收过滤器表格 FILT1 和发送过滤器表格 FILT2 的一个中设定与同一流相关联的过滤器条目。在该情况下，在接收侧和发送侧之间，路线模式变成非对称的。作为一个示例，让我们考虑只在如上所述的图 9 和图 11 中的发送过滤器表格 FILT2 中设定与流“流 1”相关联的过滤器条目的情况。在该情况下，对于流“流 1”，发送分组的传送路线被设定成发送分组不经过虚拟交换机 200 的第一路线模式，而接收分组的传送路线被设定成接收分组经过虚拟交换机 200 的第二路线模式。

[0082] 路线切换器 60 被结合在例如虚拟交换机 200 中。图 13 是示出该情况下的虚拟交换机 200 的示例性功能配置的框图。虚拟交换机 200 设有流识别功能 210、分组交换功能 220、VM 识别功能 230、队列判定功能 240 和 NIC 设定功能 250。

[0083] 虚拟交换机 200 接收来自网络适配器 100 和虚拟机 300 的分组。流识别功能 210 基于接收到的分组的头部信息来识别每个接收到的分组所属的流。另外，流识别功能 210 参考指示流标识信息（键）与动作 (Action) 之间的关系的流表格 TBL 来获得要对分组执行的动作。分组交换功能 220 根据该动作来处理分组。通常，流表格 TBL 的动作描述分组的输出端口（传送目的地）。分组交换功能 220 从该动作所指定的输出端口输出分组。所输出的分组被发送到网络适配器 100 或虚拟机 300。

[0084] 应当注意，如果在流表格 TBL 中没有与分组的过滤器条目匹配，则流识别功能 210

对分组执行预定的处理。例如，流识别功能 210 将分组传送到开放流控制器 (OFC) 并且请求路线设定。

[0085] VM 识别功能 230 指定要用来发送或接收属于指定的流的分组的虚拟机 300。这里，“指定的流”意味着如下流：在网络适配器 100 上希望对该流执行过滤器表格 FILT 中的条目设定。队列判定功能 240 判定与 VM 识别功能 230 指定的虚拟机 300 相关的发送 / 接收队列 (101、102)。NIC 设定功能 250 通过适当地参考发送 / 接收队列来准备要对过滤器表格 FILT 设定的过滤器条目。NIC 设定功能 250 随后将准备的过滤器条目通知给网络适配器 100，然后设定或修改过滤器表格 FILT。

[0086] 路线切换器 60 包含上述 VM 识别功能 230、队列判定功能 240 和 NIC 设定功能 250。

[0087] 2-3. 缓存控制

[0088] 还可实现过滤器表格 FILT 的缓存控制。这对于在网络适配器 100 中只能安装相对较小的存储单元 130 的情况是优选的。以下将参考图 14 来描述缓存控制。

[0089] 如图 14 中所示，过滤器表格 FILT 的主体被存储在服务器 10 中的主存储器 30（参考图 4）中。NIC 设定功能 250（或路线切换器 60）设定或修改主存储器 30 上的过滤器表格 FILT。

[0090] 网络适配器 100 中的存储单元 130 是相对较小容量（例如数十千字节）的缓存存储器。缓存在缓存存储器 130 中的过滤器表格 FILT（缓存）是存储在主存储器 30 中的过滤器表格 FILT 的一部分。

[0091] 网络适配器 100 中的接收过滤器 110 和发送过滤器 120 中的每一个设有检索功能 115。当接收到分组时，检索功能 115 首先检查缓存存储器 130 中缓存的条目。当其结果是缓存命中时，检索功能 115 根据命中的条目如上所述处理分组。另一方面，当发生缓存未命中时，检索功能 115 访问主存储器 30 并且搜索过滤器表格 FILT 的主体以获得必要的条目。检索功能 115 随后将所获得的条目存储在缓存存储器 130 中并且根据这些条目来处理分组。如果没有空条目，则检索功能 115 还执行缓存条目的交换。

[0092] 应当注意，过滤器表格 FILT 的每个条目可包括每次处理分组时更新的统计信息。在图 14 的示例中，每个条目包括该条目的匹配的数目。检索功能 115 在预定的定时将该统计信息从缓存存储器 130 写回到主存储器 30。该预定的定时可以是路线切换器 60 要求统计信息的信息、条目被从缓存存储器 130 去除的定时，等等。

[0093] 3. 实施例的变化

[0094] 如上所述，第一路线模式流是不通过虚拟交换机 200 处理的，即对第一路线模式流执行 NIC 卸荷。此 NIC 卸荷抑制了流量在虚拟交换机 200 上的集中。存在要被执行 NIC 卸荷的第一路线模式流的各种候选。另外，存在各种允许的 NIC 卸荷的设定定时。以下将描述若干实施例。

[0095] 3-1. 第一实施例

[0096] 在第一实施例中，要被执行 NIC 卸荷的第一路线模式流是负荷超过了预定阈值的“超负荷流”。另一方面，第二路线模式流是负荷等于或小于该预定阈值的“通常负荷流”。NIC 卸荷的开始定时是特定流从通常负荷流变成超负荷流的定时，并且 NIC 卸荷的结束定时是当该特定流从超负荷流返回到通常负荷流的定时。

[0097] 为了实现这一点，路线切换器 60 基于由虚拟机 300 发送或接收的分组来测量每个

流的负荷。路线切换器 60 将测量到的负荷与预定阈值相比较，并且判定每个流是通常负荷流还是超负荷流。当特定的流从通常负荷流变成超负荷流时，路线切换器 60 将该超负荷流切换成第一路线模式流。结果，对该超负荷流执行了 NIC 卸荷以绕开虚拟交换机 200。另外，当该特定流从超负荷流返回到通常负荷流时，路线切换器 60 使该流从第一路线模式流返回到第二路线模式流。结果，从那时起由虚拟交换机 200 处理该流。

[0098] 这样，在第一实施例中，仅对超负荷流执行 NIC 卸荷。这高效地减轻了虚拟交换机 200 上的流量集中。另外，对过滤器表格 FILT 设定的条目的数目相对较小。因此，即使当在网络适配器 100 中只能结合相对较小的存储单元 130 时，第一实施例也是可用的。应当注意，当负荷在发送侧和接收侧之间不平均时，可以使路线模式在发送侧和接收侧之间不对称。

[0099] 以下，将对根据第一实施例的具体配置和操作的示例给出描述。在此示例中，路线切换器 60 被结合在虚拟交换机 200 中。

[0100] 图 15 是示出第一实施例中的虚拟交换机 200 的配置的框图。在第一实施例中，除了如上所述的图 13 中所示的配置以外，虚拟交换机 200 还设有处理负荷测量功能 260、路线改变判定功能 270 和地址附加数据生成功能 280。处理负荷测量功能 260 按预定的频率对发送和接收分组采样，并且基于发送和接收分组测量每个流的负荷（分组处理量和处理负荷）。另外，处理负荷测量功能 260 保存指示测量结果的负荷信息。路线改变判定功能 270 通过参考负荷信息来判定每个流是超负荷流（第一路线模式流）还是通常负荷流（第二路线模式流）。即，路线改变判定功能 270 根据负荷信息动态地改变是属于第一路线模式流还是第二路线模式流。然后，路线改变判定功能 270 对于 VM 识别功能 230 指定路线模式应当被改变的（一个或多个）流。稍后将描述地址附加数据生成功能 280。

[0101] 图 16 是示出第一实施例中的示例性处理的流程图。首先，虚拟交换机 200 接收来自网络适配器 100 或虚拟机 300 的分组（步骤 A10）。流识别功能 210 根据接收到的分组的头部信息来识别分组所属的流。另外，流识别功能 210 参考流表格 TBL 并且获得应当对分组执行的动作（步骤 A20）。

[0102] 图 17 示出了流表格 TBL 的一个示例。流表格 TBL 具有多个表格条目。每个表格条目指示：用于标识每个流的键，以及对流的分组执行的动作。键是流标识信息并且由分组的头部信息中的预定协议头部字段的组合构成。动作通常指示分组的输出端口（传送目的地）。这样的流表格 TBL 被存储在预定的存储设备（通常是主存储器 30）中。另外，在图 17 的示例中，每个表格条目具有指示出网络适配器 100 上相应条目的存在与否的标志。此标志为了使得虚拟交换机 200 可以知道网络适配器 100 保存的过滤器条目的类型而准备的。

[0103] 分组交换功能 220 根据在步骤 A20 获得的动作来执行交换处理（步骤 A30）。通常，分组交换功能 220 从该动作所指定的输出端口输出分组。输出的分组被发送到网络适配器 100 或虚拟机 300。

[0104] 另一方面，处理负荷测量功能 260 响应于分组处理而更新负荷信息（步骤 A40）。另外，路线改变判定功能 270 通过参考负荷信息将与所处理的分组的流有关的负荷与预定的阈值相比较（步骤 A50）。如果负荷超过预定阈值（步骤 A50：是），则路线改变判定功能 270 判定流为超负荷流，并且指派到第一路线模式流。路线改变判定功能 270 随后判定要对有关流执行 NIC 卸荷并且向 VM 识别功能 230 报告。

[0105] 随后，虚拟交换机 200 执行卸荷设定处理（步骤 A60）。具体而言，对于由路线改变判定功能 270 指定的流，VM 识别功能 230 指定发送或接收属于指定的流的分组的虚拟机 300（步骤 A61）。这里，VM 识别功能 230 可通过参考如图 18 中所示的端口到 VM 关联表格来指定虚拟机 300。队列判定功能 240 判定与 VM 识别功能 230 指定的虚拟机 300 相关的发送或接收队列（步骤 A62）。NIC 设定功能 250 通过适当地参考该发送或接收队列来准备应当对过滤器表格 FILT 设定的过滤器条目。NIC 设定功能 250 随后将所准备的过滤器条目通知给网络适配器 100 并且设定过滤器表格 FILT（步骤 A63）。另外，NIC 设定功能 250 把图 17 中所示的相应条目的标示设定为“存在”。

[0106] 这样，对被判定为超负荷流的流执行 NIC 卸荷。图 19 是示出第一实施例中的处理映像的概念图。应当注意，当流从超负荷流返回到通常负荷流时，解除卸荷设定。当卸荷设定被解除时，关于该流的过滤器条目可被从过滤器表格 FILT 中去除。另外，图 17 中所示的相应条目的标志被设定为“不存在”。

[0107] 以下将参考图 20 来描述地址附加数据生成功能 280 的作用。存在这样的情况，即在网络适配器 100 中不存在相应的过滤器条目的情形下从数据链路通过虚拟交换机 200 到虚拟机 300 执行分组分发。这里，当没有信息指示出从虚拟交换机 200 输出的分组是寻址到“外界”还是“虚拟机 (VM)”时，网络适配器 100 不能识别分组的目的地。为了解决此问题，地址数据被附加到分组本身。具体而言，虚拟交换机 200 中的地址附加数据生成功能 280 将指示分组是寻址到“外界”还是“VM”的地址数据附加到由虚拟交换机 200 输出的分组。网络适配器 100 设有分组发送判定功能 150，该分组发送判定功能 150 通过参考地址数据来判定分组分发目的地。

[0108] 作为示例，让我们考虑图 20 中的寻址到虚拟机 VM1 的流“流 3”。在接收到流“流 3”的分组时，虚拟交换机 200 参考流表格 TBL 并因此认识到该分组是寻址到虚拟机 VM1 的。从而，地址附加数据生成功能 280 向分组附加指示出分组寻址到“VM1”的地址数据。当分组到达网络适配器 100 时，分组发送判定功能 150 通过参考附加到分组的地址数据判定分组要被发送到虚拟机 VM1。另一方面，对于图 20 中的流“流 1”和“流 2”，对于来自虚拟机 VM1 的发送分组没有附加地址数据。在该情况下，如上所述，根据发送过滤器表格 FILT2 中的过滤器条目来处理发送分组。

[0109] 3-2. 第二实施例

[0110] 在第二实施例中，在接收到“预定分组”时执行 NIC 卸荷设定。当接收到特定流的“预定分组”时，路线切换器 60 将该流指派到被执行 NIC 卸荷的第一路线模式流。从那时起，对该特定流执行 NIC 卸荷，并且属于该流的分组绕开虚拟交换机 200。另外，存在这样的情况，即，没有处理第一路线模式流的分组的时段持续了特定时间以上，即对于第一路线模式流发生超时的情况。在该情况下，路线切换器 60 可将该流从第一路线模式流切换成第二路线模式流。

[0111] “预定分组”的一个示例是第一分组，这是在属于特定流的分组之中首先接收到的分组，即在仍未准备该流的条目的情形中接收到的分组。在此情况下，对该流的第一分组和第一分组以后的分组执行 NIC 卸荷。另外，作为“预定分组”的另一示例的是包括 HTTP 请求 URL 的分组。在此情况下，在虚拟交换机 200 中执行 DPI（深度分组检查）处理之后，对剩余分组执行 NIC 卸荷。这里，DPI 处理是用于利用分组中包括的与高于传输层的层有关的信

息来判定分组所属的流的目的地或处理方法的操作，所述信息例如是该分组包括的 URL 的内容。

[0112] 如这样所述，在第二实施例中，对数据平面中的流量的大多数执行 NIC 卸荷。与第一实施例相比，这允许了进一步减轻虚拟交换机 200 上的流量集中。另外，不对控制平面执行 NIC 卸荷，而数据平面的流量大多数被 NIC 卸荷。因此，保留了通过使用虚拟交换机 200 实现的灵活性。

[0113] 以下，将对第二实施例的配置和操作的具体示例给出描述。在此示例中，路线切换器 60 被结合在虚拟交换机 200 中。另外，“预定分组”被定义为第一分组。

[0114] 图 21 是示出第二实施例中的网络适配器 100 和虚拟交换机 200 的配置示例的框图。网络适配器 100 的配置与图 9 中所示的类似。应当注意，省略了对接收过滤器表格 FILT1 和发送过滤器表格 FILT2 的图示。虚拟交换机 200 的配置与上述图 13 中所示的类似。

[0115] 图 22 是示出第二实施例中的处理示例的流程图。网络适配器 100 中的接收过滤器 110 接收来自数据链路的分组（步骤 B10）。接收过滤器 110 使用接收到的分组的头部信息并且在接收过滤器表格 FILT1 中检索精确匹配条目（步骤 B20）。如果存在与接收到的分组的流匹配的精确匹配条目（步骤 B20 ；是），则接收过滤器 110 将接收到的分组存储在与相应的虚拟机 300 相关联的接收队列 101 中。接收到的分组被直接数据传送功能 140 直接发送到相应的虚拟机 300（步骤 B30）。

[0116] 另一方面，当没有与接收到的分组的流匹配的精确匹配条目时（步骤 B20 ；否），接收过滤器 110 将接收到的分组存储在与虚拟交换机 200 相关联的接收队列 101-S 中。存储在接收队列 101-S 中的接收到的分组被发送到虚拟交换机 200（步骤 B40）。

[0117] 虚拟交换机 200 接收该接收到的分组。根据接收到的分组的头部信息，流识别功能 210 识别分组所属的流，并且搜索流表格 TBL（步骤 B50）。其结果是在流表格 TBL 中不存在与接收到的分组匹配的流条目（精确匹配条目）。从而，流识别功能 210 将接收到的分组识别为第一分组并且判定要对接收到的分组的流执行 NIC 卸荷并向 VM 识别功能 230 报告。

[0118] 接着，虚拟交换机 200 执行卸荷设定处理（步骤 B60）。具体而言，对于由流识别功能 210 指定的流，VM 识别功能 230 指定发送或接收属于该流的分组的虚拟机 300（步骤 B61）。队列判定功能 240 判定与 VM 识别功能 230 所指定的虚拟机 300 相关的发送 / 接收队列（步骤 B62）。NIC 设定功能 250 适当地参考发送 / 接收队列并因此准备要对过滤器表格 FILT 设定的过滤器条目。然后，NIC 设定功能 250 将所准备的过滤器条目通知给网络适配器 100 以设定过滤器表格 FILT（步骤 B63）。另外，NIC 设定功能 250 将该过滤器条目的拷贝也存储在流表格 TBL 中。

[0119] 分组交换功能 220 将第一分组返回到网络适配器 100（步骤 B70）。此时，命中到接收过滤器表格 FILT1 中的精确匹配条目（步骤 B20 ；是）。从而，第一分组被直接数据传送功能 140 直接发送到相应的虚拟机 300（步骤 B30）。对于第一分组以后的分组也同样进行。这样，对该流执行 NIC 卸荷。

[0120] 当对于被执行 NIC 卸荷的流发生超时时，卸荷设定可被解除。例如，网络适配器 100 中的接收过滤器 110 或发送过滤器 120 在过滤器表格 FILT 的过滤器条目中记录最终匹配时刻。虚拟交换机 200 中的流识别功能 210 按给定时段的间隔检查该最终匹配时刻以检测超时。当在特定流中发生超时时，流识别功能 210 指令解除对该流的卸荷设定。NIC 设定

功能 250 从接收过滤器表格 FILT1 中去除与有关流相关的过滤器条目。另外，NIC 设定功能 250 还从流表格 TBL 中去除有关过滤器条目。

[0121] 3-3. 第三实施例

[0122] 图 23 是示出第三实施例中的网络适配器 100 和虚拟交换机 200 的配置示例的框图。以下，将主要对与第二实施例的差异给出描述。

[0123] 网络适配器 100 除了图 21 中所示的配置以外还设有流识别功能 160 和流设定功能 170。流识别功能 160 与虚拟交换机 200 中的流识别功能 210 类似。在图 22 中的流程图中，当没有与接收到的分组的流匹配的精确匹配条目时（步骤 B20 ；否），接收过滤器 110 将接收到的分组传送到流识别功能 160（步骤 B40）。流设定功能 170 向流表格 TBL 设定关于由流识别功能 160 指定的流的过滤器条目。这些功能可通过部署在网络适配器 100 中的通用处理器来实现。

[0124] 如这样所述，路线切换器 60 在第三实施例中被结合在网络适配器 100 中。即，除了数据平面以外，还对过滤器表格 FILT 的设定执行 NIC 卸荷。

[0125] 更一般而言，在第三实施例中，是对诸如过滤器表格 FILT 的设定之类的“标准处理”执行 NIC 卸荷的。换言之，虚拟交换机 200 把用于执行这些标准处理的程序委托给网络适配器 100。这种用于执行标准处理的程序可实现为流表格 TBL 中的通配匹配条目 (wildcard match entry) 的动作。该动作例如可以是用于设定 NAPT（网络地址 / 端口转化）的程序。这允许了在网络适配器 100 中对每个流执行 NAPT 处理。虚拟交换机 200 设有流处理规则卸荷功能 290。流处理规则卸荷功能 290 把自身的流表格 TBL（精确匹配条目和通配匹配条目）的一部分或全部的内容设定到网络适配器 100 上的流表格 TBL。

[0126] 如这样所述，在第三实施例中对标准处理执行 NIC 卸荷。不能在短时间中完成的处理和高级扩展处理在虚拟交换机 200 中像通常操作中那样被处理。

[0127] 4. 路线切换处理的另一示例

[0128] 路线切换的含义不限于以上第 2 节中描述的那些。下面将描述路线切换处理的另一示例。在此处理示例中，发送分组的路线在虚拟机 300 中被分支。

[0129] 图 24 是示出此处理示例中的虚拟机 300 的配置的框图。虚拟机 300 设有 NIC 分组发送 / 接收功能 310、虚拟交换机分组发送 / 接收功能 320、分支功能 330 和协议处理功能 340。协议处理功能 340 是由执行协议处理的程序（通常是 TCP/IP 栈）实现的。NIC 分组发送 / 接收功能 310（第一发送 / 接收功能）是由网络适配器 100 中的直接数据传送功能 140 和适合于发送和接收分组的设备驱动器实现的。虚拟交换机分组发送 / 接收功能 320（第二发送 / 接收功能）是由虚拟交换机 200 和适合于发送和接收分组的设备驱动器实现的。

[0130] 分支功能 330 被部署在协议处理功能 340 与 NIC 分组发送 / 接收功能 310、320 之间。分支功能 330 接收来自协议处理功能 340 的发送分组并将发送分组传送到 NIC 分组发送 / 接收功能 310 和虚拟交换机分组发送 / 接收功能 320 之一。为了执行此传送处理（即发送分组的分类），分支功能 330 参考指示流与分组传送目的地之间的关系的流表格 TBL2。分组传送目的地是从 NIC 分组发送 / 接收功能 310（第一分组传送目的地）或虚拟交换机分组发送 / 接收功能 320（第二分组传送目的地）中选择的。

[0131] 图 25 是示出流表格 TBL2 的概念图。流表格 TBL2 具有多个表格条目。每个表格

条目指示：用于标识流的键和要对该流的发送分组执行的动作。键是流标识信息并且是由发送分组的头部信息中的预定协议头部字段的组合构成的。动作指示发送分组的传送目的地。例如，动作“NIC”指示发送分组的传送目的地是 NIC 分组发送 / 接收功能 310（第一分组传送目的地）。另外，动作“vswitch”指示发送分组的传送目的地是虚拟交换机分组发送 / 接收功能 320（第二分组传送目的地）。流表格 TBL2 被存储在预定的存储设备中（通常是主存储器 30 中）。

[0132] 分支功能 330 参考流表格 TBL2 并从而将发送分组从虚拟机 300 传送到与该发送分组的流相关的分组传送目的地。具体地，分支功能 330 包含流识别功能 331 和附加信息改写功能 332。流识别功能 331 基于发送分组的头部信息来识别发送分组的流。另外，流识别功能 331 参考流表格 TBL2 并判定与有关流相关的分组传送目的地。另外，附加信息改写功能 332 把发送分组的附加信息内的发送接口改写成作为上述判定的结果的分组传送目的地。然后，分支功能 330 将发送分组传送到相应的分组传送目的地。

[0133] 当分组传送目的地是 NIC 分组发送 / 接收功能 310 时，NIC 分组发送 / 接收功能 310 接收到来自分支功能 330 的发送分组。NIC 分组发送 / 接收功能 310 将接收到的发送分组存储在缓冲器中并且指令网络适配器 100 发送这些分组。网络适配器 100 中的直接数据传送功能 140 从缓冲器获得发送分组并将发送分组存储在与发送源的虚拟机 300 相对应的发送队列 102 中。

[0134] 当分组传送目的地是虚拟交换机分组发送 / 接收功能 320 时，虚拟交换机分组发送 / 接收功能 320 接收到来自分支功能 330 的发送分组。虚拟交换机分组发送 / 接收功能 320 将接收到的发送分组存储在缓冲器中并且请求管理程序 50 传送这些分组。管理程序 50 指令虚拟交换机 200 处理发送分组。虚拟交换机 200 从缓冲器获得发送分组并执行交换处理。

[0135] 应当注意，虚拟机 300 接收来自网络适配器 100 或虚拟交换机 200 的接收分组。对于网络适配器 100，NIC 分组发送 / 接收功能 310 接收这些接收分组并将接收分组转发到分支功能 330。另一方面，对于虚拟交换机 200，虚拟交换机分组发送 / 接收功能 320 接收这些接收分组并将接收分组转发到分支功能 330。对于两种情况，分支功能 330 都假装是从同一接口接收到接收分组的。为了实现这一点，附加信息改写功能 332 将每个接收分组的附加信息所指示的接收接口改写到分支功能 330。分支功能 330 随后将接收分组发送到协议处理功能 340。因此，从协议栈不能察觉到多个接收路线。

[0136] 应当注意，分组的附加信息指的是包括与分组的数据相关联地保存的分组的属性和分组的额外信息。除了接收接口以外，附加信息通常还包括分组的长度、有效载荷和每个协议的头部数据的开头地址。诸如发送接口和接收接口之类的“接口”指的是虚拟机 300 与网络之间的虚拟连接点。在图 24 的示例中，去往 / 来自网络适配器 100 的传送线路的输出 / 输入端口和去往 / 来自虚拟交换机 200 的传送线路的输出 / 输入端口是“接口”。

[0137] 在此处理示例中，在网络适配器 100 中未设有发送过滤器表格 FILT2。而是在虚拟机 300 中设有流表格 TBL2。取代动态改变网络适配器 100 中的发送过滤器表格 FILT2 的设定，路线切换器 60 动态地改变每个虚拟机 300 中的流表格 TBL2 的设定。具体而言，路线切换器 60 设定流表格 TBL2，以使得第一路线模式流与第一分组传送目的地相关，并且第二路线模式流与第二分组传送目的地相关。结果，第一路线模式流被不通过虚拟交换机 200 地

处理，即对第一路线模式流执行 NIC 卸荷。另一方面，第二路线模式流被虚拟交换机 200 处理。

[0138] 应当注意，与以上第 2 节的情况类似，路线切换器 60 设定网络适配器 100 中的接收过滤器表格 FILT1。另外，在接收侧与发送侧之间，路线模式可以是非对称的。

[0139] 路线切换器 60 例如被结合在虚拟交换机 200 中。图 26 是示出该情况下的虚拟交换机 200 的功能配置示例的框图。除了上述图 13 或图 15 中所示的配置以外，图 26 中所示的虚拟交换机 200 还包含分支功能设定功能 255。NIC 设定功能 250 设定网络适配器 100 中的接收过滤器表格 FILT1。另一方面，分支功能设定功能 255 设定虚拟机 300 中的流表格 TBL2。虚拟机 300 基于分支功能设定功能 255 的设定来添加、去除和更新流表格 TBL2 的条目。

[0140] 此处理示例可与上述第一至第三实施例的任何一个相组合。在这种情况下，用以上描述中的“虚拟机 300 中的流表格 TBL2 的设定”来替换“网络适配器 100 中的发送过滤器表格 FILT2 的设定”。

[0141] 在上文中，本发明的实施例是参考附图的。然而，应当注意，本发明不限于上述实施例，而是可由本领域技术人员在不脱离其概念的范围内适当地改变。

[0142] 本申请基于 2009 年 12 月 4 日提交的日本专利申请 No. 2009-276679 要求优先权，这里通过引用并入该日本专利申请的全部公开内容。

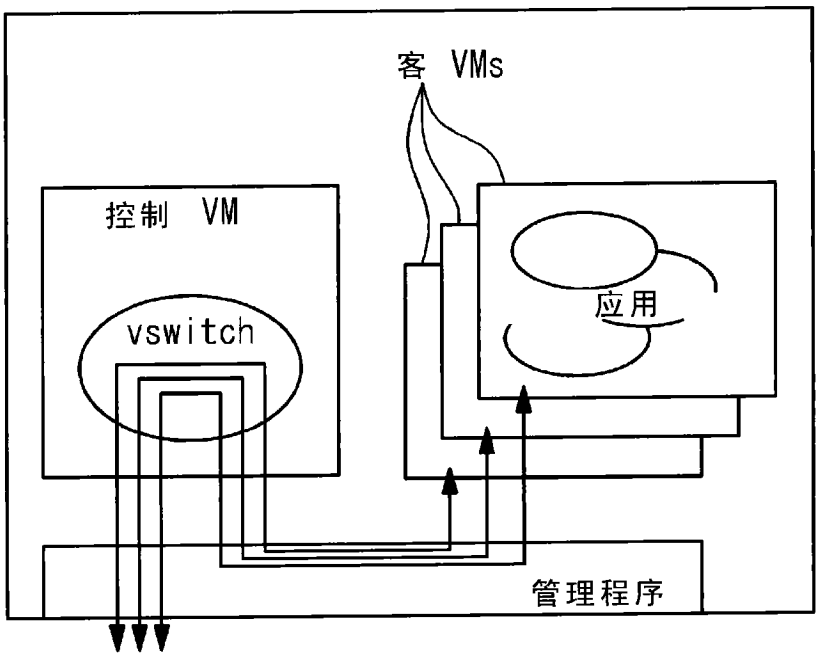


图 1A

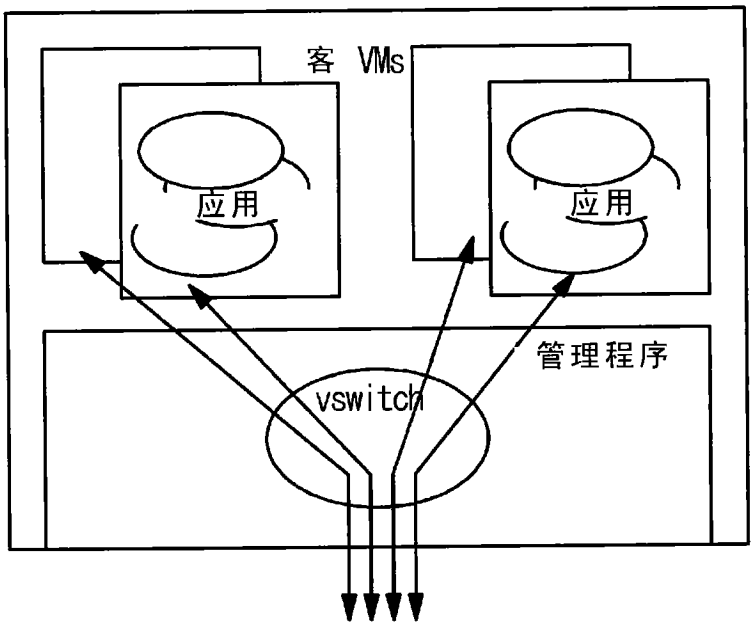


图 1B

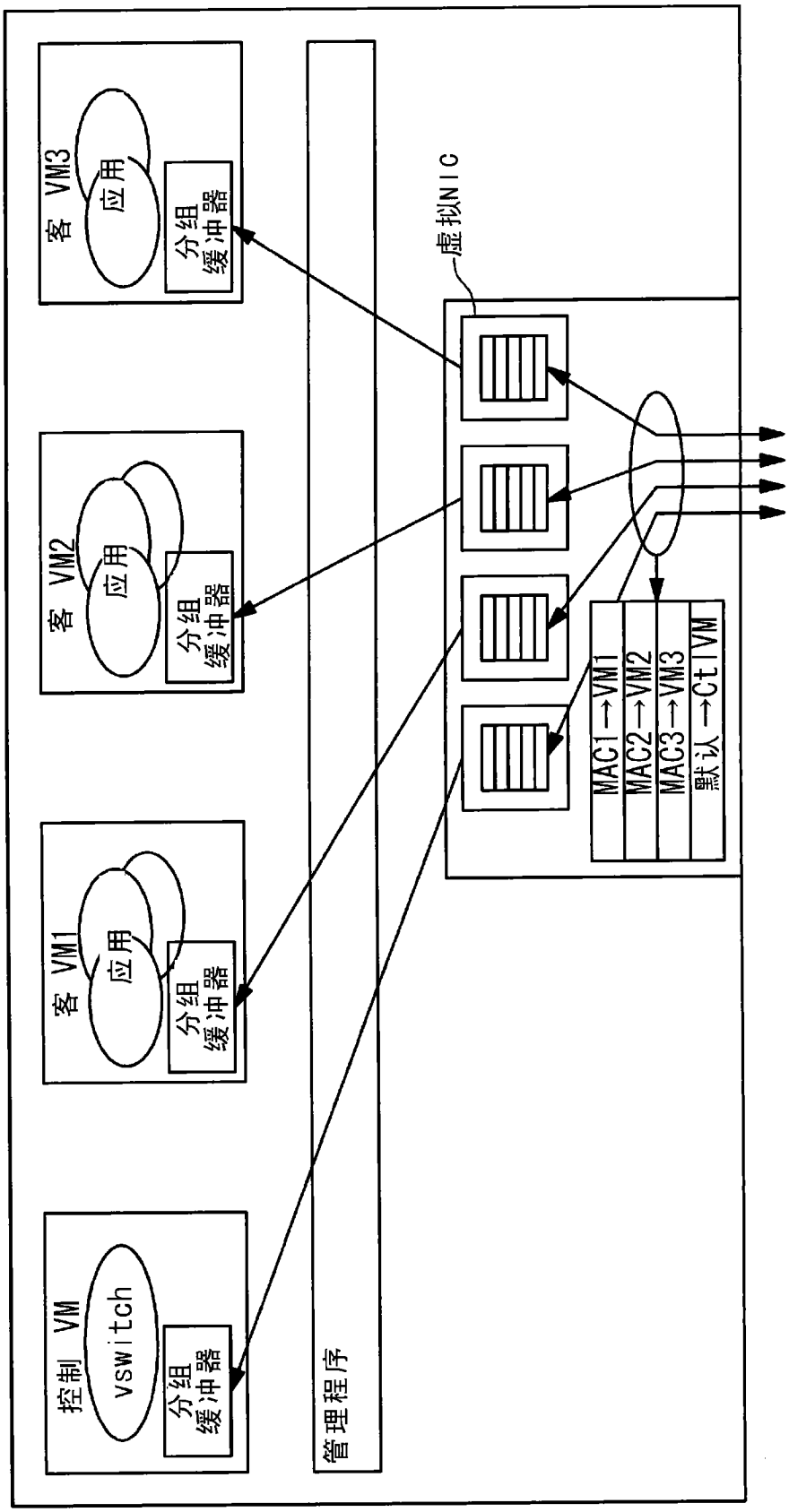

客 VM3
应用
分组缓冲器
客 VM2
应用
分组缓冲器
客 VM1
应用
分组缓冲器
控制 VM
vswitch
分组缓冲器
管理程序
虚拟NIC
MAC1→VM1
MAC2→VM2
MAC3→VM3
默认→CtlVM

图 2

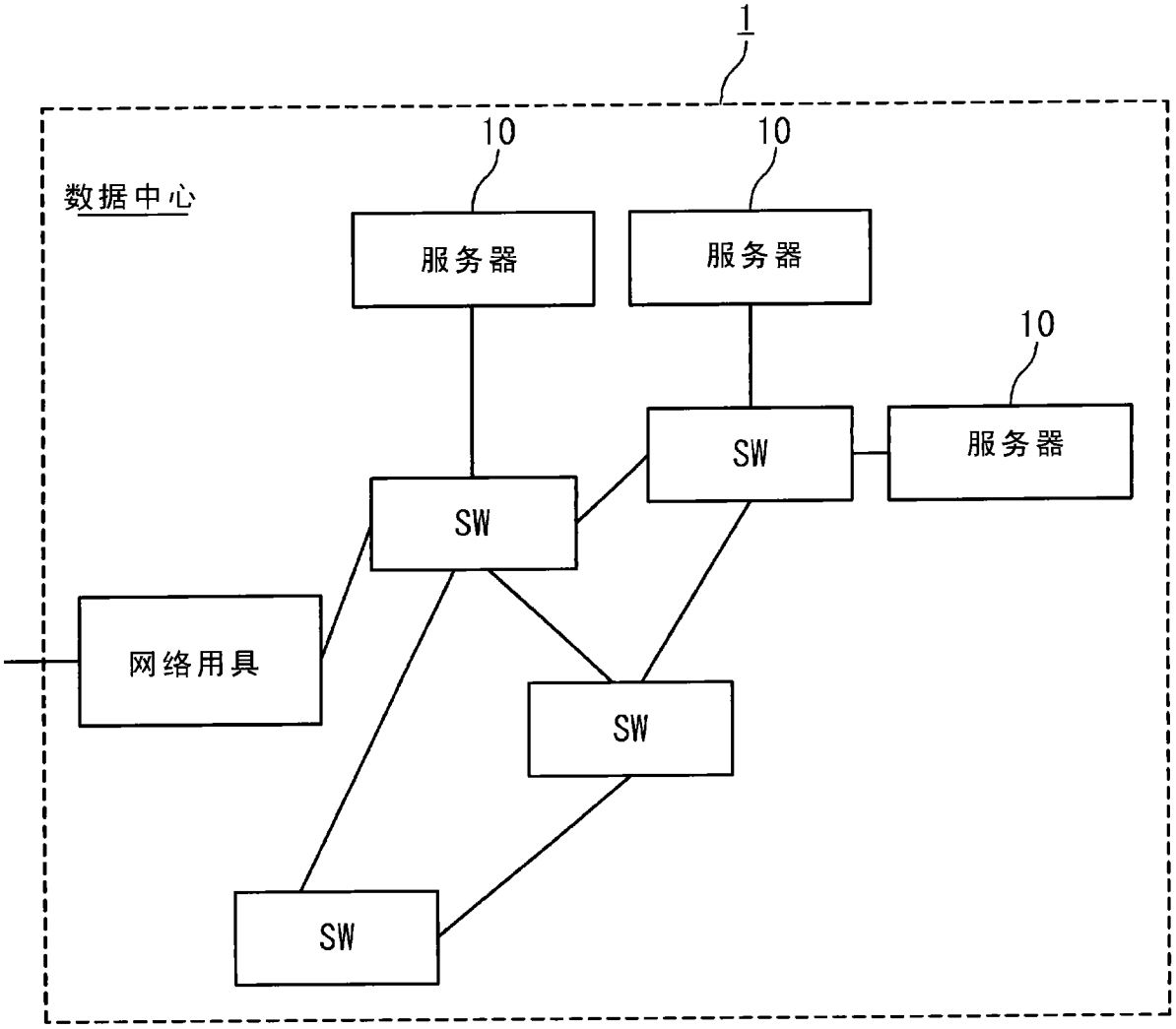


图 3

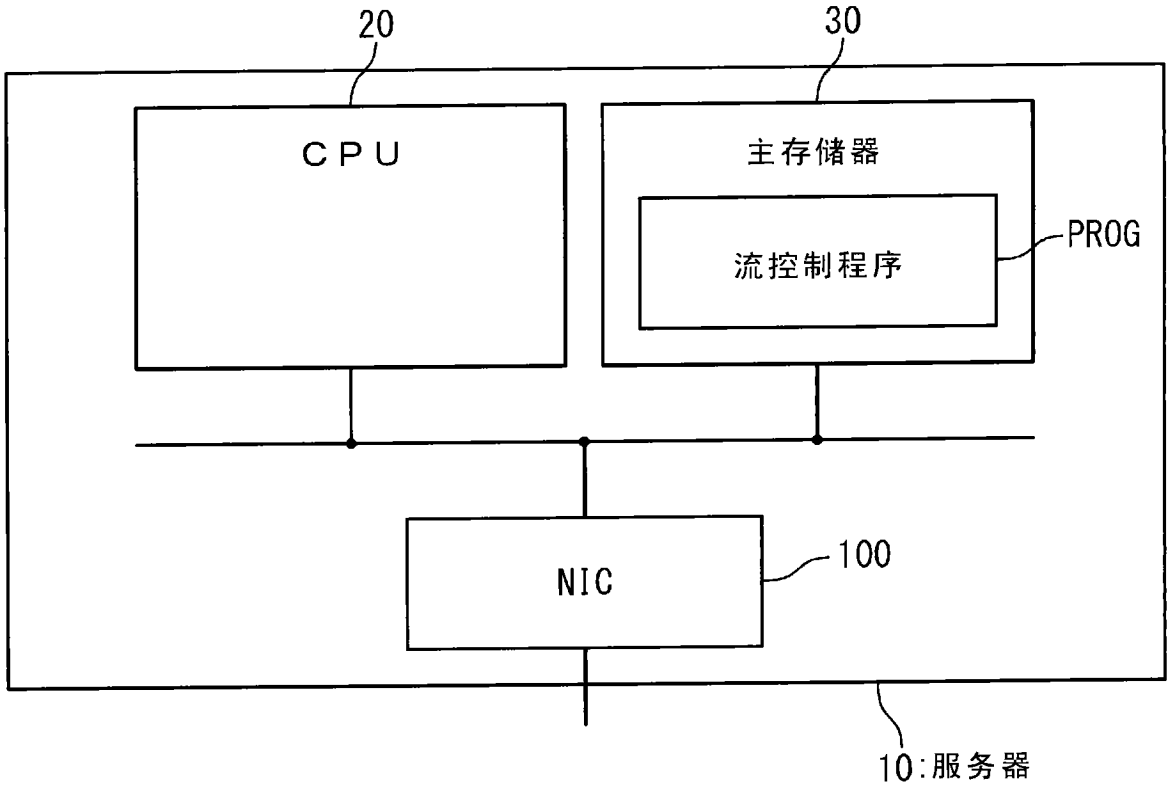


图 4

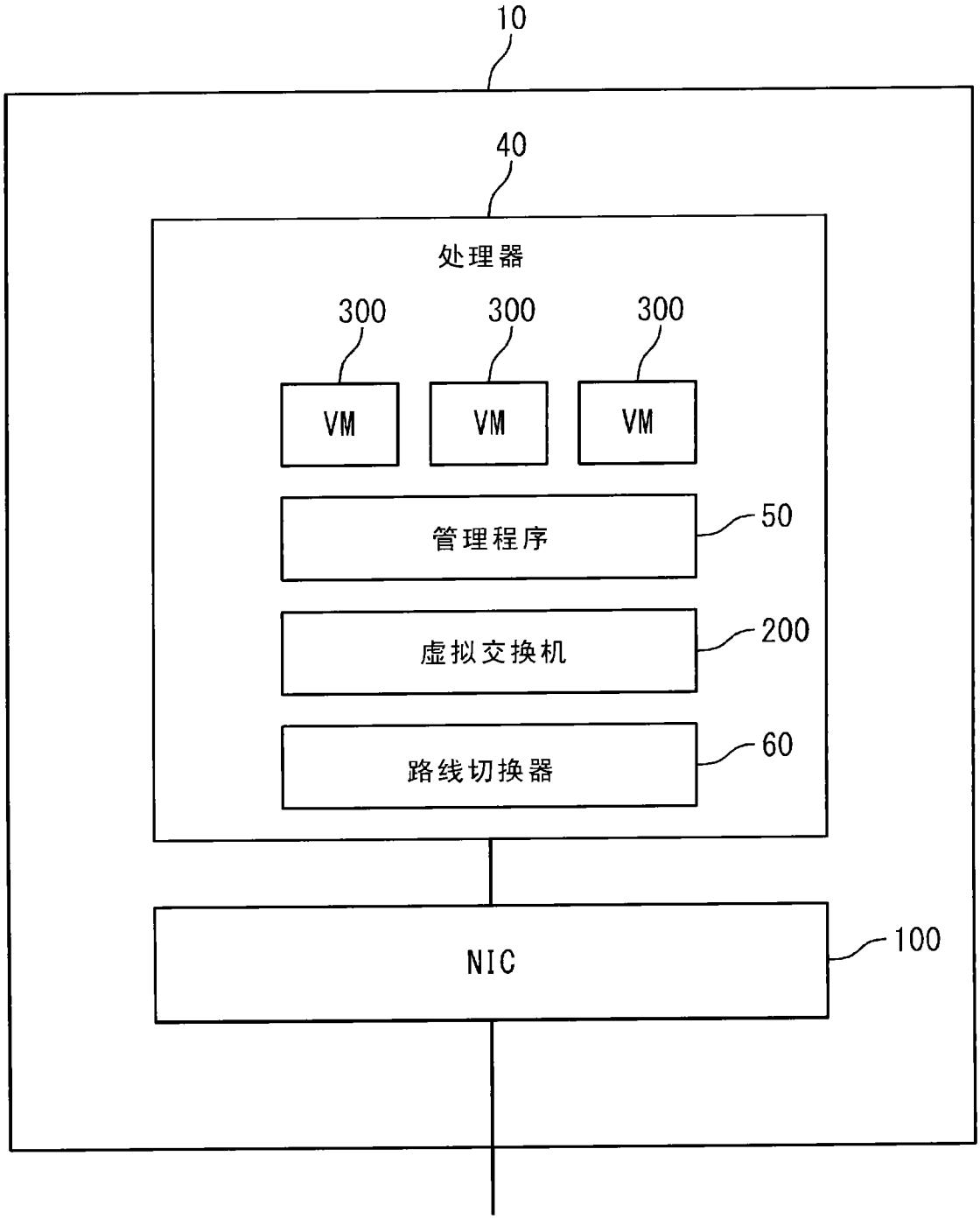


图 5

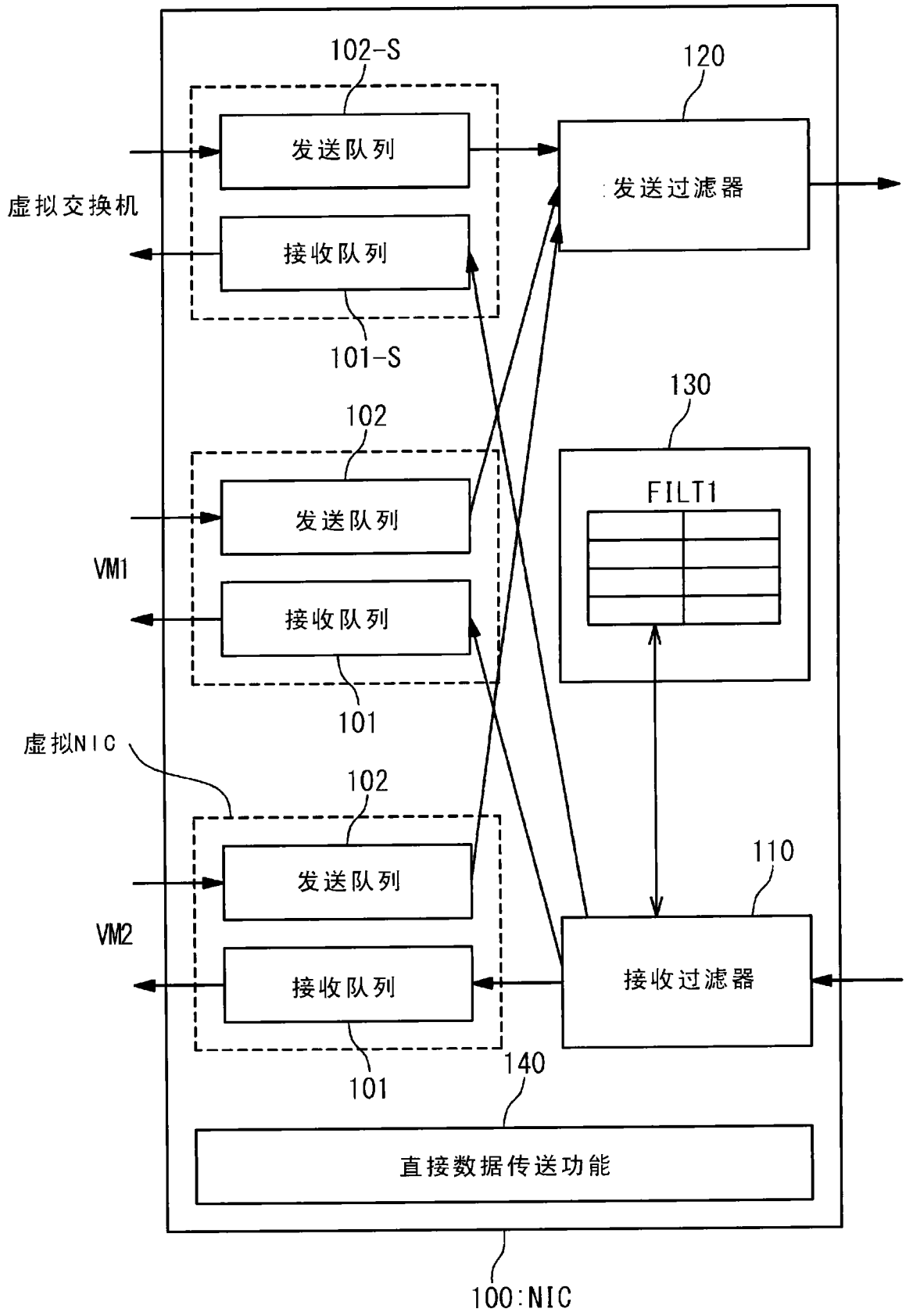

102-S
120
发送队列
发送过滤器
虚拟交换机
接收队列
101-S
130
102
FILT1
发送队列
VM1
接收队列
101
虚拟NIC
102
110
发送队列
VM2
接收过滤器
接收队列
101
140
直接数据传送功能
100:NIC

图 6

FILT1：接收过滤器表格

键	动作
流1	VM1
流2	VM2
流3	
⋮	⋮
*	vswitch

图 7

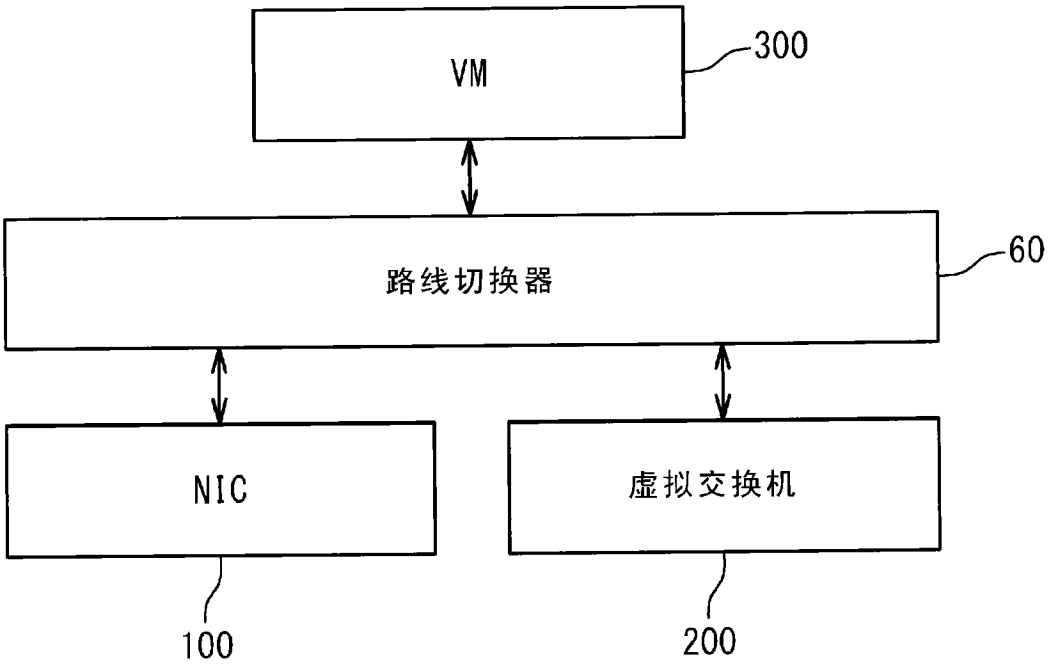


图 8

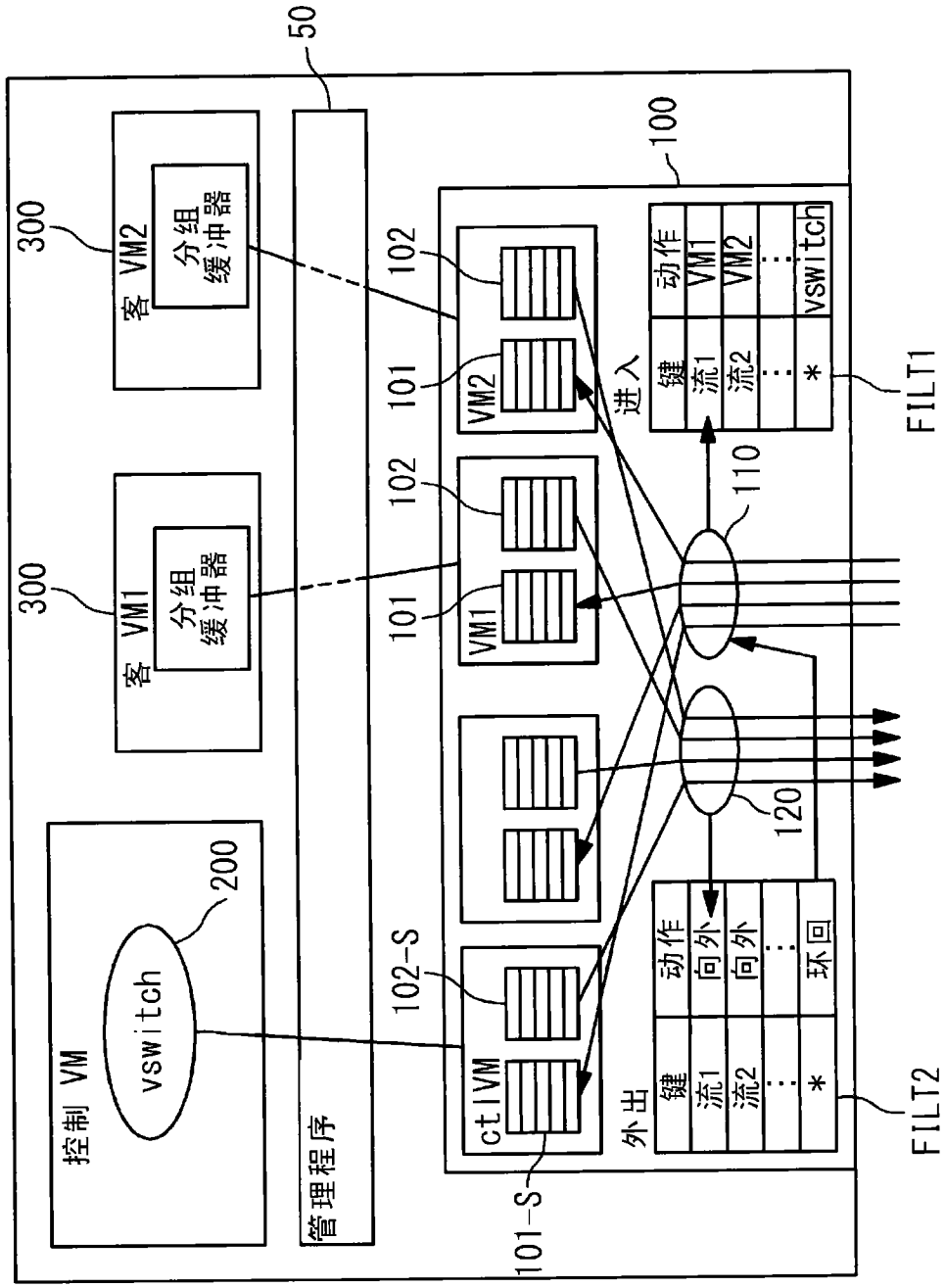

50
100
300
客 VM2
分组缓冲器
300
客 VM1
分组缓冲器
控制 VM
vswitch
200
管理程序
102
101
VM2
102
101
VM1
102-S
ctlVM
101-S
进入
键
流1
流2
*
动作
VM1
VM2
vswitch
FILT1
110
120
外出
键
流1
流2
*
动作
向外
向外
环回
FILT2

图 9

FILT2:接收过滤器表格

键	动作
192.168.0.1:1025→192.168.10.5:80	向外
192.168.1.2:1025→192.168.11.3:22	向外
192.168.1.2:1026→192.168.11.3:22	向外
⋮	⋮
*	环回

图 10

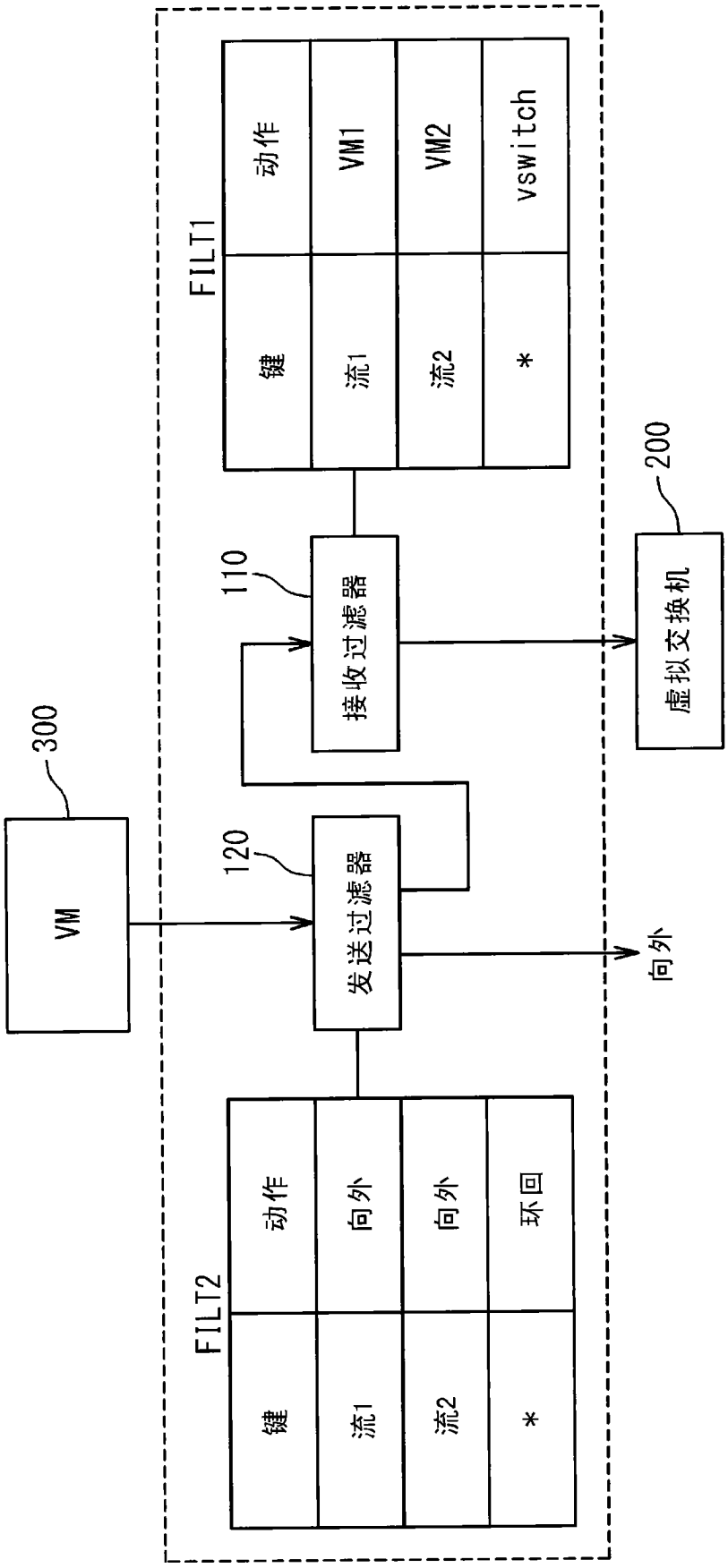

FILT1
键
流1
流2
*
动作
VM1
VM2
vswitch
110
接收过滤器
200
虚拟交换机
300
VM
120
发送过滤器
向外
FILT2
键
流1
流2
*
动作
向外
向外
环回

图 11

键	动作
192.168.0.1:1025→192.168.10.5:80	向外
192.168.10.5:80→192.168.0.1:1025	VM1
192.168.1.2:1025→192.168.11.3:22	向外
192.168.11.3:22→192.168.1.2:1025	VM2
192.168.1.2:1026→192.168.11.3:22	向外
192.168.11.3:22→192.168.1.2:1026	VM2
⋮	⋮
*	vswitch

图 12

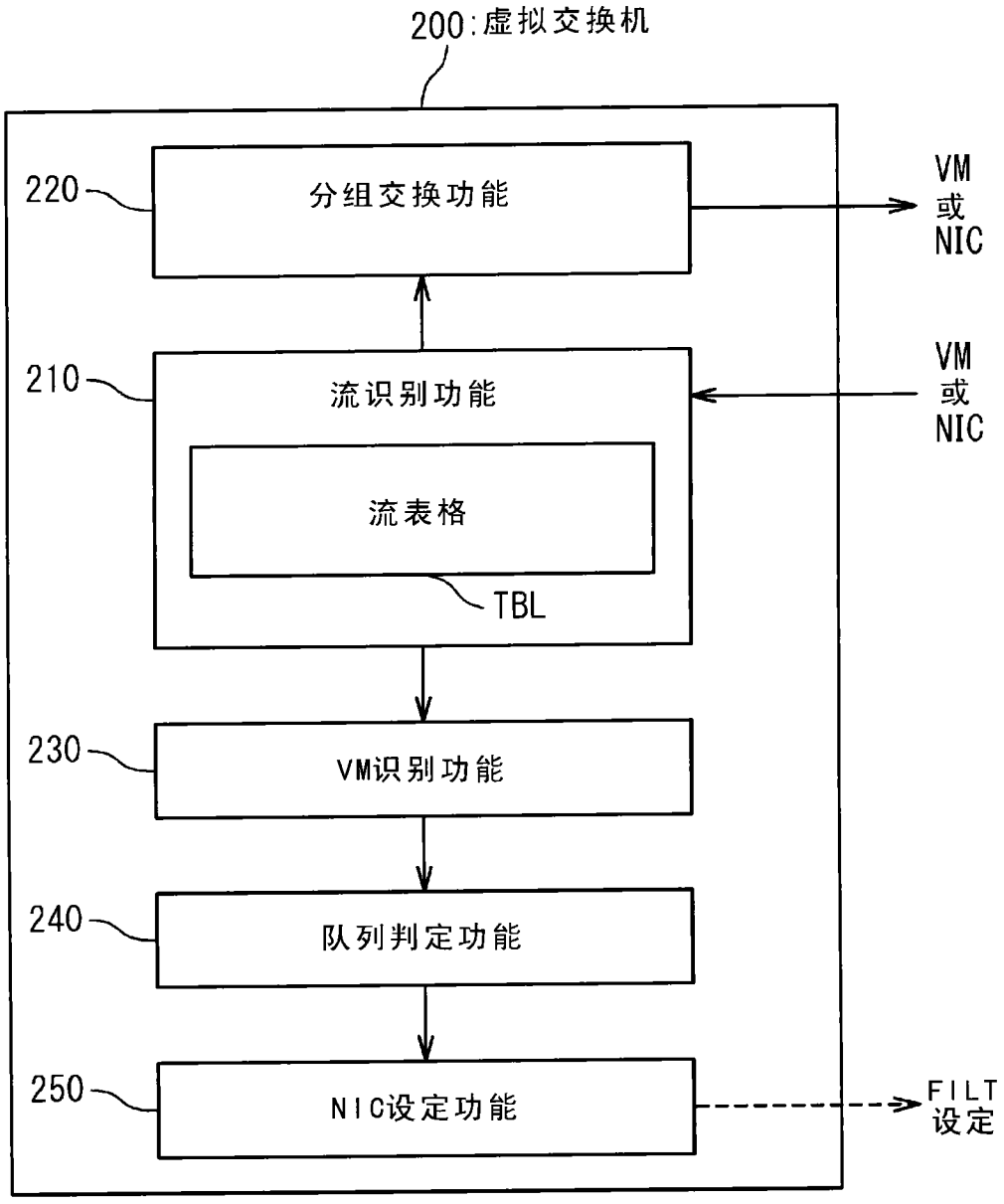

200:虚拟交换机
220
分组交换功能
VM
或
NIC
210
流识别功能
流表格
TBL
VM
或
NIC
230
VM识别功能
240
队列判定功能
250
NIC设定功能
FILT
设定

图 13

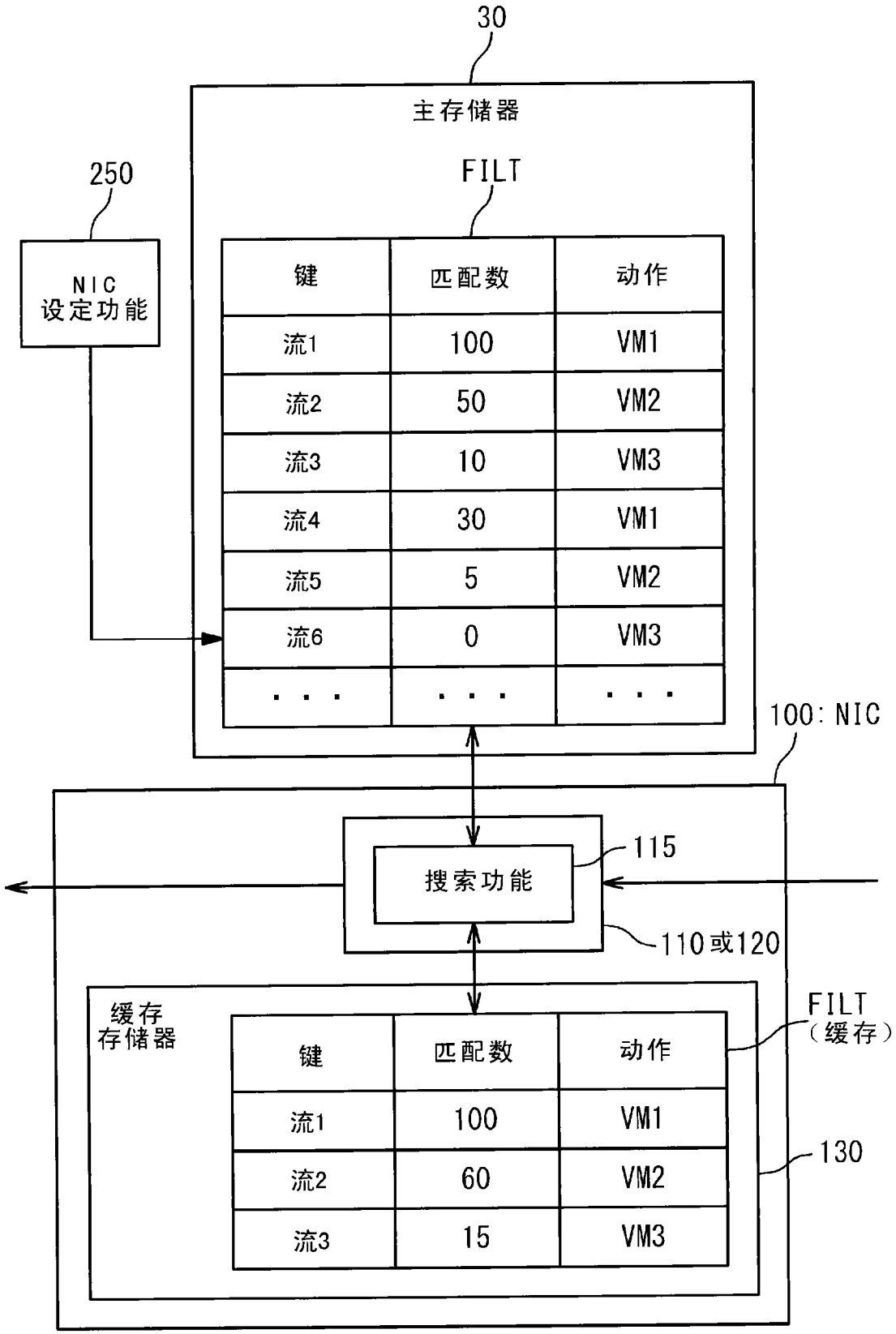

30
主存储器
250
NIC
设定功能
FILT
键
匹配数
动作
流1
100
VM1
流2
50
VM2
流3
10
VM3
流4
30
VM1
流5
5
VM2
流6
0
VM3
100: NIC
搜索功能
115
110或120
缓存
存储器
FILT
（缓存）
键
匹配数
动作
流1
100
VM1
流2
60
VM2
流3
15
VM3
130

图 14

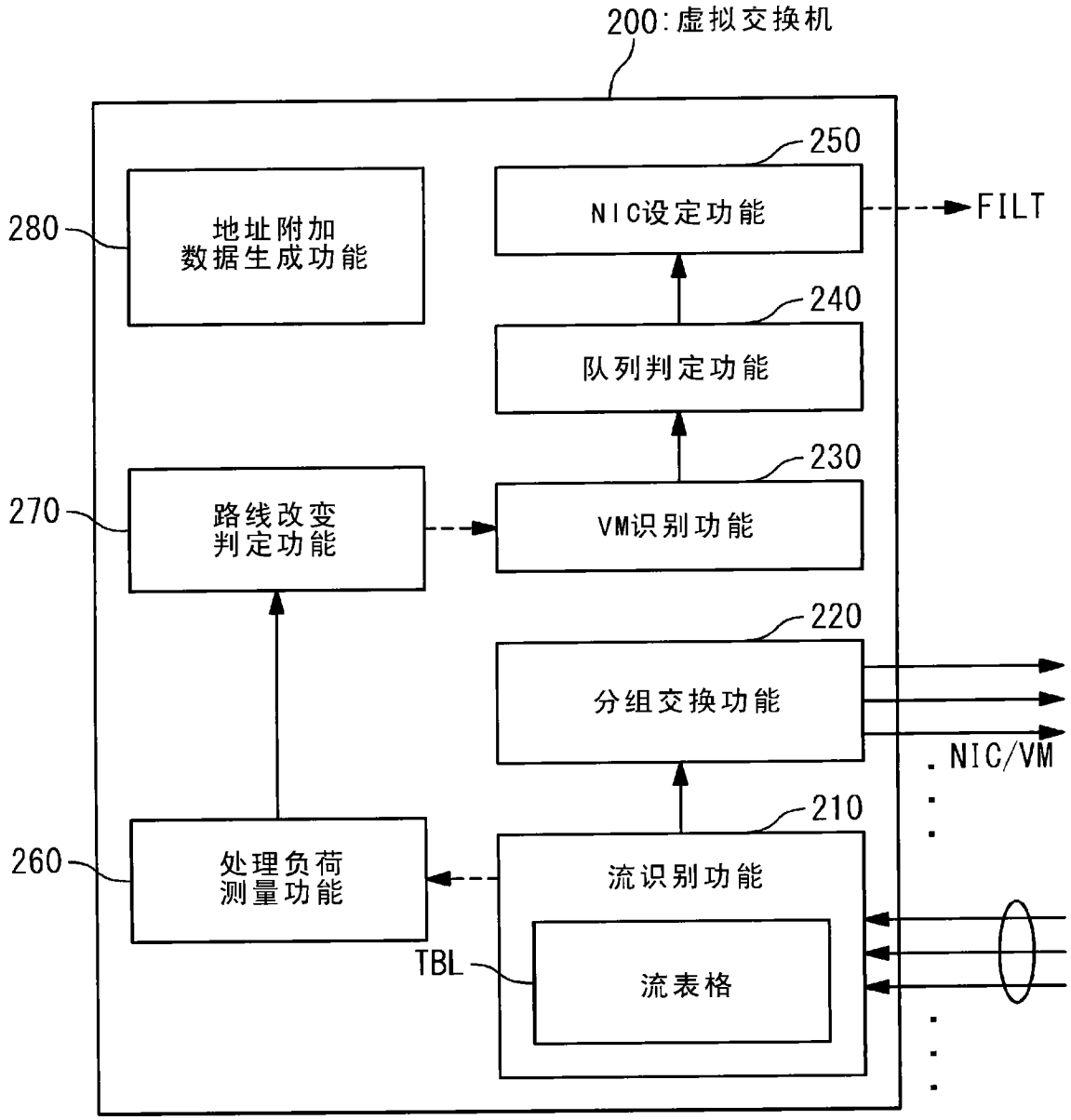

200:虚拟交换机
250
NIC设定功能
FILT
280
地址附加
数据生成功能
240
队列判定功能
230
270
路线改变
判定功能
VM识别功能
220
分组交换功能
NIC/VM
210
260
处理负荷
测量功能
流识别功能
TBL
流表格

图 15

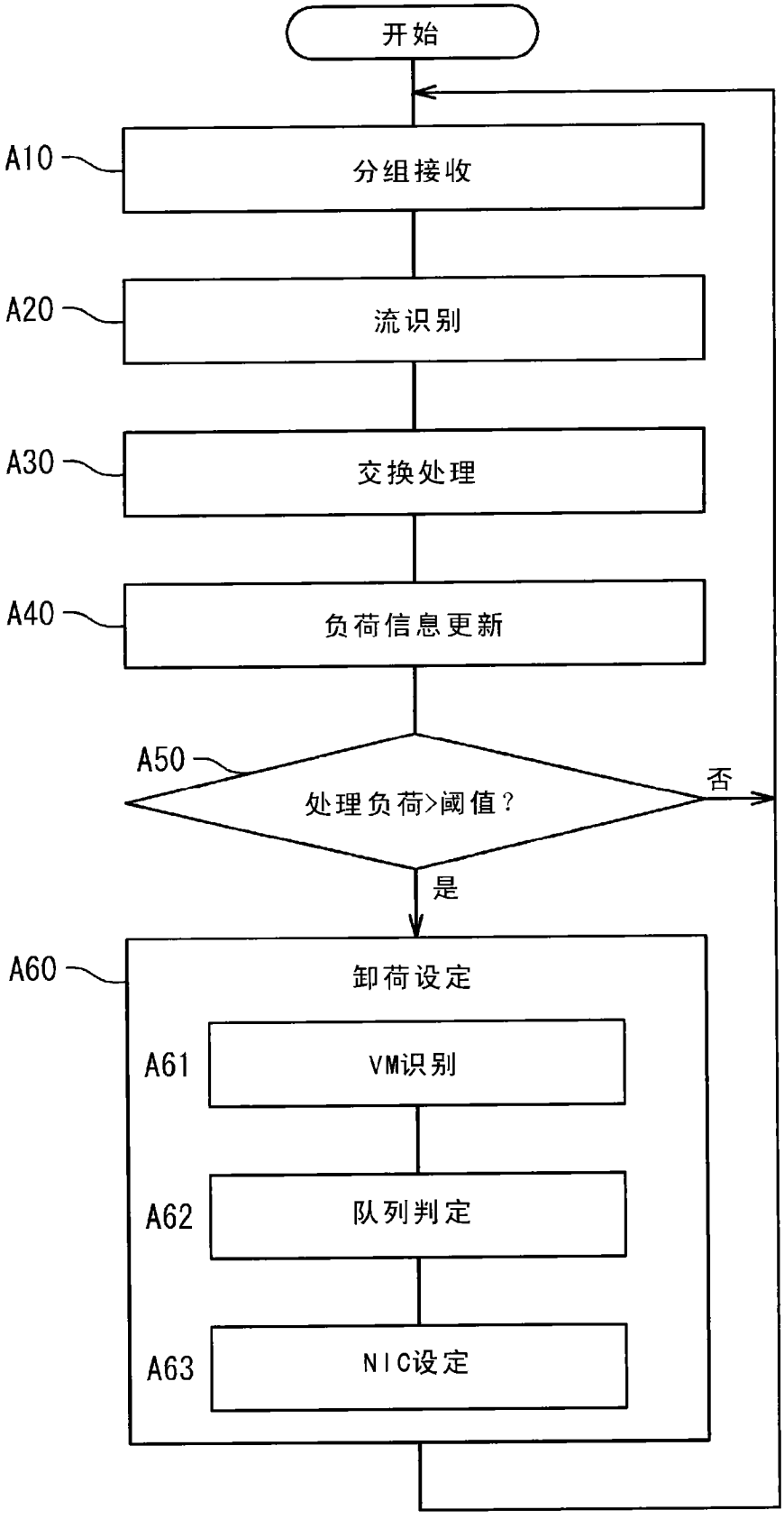

开始
A10
分组接收
A20
流识别
A30
交换处理
A40
负荷信息更新
A50
处理负荷>阈值？
否
是
A60
卸荷设定
A61
VM识别
A62
队列判定
A63
NIC设定

图 16

TBL：流表格

键	动作	NIC上的条目
192.168.0.1:1025→192.168.10.5:80	端口1	存在
192.168.10.5:80→192.168.0.1:1025	端口0	存在
192.168.1.2:1025→192.168.11.3:22	端口2	存在
192.168.11.3:22→192.168.1.2:1025	端口0	存在
192.168.1.2:1026→192.168.11.3:22	端口2	不存在
192.168.11.3:22→192.168.1.2:1026	端口0	不存在
0/0→192.168.11.1	丢弃	不存在
⋮	⋮	⋮

图 17

端口0	外部
端口1	VM1
端口2	VM2
⋮	⋮

图 18

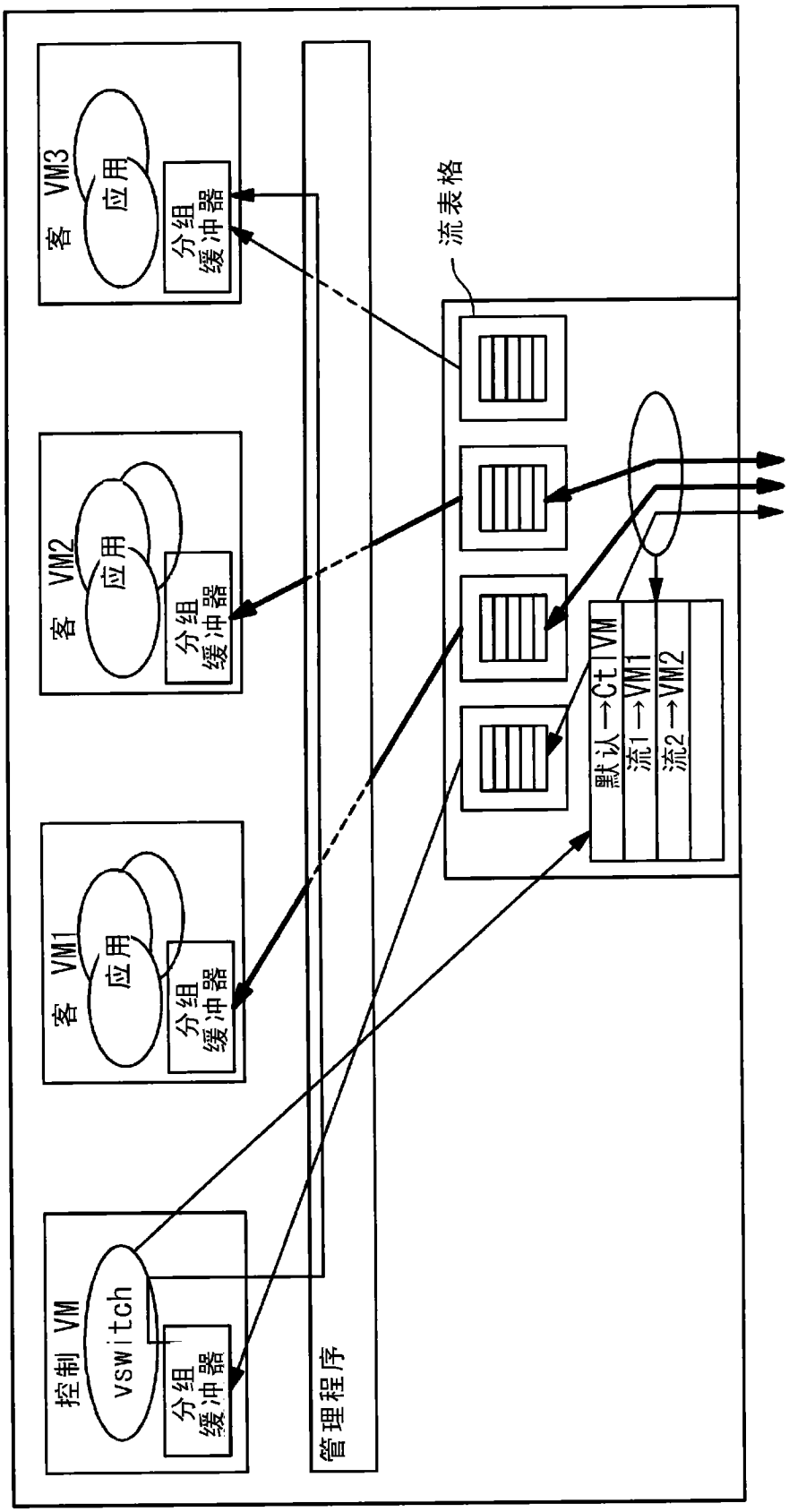

客 VM3
应用
分组缓冲器
客 VM2
应用
分组缓冲器
客 VM1
应用
分组缓冲器
控制 VM
vswitch
分组缓冲器
管理程序
流表格
默认→Ctl VM
流1→VM1
流2→VM2

图 19

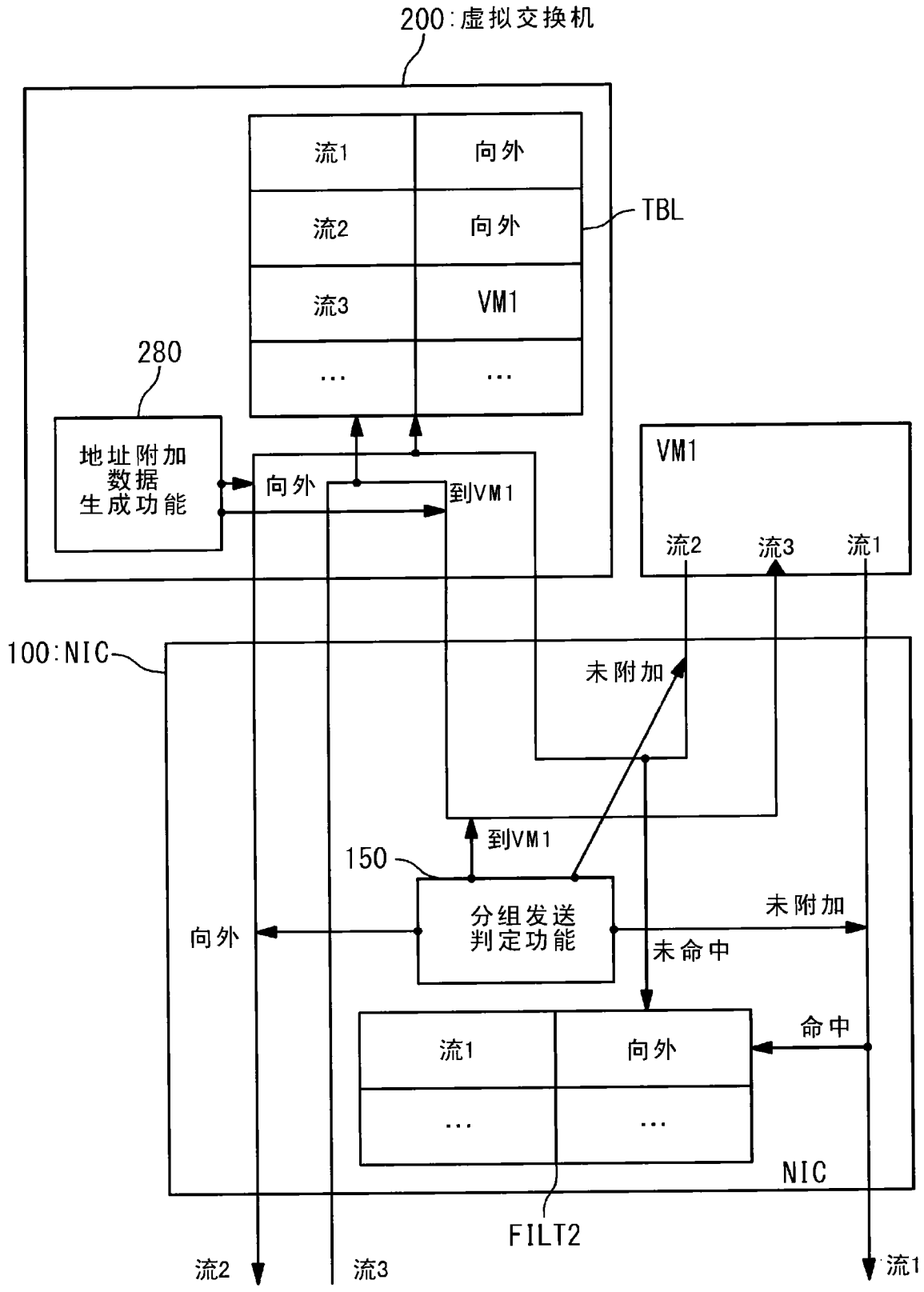

200:虚拟交换机
流1
向外
流2
向外
流3
VM1
...
...
TBL
280
地址附加数据生成功能
向外
到VM1
VM1
流2
流3
流1
100:NIC
未附加
到VM1
150
分组发送判定功能
向外
未附加
未命中
命中
流1
向外
...
...
NIC
FILT2
流2
流3
流1

图 20

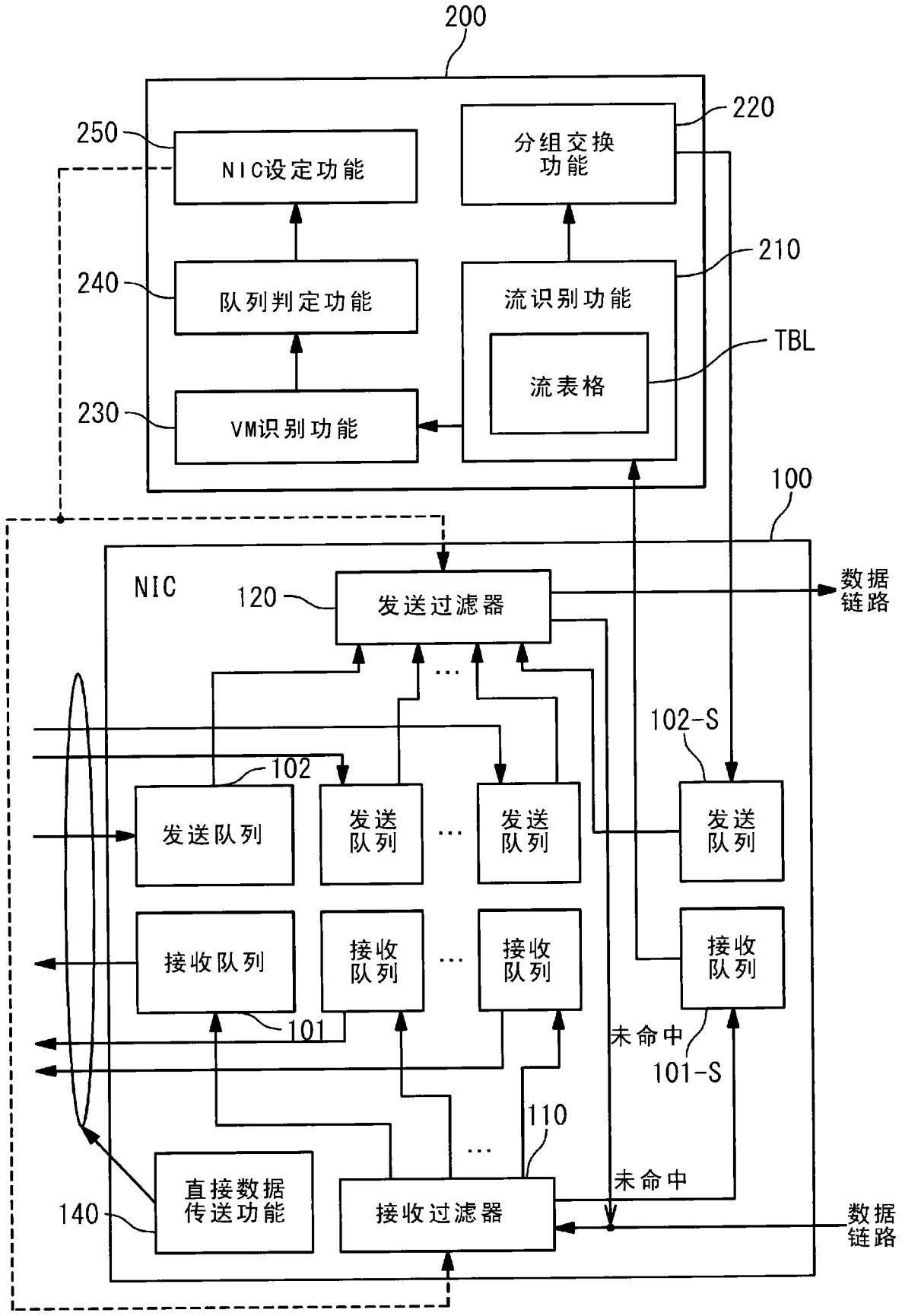

200
250
NIC设定功能
220
分组交换功能
240
队列判定功能
210
流识别功能
TBL
流表格
230
VM识别功能
100
NIC
120
发送过滤器
数据链路
102-S
102
发送队列
发送队列
发送队列
发送队列
接收队列
接收队列
接收队列
接收队列
101
未命中
101-S
110
未命中
直接数据传送功能
140
接收过滤器
数据链路

图 21

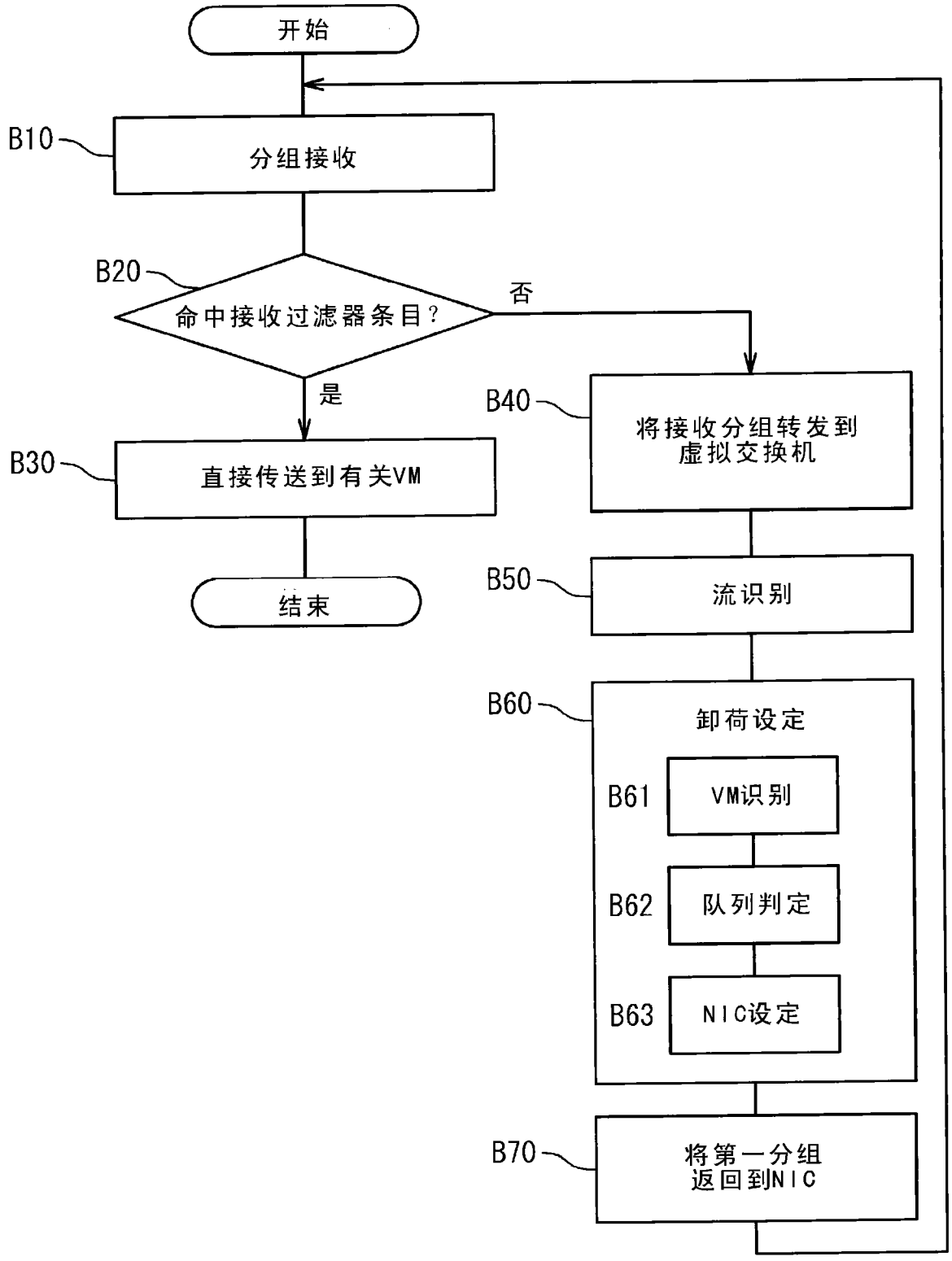

开始
B10
分组接收
B20
命中接收过滤器条目？
否
是
B30
直接传送到有关VM
结束
B40
将接收分组转发到
虚拟交换机
B50
流识别
B60
卸荷设定
B61
VM识别
B62
队列判定
B63
NIC设定
B70
将第一分组
返回到NIC

图 22

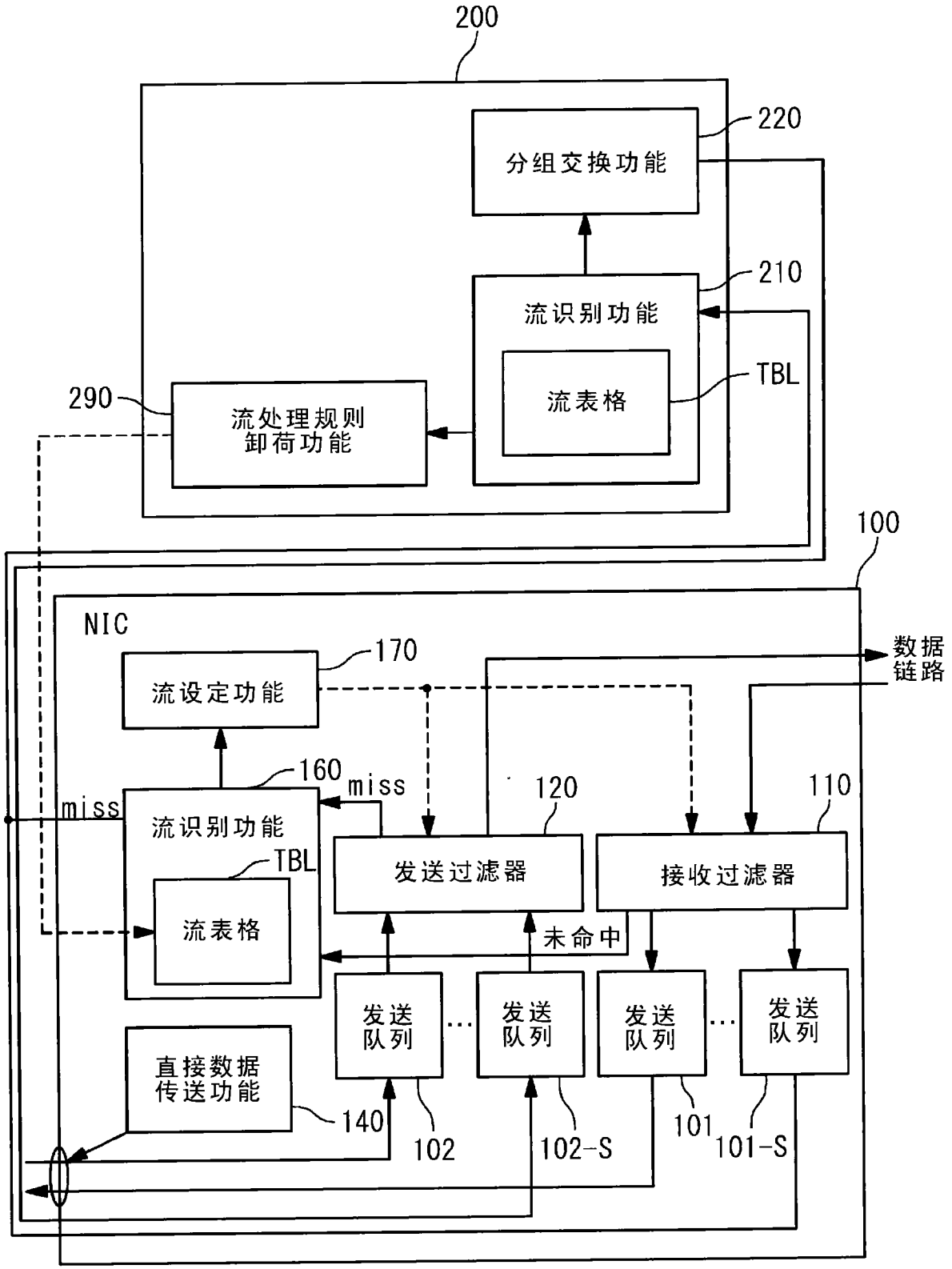

200
220
分组交换功能
210
流识别功能
TBL
流表格
290
流处理规则
卸荷功能
100
NIC
170
流设定功能
数据
链路
160
miss
miss
流识别功能
TBL
流表格
120
110
发送过滤器
接收过滤器
未命中
发送
队列
发送
队列
发送
队列
发送
队列
直接数据
传送功能
140
102
102-S
101
101-S

图 23

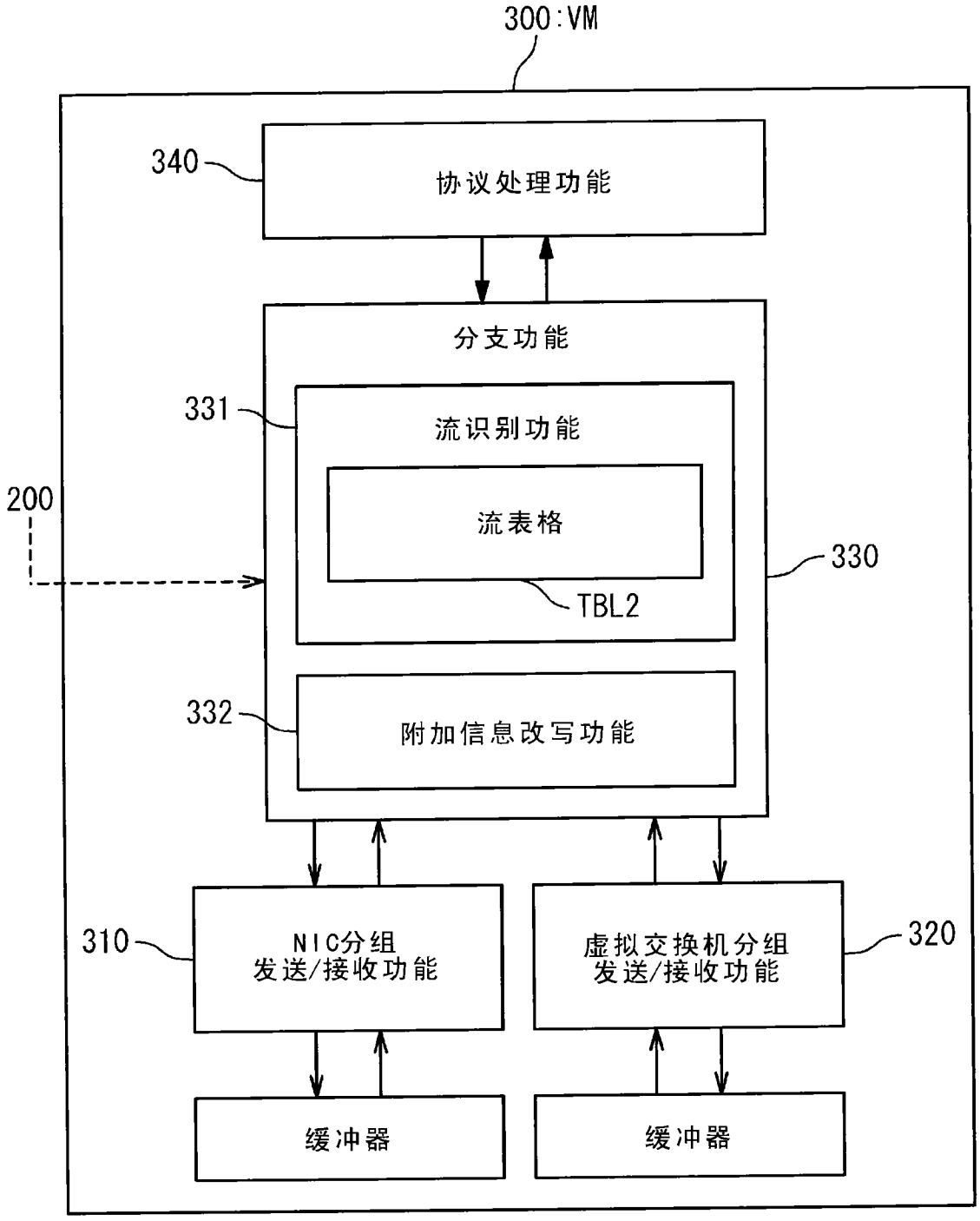

300:VM
340
协议处理功能
分支功能
331
流识别功能
流表格
TBL2
330
200
332
附加信息改写功能
310
NIC分组
发送/接收功能
虚拟交换机分组
发送/接收功能
320
缓冲器
缓冲器

图 24

TBL2：流表格

键	动作
流1	NIC
流2	NIC
流3	vswitch
. . .	. . .

图 25

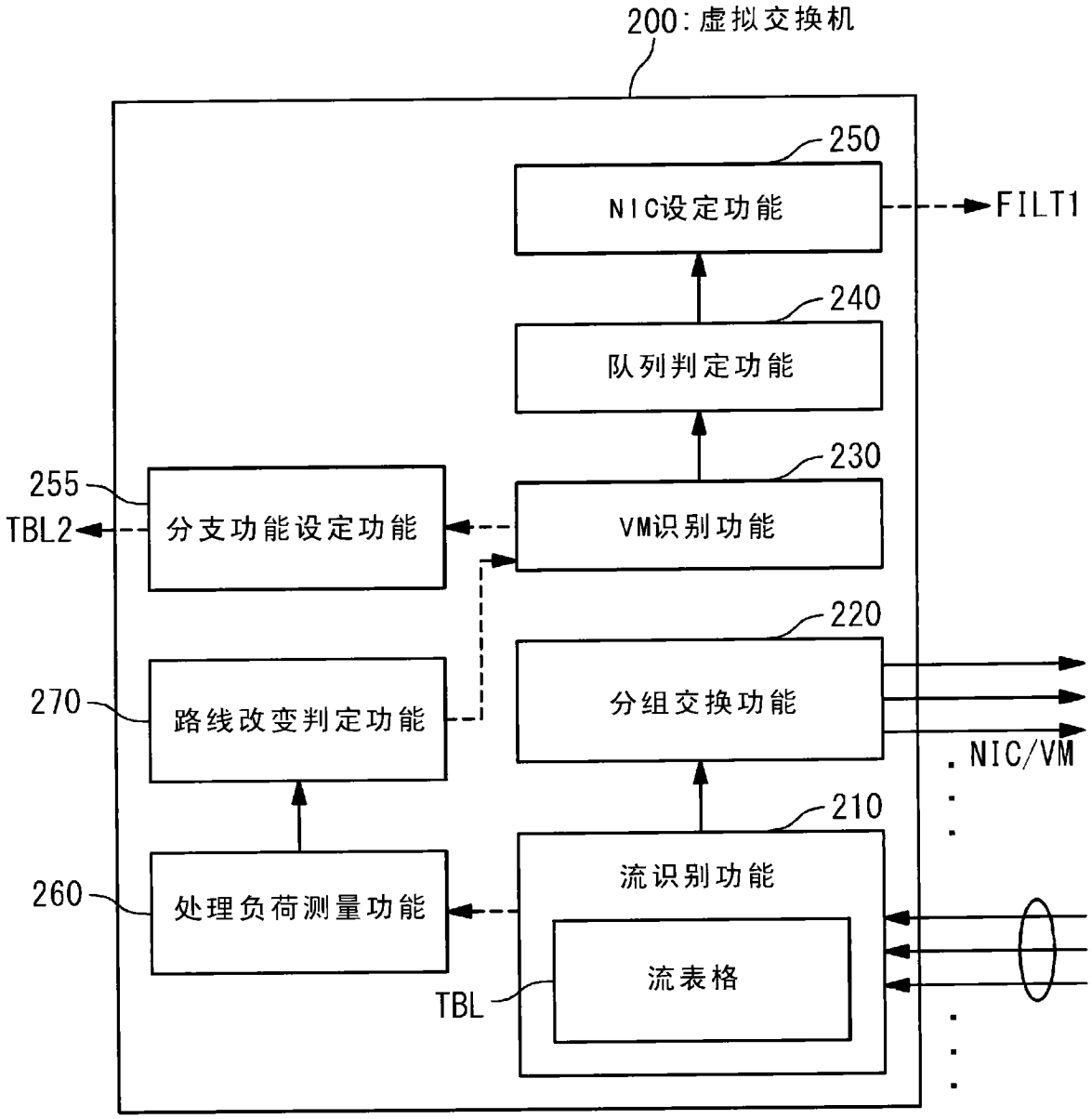

200:虚拟交换机
250
NIC设定功能
FILT1
240
队列判定功能
230
VM识别功能
255
TBL2
分支功能设定功能
220
分组交换功能
270
路线改变判定功能
NIC/VM
210
流识别功能
260
处理负荷测量功能
流表格
TBL

图 26