(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 103384551 B
(45)授权公告日 2015.12.09

(21)申请号 201380000042.9
(22)申请日 2013.01.04
(85)PCT国际申请进入国家阶段日
2013.04.01
(86)PCT国际申请的申请数据
PCT/CN2013/070016 2013.01.04
(87)PCT国际申请的公布数据
WO2014/106321 ZH 2014.07.10
(73)专利权人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼
(72)发明人 李思聪 邱军
(51)Int.Cl.
H04L 29/06(2006.01)
H04L 29/12(2006.01)

(56)对比文件
CN 101299757 A,2008.11.05, 全文.
US 2011/0202701 A1,2011.08.18, 说明书[0005]-[0013]、[0027]-[0067],附图3、6-13.
Jack Regula.Using Non-transparent Bridging in PCI Express Systems.《PLX Technology, Inc.》.2004, 全文.

审查员 黄苏一

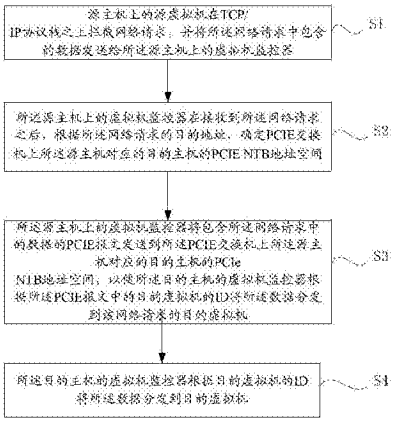
权利要求书6页 说明书17页 附图6页

(54)发明名称

一种基于PCIE网络的虚拟机通信方法、服务器及系统

(57)摘要

本发明提供一种基于PCIE网络的虚拟机通信系统及方法，该系统包括：源主机，源主机包括虚拟机监控器和至少一个虚拟机；源主机上的虚拟机监控器与源主机上的虚拟机通信相连；其中，源主机上的一个虚拟机作为源虚拟机将在TCP/IP协议栈之上拦截的网络请求进行处理并发送给源主机上的虚拟机监控器；源主机上的虚拟机监控器根据网络请求的目的虚拟机的ID，将网络请求中包含的数据发送到目的虚拟机所属的目的主机对应的PCIENTB地址空间中对应于源主机的PCIE NTB地址空间。本发明将PCIE网络作为虚拟机间的通信机制，实现了虚拟机间的快速通信；且还突破PCIE网络的限制，基于PCIE网络的子网管理共享子网管理服务器中的PCIE NTB地址空间，提高了虚拟机间的通信速度。



CN 103384551 B

1. 一种基于高速外围组件互连 PCIE 网络的虚拟机通信方法，其特征在于，包括：

源主机上的源虚拟机在 TCP/IP 协议栈之上拦截网络请求，并将所述网络请求中包含的数据发送给所述源主机上的虚拟机监控器；

所述源主机上的虚拟机监控器根据所述网络请求的目的地址，确定 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间，其中，所述目的主机为所述网络请求的目的虚拟机所属的主机，所述 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间为所述 PCIE 交换机上分配给所述目的主机的 PCIE NTB 地址子空间中对应于所述源主机的 PCIE NTB 地址微空间，其中，PCIE NTB 地址空间分成与 PCIE 交换机连接的主机个数相同份数的 PCIE NTB 地址子空间，在每个主机对应的地址子空间中为每个其他主机分配一个地址微空间，每个所述地址微空间对应一个其他主机；

所述源主机上的虚拟机监控器将包含所述数据的 PCIE 报文发送到所述 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间，以便所述目的主机的虚拟机监控器根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机。

2. 根据权利要求 1 所述的方法，其特征在于，所述源主机上的虚拟机监控器根据所述网络请求的目的地址，确定 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间，具体为：

所述源主机上的虚拟机监控器根据所述网络请求的目的地址，确定所述网络请求的目的主机，并根据本地缓存的所述源主机对应的所述目的主机的 PCIE NTB 地址空间信息，确定所述 PCIE 交换机上所述源主机对应的所述目的主机的 PCIE NTB 地址空间；

其中，所述源主机对应的所述目的主机的 PCIE NTB 地址空间信息为所述源主机上的虚拟机监控器在所述源主机上的源虚拟机发送地址解析协议请求进行地址解析时缓存的。

3. 根据权利要求 2 所述的方法，其特征在于，所述源主机上的源虚拟机发送地址解析协议请求进行地址解析时缓存，包括：

所述源虚拟机监控器截获到所述源虚拟机向所述网络请求的目的虚拟机发送的所述地址解析协议请求后，发起查询所述源虚拟机对应的所述目的主机的 PCIE NTB 地址空间信息的请求消息到子网管理服务器；

所述源主机上的虚拟机监控器接收所述子网管理服务器返回的响应消息，将所述响应消息中包含的所述源虚拟机对应的所述目的主机的 PCIE NTB 地址空间信息缓存在本地，其中，所述源虚拟机对应的所述目的主机的 PCIE NTB 地址空间与所述源主机对应的所述目的主机的 PCIE NTB 地址空间相同。

4. 根据权利要求 1 至 3 任意一项所述的方法，其特征在于，所述源主机上的虚拟机监控器将包含所述网络请求中的数据的 PCIE 报文发送到所述 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间，包括：

所述源主机上的虚拟机监控器将所述网络请求包含的数据、所述源虚拟机的 ID 和源 NTB 端口 ID 放入所述 PCIE 报文的数据荷载，将所述网络请求的目的虚拟机的 ID 作为所述 PCIE 报文的目的地址，发送所述 PCIE 报文给所述 PCIE 交换机，使所述 PCIE 交换机将所述 PCIE 报文写入目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间；

其中，所述源 NTB 端口 ID 为所述 PCIE 交换机中与所述源主机上的虚拟机监控器相连

的 NTB 端口的 ID。

5. 根据权利要求 3 所述的方法，其特征在于，还包括：

全部 PCIE NTB 地址空间由所述子网管理服务器进行维护管理，所述全部 PCIE NTB 地址空间分成与所述 PCIE 交换机连接的主机个数相同的份数的 PCIE NTB 地址子空间；

每个所述地址子空间对应一个主机，用于接收发送给所述对应主机的数据；

在每个所述地址子空间中为每个其他主机分配一个地址微空间，每个所述地址微空间对应一个所述其他主机，用于接收来自于所述其他主机发送给所述地址子空间对应主机的数据，其中，所述其他主机是除所述地址子空间对应主机之外，与所述 PCIE 交换机连接的主机。

6. 根据权利要求 5 所述的方法，其特征在于，还包括：

所述子网管理服务器在所述 PCIE 网络初始化时分配所述全部 PCIE NTB 地址空间，并将自身的 PCIE NTB 地址空间信息通知所有主机，各主机缓存所对应的所述子网管理服务器的 PCIE NTB 地址空间信息，以便能向所述子网管理服务器发送所述查询目的主机 PCIE NTB 地址空间的请求消息。

7. 根据权利要求 3 所述的方法，其特征在于，还包括：

所述子网管理服务器上维护有主机与虚拟机的对应关系，并根据虚拟机的创建、更改、销毁实时更新所述对应关系。

8. 根据权利要求 1 所述的方法，其特征在于，所述目的主机的虚拟机监控器根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机，包括：

所述目的主机的虚拟机监控器读取所述目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间中的所述 PCIE 报文，根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机。

9. 一种基于高速外围组件互连网络的虚拟机通信系统，包括：至少两台主机、一个 PCIE 交换机，所述至少两台主机通过所述 PCIE 交换机通信相连；

所述至少两台主机中的每个主机，用于运行虚拟机和虚拟机监控模块，其中，所述至少两台主机中的源主机上的源虚拟机在 TCP/IP 协议栈之上拦截网络请求，并将所述网络请求中包含的数据发送给所述源主机的虚拟机监控器，所述源主机的虚拟机监控器根据所述源虚拟机发送的所述网络请求的目的地址，确定所述 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间，将包含所述网络请求中的数据的 PCIE 报文发送到所述 PCIE 交换机上所述源主机对应的所述目的主机的 PCIE NTB 地址空间，以便所述目的主机的虚拟机监控器根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机；

其中，所述目的主机为所述网络请求的目的虚拟机所属的主机，所述 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间为所述 PCIE 交换机上分配给所述目的主机的 PCIE NTB 子地址空间中对应于所述源主机的 PCIE NTB 地址微空间，其中，PCIE NTB 地址空间分成与 PCIE 交换机连接的主机个数相同份数的 PCIE NTB 地址子空间，在每个主机对应的地址子空间中为每个其他主机分配一个地址微空间，每个所述地址微空间对应一个其他主机。

10. 根据权利要求 9 所述的系统，其特征在于，所述每个主机包括：

虚拟机运行模块，用于运行所述虚拟机；

虚拟机监控模块，用于运行所述虚拟机监控器。

11. 根据权利要求 9 所述的系统，其特征在于，所述每个主机包括：至少一个处理器，存储器；

所述存储器，用于存储执行指令；

所述处理器，用于当所述主机运行时，与所述存储器之间通信，执行所述执行指令使得所述主机运行所述虚拟机和所述虚拟机监控器。

12. 根据权利要求 9 至 11 任一所述的系统，其特征在于，所述源主机的虚拟机监控器根据所述源虚拟机发送的所述网络请求的目的地址，确定所述 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间，具体为：

所述源主机上的虚拟机监控器根据源虚拟机发送的所述网络请求的目的地址，确定所述网络请求的目的主机，并根据本地缓存的所述源主机对应的所述目的主机的 PCIE NTB 地址空间信息，确定所述 PCIE 交换机上所述源主机对应的所述目的主机的 PCIE NTB 地址空间；

其中，所述源主机对应的所述目的主机的 PCIE NTB 地址空间信息为所述源主机上的虚拟机监控器在所述源主机上的源虚拟机发送地址解析协议请求进行地址解析时缓存的。

13. 根据权利要求 12 所述的系统，其特征在于，还包括子网管理服务器；

则，所述源主机上的虚拟机监控器在所述源主机上的源虚拟机发送地址解析协议请求进行地址解析时缓存，具体为：当所述源虚拟机上没有所述网络请求的目的虚拟机的 MAC 地址信息时，所述源虚拟机向所述网络请求的目的虚拟机发送所述地址解析协议请求，所述源主机上的虚拟机监控器截获到所述虚拟机发送的所述地址解析协议请求后，向所述子网管理服务器发起查询所述目的虚拟机对应的 PCIE NTB 地址空间的请求信息，并在接收到所述子网管理服务器返回的响应消息后，将所述响应消息中包含的所述源虚拟机对应的目的主机的 PCIE NTB 地址空间信息缓存在本地，其中，所述源虚拟机对应的所述目的主机的 PCIE NTB 地址空间与所述源主机对应的所述目的主机的 PCIE NTB 地址空间相同；

所述子网管理服务器，用于接收到所述源主机的虚拟机监控器发送的查询所述目的虚拟机对应的 PCIE NTB 地址空间的请求消息后，根据主机与虚拟机的对应关系确定所述目的虚拟机所属的目的主机，获取所述源主机对应的所述目的主机的 PCIE NTB 地址空间，发送所述响应消息给所述源主机的虚拟机监控器。

14. 根据权利要求 9 所述的系统，其特征在于，所述源主机上的虚拟机监控器将包含所述网络请求中的数据的 PCIE 报文发送到所述 PCIE 交换机上所述源主机对应的所述目的主机的 PCIE NTB 地址空间，包括：

所述源主机上的虚拟机监控器将所述网络请求包含的数据、所述源虚拟机的 ID 和源 NTB 端口 ID 放入所述 PCIE 报文的数据荷载，将所述网络请求的目的虚拟机的 ID 作为所述 PCIE 报文的目的地址，发送所述 PCIE 报文给所述 PCIE 交换机，使所述 PCIE 交换机将所述 PCIE 报文写入目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间，其中，所述源 NTB 端口 ID 为所述 PCIE 交换机中与所述源主机上的虚拟机监控器相连的 NTB 端口的 ID；

则，所述 PCIE 交换机，还用于将所述写入所述目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间。

15. 根据权利要求 13 所述的系统，其特征在于：

所述子网管理服务器，还用于维护管理全部 PCIE NTB 地址空间，所述全部 PCIE NTB 地址空间分成与所述 PCIE 交换机连接的主机个数相同的份数的 PCIE NTB 地址子空间，每个所述地址子空间对应一个主机，用于接收发送给所述对应主机的数据，在每个所述地址子空间中为每个其他主机分配一个地址微空间，每个所述地址微空间对应一个所述其他主机，用于接收来自于所述其他主机发送给所述地址子空间对应主机的数据，其中，所述其他主机是除所述地址子空间对应主机之外，与所述 PCIE 交换机连接的主机。

16. 根据权利要求 15 所述的系统，其特征在于：

所述子网管理服务器，还用于在所述 PCIE 网络初始化时分配所述全部 PCIE NTB 地址空间，并将自身的 PCIE NTB 地址空间信息通知所有主机，以便各主机缓存所对应的所述子网管理服务器的 PCIE NTB 地址空间信息从而能向所述子网管理服务器发送所述查询目的主机 PCIE NTB 地址空间的请求消息。

17. 根据权利要求 13 所述的系统，其特征在于：

所述子网管理服务器，还用于维护主机与虚拟机的对应关系，并根据虚拟机的创建、更改、销毁实时更新所述对应关系。

18. 根据权利要求 9 所述的系统，其特征在于，所述目的主机的虚拟机监控器根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机，包括：

所述目的主机的虚拟机监控器读取所述目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间中的所述 PCIE 报文，根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机。

19. 一种服务器，包括：

虚拟机运行模块，用于运行虚拟机；

虚拟机监控模块，用于运行述虚拟机监控器；

所述虚拟机，用于在 TCP/IP 协议栈之上拦截网络请求，并将所述网络请求中包含的数据发送给虚拟机监控器；

所述虚拟机监控器，用于根据所述虚拟机发送的所述网络请求的目的地址，确定 PCIE 交换机上所述服务器对应的目的主机的 PCIE NTB 地址空间，将包含所述网络请求中的数据的 PCIE 报文发送到所述 PCIE 交换机上所述服务器对应的所述目的主机的 PCIE NTB 地址空间，以便所述目的主机的虚拟机监控器根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机；

其中，所述目的主机为所述网络请求的目的虚拟机所属的主机，所述 PCIE 交换机上所述服务器对应的目的主机的 PCIE NTB 地址空间为所述 PCIE 交换机上分配给所述目的主机的 PCIE NTB 地址子空间中对应于所述服务器的 PCIE NTB 地址微空间，其中，PCIE NTB 地址空间分成与 PCIE 交换机连接的主机个数相同份数的 PCIE NTB 地址子空间，在每个主机对应的地址子空间中为每个其他主机分配一个地址微空间，每个所述地址微空间对应一个其他主机。

20. 根据权利要求 19 所述的服务器，其特征在于，所述虚拟机监控器用于根据所述虚拟机发送的所述网络请求的目的地址，确定 PCIE 交换机上所述服务器对应的目的主机的 PCIE NTB 地址空间，具体为：

所述虚拟机监控器，用于根据所述虚拟机发送的所述网络请求的目的地址，确定所述网络请求的目的主机，并根据本地缓存的所述服务器对应的所述目的主机的PCIE NTB地址空间信息，确定所述PCIE交换机上所述服务器对应的所述目的主机的PCIE NTB地址空间；

其中，所述服务器对应的所述目的主机的PCIE NTB地址空间信息为所述虚拟机监控器在所述虚拟机发送地址解析协议请求进行地址解析时缓存的。

21. 根据权利要求 20 所述的服务器，其特征在于，所述虚拟机监控器在所述虚拟机发送地址解析协议请求进行地址解析时缓存，具体为：

所述虚拟机，还用于，当所述虚拟机上没有所述网络请求的目的虚拟机的 MAC 地址信息时，向所述网络请求的目的虚拟机发送所述地址解析协议请求；

所述虚拟机监控器，还用于截获到所述虚拟机发送的所述地址解析协议请求后，向子网管理服务器发起查询所述目的虚拟机对应的PCIE NTB地址空间的请求信息，并在接收到所述子网管理服务器返回的响应消息后，将所述响应消息中包含的所述虚拟机对应的目的主机的 PCIE NTB 地址空间信息缓存在本地，其中，所述虚拟机对应的所述目的主机的 PCIE NTB 地址空间与所述服务器对应的所述目的主机的 PCIE NTB 地址空间相同。

22. 根据权利要求 19 至 21 任意一项所述的服务器，其特征在于，所述虚拟机监控器将包含所述网络请求中的数据的PCIE报文发送到所述PCIE交换机上所述服务器对应的所述目的主机的 PCIE NTB 地址空间，包括：

所述虚拟机监控器，用于将所述网络请求包含的数据、所述虚拟机的 ID 和源 NTB 端口 ID 放入所述 PCIE 报文的数据荷载，将所述网络请求的目的虚拟机的 ID 作为所述 PCIE 报文的目的地址，发送所述 PCIE 报文给所述 PCIE 交换机，使所述 PCIE 交换机将所述 PCIE 报文写入目的主机的 PCIE NTB 地址空间中对应于所述服务器的 PCIE NTB 地址空间；

其中，所述源 NTB 端口 ID 为所述 PCIE 交换机中与所述虚拟机监控器相连的 NTB 端口的 ID。

23. 根据权利要求 21 所述的服务器，其特征在于：

所述服务器对应的所述目的主机的 PCIE NTB 地址空间由所述子网管理服务器维护管理，用于接收来自于所述服务器发送给所述目的主机的数据；

其中，全部 PCIE NTB 地址空间分成与所述 PCIE 交换机连接的主机个数相同的份数的 PCIE NTB 地址子空间，每个所述地址子空间对应一个主机，用于接收发送给所述对应主机的数据，在每个所述地址子空间中为每个其他主机分配一个地址微空间，每个所述地址微空间对应一个所述其他主机，用于接收来自于所述其他主机发送给所述地址子空间对应主机的数据，所述其他主机是除所述地址子空间对应主机之外，与所述 PCIE 交换机连接的主机。

24. 根据权利要求 23 所述的服务器，其特征在于：

所述虚拟机监控器，还用于接收并缓存所述服务器对应的子网管理服务器的 PCIE NTB 地址空间信息。

25. 根据权利要求 19 所述的服务器，其特征在于：

所述服务器与所述虚拟机的对应关系由子网管理服务器维护，并根据所述虚拟机的创建、更改、销毁实时更新所述对应关系。

26. 根据权利要求 19 所述的服务器，其特征在于，所述目的主机的虚拟机监控器根据所述 PCIE 报文中的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机，包括：

所述目的主机的虚拟机监控器读取所述目的主机的 PCIE NTB 地址空间中对应于所述服务器的 PCIE NTB 地址空间中的所述 PCIE 报文，根据所述 PCIE 报文中的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机。

一种基于 PCIE 网络的虚拟机通信方法、服务器及系统

技术领域

[0001] 本发明属于计算机互联网通信技术领域，涉及一种虚拟机通信方法，特别是涉及一种基于高速外围组件互连 (Peripheral Component Interconnect Express，PCIE) 网络的虚拟机通信方法、服务器及系统。

背景技术

[0002] 随着更多虚拟化服务器的部署，虚拟化技术不断发展，网络虚拟化占据了越来越重要的地位，虚拟机 (Virtual Machine，简称 VM) 间的通信也越来越频繁。

[0003] 目前，虚拟机间通信主要是基于以太网。两个虚拟机之间进行通信时，首先虚拟机中的用户态应用程序会调用 Socket 接口，通过传输控制协议 (Transmission Control Protocol，简称 TCP)/ 因特网互联协议 (Internet Protocol，简称 IP) 堆栈，在虚拟机监控器 (Virtual Machine Monitor，简称 VMM) 中进行实际的转发和处理。而 TCP/IP 协议较复杂，造成操作系统 (Operating System，简称 OS) 的协议处理占据了网络时延的很大部分，此外，在数据传输过程中，频繁的数据复制和中断处理导致主机中央处理器 (Central Processing Unit，简称 CPU) 资源无法专注于处理用户业务。

[0004] 业界也提出了通过共享内存的方式绕过内核中 TCP/IP 协议栈实现服务器内部虚拟机间高速通信的方法，但目前只能应用于同一个物理机内部的虚拟机之间的通信。

发明内容

[0005] 鉴于以上所述现有技术的缺点，本发明的目的在于提供一种基于 PCIE 网络的虚拟机通信方法、服务器及系统，用于解决现有技术中不同物理机上的虚拟机之间通信速度慢的问题。

[0006] 第一方面，本发明实施例提供了一种基于高速外围组件互连 PCIE 网络的虚拟机通信方法，包括：

[0007] 源主机上的源虚拟机在 TCP/IP 协议栈之上拦截网络请求，并将所述网络请求中包含的数据发送给所述源主机上的虚拟机监控器；

[0008] 所述源主机上的虚拟机监控器根据所述网络请求的目的地址，确定 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间，其中，所述目的主机为所述网络请求的目的虚拟机所属的主机，所述 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间为所述 PCIE 交换机上分配给所述目的主机的 PCIE NTB 地址子空间中对应于所述源主机的 PCIE NTB 地址微空间，其中，PCIE NTB 地址空间分成与 PCIE 交换机连接的主机个数相同份数的 PCIE NTB 地址子空间，在每个主机对应的地址子空间中为每个其他主机分配一个地址微空间，每个所述地址微空间对应一个其他主机；

[0009] 所述源主机上的虚拟机监控器将包含所述数据的 PCIE 报文发送到所述 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间，以便所述目的主机的虚拟机监控器根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求

的目的虚拟机。

[0010] 在第一方面的第一种可能的实现方式中，所述源主机上的虚拟机监控器根据所述网络请求的目的地址，确定 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间，具体为：

[0011] 所述源主机上的虚拟机监控器根据所述网络请求的目的地址，确定所述网络请求的目的主机，并根据本地缓存的所述源主机对应的所述目的主机的 PCIE NTB 地址空间信息，确定所述 PCIE 交换机上所述源主机对应的所述目的主机的 PCIE NTB 地址空间；

[0012] 其中，所述源主机对应的所述目的主机的PCIE NTB地址空间信息为所述源主机上的虚拟机监控器在所述源主机上的源虚拟机发送地址解析协议请求进行地址解析时缓存的。

[0013] 结合第一方面的第一种可能的实现方式，在第二种可能的实现方式中，所述源主机上的源虚拟机发送地址解析协议请求进行地址解析时缓存，包括：

[0014] 所述源虚拟机监控器截获到所述源虚拟机向所述网络请求的目的虚拟机发送的所述地址解析协议请求后，发起查询所述源虚拟机对应的所述目的主机的PCIE NTB地址空间信息的请求消息到子网管理服务器；

[0015] 所述源主机上的虚拟机监控器接收所述子网管理服务器返回的响应消息，将所述响应消息中包含的所述源虚拟机对应的所述目的主机的 PCIE NTB 地址空间信息缓存在本地，其中，所述源虚拟机对应的所述目的主机的 PCIE NTB 地址空间与所述源主机对应的所述目的主机的 PCIE NTB 地址空间相同。

[0016] 结合第一方面或第一方面的第一种可能的实现方式或第一方面的第二种可能的实现方式，在第三种可能的实现方式中，所述源主机上的虚拟机监控器将包含所述网络请求中的数据的PCIE报文发送到所述PCIE交换机上所述源主机对应的目的主机的PCIE NTB地址空间，包括：

[0017] 所述源主机上的虚拟机监控器将所述网络请求包含的数据、所述源虚拟机的 ID 和源 NTB 端口 ID 放入所述 PCIE 报文的数据荷载，将所述网络请求的目的虚拟机的 ID 作为所述 PCIE 报文的目的地址，发送所述 PCIE 报文给所述 PCIE 交换机，使所述 PCIE 交换机将所述 PCIE 报文写入目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间；

[0018] 其中，所述源 NTB 端口 ID 为所述 PCIE 交换机中与所述源主机上的虚拟机监控器相连的 NTB 端口的 ID。

[0019] 结合第一方面的第二种可能的实现方式或第一方面的第三种可能的实现方式，在第四种可能的实现方式中，还包括：

[0020] 全部PCIE NTB地址空间由所述子网管理服务器进行维护管理，所述全部PCIE NTB地址空间分成与所述 PCIE 交换机连接的主机个数相同的份数的 PCIE NTB 地址子空间；

[0021] 每个所述地址子空间对应一个主机，用于接收发送给所述对应主机的数据；

[0022] 在每个所述地址子空间中为每个其他主机分配一个地址微空间，每个所述地址微空间对应一个所述其他主机，用于接收来自于所述其他主机发送给所述地址子空间对应主机的数据，其中，所述其他主机是除所述地址子空间对应主机之外，与所述 PCIE 交换机连接的主机。

[0023] 结合第一方面的第四种可能的实现方式，在第五种可能的实现方式中，还包括：

[0024] 所述子网管理服务器在所述 PCIE 网络初始化时分配所述全部 PCIE NTB 地址空间，并将自身的 PCIE NTB 地址空间信息通知所有主机，各主机缓存所对应的所述子网管理服务器的 PCIE NTB 地址空间信息，以便能向所述子网管理服务器发送所述查询目的主机 PCIE NTB 地址空间的请求消息。

[0025] 结合第一方面的第二至第五种任一可能的实现方式，在第六种可能的实现方式中，还包括：

[0026] 所述子网管理服务器上维护有主机与虚拟机的对应关系，并根据虚拟机的创建、更改、销毁实时更新所述对应关系。

[0027] 结合第一方面或第一方面的第一至第六种任一可能的实现方式，在第七种可能的实现方式中，所述目的主机的虚拟机监控器根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机，包括：

[0028] 所述目的主机的虚拟机监控器读取所述目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间中的所述 PCIE 报文，根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机。

[0029] 第二方面，本发明实施例提供了一种基于高速外围组件互连网络的虚拟机通信系统，包括：至少两台主机、一个 PCIE 交换机，所述至少两台主机通过所述 PCIE 交换机通信相连；

[0030] 所述至少两台主机中的每个主机，用于运行虚拟机和虚拟机监控模块，其中，所述至少两台主机中的源主机上的源虚拟机在 TCP/IP 协议栈之上拦截网络请求，并将所述网络请求中包含的数据发送给所述源主机的虚拟机监控器，所述源主机的虚拟机监控器根据所述源虚拟机发送的所述网络请求的目的地址，确定所述 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间，将包含所述网络请求中的数据的 PCIE 报文发送到所述 PCIE 交换机上所述源主机对应的所述目的主机的 PCIE NTB 地址空间，以便所述目的主机的虚拟机监控器根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机；

[0031] 其中，所述目的主机为所述网络请求的目的虚拟机所属的主机，所述 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间为所述 PCIE 交换机上分配给所述目的主机的 PCIE NTB 地址子空间中对应于所述源主机的 PCIE NTB 地址微空间，其中，PCIE NTB 地址空间分成与 PCIE 交换机连接的主机个数相同份数的 PCIE NTB 地址子空间，在每个主机对应的地址子空间中为每个其他主机分配一个地址微空间，每个所述地址微空间对应一个其他主机。

[0032] 在第二方面的第一种可能的实现方式中，所述每个主机包括：

[0033] 虚拟机运行模块，用于运行所述虚拟机；

[0034] 虚拟机监控模块，用于运行所述虚拟机监视器。

[0035] 在第二方面的第二种可能的实现方式中，所述每个主机包括：至少一个处理器，存储器；

[0036] 所述存储器，用于存储执行指令；

[0037] 所述处理器，用于当所述主机运行时，与所述存储器之间通信，执行所述执行指令

使得所述主机运行所述虚拟机和所述虚拟机监控器。

[0038] 结合第二方面或第二方面的第一种或第二方面的第二种可能的实现方式，在第三种可能的实现方式中，所述源主机的虚拟机监控器根据所述源虚拟机发送的所述网络请求的目的地址，确定所述 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间，具体为：

[0039] 所述源主机上的虚拟机监控器根据源虚拟机发送的所述网络请求的目的地址，确定所述网络请求的目的主机，并根据本地缓存的所述源主机对应的所述目的主机的 PCIE NTB 地址空间信息，确定所述 PCIE 交换机上所述源主机对应的所述目的主机的 PCIE NTB 地址空间；

[0040] 其中，所述源主机对应的所述目的主机的 PCIE NTB 地址空间信息为所述源主机上的虚拟机监控器在所述源主机上的源虚拟机发送地址解析协议请求进行地址解析时缓存的。

[0041] 结合第二方面的第三种可能的实现方式，在第四种可能的实现方式中，还包括子网管理服务器；

[0042] 则，所述源主机上的虚拟机监控器在所述源主机上的源虚拟机发送地址解析协议请求进行地址解析时缓存，具体为：当所述源虚拟机上没有所述网络请求的目的虚拟机的 MAC 地址信息时，所述源虚拟机向所述网络请求的目的虚拟机发送所述地址解析协议请求，所述源主机上的虚拟机监控器截获到所述虚拟机发送的所述地址解析协议请求后，向所述子网管理服务器发起查询所述目的虚拟机对应的 PCIE NTB 地址空间的请求信息，并在接收到所述子网管理服务器返回的响应消息后，将所述响应消息中包含的所述源虚拟机对应的目的主机的 PCIE NTB 地址空间信息缓存在本地，其中，所述源虚拟机对应的所述目的主机的 PCIE NTB 地址空间与所述源主机对应的所述目的主机的 PCIE NTB 地址空间相同；

[0043] 所述子网管理服务器，用于接收到所述源主机的虚拟机监控器发送的查询所述目的虚拟机对应的 PCIE NTB 地址空间的请求消息后，根据主机与虚拟机的对应关系确定所述目的虚拟机所属的目的主机，获取所述源主机对应的所述目的主机的 PCIE NTB 地址空间，发送所述响应消息给所述源主机的虚拟机监控器。

[0044] 结合第二方面或第二方面的第一至第四种任一可能的实现方式，在第五种可能的实现方式中，所述源主机上的虚拟机监控器将包含所述网络请求中的数据的 PCIE 报文发送到所述 PCIE 交换机上所述源主机对应的所述目的主机的 PCIE NTB 地址空间，包括：

[0045] 所述源主机上的虚拟机监控器将所述网络请求包含的数据、所述源虚拟机的 ID 和源 NTB 端口 ID 放入所述 PCIE 报文的数据荷载，将所述网络请求的目的虚拟机的 ID 作为所述 PCIE 报文的目的地址，发送所述 PCIE 报文给所述 PCIE 交换机，使所述 PCIE 交换机将所述 PCIE 报文写入目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间，其中，所述源 NTB 端口 ID 为所述 PCIE 交换机中与所述源主机上的虚拟机监控器相连的 NTB 端口的 ID；

[0046] 则，所述 PCIE 交换机，还用于将所述写入所述目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间。

[0047] 结合第二方面的第四种或第五种可能的实现方式，在第六种可能的实现方式中：

[0048] 所述子网管理服务器，还用于维护管理全部 PCIE NTB 地址空间，所述全部 PCIE

NTB地址空间分成与所述PCIE交换机连接的主机个数相同的份数的PCIE NTB地址子空间，每个所述地址子空间对应一个主机，用于接收发送给所述对应主机的数据，在每个所述地址子空间中为每个其他主机分配一个地址微空间，每个所述地址微空间对应一个所述其他主机，用于接收来自于所述其他主机发送给所述地址子空间对应主机的数据，其中，所述其他主机是除所述地址子空间对应主机之外，与所述 PCIE 交换机连接的主机。

[0049] 结合第二方面的第六种可能的实现方式，在第七种可能的实现方式中：

[0050] 所述子网管理服务器，还用于在所述 PCIE 网络初始化时分配所述全部 PCIE NTB 地址空间，并将自身的 PCIE NTB 地址空间信息通知所有主机，以便各主机缓存所对应的所述子网管理服务器的 PCIE NTB 地址空间信息从而能向所述子网管理服务器发送所述查询目的主机 PCIE NTB 地址空间的请求消息。

[0051] 结合第二方面的第四种至第七种任一可能的实现方式，在第八种可能的实现方式中：

[0052] 所述子网管理服务器，还用于维护主机与虚拟机的对应关系，并根据虚拟机的创建、更改、销毁实时更新所述对应关系。

[0053] 结合第二方面或第二方面的第一种至第八种任一可能的实现方式，在第九种可能的实现方式中，所述目的主机的虚拟机监控器根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机，包括：

[0054] 所述目的主机的虚拟机监控器读取所述目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间中的所述 PCIE 报文，根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机。

[0055] 第三方面，本发明实施例提供了一种服务器，包括：

[0056] 虚拟机运行模块，用于运行虚拟机；

[0057] 虚拟机监控模块，用于运行述虚拟机监视器；

[0058] 所述虚拟机，用于在 TCP/IP 协议栈之上拦截网络请求，并将所述网络请求中包含的数据发送给虚拟机监控器；

[0059] 所述虚拟机监控器，用于根据所述虚拟机发送的所述网络请求的目的地址，确定 PCIE 交换机上所述服务器对应的目的主机的 PCIE NTB 地址空间，将包含所述网络请求中的数据的PCIE报文发送到所述PCIE交换机上所述服务器对应的所述目的主机的PCIE NTB 地址空间，以便所述目的主机的虚拟机监控器根据所述 PCIE 报文中的所述网络请求的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机；

[0060] 其中，所述目的主机为所述网络请求的目的虚拟机所属的主机，所述 PCIE 交换机上所述服务器对应的目的主机的PCIE NTB地址空间为所述PCIE交换机上分配给所述目的主机的PCIE NTB地址子空间中对应于所述服务器的PCIE NTB地址微空间，其中，PCIE NTB 地址空间分成与 PCIE 交换机连接的主机个数相同份数的 PCIE NTB 地址子空间，在每个主机对应的地址子空间中为每个其他主机分配一个地址微空间，每个所述地址微空间对应一个其他主机。

[0061] 在第三方面的第一种可能的实现方式中，所述虚拟机监控器用于根据所述虚拟机发送的所述网络请求的目的地址，确定 PCIE 交换机上所述服务器对应的目的主机的 PCIE NTB 地址空间，具体为：

[0062] 所述虚拟机监控器，用于根据所述虚拟机发送的所述网络请求的目的地址，确定所述网络请求的目的主机，并根据本地缓存的所述服务器对应的所述目的主机的 PCIE NTB 地址空间信息，确定所述 PCIE 交换机上所述服务器对应的所述目的主机的 PCIE NTB 地址空间；

[0063] 其中，所述服务器对应的所述目的主机的PCIE NTB地址空间信息为所述虚拟机监控器在所述虚拟机发送地址解析协议请求进行地址解析时缓存的。

[0064] 结合第三方面的第一种可能的实现方式，在第二种可能的实现方式中，所述虚拟机监控器在所述虚拟机发送地址解析协议请求进行地址解析时缓存，具体为：

[0065] 所述虚拟机，还用于，当所述虚拟机上没有所述网络请求的目的虚拟机的 MAC 地址信息时，向所述网络请求的目的虚拟机发送所述地址解析协议请求；

[0066] 所述虚拟机监控器，还用于截获到所述虚拟机发送的所述地址解析协议请求后，向子网管理服务器发起查询所述目的虚拟机对应的PCIE NTB地址空间的请求信息，并在接收到所述子网管理服务器返回的响应消息后，将所述响应消息中包含的所述虚拟机对应的目的主机的 PCIE NTB 地址空间信息缓存在本地，其中，所述虚拟机对应的所述目的主机的 PCIE NTB 地址空间与所述服务器对应的所述目的主机的 PCIE NTB 地址空间相同。

[0067] 结合第三方面或第三方面的第一种可能的实现方式或第三方面的第二种可能的实现方式，在第三种可能的实现方式中，所述源主机上的虚拟机监控器将包含所述网络请求中的数据的PCIE报文发送到所述PCIE交换机上所述服务器对应的所述目的主机的PCIE NTB 地址空间，包括：

[0068] 所述虚拟机监控器，用于将所述网络请求包含的数据、所述虚拟机的 ID 和源 NTB 端口 ID 放入所述PCIE报文的数据荷载，将所述网络请求的目的虚拟机的 ID 作为所述PCIE报文的目的地址，发送所述 PCIE 报文给所述 PCIE 交换机，使所述 PCIE 交换机将所述 PCIE 报文写入目的主机的 PCIE NTB 地址空间中对应于所述服务器的 PCIE NTB 地址空间；

[0069] 其中，所述源 NTB 端口 ID 为所述 PCIE 交换机中与所述虚拟机监控器相连的 NTB 端口的 ID。

[0070] 结合第三方面的第二或第三种可能的实现方式，在第四种可能的实现方式中：

[0071] 所述服务器对应的所述目的主机的 PCIE NTB 地址空间由所述子网管理服务器维护管理，用于接收来自于所述服务器发送给所述目的主机的数据；

[0072] 其中，全部 PCIE NTB 地址空间分成与所述 PCIE 交换机连接的主机个数相同的份数的 PCIE NTB 地址子空间，每个所述地址子空间对应一个主机，用于接收发送给所述对应主机的数据，在每个所述地址子空间中为每个其他主机分配一个地址微空间，每个所述地址微空间对应一个所述其他主机，用于接收来自于所述其他主机发送给所述地址子空间对应主机的数据，所述其他主机是除所述地址子空间对应主机之外，与所述 PCIE 交换机连接的主机。

[0073] 结合第三方面的第四种可能的实现方式，在第五种可能的实现方式中：

[0074] 所述虚拟机监控器，还用于接收并缓存所述服务器对应的子网管理服务器的 PCIE NTB 地址空间信息。

[0075] 结合第三方面或第三方面的第一至第五种任一可能的实现方式，在第六种可能的实现方式中，所述目的主机的虚拟机监控器根据所述 PCIE 报文中的目的虚拟机的 ID 将所

述数据分发到所述网络请求的目的虚拟机，包括：

[0076] 所述目的主机的虚拟机监控器读取所述目的主机的 PCIE NTB 地址空间中对应于所述服务器的 PCIE NTB 地址空间中的所述 PCIE 报文，根据所述 PCIE 报文中的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机。

[0077] 如上所述，本发明所述的基于 PCIE 网络的虚拟机通信方法、服务器及系统，具有以下有益效果：

[0078] 本发明将 PCIE 网络作为虚拟机间的通信机制，利用 PCIE 网络的高带宽、低延时的优势，并通过基于 PCIE 网络的子网管理，使得网络中的所有主机上的虚拟机绕过 OS 中的 TCP/IP 协议栈，共享子网管理服务器维护管理的 PCIE 非透明桥 (Non-Transparent Bridge，简称 NTB) 地址空间，更进一步提高了虚拟机间的通信速度。

附图说明

[0079] 图 1 为本发明所述的基于 PCIE 网络的虚拟机通信系统的主要结构示意图。

[0080] 图 2 为本发明所述的基于 PCIE 网络的虚拟机通信系统中主机的结构示意图。

[0081] 图 3 为本发明所述的基于 PCIE 网络的虚拟机通信系统中主机上运行的虚拟机的结构示意图。

[0082] 图 4 为本发明所述的基于 PCIE 网络的虚拟机通信系统中主机上运行的虚拟机监控器的结构示意图。

[0083] 图 5 为本发明所述的基于 PCIE 网络的虚拟机通信系统中 PCIE 交换机的结构示意图。

[0084] 图 6 为本发明所述的基于 PCIE 网络的虚拟机通信系统中子网管理服务器的结构示意图。

[0085] 图 7 为本发明所述的基于 PCIE 网络的虚拟机通信方法的流程示意图。

[0086] 图 8 为本发明所述的基于 PCIE 网络的虚拟机通信方法中步骤 S1 的详细流程示意图。

[0087] 图 9 为本发明所述的基于 PCIE 网络的虚拟机通信方法中地址解析的详细流程示意图。

[0088] 图 10 为本发明另一实施例的服务器结构示意图。

具体实施方式

[0089] 以下通过特定的具体实例说明本发明的实施方式，本领域技术人员可由本说明书所揭露的内容轻易地了解本发明的其他优点与功效。本发明还可以通过另外不同的具体实施方式加以实施或应用，本说明书中的各项细节也可以基于不同观点与应用，在没有背离本发明的精神下进行各种修饰或改变。

[0090] 针对本具体实施方式部分涉及的附图，需要说明的是，本实施例中所提供的图示仅以示意方式说明本发明的基本构想，遂图示中仅显示与本发明中有关的组件而非按照实际实施时的组件数目、形状及尺寸绘制，其实际实施时各组件的形态、数量及比例可为一种随意的改变，且其组件布局形态也可能更为复杂。

[0091] 随着 PCIE 技术的发展，PCIE 总线进一步拓展为两台主机之间的数据通信总线。

PCIE 传输数据包是通过地址范围来寻址的，即处理器访问外围设备时会读写分配给外围设备的地址空间，最终 PCIE 总线通过这个地址把数据报文路由给外围设备。这个路由机制要求在一个主机系统中分配给外围设备的地址空间是不能重复的，但在两个主机之间通过 PCIE 来交互数据时无法保证地址空间不重复，此时 PCIE 总线引入了非透明桥 NTB 机制。NTB 机制是利用预先设置好的地址转换规则将一个主机发往另一个主机的数据报文的地址进行转换，转换为目标主机中不会冲突的地址范围，从而实现 PCIE 总线在主机之间的数据路由。NTB 一般集成在 PCIE 交换芯片当中，提供高宽带和低延时的主机间通信链路，并且具有 PCIE 设备的功能，例如中断和错误处理等。此外，PCIE 交换机对于 Host2Host（主机对主机）通信的实现主要起到以下两方面作用 :1) 对于 PIO(Programming Input Output，过程通道）的读与写通过 NTB 进行标识 (Identity，简称 ID) 路由和地址路由 ;2) 对于 DMA(Direct memory access，直接内存访问) 的发送和接收通过 ID 路由。

[0092] 因为 PCIE 总线的延迟时间是 1 微妙，以太网的延迟则接近于 10 微秒，同时 TCP/IP 协议极其复杂，造成 OS 的协议处理占据了网络时延的很大部分，此外，在数据传输过程中，TCP/IP 协议栈处理的频繁的数据复制和中断处理导致主机 CPU 资源无法专注于处理用户业务，所以本发明提出一种基于 PCIE 网络的虚拟机通信方法，该方法构建了一个为所有 VM 的数据交换服务的 PCIE 网络，主机的 VMM 使用 PCIE NTB 的地址窗口和中断机制进行通信，从而加速了 VM 之间的通信。

[0093] 下面结合实施例和附图对本发明进行详细说明。

[0094] 实施例

[0095] 本实施例提供一种基于 PCIE 网络的虚拟机通信系统，如图 1 所示，包括至少 2 台主机 10(Host)、一个 PCIE 交换机 20(PCIE Switch) 以及一个子网管理服务器 30 ;各主机 10 上运行有一个虚拟机监控器 (VMM)111 和至少一个虚拟机 (VM)101 ;PCIE 交换机 20 与所有主机相连，实现主机之间的数据转发。所述子网管理服务器 30 与所述 PCIE 交换机 20 相连，用于维护系统中主机与虚拟机的对应关系。本发明实施例中提供的系统的主机具体可以为普通的计算机、移动终端、工作站或服务器、专用服务器等，子网管理服务器可以是系统中的一台普通主机，也可以是一台专用主机，本发明不作具体限定。

[0096] 下面对主机、PCIE 交换机以及子网管理服务器作进一步详细描述。

[0097] 如图 2 所示，每个所述主机 10 包括 :虚拟机运行模块 100、虚拟机监控模块 110 以及硬件平台 120。虚拟机运行模块 100 和虚拟机监控模块 110 分别提供了虚拟机的硬件平台和软件平台，虚拟机监控模块 110 包含了虚拟机监视器 111，是物理硬件和虚拟机之间的中间层，负责协调各个虚拟机对硬件平台的访问。虚拟机运行模块 100 中运行了多个虚拟机 101。每台主机上的虚拟机监控器 111 与该主机上的所有虚拟机 101 通信相连，将虚拟机 101 发出的数据以 PCIE 协议发送给所述 PCIE 交换机 20。

[0098] 此外，主机的硬件平台 120 可以包括 CPU、内存、网卡、存储和互联设备等高速 I/O 设备以及基本输入输出等低速设备。虚拟机运行在主机的硬件平台上，且感知不到网络集群内部的通信是通过 PCIE 网络完成的。

[0099] 虚拟机 101，用于在 TCP/IP 协议栈之上拦截网络请求，并将所述网络请求中包含的数据发送给虚拟机监控器 111。优选的，虚拟机 101，用于在 TCP/IP 协议栈之上拦截所述网络请求，获取所述网络请求中包含的数据，将所述网络请求中包含的数据放入自己对应

的虚拟机数据包缓冲队列中，并通知虚拟机监控器 111。优选的，虚拟机 101 可以仅针对数据类型报文的网络请求进行拦截并进行后续处理，非数据类型报文的网络请求还按照原来的 TCP/IP 协议栈方式进行后续处理。

[0100] 虚拟机监控器 111，用于在接收到虚拟机 101 发送的所述网络请求之后，根据所述网络请求的目的地址，确定 PCIE 交换机上源主机对应的目的主机的 PCIE NTB 地址空间，将包含所述网络请求中的数据的 PCIE 报文发送到所述 PCIE 交换机上所述源主机对应的所述目的主机的 PCIE NTB 地址空间，以便所述目的主机的虚拟机监控器根据所述 PCIE 报文中的目的虚拟机的 ID 将所述数据分发到所述网络请求的目的虚拟机。其中，所述目的主机为所述网络请求的目的虚拟机所属的主机，所述 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间为所述 PCIE 交换机上分配给所述目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间。

[0101] 优选的，VMM 111，用于接收到虚拟机 101 发送的网络请求后，根据网络请求中的目的地址（即网络请求的目的虚拟机的 IP 地址），确定源主机对应的目的主机的 PCIE NTB 地址空间，即确定该源主机与目的主机通信（此处是源主机上的源 VM 与目的主机上的目的 VM 进行通信）时，需要写入的目的主机的 PCIE NTB 地址空间。为描述方便，在本文中，将发出该网络请求的虚拟机所在的主机称为源主机，将该网络请求的目的 VM 所在的主机称为目的主机。

[0102] 优选的，当 VMM 111 上缓存有该网络请求所对应的目的主机（即网络请求的目的虚拟机的 IP 地址所对应的物理主机）的 PCIE NTB 地址空间时，则 VMM 111 可以确定所述网络请求对应的目的主机对应的 PCIE NTB 地址空间，即该源主机应写入的目的主机的 PCIE NTB 地址空间。优选的，在虚拟机 101 发送地址解析协议 (Address Resolution Protocol，简称 ARP) 请求进行地址解析流程时，VMM 111 会缓存该源主机所对应的目的主机的 PCIE NTB 地址空间。进一步的，当虚拟机 101 要发出 TCP/IP 网络请求时，如果虚拟机 101 上没有所述网络请求的目的虚拟机的 MAC 地址信息，则虚拟机 111 会按照 TCP/IP 协议发出 ARP 请求进行地址解析。具体为：

[0103] 虚拟机 101，还用于当虚拟机 101 上没有网络请求的目的虚拟机的 MAC 地址信息时，向网络请求的目的虚拟机发送以太网 ARP 请求。ARP 请求中包含目的虚拟机的 IP 地址。则，VMM 111，还用于接收到所述虚拟机 101 发送的 ARP 请求后，向子网管理服务器 30 发起查询目的虚拟机对应的 PCIE NTB 地址空间的请求消息，该请求消息带有目的虚拟机的 IP 地址，以及接收到所述子网管理服务器 30 返回的包含所述源 VM 对应的目的主机的 PCIE NTB 地址空间信息的响应消息后，将所述源 VM 对应的目的主机的 PCIE NTB 地址空间信息缓存在本地。

[0104] 子网管理服务器 30，用于接收到 VMM 发起的查询目的虚拟机对应的 PCIE NTB 地址空间的请求消息后，根据主机与虚拟机的对应关系确定所述目的虚拟机所属的目的主机，获知源主机对应的目的主机的 PCIE NTB 地址空间，发送响应消息给虚拟机监控器 111。

[0105] 优选的，子网管理服务器 30，用于维护管理所有 PCIE NTB 地址空间，PCIE NTB 地址空间配置在与所述源主机和目的主机通信相连的 PCIE 交换机上，所述 PCIE NTB 地址空间分成与 PCIE 交换机连接的主机个数相同份数的 PCIE NTB 地址子空间，每个所述地址子空间对应一个主机，用于接收发送给该主机的数据，在每个主机对应的地址子空间中为每

个其他主机分配一个地址微空间，每个所述地址微空间对应一个其他主机，用于接收来自于这个其他主机发送给所述地址子空间对应的该主机的数据，这里所说的其他主机是除所述地址子空间对应的主机之外，与 PCIE 交换机连接的主机。例如，PCIE 交换机最大能接入 N 个主机（主机 1—主机 N），那么子网管理服务器维护的 PCIE NTB 地址空间也将被均分成 N 个 PCIE NTB 地址子空间，一个 PCIE NTB 地址子空间对应着一个主机。进一步的，一个 PCIE NTB 地址子空间又可以分出 N-1 个 PCIE NTB 地址微空间，每个 PCIE NTB 地址微空间对应一个其他主机，主机和主机之间通过写指定的 PCIE NTB 地址空间的方式就可以实现互相通信，例如，整个 PCIE NTB 地址空间分成 N 个 PCIE NTB 地址子空间，主机 1 对应的 PCIE NTB 地址子空间分为 N-1 个地址微空间，分别标识为 H2、H3、……HN，当主机 2 与主机 1 进行通信时，主机 2 向主机 1 对应的 PCIE NTB 地址空间（地址子空间）中的 H2 区域（地址微空间）写入数据即可以完成数据由主机 2 向主机 1 的传递。子网管理服务器 30 会在 PCIE 网络初始化时通知网络集群中的所有 Host 自己所对应的地址空间。本发明实施例可以通过地址空间的这种机制实现多个主机与一个主机并发进行通信

[0106] 进一步的，子网管理服务器 30，还用于维护各主机与其上运行的虚拟机的对应关系，并根据虚拟机的创建、更改、销毁等实时更新该对应关系。所述子网管理服务器维护的对应关系可以为 IP 地址的映射关系或虚拟机 ID 的映射关系，或者其他类型的映射关系，例如，可以为主机的 IP 地址与虚拟机的 ID 之间的对应关系，或主机对应的 PCIE NTB 地址空间与虚拟机的 ID 之间的对应关系。每个 VM 的创建、销毁、更改都会通知子网管理服务器刷新其映射关系。所述子网管理服务器可以为一个物理主机，也可以是一个虚拟机。

[0107] 具体的，子网管理服务器 30，用于接收到源 VMM 发送的请求消息，根据消息中携带的目的虚拟机的 IP 地址，以及本地维护保存的目的主机与目的虚拟机的对应关系确定所述目的虚拟机所属的目的主机，并进一步确定源 VM 对应的目的主机的 PCIE NTB 地址空间，即源 VM 对应的目的 VM 的 PCIE NTB 地址空间，发送响应消息给源虚拟机监控器，将源 VM 对应的目的主机的 PCIE NTB 地址空间信息返回给源 VMM。

[0108] 则，所述虚拟机监控器 111 在得到所述源 VM 对应的目的主机的 PCIE NTB 地址空间消息后，还用于继续进行源 VM 的 ARP 流程，后续的 ARP 流程与现有的 ARP 流程类似：虚拟机监控器将以太网的 ARP 请求发送给所述网络请求的目的虚拟机，即，源 VMM 将 ARP 请求发送给目的虚拟机所在的主机的 VMM，所述目的主机上的虚拟机监控器收到所述 ARP 请求后，将所述 ARP 请求发送给所述目的虚拟机；目的虚拟机返回 ARP 响应，其中包含目的虚拟机的 MAC 地址；所述源虚拟机接收到所述目的虚拟机发送的 ARP 响应，至此完成了目的虚拟机的 ARP 地址解析过程。

[0109] 优选的，虚拟机监控器 111，用于将所述网络请求发送给目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间，具体为：虚拟机监控器 111，用于将所述网络请求包含的数据、源虚拟机 ID 和源 NTB 端口 ID 放入 PCIE 报文的数据荷载，将所述网络请求的目的虚拟机的 ID 作为所述 PCIE 报文的目的地址，发送所述 PCIE 报文给 PCIE 交换机 20，使所述 PCIE 交换机 20 将所述数据写入目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间。其中，所述源虚拟机 ID 为发出所述数据的虚拟机的 ID，所述源 NTB 端口 ID 为所述 PCIE 交换机中与所述源主机上的虚拟机监控器相连的 NTB 端口的 ID。

[0110] 进一步的，目的主机的虚拟机监控器，还用于根据目的虚拟机的 ID 将所述数据分发到目的虚拟机。具体地，目的主机的虚拟机监控器读取所述目的主机对应的 PCIE NTB 地址空间中对应于所述源主机的PCIE NTB地址空间中的数据，根据所述数据的目的虚拟机的 ID 将数据分发到所述数据的目的虚拟机的缓冲队列中，并通知所述数据的目的虚拟机从所述缓冲队列中接收数据。

[0111] 进一步的，如图 3 所示，所述虚拟机 101 中包括 ：网络请求拦截模块 1011、内核协议模块 1012、虚拟设备驱动模块 1013 ；

[0112] 所述网络请求拦截模块 1011，用于在 TCP/IP 协议栈之上拦截所述网络请求。

[0113] 所述内核协议模块 1012，与所述网络请求拦截模块 1011 通信相连，用于对所述网络拦截模块 1011 拦截的所述网络请求进行处理。具体的处理过程为 ：内核协议模块 1012 接收所述网络请求，解析所述网络请求，获取所述网络请求包含的数据。

[0114] 所述虚拟设备驱动模块 1013，与所述内核协议模块 1012 通信相连，用于将所述网络请求包含的数据放入所述虚拟机对应的数据包缓冲队列中，并通知与所述虚拟机运行在同一台主机上的虚拟机监控器。具体地，虚拟机中的虚拟设备驱动模块 1013 将所述网络请求包含的数据放入由虚拟机监控器监管且属于所述虚拟机私有的数据包缓冲队列（以下简称缓冲队列）中，同时通知虚拟机监控器有新请求要发送。其中，每个虚拟机均私有一个缓冲队列，由于每个主机上可设有多个虚拟机，故每个主机上的虚拟机监控器监管该主机上的全部虚拟机私有的缓冲队列。

[0115] 进一步的，如图 4 所示，所述虚拟机监控器 111，包括与虚拟机 101 通信相连的 PCIE 数据包分发器 1111、PCIE 通信模块 1112。

[0116] 所述 PCIE 数据包分发器 1111，用于将所述网络请求中包含的数据加上源虚拟机 ID 和源 NTB 端口 ID 放入 PCIE 报文的数据荷载，将所述网络请求的目的虚拟机的 ID 作为所述 PCIE 报文的目的地址。其中，拦截网络请求的虚拟机称为源虚拟机，接收网络请求的虚拟机为目的虚拟机，源虚拟机所属的主机为源主机，目的虚拟机所属的主机为目的主机，所述源虚拟机 ID 为所述发出数据的虚拟机的 ID，所述源 NTB 端口 ID 为与所述源主机上的虚拟机监控器相连的 NTB 端口的 ID。由于一台主机上可能运行着多台 VM，因此 PCIE 数据包分发器 1111 的功能是将某个 VM 的网络请求加上源虚拟机 ID 和源 NTB 端口 ID，区分每个 VM 的数据流。

[0117] 所述 PCIE 通信模块 1112，用于将 PCIE 数据包分发器 1111 生成的所述 PCIE 报文发送给所述 PCIE 交换机 20，使所述 PCIE 交换机 20 将所述数据写入目的主机对应的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间。

[0118] 目的主机上的虚拟机监控器 111 上的 PCIE 通信模块 1112，还用于从所述 PCIE 交换机 20 接收 PCIE 报文，则所述 PCIE 数据包分发器 1111，还用于将 PCIE 通信模块 1112 接收的 PCIE 报文根据报文中的目的虚拟机 ID 将 PCIE 报文中的数据分发到报文所对应的目的虚拟机的缓冲队列中。

[0119] 如图 5 所示，所述 PCIE 交换机 20 主要包括与主机 Host 相连的非透明桥 NTB，NTB 提供地址窗口机制用来在主机间收发数据。在本发明所述的基于 PCIE 网络的虚拟机通信系统中，所述 PCIE 交换机 20 包括至少 2 个 NTB 端口，NTB 端口与主机以及子网管理服务器一一对应相连。例如在图 5 中，主机 1 写主机 2 的 NTB 所在的地址窗口，那么数据会被 PCIE

交换机 20 以数据报文的形式从主机 1 的 NTB 发送到主机 2 的 NTB，最终发送到主机 2 的内存中；主机 1 也可以写主机 2 对应的门铃寄存器 (doorbell)，使得主机 2 连接的 NTB 向主机 2 发送中断；最后，PCIE 交换机 20 还有更高级的远程直接数据存取 (Remote direct memory access，简称 RDMA) 功能，利用这个功能，PCIE 交换机 20 可以实现利用 NTB 相连的主机间拷贝内存。

[0120] 如图 6 所示，所述子网管理服务器 30，包括主机与虚拟机对应关系管理模块 301、PCIE NTB 地址空间管理模块 302、PCIE 通信模块 303。

[0121] 主机与虚拟机对应关系管理模块 301，用于维护系统中主机与虚拟机的对应关系。具体地，所述主机与虚拟机对应关系管理模块 301 维护了当前系统中每个 Host 与运行于其上的 VM 的对应关系。所述主机与虚拟机对应关系管理模块 301 维护的对应关系可以为 IP 映射关系或 ID 映射关系，或者其他类型的映射关系，例如主机的 IP 地址与虚拟机的 ID 之间的归属关系或主机对应的 PCIE NTB 地址空间与虚拟机的 ID 之间的对应关系。主机与虚拟机对应关系管理模块 301 会根据虚拟机的创建、更改、销毁等实时更新主机与虚拟机的对应关系。

[0122] 所述 PCIE NTB 地址空间管理模块 302，用于维护管理为每个主机预先配置的 PCIE NTB 地址空间。所有 PCIE NTB 地址空间分成与所述 PCIE 交换机 20 连接的主机个数相同的份数的 PCIE NTB 地址子空间，每个所述地址子空间对应一个主机，用于接收发送给所述地址子空间对应的主机的数据，进一步的，在每个所述地址子空间中为其他主机分配一个地址微空间，每个所述地址微空间对应一个所述其他主机，用于接收来自于所述其他主机发送给所述地址子空间对应的主机的数据，其中，所述其他主机是除所述地址子空间对应的主机之外，与所述 PCIE 交换机连接的主机。例如，PCIE 交换机最大能接入 N 个主机，主机 1—主机 N，那么子网管理服务器维护的 PCIE NTB 地址空间也将被均分成 N 个 PCIE NTB 地址子空间，一个 PCIE NTB 地址子空间对应着一个主机。进一步的，一个 PCIE NTB 地址子空间又可以分出 N-1 个 PCIE NTB 地址微空间，每个 PCIE NTB 地址微空间对应一个其他主机，主机和主机之间通过写指定的 PCIE NTB 地址空间的方式就可以实现互相通信，例如，整个 PCIE NTB 地址空间分成 N 个 PCIE NTB 地址子空间，主机 1 对应的 PCIE NTB 地址子空间分为 N-1 个地址微空间，分别标识为 H2、H3、……HN，当主机 2 与主机 1 进行通信时，主机 2 向主机 1 对应的 PCIE NTB 地址空间（地址子空间）中的 H2 区域（地址微空间）写入数据即可以完成数据由主机 2 向主机 1 的传递。PCIE NTB 地址空间管理模块 302，会在 PCIE 网络初始化时通知网络集群中的所有 Host 子网管理服务器 30 所对应的地址空间，各主机上的 VMM 会缓存本主机所对应的子网管理服务器的地址空间信息，以便 VMM 在收到源 VM 的地址解析请求时能够正确向子网管理服务器发送查询源 VM 对应的目的主机的 PCIE NTB 地址空间的请求消息。

[0123] 所述 PCIE 通信模块 303，用于与所述 PCIE 交换机 20 通信，收发 PCIE 报文。

[0124] 进一步的，PCIE 通信模块 303，用于接收源 VMM 通过 PCIE 交换机发送的请求消息；所述主机与虚拟机对应关系管理模块 301，用于根据该请求消息中携带的目的虚拟机的 IP 地址，以及维护保存的目的主机与目的虚拟机的对应关系确定所述目的虚拟机所属的目的主机；PCIE NTB 地址空间管理模块 302，用于根据所述目的主机以及源主机的信息，确定源 VM 对应的目的主机的 PCIE NTB 地址空间，即源 VM 对应的目的 VM 的 PCIE NTB 地址空间；

通过所述 PCIE 通信模块 303 发送响应消息给源虚拟机监控器，将源 VM 对应的目的主机的 PCIE NTB 地址空间信息返回给源 VMM。

[0125] 本发明将 PCIE 网络作为虚拟机间的通信机制，跳过了以太网的传输，高效利用 PCIE 网络的高带宽、低延时的优势，大幅度降低了传输时延，实现了虚拟机间的快速通信。而且，本发明还突破了 PCIE 网络的限制，实现了基于 PCIE 网络的子网管理，使得网络中的所有主机上的虚拟机绕过 OS 中的 TCP/IP 协议栈，维护了系统中主机与虚拟机的对应关系，实现了整个系统的统一管理维护，更进一步提高了基于 PCIE 网络的虚拟机通信系统的完整性。此外，本发明通过子网管理服务器实现了虚拟机对 PCIE NTB 地址空间的共享，解决了 PCIE 网络的并发读写问题，更进一步提高了虚拟机间数据接收的速度。

[0126] 本实施例还提供一种基于 PCIE 网络的虚拟机通信方法，该方法可以由所述的基于 PCIE 网络的虚拟机通信系统实现，但其实现结构不限于实施例所述的系统。

[0127] 如图 7 所示，所述基于 PCIE 网络的虚拟机通信方法包括：

[0128] S1，源主机上的源虚拟机在 TCP/IP 协议栈之上拦截网络请求，并将所述网络请求中包含的数据发送给所述源主机上的虚拟机监控器。

[0129] 具体地，可以如图 8 所示，所述步骤 S1 包括：

[0130] S11，源虚拟机在 TCP/IP 协议栈之上拦截网络请求；

[0131] 优选的，本发明实施例中可以仅针对数据类型报文的网络请求进行拦截并进行后续处理，非数据类型报文的网络请求还按照原来的 TCP/IP 协议栈方式进行后续处理。因为非数据类型报文的网络请求对主机性能的影响不大，且实现逻辑复杂，采用本发明实施例中的方法并无明显优势，即，优选的，本发明实施例将控制报文仍全部转交给现有的 TCP/IP 协议栈进行后续处理，而拦截数据类型报文进行本发明实施例中的后续处理。

[0132] S12，所述源虚拟机解析所述网络请求，获取所述网络请求包含的数据；

[0133] S13，所述源虚拟机将所述数据放入所述源虚拟机对应的数据包缓冲队列中，并通知所述源主机上的虚拟机监控器。

[0134] S2，所述源主机上的虚拟机监控器在接收到所述网络请求之后，根据所述网络请求的目的地址，确定 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间，其中，目的主机的 PCIE NTB 地址空间为 PCIE 交换机上分配给所述目的主机的 PCIE NTB 地址空间，PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间为所述 PCIE 交换机上分配给所述目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间。源主机上的 VMM 接收到所述网络请求后，根据网络请求中的目的地址（即网络请求的目的虚拟机的 IP 地址）确定所述网络请求对应的目的主机（目的虚拟机所属的物理主机），并进一步确定所述源主机对应的目的主机的PCIE NTB地址空间，即确定该源主机与目的主机通信（此处也是源主机上的源 VM 与目的主机上的目的 VM 进行通信）时，需要写入的目的主机的 PCIE NTB 地址空间。PCIE NTB 地址空间配置在与所述源主机和目的主机通信相连的一个 PCIE 交换机上。

[0135] 优选的，当源主机上的 VMM 上缓存有该网络请求所对应的目的主机（即网络请求的目的虚拟机的 IP 地址所对应的物理主机）的 PCIE NTB 地址空间时，则源主机上的 VMM 可以确定所述网络请求对应的目的主机对应的PCIE NTB地址空间，即该源主机对应应写入的目的主机的 PCIE NTB 地址空间。在源虚拟机发送地址解析协议请求进行地址解析流程

时，源主机上的 VMM 会缓存该源主机所对应的目的主机的 PCIE NTB 地址空间，所述目的主机就是该网络请求的目的虚拟机的 IP 地址所对应的物理主机。具体的，当源虚拟机要发出 TCP/IP 网络请求时，如果源虚拟机上没有所述网络请求的目的虚拟机的 MAC 地址信息，则源虚拟机会按照 TCP/IP 协议发出 ARP 请求进行地址解析，可以参见图 9，其中源 VMM 指源主机上的虚拟机监控器，目的 VMM 指目的主机上的虚拟机监控器。地址解析的具体实现过程为：

[0136] S21，源虚拟机通过源主机上的虚拟机监控器向网络请求的目的虚拟机发送以太网 ARP 请求，这里的 ARP 请求是 TCP/IP 协议中的标准请求，里面会包含目的虚拟机的 IP 地址。

[0137] S22，源 VMM 接收到带有目的虚拟机 IP 地址的 ARP 请求后，发起查询目的虚拟机对应的 PCIE NTB 地址空间的请求消息到子网管理服务器，该请求消息带有目的虚拟机的 IP 地址。源 VMM 上保存有子网管理服务器的 PCIE NTB 地址空间信息，发送查询源 VM 对应的目的主机的 PCIE NTB 地址空间的请求消息。

[0138] S23，所述子网管理服务器根据其维护管理的主机与虚拟机的对应关系确定所述目的虚拟机所属的目的主机，获知源主机对应的目的主机的 PCIE NTB 地址空间。

[0139] 优选的，子网管理服务器维护管理着所有 PCIE NTB 地址空间，所有 PCIE NTB 地址空间分成与 PCIE 交换机连接的主机个数相同份数的 PCIE NTB 地址子空间，每个所述地址子空间对应一个主机，用于接收发送给该主机的数据，在每个主机对应的地址子空间中为每个其他主机分配一个地址微空间，每个所述地址微空间对应一个其他主机，用于接收来自于这个其他主机发送给所述地址子空间对应的该主机的数据，这里所说的其他主机是除所述地址子空间对应的主机之外，与 PCIE 交换机连接的主机。

[0140] 例如，PCIE 交换机最大能接入 N 个主机（主机 1—主机 N），那么子网管理服务器维护的 PCIE NTB 地址空间也将被均分成 N 个 PCIE NTB 地址子空间，一个 PCIE NTB 地址子空间对应着一个主机。进一步的，一个 PCIE NTB 地址子空间又可以分出 N-1 个 PCIE NTB 地址微空间，每个 PCIE NTB 地址微空间对应一个其他主机，主机和主机之间通过写指定的 PCIE NTB 地址空间的方式就可以实现互相通信，例如，整个 PCIE NTB 地址空间分成 N 个 PCIE NTB 地址子空间，主机 1 对应的 PCIE NTB 地址子空间分为 N-1 个地址微空间，分别标识为 H2、H3、……HN，当主机 2 与主机 1 进行通信时，主机 2 向主机 1 对应的 PCIE NTB 地址空间（地址子空间）中的 H2 区域（地址微空间）写入数据即可以完成数据由主机 2 向主机 1 的传递。本发明实施例可以通过地址空间的这种机制实现多个主机与一个主机并发进行通信。

[0141] 优选的，子网管理服务器还维护有各主机与其上运行的虚拟机的对应关系，并根据虚拟机的创建、更改、销毁等实时更新该对应关系。所述子网管理服务器维护的对应关系可以为 IP 地址的映射关系或虚拟机 ID 的映射关系，或者其他类型的映射关系，例如，可以为主机的 IP 地址与虚拟机的 ID 之间的对应关系，或主机对应的 PCIE NTB 地址空间与虚拟机的 ID 之间的对应关系。

[0142] 具体的，子网管理服务器接收到源 VMM 发送的请求消息，子网管理服务器根据消息中携带的目的虚拟机的 IP 地址，以及本地维护保存的目的主机与目的虚拟机的对应关系确定所述目的虚拟机所属的目的主机，并进一步确定源 VM 对应的目的主机的 PCIE NTB

地址空间，即源 VM 对应的目的 VM 的 PCIE NTB 地址空间。

[0143] S24，子网管理服务器发送响应消息给源虚拟机监控器，将源 VM 对应的目的主机的 PCIE NTB 地址空间信息返回给源 VMM。

[0144] S25，所述源主机上的虚拟机监控器得到所述源 VM 对应的目的主机的 PCIE NTB 地址空间信息后，将所述源 VM 对应的目的主机的 PCIE NTB 地址空间信息缓存在本地。也就是缓存了源主机对应的目的主机的 PCIE NTB 地址空间信息。

[0145] 所述源主机上的虚拟机监控器得到所述源 VM 对应的目的主机的 PCIE NTB 地址空间消息后，继续进行源 VM 的 ARP 流程，后续的 ARP 流程与现有的 ARP 流程类似：

[0146] S26，所述源主机上的虚拟机监控器将以太网的 ARP 请求发送给所述网络请求的目的虚拟机，即，源 VMM 将 ARP 请求发送给目的虚拟机所在的主机的 VMM，所述目的主机上的虚拟机监控器收到所述 ARP 请求后，将所述 ARP 请求发送给所述目的虚拟机；

[0147] S27，目的虚拟机返回 ARP 响应，其中包含目的虚拟机的 MAC 地址所述源虚拟机接收到所述目的虚拟机发送的 ARP 响应，至此完成了目的虚拟机的 ARP 地址解析过程。

[0148] 优选的，子网管理服务器会在 PCIE 网络初始化时预配置好 PCIE NTB 地址空间，并将子网管理服务器所对应的地址空间信息通知子网中的所有主机，各主机收到子网管理服务器的信息后，会缓存自己所对应的子网管理服务器的地址空间信息，以便能正确地向子网管理服务器发送查询对应的 PCIE NTB 地址空间的请求消息。

[0149] S3，所述源主机上的虚拟机监控器将包含所述网络请求中的数据的 PCIE 报文发送到所述 PCIE 交换机上所述源主机对应的目的主机的 PCIE NTB 地址空间，以便所述目的主机的虚拟机监控器根据所述 PCIE 报文中的目的虚拟机的 ID 将所述数据分发到该网络请求的目的虚拟机。

[0150] 具体地，所述源主机上的虚拟机监控器将所述网络请求发送给目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间，详细实现过程为：所述源主机上的虚拟机监控器将所述网络请求包含的数据、所述源虚拟机的 ID 和源 NTB 端口 ID 放入 PCIE 报文的数据荷载，将所述网络请求的目的虚拟机的 ID 作为所述 PCIE 报文的目的地址，发送所述 PCIE 报文给所述 PCIE 交换机，使所述 PCIE 交换机将所述数据写入目的主机的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间。其中，所述源 NTB 端口 ID 为所述 PCIE 交换机中与所述源主机上的虚拟机监控器相连的 NTB 端口的 ID。

[0151] 进一步的，所述基于 PCIE 网络的虚拟机通信方法还包含步骤 S4，所述目的主机的虚拟机监控器根据所述目的虚拟机的 ID 将所述数据分发到目的虚拟机。

[0152] 具体地，本步骤的详细实现过程为：所述目的主机的虚拟机监控器读取所述目的主机对应的 PCIE NTB 地址空间中对应于所述源主机的 PCIE NTB 地址空间中的数据，根据所述数据的目的虚拟机的 ID 将数据分发到所述数据的目的虚拟机的缓冲队列中，并通知所述数据的目的虚拟机从所述缓冲队列中接收数据。即，在收到 doorbell 中断后，目的主机的 VMM 会迅速解析缓冲区的数据，并根据目标 VM 的区域 (domain) ID 将数据分发到目标 VM 的缓冲队列中，同时目标 VMM 会注入中断通知目标 VM 有新数据到来。目标 VM 最终会通过其内核中的虚拟设备驱动模块从缓冲队列中接收到所述数据。每个虚拟机均私有一个缓冲队列，由于每个主机上可设有多个虚拟机，故每个主机内的虚拟机监控器监管该主机上的全部虚拟机私有的缓冲队列。

[0153] 本发明实施例利用子网管理服务器维护了 PCIE 网络中主机与虚拟机的对应关系，实现了目标 VM 的地址解析功能，更进一步实现了整个 PCIE 网络的统一管理维护，提高了 PCIE 网络的集成性。

[0154] 图 10 为本发明实施例的服务器结构图，如图所示，本实施例的服务器 40 包括至少一个处理器 401，存储器 405、至少一个网络接口卡 404，和至少一通信总线 402。该服务器 40 可选的包含用户接口 403，包括显示器，键盘或者点击设备。存储器 405 存储执行指令，当服务器 40 运行时，处理器 401 与存储器 405 之间通信，处理器 401 执行执行指令使得服务器 40 执行本发明前面实施例中所描述的主机实施的方法，具体为：可以如图 7 所示，服务器上的源虚拟机在 TCP/IP 协议栈之上拦截网络请求，并将所述网络请求中包含的数据发送给服务器上的虚拟机监控器；服务器上的虚拟机监控器在接收到所述网络请求之后，根据所述网络请求的目的地址，确定 PCIE 交换机上所述服务器对应的目的主机的 PCIE NTB 地址空间；服务器上的虚拟机监控器将包含所述网络请求中的数据的 PCIE 报文发送到所述 PCIE 交换机上所述服务器对应的目的主机的 PCIE NTB 地址空间，以便所述目的主机的虚拟机监控器根据所述 PCIE 报文中的目的虚拟机的 ID 将所述数据分发到该网络请求的目的虚拟机。实现原理和技术效果与前面实施例中所述的方法类似，此处不再赘述。操作系统 406，包含各种程序，用于实现各种基础业务以及处理基于硬件的任务。该服务器 40 还包括显示接口卡。另外，本发明实施例中还提供一种计算机可读介质，包含计算机执行指令，计算机执行指令能够使服务器 40 执行本发明前面实施例中所描述的主机实施的方法，其实现原理和技术效果类似，此处不再赘述。

[0155] 进一步的，本实施例中的服务器 40，还可以替代前面实施例中 PCIE 网络的虚拟机通信系统中的主机 10，实现上述实施例中主机 10 的功能，作为组成 PCIE 网络的虚拟机通信系统中的一部分，与 PCIE 网络的虚拟机通信系统中的其他组成部分通信配合，使所述 PCIE 网络的虚拟机通信系统实现将 PCIE 网络作为虚拟机间的通信机制，实现虚拟机间的快速通信。其实现原理和技术效果与前面实施例中所述的通信系统和通信方法类似，此处不再赘述。

[0156] 本发明规避了 TCP/IP 协议栈导致的时延和主机性能损耗，利用了 PCIE 网络的低延迟，高带宽的优势，大幅度降低了传输时延，实现了虚拟机间的快速通信；同时还基于标准的 PCIE 实现，解决了 PCIE 网络的并发读写问题，更进一步提高了虚拟机间数据接收的速度。

[0157] 综上所述，本发明有效克服了现有技术中的种种缺点而具高度产业利用价值。

[0158] 通过以上的实施方式的描述，所属领域的技术人员可以清楚地了解到本发明可以用硬件实现，或固件实现，或它们的组合方式来实现。当使用软件实现时，可以将上述功能存储在计算机可读介质中或作为计算机可读介质上的一个或多个指令或代码进行传输。计算机可读介质包括计算机存储介质和通信介质，其中通信介质包括便于从一个地方向另一个地方传送计算机程序的任何介质。存储介质可以是计算机能够存取的任何可用介质。以此为例但不限于：计算机可读介质可以包括 RAM、ROM、EEPROM、CD-ROM 或其他光盘存储、磁盘存储介质或者其他磁存储设备、或者能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其他介质。此外。任何连接可以适当的成为计算机可读介质。例如，如果软件是使用同轴电缆、光纤光缆、双绞线、数字用户线 (DSL) 或者诸

如红外线、无线电和微波之类的无线技术从网站、服务器或者其他远程源传输的，那么同轴电缆、光纤光缆、双绞线、DSL 或者诸如红外线、无线和微波之类的无线技术包括在所属介质的定影中。如本发明所使用的，盘 (Disk) 和碟 (disc) 包括压缩光碟 (CD)、激光碟、光碟、数字通用光碟 (DVD)、软盘和蓝光光碟，其中盘通常磁性的复制数据，而碟则用激光来光学的复制数据。上面的组合也应当包括在计算机可读介质的保护范围之内。

[0159] 上述实施例仅例示性说明本发明的原理及其功效，而非用于限制本发明。任何熟悉此技术的人士皆可在不违背本发明的精神及范畴下，对上述实施例进行修饰或改变。因此，举凡所属技术领域中具有通常知识者在未脱离本发明所揭示的精神与技术思想下所完成的一切等效修饰或改变，仍应由本发明的权利要求所涵盖。

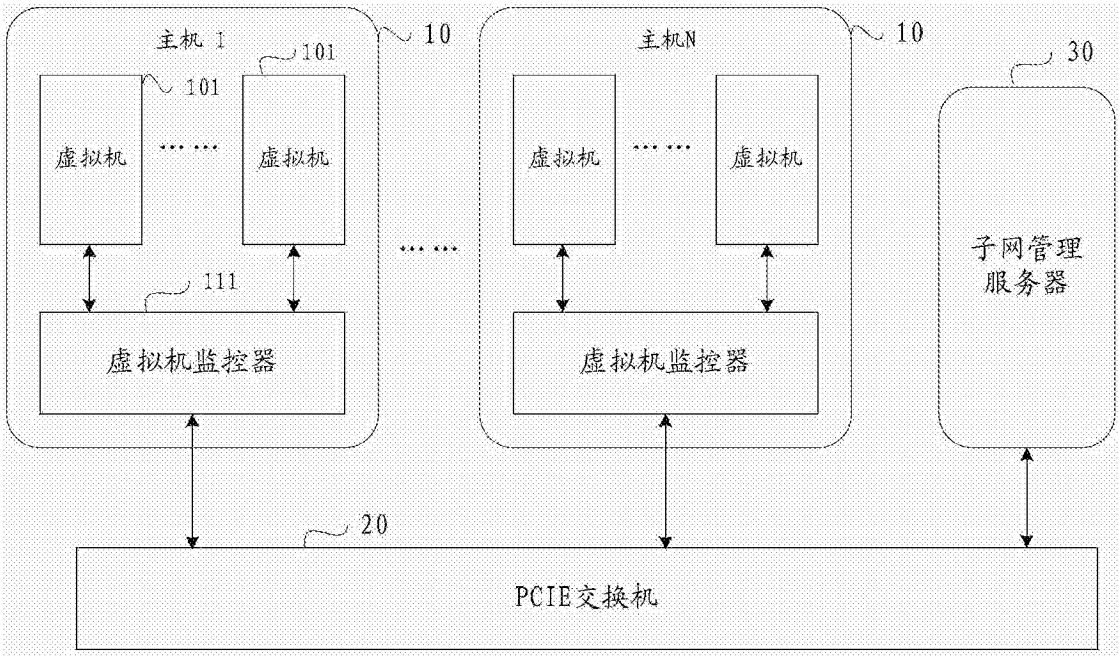

主机 I
10
101
101
虚拟机
虚拟机
111
虚拟机监控器
主机N
10
虚拟机
虚拟机
虚拟机监控器
30
子网管理
服务器
20
PCIE交换机

图 1

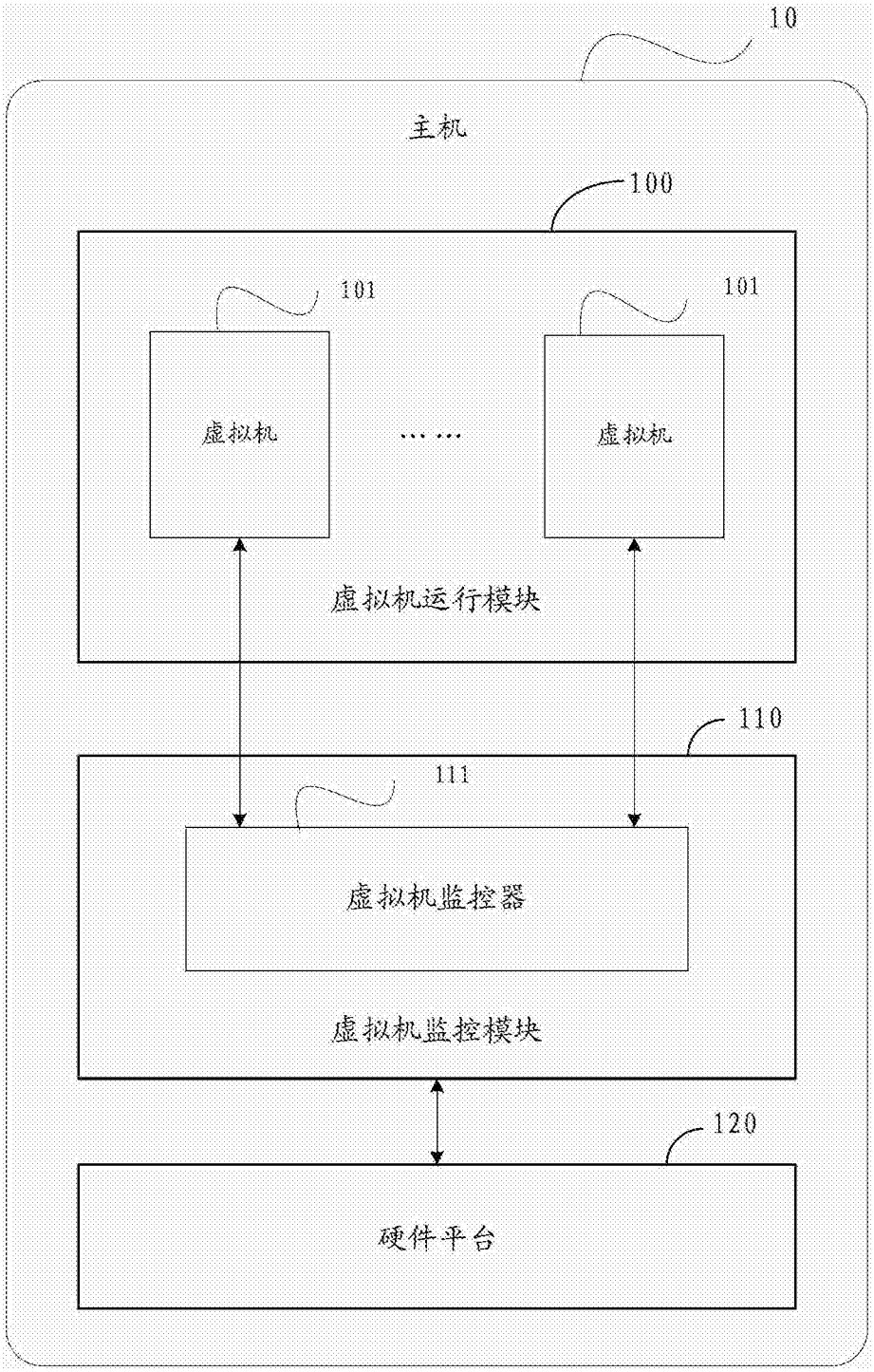

10
主机
100
101
101
虚拟机
……
虚拟机
虚拟机运行模块
110
111
虚拟机监控器
虚拟机监控模块
120
硬件平台

图 2

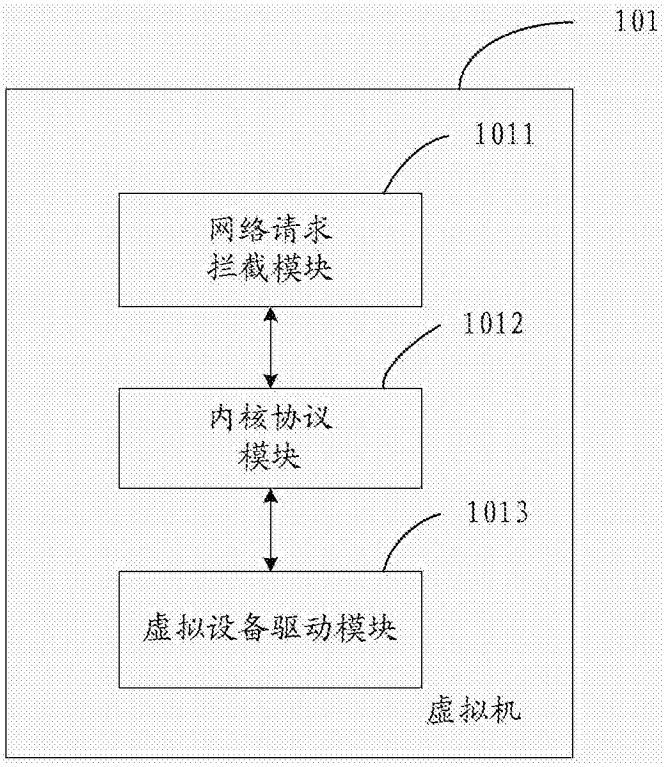

101
1011
网络请求
拦截模块
1012
内核协议
模块
1013
虚拟设备驱动模块
虚拟机

图 3

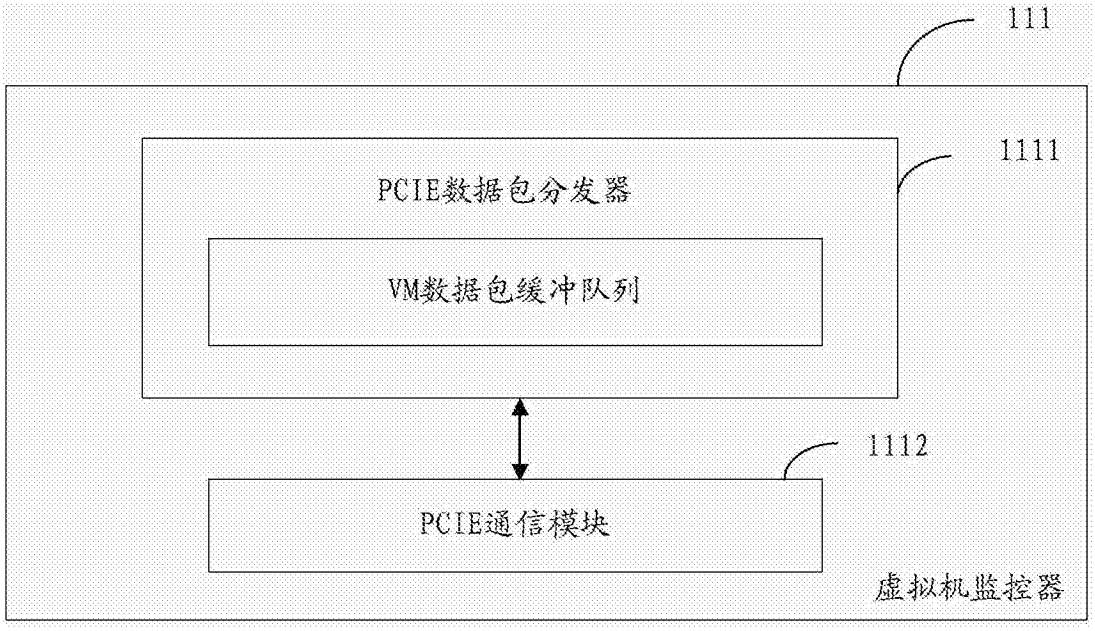

111
1111
PCIE数据包分发器
VM数据包缓冲队列
1112
PCIE通信模块
虚拟机监控器

图 4

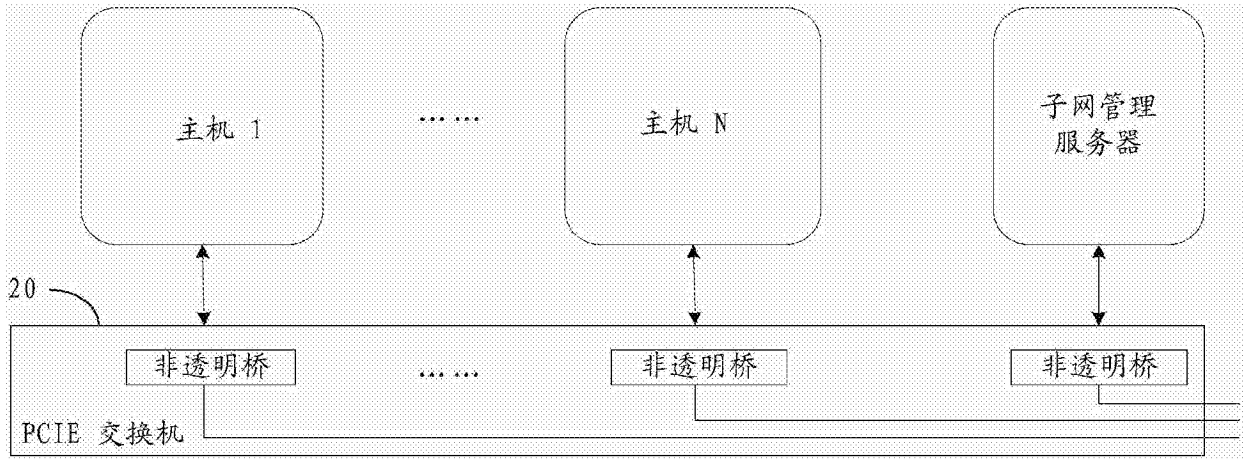

主机 1
……
主机 N
子网管理
服务器
20
非透明桥
……
非透明桥
非透明桥
PCIE 交换机

图 5

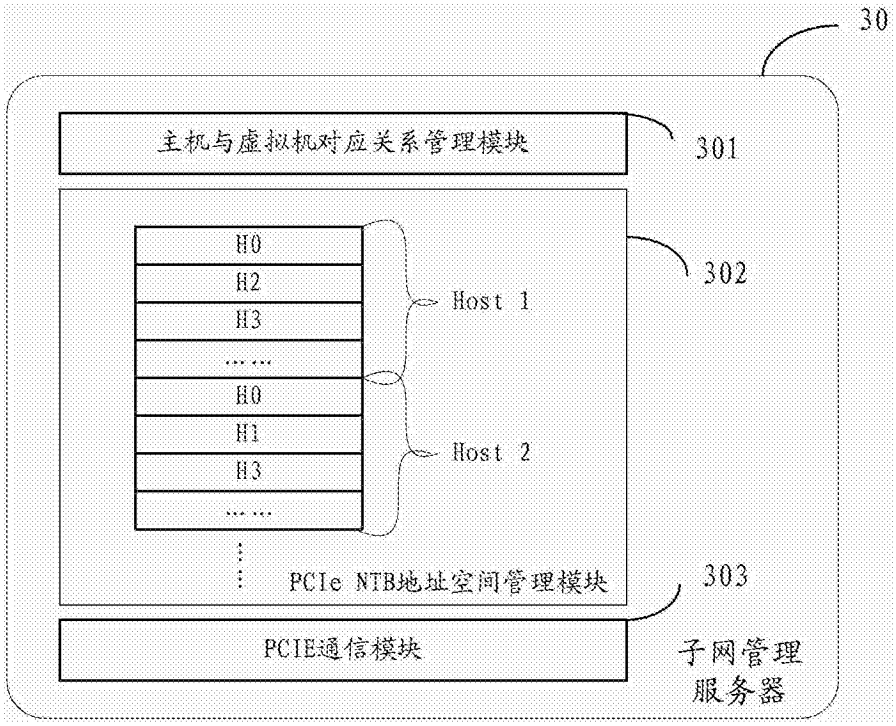

30
主机与虚拟机对应关系管理模块
301
H0
H2
H3
……
H0
H1
H3
……
Host 1
Host 2
302
PCIe NTB地址空间管理模块
303
PCIE通信模块
子网管理
服务器

图 6

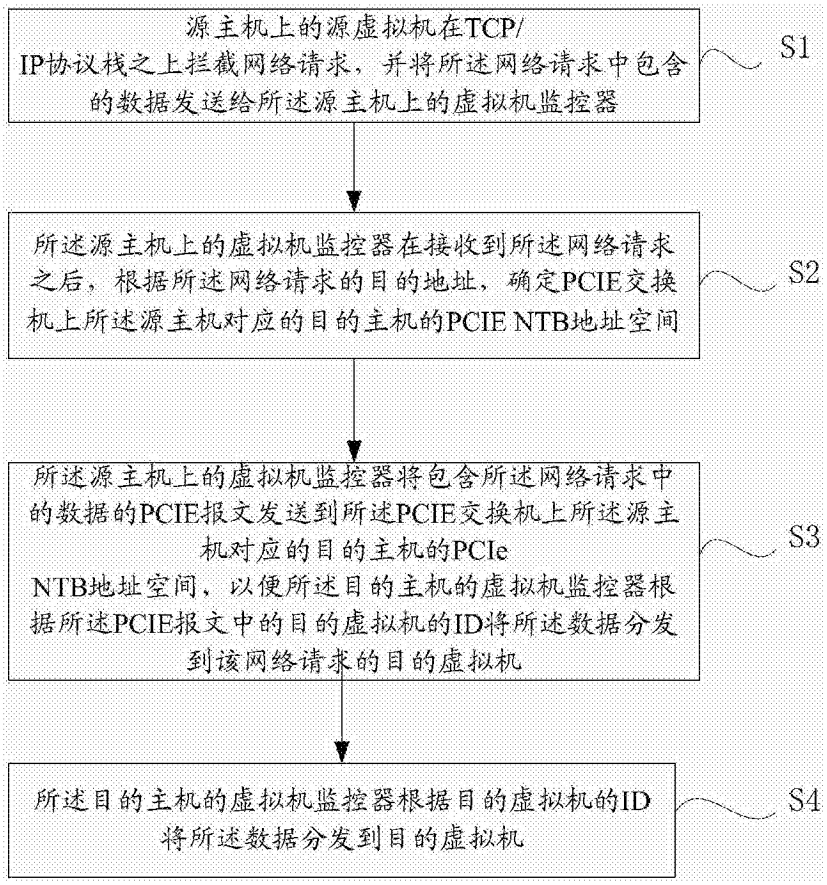

源主机上的源虚拟机在TCP/IP协议栈之上拦截网络请求，并将所述网络请求中包含的数据发送给所述源主机上的虚拟机监控器
S1
所述源主机上的虚拟机监控器在接收到所述网络请求之后，根据所述网络请求的目的地址，确定PCIE交换机上所述源主机对应的目的主机的PCIE NTB地址空间
S2
所述源主机上的虚拟机监控器将包含所述网络请求中的数据的PCIE报文发送到所述PCIE交换机上所述源主机对应的目的主机的PCIe NTB地址空间，以便所述目的主机的虚拟机监控器根据所述PCIE报文中的目的虚拟机的ID将所述数据分发到该网络请求的目的虚拟机
S3
所述目的主机的虚拟机监控器根据目的虚拟机的ID将所述数据分发到目的虚拟机
S4

图 7

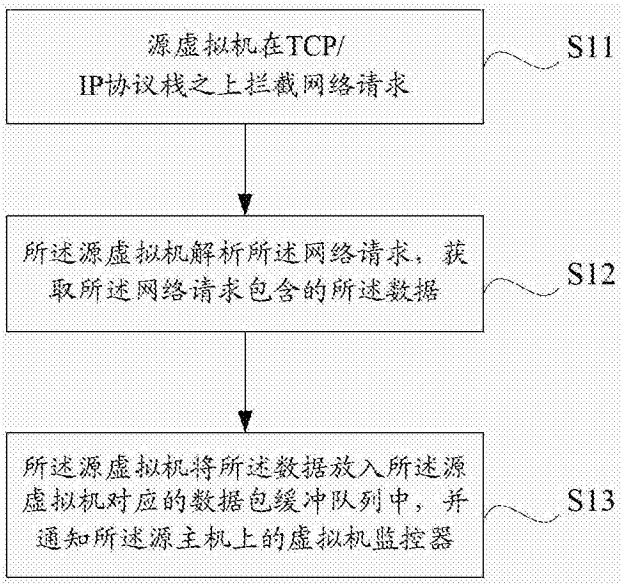

源虚拟机在TCP/IP协议栈之上拦截网络请求
S11
所述源虚拟机解析所述网络请求，获取所述网络请求包含的所述数据
S12
所述源虚拟机将所述数据放入所述源虚拟机对应的数据包缓冲队列中，并通知所述源主机上的虚拟机监控器
S13

图 8

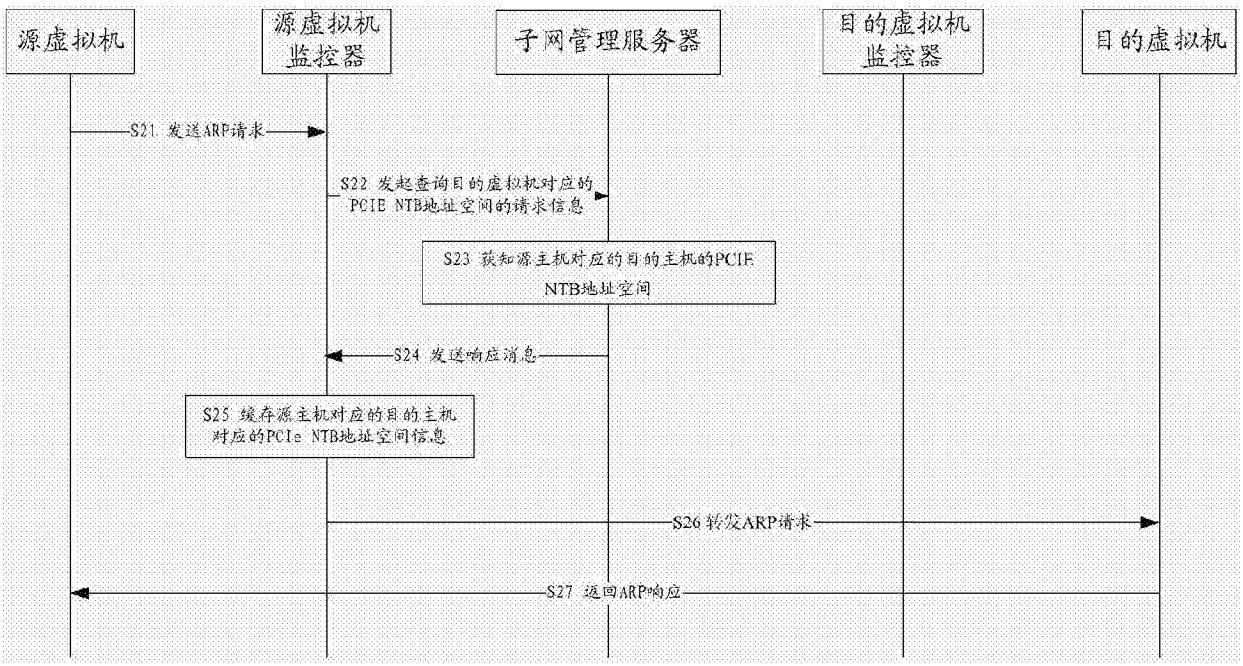

源虚拟机
源虚拟机监控器
子网管理服务器
目的虚拟机监控器
目的虚拟机
S21 发送ARP请求
S22 发起查询目的虚拟机对应的PCIE NTB地址空间的请求信息
S23 获知源主机对应的目的主机的PCIE NTB地址空间
S24 发送响应消息
S25 缓存源主机对应的目的主机对应的PCIe NTB地址空间信息
S26 转发ARP请求
S27 返回ARP响应

图 9

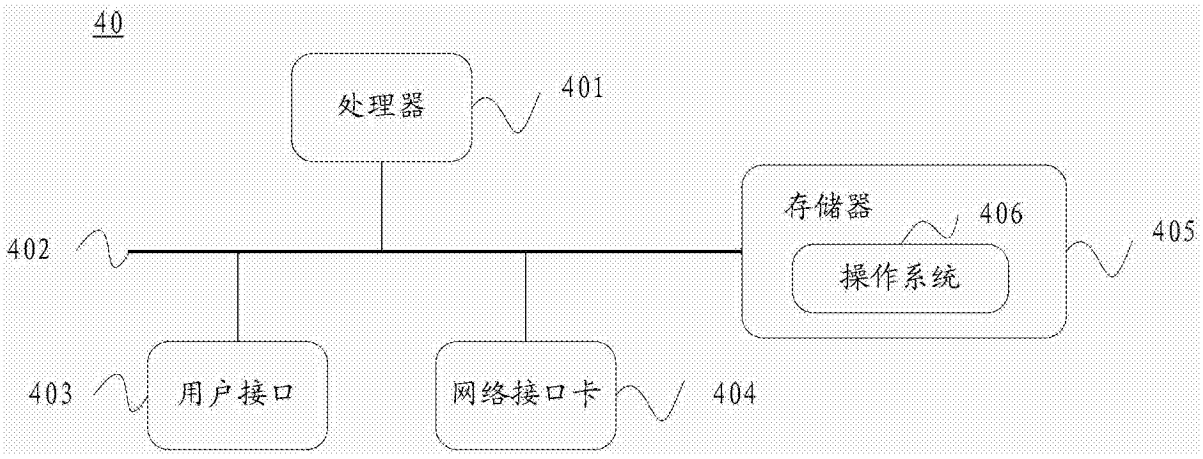

40
处理器
401
存储器
406
操作系统
405
402
403
用户接口
网络接口卡
404

图 10