(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 103530167 B
(45)授权公告日 2017.04.05

(21)申请号 201310462273.3

(22)申请日 2013.09.30

(65)同一申请的已公布的文献号
申请公布号 CN 103530167 A

(43)申请公布日 2014.01.22

(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 杨杰 章晓峰

(74)专利代理机构 北京中博世达专利商标代理有限公司 11274
代理人 张娜

(51)Int.Cl.
G06F 9/455(2006.01)
G06F 15/16(2006.01)

(56)对比文件
CN 102576343 A,2012.07.11,
CN 102081552 A,2011.06.01,
CN 101464812 A,2009.06.24,
US 2007/0288921 A1,2007.12.13,
US 2012/0259940 A1,2012.10.11,
CN 102081552 A,2011.06.01,
Wei Huang 等.High performance virtual machine migration with RDMA over modern interconnects.《2007 IEEE International Conference on Cluster Computing》.2007,

审查员 黄超

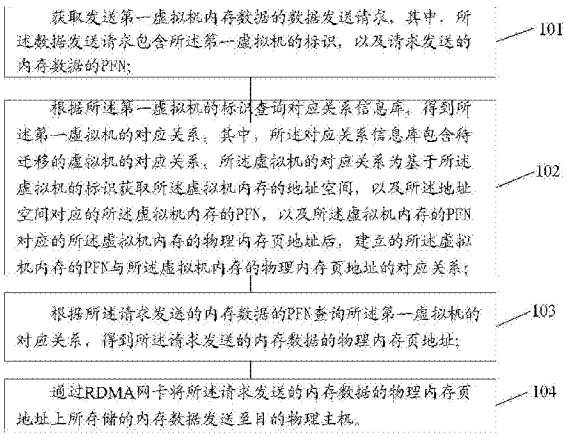
权利要求书3页 说明书22页 附图6页

(54)发明名称

一种虚拟机内存数据的迁移方法及相关装置和集群系统

(57)摘要

本发明实施例公开了一种虚拟机内存的迁移方法及相关装置和集群系统。涉及通信领域，降低了虚拟机内存迁移时的处理器利用率及时间开销。本发明实施例提供的方法包括：获取发送第一虚拟机内存数据的数据发送请求，其中，数据发送请求包含第一虚拟机的标识，以及请求发送的内存数据的PFN；根据第一虚拟机的标识查询对应关系信息库，得到第一虚拟机的对应关系，根据请求发送的内存数据的PFN查询第一虚拟机的对应关系，得到请求发送的内存数据的物理内存页地址；通过RDMA网卡将所述请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。



CN 103530167 B

1.一种虚拟机内存数据的迁移方法，其特征在于，包括：

获取发送第一虚拟机内存数据的数据发送请求，其中，所述数据发送请求包含所述第一虚拟机的标识，以及请求发送的内存数据的物理页框号PFN；

根据所述第一虚拟机的标识查询对应关系信息库，得到所述第一虚拟机的对应关系，其中，所述对应关系信息库包含待迁移的虚拟机的对应关系，所述第一虚拟机的对应关系为基于所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间，以及所述地址空间对应的所述第一虚拟机内存的PFN，以及所述第一虚拟机内存的PFN对应的所述第一虚拟机内存的物理内存页地址后，建立的所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址的对应关系；

根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址；

通过远程直接内存读取RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机。

2.根据权利要求1所述的方法，其特征在于，在所述获取发送第一虚拟机内存数据的数据发送请求之前，所述方法还包括：

获取特权虚拟机用户态进程触发的迁移所述第一虚拟机的内存数据的迁移请求，其中，所述迁移请求中包含所述第一虚拟机的标识；

根据所述迁移请求中的所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间；

根据所述地址空间获取所述第一虚拟机内存的PFN；

根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址；

建立所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址之间的对应关系。

3.根据权利要求2所述的方法，其特征在于，在所述根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址之后，所述获取发送第一虚拟机内存数据的数据发送请求之前，所述方法还包括：

将所述第一虚拟机内存的物理内存页地址注册至所述RDMA网卡。

4.根据权利要求1或2所述的方法，其特征在于，在所述根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址之后，通过RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机之前，所述方法还包括：

将所述请求发送的内存数据的物理内存页地址注册至所述RDMA网卡。

5.根据权利要求1所述方法，其特征在于，所述获取发送第一虚拟机内存数据的数据发送请求包括：

依序提取数据发送请求队列中的数据发送请求，其中，所述数据发送请求包含不同的待迁移的虚拟机的数据发送请求，所述数据发送请求队列中的数据发送请求按照时间先后或者优先级高低进行排序。

6.一种宿主机Host，其特征在于，包括：

获取单元，用于获取发送第一虚拟机内存数据的数据发送请求，其中，所述数据发送请求包含所述第一虚拟机的标识，以及请求发送的内存数据的物理页框号PFN；

第一查询单元，用于根据所述第一虚拟机的标识查询对应关系信息库，得到所述第一虚拟机的对应关系，其中，所述对应关系信息库包含待迁移的虚拟机的对应关系，所述第一虚拟机的对应关系为基于所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间，以及所述地址空间对应的所述第一虚拟机内存的PFN，以及所述第一虚拟机内存的PFN对应的所述第一虚拟机内存的物理内存页地址后，建立的所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址的对应关系；

第二查询单元，用于根据所述请求发送的内存数据的PFN查询所述第一查询单元得到的所述第一虚拟机的对应关系，获取所述请求发送的内存数据的物理内存页地址；

驱动发送单元，用于通过远程直接内存读取RDMA网卡将所述第二查询单元得到的请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机。

7.根据权利要求6所述的Host，其特征在于，所述Host还包括

建立单元，用于在所述获取单元获取发送第一虚拟机内存数据的数据发送请求之前，

获取特权虚拟机用户态进程触发的迁移所述第一虚拟机的内存数据的迁移请求，其中，所述迁移请求中包含所述第一虚拟机的标识；

根据所述迁移请求中的所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间；

根据所述地址空间获取所述第一虚拟机内存的PFN；

根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址；

建立所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址之间的对应关系。

8.根据权利要求7所述的Host，其特征在于，所述Host还包括：第一注册单元，用于在所述建立单元根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址之后，所述获取单元获取发送第一虚拟机内存数据的数据发送请求之前，

将所述建立单元获取的第一虚拟机内存的物理内存页地址注册至所述RDMA网卡。

9.根据权利要求6或7所述的Host，其特征在于，所述Host还包括：第二注册单元，用于在所述第二查询单元根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址之后，所述驱动发送单元通过RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机之前，

将所述第二查询单元得到的请求发送的内存数据的物理内存页地址注册至所述RDMA网卡。

10.根据权利要求6所述Host，其特征在于，所述获取单元具体用于，依序提取数据发送请求队列中的数据发送请求，其中，所述数据发送请求包含不同的待迁移的虚拟机的数据发送请求，所述数据发送请求队列中的数据发送请求按照时间先后或者优先级高低进行排序。

11.一种物理主机，其特征在于，包括：硬件层、运行在所述硬件层之上的宿主机Host、以及运行在所述Host之上的至少一个虚拟机VM，以及所述硬件层包括远程直接内存读取RDMA网卡；所述至少一个虚拟机包括第一虚拟机，其中，

所述Host用于：

获取发送第一虚拟机内存数据的数据发送请求，其中，所述数据发送请求包含所述第

一虚拟机的标识，以及请求发送的内存数据的物理页框号PFN；

根据所述第一虚拟机的标识查询对应关系信息库，得到所述第一虚拟机的对应关系，其中，所述对应关系信息库包含待迁移的虚拟机的对应关系，所述第一虚拟机的对应关系为基于所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间，以及所述地址空间对应的所述第一虚拟机内存的PFN，以及所述第一虚拟机内存的PFN对应的所述第一虚拟机内存的物理内存页地址后，建立的所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址的对应关系；

根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址；

通过所述远程直接内存读取RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机。

12.根据权利要求11所述的物理主机，其特征在于，所述Host还用于：

在获取发送第一虚拟机内存数据的数据发送请求之前，

获取特权虚拟机用户态进程触发的迁移所述第一虚拟机的内存数据的迁移请求，其中，所述迁移请求中包含所述第一虚拟机的标识；

根据所述迁移请求中的所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间；

根据所述地址空间获取所述第一虚拟机内存的PFN；

根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址；

建立所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址之间的对应关系。

13.根据权利要求12所述物理主机，其特征在于，所述Host还用于：

在根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址之后，以及获取发送第一虚拟机内存数据的数据发送请求之前，

将所述第一虚拟机内存的物理内存页地址注册至所述RDMA网卡。

14.根据权利要求11或12所述物理主机，其特征在于，所述Host还用于：在根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址之后，通过RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机之前，

将所述请求发送的内存数据的物理内存页地址注册至所述RDMA网卡。

15.一种集群系统，其特征在于，包括：多台如权利要求11至14任一项所述的物理主机，所述多台物理主机包括源物理主机和目的物理主机，其中每台物理主机包括硬件层、运行在所述硬件层之上的宿主机Host、以及运行在所述Host之上的至少一个虚拟机VM，以及所述硬件层包括RDMA网卡。

一种虚拟机内存数据的迁移方法及相关装置和集群系统

技术领域

[0001] 本发明涉及通信领域，尤其涉及一种虚拟机数据内存的迁移方法及相关装置和集群系统。

背景技术

[0002] 虚拟机迁移是将虚拟机的内存从源物理主机发送至目的物理主机。虚拟机迁移是虚拟机实时迁移的主要部分，虚拟机实时迁移可以在保持虚拟机运行的同时，将虚拟机从源物理主机迁移至目的物理主机，并在目的物理主机上恢复运行，实现服务整合。通过虚拟机实时迁移可以实现服务器的在线维护、在线升级、负载均衡，并为灾难恢复提供了一种解决方案。

[0003] 在现有技术中，xen虚拟化平台下，虚拟机迁移采用迭代发送虚拟机内存的方法，每次迭代时分批次地选取当次迭代的脏页并将这些页面映射至特权虚拟机的用户态进程的线性地址空间，映射完成后，再调用超级调用进行页表的更新从而建立线性地址空间与物理地址的联系，取得映射的线性地址后，通过用户态的远程直接内存读取(Remote Direct Memory Access，简称RDMA)接口注册已映射的虚拟机内存对应的物理地址至RDMA网卡，并通过其他RDMA系统调用按照RDMA通信协议进行后续的数据传输。在数据成功发送后，将虚拟机的内存从用户态进程空间中解映射，同时注销已注册的物理内存。

[0004] 发明人发现现有技术至少存在以下问题：在虚拟化平台下，虚拟机迁移需要映射虚拟机的内存并通过超级调用更新页表，导致虚拟机所在的物理机中的中央处理器(Central Processing Unit，简称CPU)利用率及时间开销较大。

发明内容

[0005] 本发明实施例提供一种虚拟机内存数据的迁移方法及相关装置和集群系统，以降低虚拟机所在的物理主机中的处理器利用率及时间开销。

[0006] 本发明实施例采用的技术方案是：

[0007] 第一方面，提供了一种虚拟机内存数据的迁移方法，包括：

[0008] 获取发送第一虚拟机内存数据的数据发送请求，其中，所述数据发送请求包含所述第一虚拟机的标识，以及请求发送的内存数据的物理页框号(Physical Frame Number，简称PFN)；

[0009] 根据所述第一虚拟机的标识查询对应关系信息库，得到所述第一虚拟机的对应关系，其中，所述对应关系信息库包含待迁移的虚拟机的对应关系，所述第一虚拟机的对应关系为基于所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间，以及所述地址空间对应的所述第一虚拟机内存的PFN，以及所述第一虚拟机内存的PFN对应的所述第一虚拟机内存的物理内存页地址后，建立的所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址的对应关系；

[0010] 根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述

请求发送的内存数据的物理内存页地址；

[0011] 通过RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机。

[0012] 在第一种可能的实现方式中，根据第一方面，在所述获取发送第一虚拟机内存数据的数据发送请求之前，所述方法还包括：

[0013] 获取特权虚拟机用户态进程触发的迁移所述第一虚拟机的内存数据的迁移请求，其中，所述迁移请求中包含所述第一虚拟机的标识；

[0014] 根据所述迁移请求中的所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间；

[0015] 根据所述地址空间获取所述第一虚拟机内存的PFN；

[0016] 根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址；

[0017] 建立所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址之间的对应关系。

[0018] 在第二种可能的实现方式中，结合第一种可能的实现方式，在所述根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址之后，所述获取发送第一虚拟机内存数据的数据发送请求之前，所述方法还包括：

[0019] 将所述第一虚拟机内存的物理内存页地址注册至所述RDMA网卡。

[0020] 在第三种可能的实现方式中，结合第一方面或者第一种可能的实现方式，根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址之后，通过RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机之前，所述方法还包括：

[0021] 将所述请求发送的内存数据的物理内存页地址注册至所述RDMA网卡。

[0022] 在第四种可能的实现方式中，结合第一方面，第一种可能的实现方式至第三种可能的实现方式中的任一种，所述获取发送第一虚拟机内存数据的数据发送请求包括：

[0023] 依序提取数据发送请求队列中的数据发送请求，其中，所述数据发送请求包含不同的待迁移的虚拟机的数据发送请求，所述数据发送请求队列中的数据发送请求按照时间先后或者优先级高低进行排序。

[0024] 第二方面，提供了一种宿主机，包括：

[0025] 获取单元，用于获取发送第一虚拟机内存数据的数据发送请求，其中，所述数据发送请求包含所述第一虚拟机的标识，以及请求发送的内存数据的PFN；

[0026] 第一查询单元，用于根据所述第一虚拟机的标识查询对应关系信息库，得到所述第一虚拟机的对应关系，其中，所述对应关系信息库包含待迁移的虚拟机的对应关系，所述第一虚拟机的对应关系为基于所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间，以及所述地址空间对应的所述第一虚拟机内存的PFN，以及所述第一虚拟机内存的PFN对应的所述第一虚拟机内存的物理内存页地址后，建立的所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址的对应关系；

[0027] 第二查询单元，用于根据所述请求发送的内存数据的PFN查询所述第一查询单元得到的所述第一虚拟机的对应关系，获取所述请求发送的内存数据的物理内存页地址；

[0028] 驱动发送单元，用于通过RDMA网卡将所述第二查询单元得到的请求发送的内存数

据的物理内存页地址上所存储的内存数据发送至目的物理主机。

[0029] 在第一种可能的实现方式中，结合第二方面，所述Host还包括建立单元，用于在所述获取单元获取发送第一虚拟机内存数据的数据发送请求之前，获取特权虚拟机用户态进程触发的迁移所述第一虚拟机的内存数据的迁移请求，其中，所述迁移请求中包含所述第一虚拟机的标识；

[0030] 根据所述迁移请求中的所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间；

[0031] 根据所述地址空间获取所述第一虚拟机内存的PFN；

[0032] 根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址；

[0033] 建立所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址之间的对应关系。

[0034] 在第二种可能的实现方式中，结合第一种可能的实现方式，所述Host还包括：第一注册单元，用于在所述建立单元根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址之后，所述获取单元获取发送第一虚拟机内存数据的数据发送请求之前，

[0035] 将所述建立单元获取的第一虚拟机内存的物理内存页地址注册至所述RDMA网卡。

[0036] 在第三种可能的实现方式中，结合第二方面或者第一种可能的实现方式，所述Host还包括：第二注册单元，用于在所述第二查询单元根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址之后，所述驱动发送单元通过RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机之前，将所述第二查询单元得到的请求发送的内存数据的物理内存页地址注册至所述RDMA网卡。

[0037] 在第四种可能的实现方式中，结合第二方面，第一种可能的实现方式至第三种可能的实现方式中的任一种，所述获取单元具体用于，依序提取数据发送请求队列中的数据发送请求，其中，所述数据发送请求包含不同的待迁移的虚拟机的数据发送请求，所述数据发送请求队列中的数据发送请求按照时间先后或者优先级高低进行排序。

[0038] 第三方面，提供一种物理主机，包括：硬件层、运行在所述硬件层之上的Host、以及运行在所述Host之上至少一个虚拟机VM，以及所述硬件层包括远程直接内存读取RDMA网卡；所述至少一个虚拟机包括第一虚拟机，其中，所述Host用于：

[0039] 获取发送第一虚拟机内存数据的数据发送请求，其中，所述数据发送请求包含所述第一虚拟机的标识，以及请求发送的内存数据的物理页框号PFN；

[0040] 根据所述第一虚拟机的标识查询对应关系信息库，得到所述第一虚拟机的对应关系，其中，所述对应关系信息库包含待迁移的虚拟机的对应关系，所述第一虚拟机的对应关系为基于所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间，以及所述地址空间对应的所述第一虚拟机内存的PFN，以及所述第一虚拟机内存的PFN对应的所述第一虚拟机内存的物理内存页地址后，建立的所述第一虚拟机内存的PFN与所述第一虚拟机的物理内存页地址的对应关系；

[0041] 根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址；

[0042] 通过远程直接内存读取RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机。

[0043] 在第一种可能的实现方式中，结合第三方面，所述Host还用于：

[0044] 在获取发送第一虚拟机内存数据的数据发送请求之前，获取特权虚拟机用户态进程触发的迁移所述第一虚拟机的内存数据的迁移请求，其中，所述迁移请求中包含所述第一虚拟机的标识；

[0045] 根据所述迁移请求中的所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间；

[0046] 根据所述地址空间获取所述第一虚拟机内存的PFN；

[0047] 根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址；

[0048] 建立所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址之间的对应关系。

[0049] 在第二可能的实现方式中，结合第一种可能的实现方式，所述Host还用于：

[0050] 在根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址之后，以及获取发送第一虚拟机内存数据的数据发送请求之前，将所述第一虚拟机内存的物理内存页地址注册至所述RDMA网卡。

[0051] 在第三可能的实现方式中，结合第三方面或者第一种可能的实现方式，所述Host还用于：在根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址之后，通过RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机之前，将所述请求发送的内存数据的物理内存页地址注册至所述RDMA网卡。

[0052] 第四方面，提供一种集群系统，包括：多台上述任一项所述的物理主机，所述多台物理主机包括源物理主机和目的物理主机，其中每台物理主机包括硬件层、运行在所述硬件层之上的Host、以及运行在所述Host之上的至少一个虚拟机VM，以及所述硬件层包括RDMA网卡。

[0053] 由上可见，本发明实施例的物理主机可包括：硬件层、运行在该硬件层之上的Host、以及运行在该Host之上的至少一个虚拟机VM，该硬件层包括RDMA网卡，所述至少一个虚拟机包括第一虚拟机；其中，该Host用于获取发送第一虚拟机内存数据的数据发送请求，其中，该数据发送请求包含第一虚拟机的标识，以及请求发送的内存数据的PFN；该Host根据第一虚拟机的标识查询对应关系信息库，获取第一虚拟机的对应关系，其中，该对应关系信息库包含待迁移的虚拟机的对应关系，虚拟机的对应关系为基于该虚拟机的标识获取该虚拟机内存的地址空间，以及该地址空间对应的该虚拟机内存的PFN，以及该虚拟机内存的PFN对应的该虚拟机内存的物理内存页地址后，建立的该虚拟机内存的PFN与物理内存页地址的对应关系；所以，该Host可根据请求发送的内存数据的PFN查询第一虚拟机的对应关系，获取请求发送的内存数据的物理内存页地址；并通过RDMA网卡将请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机。如此，在虚拟机迁移的过程中，不需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，通过该待迁移虚拟的对应关系以及数据发送求中的PFN，可直接查询获取请求发送的内存数据对应的物理内存页地址，进而大大降低了虚拟机所在的物理机中的处理器利用率及时间开销，一定程

度上解决了现有技术中由于在虚拟化平台下，虚拟机迁移的过程中，需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，导致虚拟机所在的物理机中的处理器利用率及时间开销较大的问题。

附图说明

[0054] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0055] 图1为本发明实施例提供的一种虚拟机内存数据迁移方法的流程示意图；

[0056] 图2为本发明实施例提供的一种应用场景的示意图；

[0057] 图3为本发明实施例提供的另一种虚拟机内存数据迁移方法的流程示意图；

[0058] 图4为本发明实施例提供的一种数据发送请求队列的示意图；

[0059] 图5为本发明实施例提供的再一种虚拟机内存数据迁移的方法的流程示意图；

[0060] 图6为本发明实施例提供的一种宿主机的结构示意图；

[0061] 图7为本发明实施例提供的另一种宿主机的结构示意图；

[0062] 图8为本发明实施例提供的另一种宿主机的结构示意图；

[0063] 图9为本发明实施例提供的一种物理主机的装置结构示意图；

[0064] 图10为本发明实施例提供的另一种物理主机的装置结构示意图；

[0065] 图11为本发明实施例提供的一种集群系统示意图。

具体实施方式

[0066] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0067] 为了方便理解本发明实施例，首先在此介绍本发明实施例描述中会引入的几个术语；

[0068] 虚拟机VM：通过虚拟机软件可以在一台物理主机上模拟出一台或者多台虚拟的计算机，而这些虚拟机就像真正的计算机那样进行工作，虚拟机上可以安装操作系统和应用程序，虚拟机还可访问网络资源。对于在虚拟机中运行的应用程序而言，虚拟机就像是在真正的计算机中进行工作。

[0069] 硬件层：虚拟化环境运行的硬件平台。其中，硬件层可包括多种硬件，例如某物理主机的硬件层可包括处理器(例如CPU)和内存，还可以包括网卡(例如RDMA网卡)、存储器等等高速/低速输入/输出(I/O，Input/Output)设备，及具有特定处理功能的其它设备。

[0070] 宿主机(Host)：作为管理层，用以完成硬件资源的管理、分配；为虚拟机呈现虚拟硬件平台；实现虚拟机的调度和隔离。其中，Host可能是虚拟机监控器(VMM)；此外，有时VMM和1个特权虚拟机配合，两者结合组成Host。其中，虚拟硬件平台对其上运行的各个虚拟机提供各种硬件资源，如提供虚拟处理器(如VCPU)、虚拟内存、虚拟磁盘、虚拟网卡等等。其

中，该虚拟磁盘可对应Host的一个文件或者一个逻辑块设备。虚拟机运行在Host为其准备的虚拟硬件平台上，Host上运行一个或多个虚拟机。

[0071] 特权虚拟机：一种特殊的虚拟机，亦可称为驱动域，例如这种特殊的虚拟机在Xen Hypervisor平台上被称作Dom0，在该虚拟机中安装了例如网卡、SCSI磁盘等真实物理设备的驱动程序，能检测和直接访问这些真实物理设备。其他虚拟机利用Hypervisor提供的相应机制通过特权虚拟机访问真实物理设备。

[0072] 应理解，本发明实施例可以应用于xen虚拟机平台中，也可以应用于可以应用于任意一个迁移虚拟机时需要将虚拟机内存进行映射的虚拟化平台中；本发明实施例对此不进行限制。

[0073] 实施例一

[0074] 参见图1，为本发明实施例提供的一种虚拟机内存数据的迁移方法，如图1所示，可以包括以下步骤：

[0075] 101：获取发送第一虚拟机内存数据的数据发送请求，其中，所述数据发送请求包含所述第一虚拟机的标识，以及请求发送的内存数据的PFN；

[0076] 本发明实施例中提供的方法，可以由第一虚拟机所在的物理主机执行，例如，可以由该物理主机上的Host执行，Host为VMM和运行在该VMM上的特权虚拟机的结合。数据发送请求由特权虚拟机上的用户态进程发送给特权虚拟机。虚拟机的标识可以为任何能够唯一表示虚拟机的参数，例如，可以为虚拟机的域名。

[0077] 102：根据所述第一虚拟机的标识查询对应关系信息库，得到所述第一虚拟机的对应关系，其中，所述对应关系信息库包含待迁移的虚拟机的对应关系，所述第一虚拟机的对应关系为基于所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间，以及所述地址空间对应的所述第一虚拟机内存的PFN，以及所述第一虚拟机内存的PFN对应的所述第一虚拟机内存的物理内存页地址后，建立的所述第一虚拟机内存的PFN与所述第一虚拟机的物理内存页地址的对应关系；

[0078] 103：根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址；

[0079] 104：通过远程直接内存读取RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机。

[0080] 可选的，在所述获取发送第一虚拟机内存数据的数据发送请求之前，所述方法还包括：

[0081] 获取特权虚拟机用户态进程触发的迁移所述第一虚拟机的内存数据的迁移请求，其中，所述迁移请求中包含所述第一虚拟机的标识；

[0082] 根据所述迁移请求中的所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间；

[0083] 根据所述地址空间获取所述第一虚拟机内存的PFN；

[0084] 根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址；

[0085] 建立所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址之间的对应关系。

[0086] 进一步的，该方法还包括将第一虚拟机内存数据的物理页内存地址注册给RDMA网

卡的过程，具体可以通过以下两种方式中任意一种方式来实现：

[0087] 1、在所述根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址之后，所述获取发送第一虚拟机内存数据的数据发送请求之前，将所述第一虚拟机内存的物理内存页地址注册至所述RDMA网卡。

[0088] 在这种方式下，将第一虚拟机全部内存的物理内存页地址一次注册给RDMA网卡，例如，第一虚拟机全部内存对应的物理内存页地址为0x00010000～0x0001ffff，则将0x00010000～0x0001ffff一次性全部注册至RDMA网卡。

[0089] 2、在根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址之后，通过RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机之前，将所述请求发送的内存数据的物理内存页地址注册至所述RDMA网卡。

[0090] 在这种方式下，根据每次的数据发送请求，分别就将每个数据发送请求请求发送的内存数据的物理内存页地址分批次的注册至RDMA网卡，例如，第一虚拟机全部内存对应的物理内存页地址为0x00010000～0x0001ffff，第一个数据发送请求请求发送的内存数据的物理内存页地址为0x00010000～0x000100ff，则在获取第一个数据发送请求之后，通过RDMA网卡将物理内存页地址0x00010000～0x000100ff上存储的内存数据发送给目的物理主机之前，将物理内存页地址0x00010000～0x0001ffff注册至RDMA网卡；再例如，第二个数据发送请求请求发送的内存数据的物理内存页地址为0x00010100～0x000101ff，则在获取第二个数据发送请求之后，通过RDMA网卡将物理内存页地址0x00010100～0x000101ff上存储的内存数据发送给目的物理主机之前，将物理内存页地址0x00010100～0x000101ff注册至RDMA网卡。

[0091] 本发明实施例的虚拟机内存数据的迁移方法，通过获取发送第一虚拟机内存数据的数据发送请求，其中，该数据发送请求包含第一虚拟机的标识，以及请求发送的内存数据的PFN；根据第一虚拟机的标识查询对应关系信息库，获取第一虚拟机的对应关系，其中，该对应关系信息库包含待迁移的虚拟机的对应关系，虚拟机的对应关系为基于该虚拟机的标识获取该虚拟机内存的地址空间，以及该地址空间对应的该虚拟机内存的PFN，以及该虚拟机内存的PFN对应的该虚拟机内存的物理内存页地址后，建立的该虚拟机内存的PFN与物理内存页地址的对应关系；所以，可根据请求发送的内存数据的PFN查询第一虚拟机的对应关系，获取请求发送的内存数据的物理内存页地址；再通过RDMA网卡将请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。如此，在虚拟机迁移的过程中，不需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，通过该待迁移虚拟的对应关系以及数据发送求中的PFN，可直接查询获取请求发送的内存数据对应的物理内存页地址，进而大大降低了虚拟机所在的物理机中的处理器利用率及时间开销，一定程度上解决了现有技术中由于在虚拟化平台下，虚拟机迁移的过程中，需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，导致虚拟机所在的物理机中的处理器利用率及时间开销较大的问题。

[0092] 下面通过具体的实施例对上述方法实施例进行说明。

[0093] 本发明实施例的虚拟机内存数据的迁移方法可以应用于集群系统，该集群系统包含多台物理主机，多台物理主机包括源物理主机和目的物理主机，其中，每台物理主机包括

硬件层、运行在该硬件层之上的Host、以及运行在该Host之上的至少一个虚拟机VM，以及该硬件层包括RDMA网卡，处理器和内存，参见图2，为本发明实施例设定的一种集群系统的应用场景，下述具体实施例在该场景下进行具体说明，包括两台物理主机100和200，以及专用光缆300。两台物理主机100和200的硬件层分别包含RDMA网卡13、23，处理器12、22，内存11、21，其中，内存11、内存21可以为共享内存，RDMA网卡13、23可以为无线宽带(InfiniBand)卡或以太网(RDMA over Converged Ethernet，简称RoCE)卡等，物理主机100和101的硬件层上分别运行有Host17、27，其中，Host17由VMM 14和VMM 14上运行的特权虚拟机15结合组成，Host27由VMM 24和VMM 24上运行的特权虚拟机25结合组成，，VMM 14、24上分别运行有至少一个虚拟机16、26，虚拟机为特权虚拟机之外的其他虚拟机。特权虚拟机15、25上分别加载了RDMA网卡13、23对应的网卡驱动。

[0094] 专用光缆300用于连接两台物理主机上的RDMA网卡，专用光缆300可以为四通道SFP接口(Quad Small Form-factor Pluggable，简称QSPF)，也可以为其它型号的光缆，本发明实施例对此不进行限制。上述共享存储可以为存储网络(Storage Area Network，简称SAN)，也可以为小型计算机系统接口(Internet Small Computer System Interface，简称iSCSI)或网络连接式存储(Network Attached Storage，简称NAS)，本发明实施例对此不进行限制。具体场景为将物理主机100(源物理主机)上的虚拟机16迁移到物理主机200(目的物理主机)上。本发明实施例仅对将虚拟机实时迁移中的虚拟机迁移部分进行详细说明。

[0095] 实施例二

[0096] 参见图3，为本发明实施例提供的另一种虚拟机内存数据的迁移方法，该方法可以由物理主机100来执行，例如，可以由物理主机100运行的Host17执行，如图3所示，可以包括以下步骤：

[0097] S301：获取特权虚拟机15用户态进程发送的虚拟机的迁移命令，其中，虚拟机的迁移命令包含虚拟机的标识；

[0098] 示例性的，虚拟机可以为一个，也可以为多个，本发明实施例对此不进行限制，但虚拟机所在的源物理主机相同。

[0099] 示例性的，Host17中的特权虚拟机15上加载了RDMA网卡13对应的网卡驱动，并封装了所有的RDMA通信接口，每个RDMA通信接口分别用于特权虚拟机内用户态与内核态之间交互，一个RDMA通信接口对应于一个虚拟机，例如，有三个虚拟机，分别为虚拟机1、虚拟机2和虚拟机3，虚拟机1对应RDMA通信接口1，虚拟机2对应RDMA通信接口2，虚拟机3对应RDMA通信接口3，RDMA通信接口1、RDMA通信接口2和RDMA通信接口3都封装在特权虚拟机的网卡驱动中。其中，RDMA通信接口用于创建保护域、队列对等。

[0100] 示例性的，一个用户态进程对应一个虚拟机，例如，有三个虚拟机，分别为虚拟机1、虚拟机2和虚拟机3，虚拟机1对应RDMA通信接口1和用户态进程1，虚拟机2对应RDMA通信接口2和用户态进程2，虚拟机3对应RDMA通信接口3和用户态进程3，RDMA通信接口1、RDMA通信接口2和RDMA通信接口3都封装在特权虚拟机15中，特权虚拟机15中的RDMA通信接口1、RDMA通信接口2和RDMA通信接口3分别接收用户态进程1、用户态进程2和用户态进程3分别发送的虚拟机1、虚拟机2和虚拟机3的迁移命令。

[0101] 示例性的，虚拟机的标识可以为虚拟机的域名(IDentity，简称ID)，也可以为可以唯一代表虚拟机的其它标识，本发明实施例对此不进行限制。

[0102] S302：根据第一虚拟机的标识建立第一虚拟机的对应关系，对应关系为第一虚拟机内存的PFN与物理内存页地址的对应关系；

[0103] 示例性的，Host17中的VMM14可以根据所述迁移请求中的所述虚拟机的标识获取所述虚拟机内存的地址空间；

[0104] 然后根据所述地址空间获取所述第一虚拟机内存的PFN；

[0105] 再根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址；

[0106] 建立所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址之间的对应关系。

[0107] 示例性的，例如，若虚拟机1、虚拟机2和虚拟机3分别对应一个虚拟机的域ID，Host17中的VMM14根据虚拟机1的域ID可以查询虚拟机1对应的地址空间，进而获取该虚拟机1对应的PFN，假设虚拟机1的PFN范围为0～65535，VMM14根据虚拟机1的PFN计算虚拟机1对应的物理内存页地址，假设为0x00010000～0x0001ffff，Host17可以基于此建立两者的对应关系，虚拟机2和虚拟机3与虚拟机1类似。

[0108] 其中，第一虚拟机的对应关系可以为第一虚拟机内存的PFN与物理内存页地址的一一对应的关系表的形式，也可以为其它可以包含第一虚拟机内存的PFN与物理内存页地址的对应关系的任意形式，本发明实施例对此不进行限制。

[0109] S303：将第一虚拟机内存的PFN对应的物理内存页地址注册至RDMA网卡；

[0110] 其中，可以根据S302建立第一虚拟机的对应关系过程中获取的第一虚拟机内存的物理内存页地址，将第一虚拟机全部内存的物理页地址注册至RDMA网卡，也可以根据第一虚拟机的标识查询对应关系库，获取该第一虚拟机的对应关系，再根据该对应关系查找该第一虚拟机内存的PFN对应的物理内存页地址，然后将第一虚拟机全部内存的物理页地址注册至RDMA网卡。其中，对应关系库中可以包含每一个待迁移的虚拟机的对应关系。

[0111] 例如，根据虚拟机1的标识查询对应关系库，得到虚拟机1的对应关系，再根据虚拟机1内存的PFN查询虚拟机1的对应关系，获取虚拟机1内存的物理内存页地址，将虚拟机1内存的物理内存页地址注册至RDMA网卡。

[0112] 示例性的，若待迁移的虚拟机包含更多个，则可以将所有虚拟机内存的物理内存页地址分别注册至RDMA网卡13。

[0113] S304：获取虚拟机对应的用户态进程触发的数据发送请求，其中，该数据发送请求包含用户态进程对应的虚拟机的标识，以及请求发送的内存数据的PFN；

[0114] 例如，如果有三个用户态进程，分别请求发送虚拟机1、虚拟机2和虚拟机3的部分内存数据，接收虚拟机1对应的用户态进程1发送的数据发送请求1，以及虚拟机2对应的用户态进程2发送的数据发送请求2，以及虚拟机3对应的用户态进程3发送的数据发送请求3。

[0115] 其中，数据发送请求1包含虚拟机1的标识和用户态进程1请求发送的虚拟机1的内存数据的PFN、数据发送请求2包含虚拟机2的标识和用户态进程2请求发送的虚拟机2的内存数据的PFN、数据发送请求3中包含虚拟机3的标识和用户态进程3请求发送的虚拟机3的内存数据的PFN。

[0116] 示例性的，若待迁移的虚拟机个数包含更多个时，获取每一个虚拟机分别对应的用户态进程发送的数据发送请求，每个用户态进程管理对应的虚拟机内存数据的迁移。

[0117] S305：将数据发送请求分别放入数据发送请求队列中；

[0118] 可选的，可以按照获取到数据发送请求的时间的先后顺序将数据发送请求分别放入数据发送请求队列中，例如，如图4，为本发明实施例提供的一种数据发送请求队列的示意图，如图所示，若依次获取到数据发送请求1，数据发送请求2，数据发送请求3，则，首先，将用户态进程1发送的数据发送请求1放入数据发送请求队列中，然后，将用户态进程2发送的数据发送请求2放入数据发送请求1后面的数据发送请求队列中，再然后，将用户态进程3发送的数据发送请求3放入数据发送请求2后面的数据发送请求队列中。

[0119] 可选的，可以按照获取的数据发送请求的优先级将数据发送请求分别放入数据发送请求队列中，例如，若获取到3个数据发送请求，优先级由高到低依次为数据发送请求1，数据发送请求2，数据发送请求3，则，首先，将用户态进程1发送的数据发送请求1放入数据发送请求队列中，然后，将用户态进程2发送的数据发送请求2放入数据发送请求1后面的数据发送请求队列中，再然后，将用户态进程3发送的数据发送请求3放入数据发送请求2后面的数据发送请求队列中。

[0120] 当然，若虚拟机的个数包含更多个，则可以按上述方式将每个虚拟机对应的用户态进程分别发送的数据发送请求按序放入数据发送请求队列中。

[0121] S306：按顺序提取数据发送请求队列中的数据发送请求；

[0122] 例如，如图4所示，首先从数据发送请求队列中提取数据发送请求1，再提取数据发送请求2，最后提取发送请求3。

[0123] S307：根据提取数据发送请求中虚拟机标识查询对应关系库，得到该虚拟机的对应关系；

[0124] S308：根据数据发送请求中的PFN查询该虚拟机的对应关系，查找数据发送请求请求发送的内存数据的物理内存页地址；

[0125] S309：通过RDMA网卡依次数据发送请求队列中的数据发送请求请求发送的内存数据的物理内存页地址存储的数据至目的物理主机200；

[0126] S310：异步通知每一个用户态进程；

[0127] 例如，通过RDMA网卡发送完数据发送请求1的内存数据后，异步通知数据发送请求1对应的用户态进程1；通过RDMA网卡发送完数据发送请求2的内存数据后，通知数据发送请求2对应的用户态进程2；通过RDMA网卡发送完数据发送请求3的内存数据后，异步通知数据发送请求3对应的用户态进程3。

[0128] S311：虚拟机16的内存数据是否全部发送至目的物理机200，若否，则执行步骤312，若是，则执行步骤313；

[0129] S312：再按序执行S305-S311步骤；

[0130] 其中，在上述发送数据的过程中，可以采用迭代的方式来发送，以保证虚拟机16在迁移过程中的正常运行。

[0131] S313：将虚拟机16的对应关系删除，并注消虚拟机16内存，将虚拟机16在源物理机100上销毁；进而，虚拟机26在目的物理主机200上被启动。

[0132] 至此，完成虚拟机16从源物理主机100迁移至目的物理主机200的过程。

[0133] 由上可见，本发明实施例中，通过获取特权虚拟机用户态进程发送的虚拟机的迁移命令，并根据每个虚拟机的迁移命令中包含的虚拟机的标识建立每个虚拟机内存的PFN与物理内存页地址的对应关系；根据每个虚拟机的对应关系，可以获取每个虚拟机内存的

PFN对应的物理内存页地址，并将每个虚拟机的物理内存页地址注册至该RDMA网卡；再通过RDMA网卡将请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。如此，在虚拟机迁移的过程中，不需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，通过该待迁移虚拟的对应关系以及数据发送求中的PFN，可直接查询获取请求发送的内存数据对应的物理内存页地址，进而大大降低了虚拟机所在的物理机中的处理器利用率及时间开销，一定程度上解决了现有技术中由于在虚拟化平台下，虚拟机迁移的过程中，需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，导致虚拟机所在的物理机中的处理器利用率及时间开销较大的问题。

[0134] 而且，因为在RDMA网卡发送数据前，把每个虚拟机的物理内存页地址一次性注册至该RDMA网卡，可以保证RDMA网卡根据用户态进程发送的数据发送请求成功传输相关的数据；

[0135] 更进一步的，可以将分别获取的每一个虚拟机对应的用户态进程的数据发送请求通过数据发送请求队列进行统一管理，可以实现针对不同的用户态进程的数据迁移设置不同优先级，进而对不同的业务进行QoS优化，提高虚拟机迁移的性能。

[0136] 实施例三

[0137] 参见图5，为本发明实施例提供的另一种虚拟机迁移的方法，该方法可以由物理主机100来执行，例如，可以由物理主机100运行的Host17执行，如图5所示，可以包括以下步骤：

[0138] S501：获取特权虚拟机用户态进程发送的虚拟机的迁移命令，其中，虚拟机的迁移命令包含虚拟机的标识；

[0139] 示例性的，虚拟机可以为一个，也可以为多个，本发明实施例对此不进行限制，但虚拟机所在的源物理主机相同。

[0140] 示例性的，Host17中的特权虚拟机15上加载了RDMA网卡13对应的网卡驱动，并封装了所有的RDMA通信接口，每个RDMA通信接口分别用于特权虚拟机内用户态与内核态之间交互，一个RDMA通信接口对应于一个虚拟机，例如，有三个虚拟机，分别为虚拟机1、虚拟机2和虚拟机3，虚拟机1对应RDMA通信接口1，虚拟机2对应RDMA通信接口2，虚拟机3对应RDMA通信接口3，RDMA通信接口1、RDMA通信接口2和RDMA通信接口3都封装在特权虚拟机的网卡驱动中。其中，RDMA通信接口用于创建保护域、队列对等。

[0141] 示例性的，一个用户态进程对应一个虚拟机，例如，有三个虚拟机，分别为虚拟机1、虚拟机2和虚拟机3，虚拟机1对应RDMA通信接口1和用户态进程1，虚拟机2对应RDMA通信接口2和用户态进程2，虚拟机3对应RDMA通信接口3和用户态进程3，RDMA通信接口1、RDMA通信接口2和RDMA通信接口3都封装在特权虚拟机15中，特权虚拟机15中的RDMA通信接口1、RDMA通信接口2和RDMA通信接口3分别接收用户态进程1、用户态进程2和用户态进程3分别发送的虚拟机1、虚拟机2和虚拟机3的迁移命令。

[0142] 示例性的，虚拟机的标识可以为虚拟机的域名(IDentity，简称ID)，也可以为可以唯一代表虚拟机的其它标识，本发明实施例对此不进行限制。

[0143] S502：根据第一虚拟机的标识建立第一虚拟机的对应关系，第一虚拟机的对应关系包含第一虚拟机内存的PFN与物理内存页地址的对应关系；

[0144] 示例性的，Host17中的VMM14可以根据所述迁移请求中的所述虚拟机的标识获取

所述虚拟机内存的地址空间；

[0145] 然后根据所述地址空间获取所述第一虚拟机内存的PFN；

[0146] 再根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址；

[0147] 建立所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址之间的对应关系。

[0148] 示例性的，例如，若虚拟机1、虚拟机2和虚拟机3分别对应一个虚拟机的域ID，Host17中的VMM14根据虚拟机1的域ID可以查询虚拟机1对应的地址空间，进而获取该虚拟机1对应的PFN，假设虚拟机1的PFN范围为0～65535，VMM14根据虚拟机1的PFN计算虚拟机1对应的物理内存页地址，假设为0x00010000～0x0001ffff，Host17可以基于此建立两者的对应关系，虚拟机2和虚拟机3与虚拟机1类似。

[0149] 其中，第一虚拟机的对应关系可以为虚拟机内存的PFN与物理内存页地址的一一对应的关系表的形式，也可以为其它可以包含第一虚拟机内存的PFN与物理内存页地址的对应关系的任意形式，本发明实施例对此不进行限制。

[0150] S503：获取虚拟机对应的用户态进程触发的数据发送请求，其中，该数据发送请求包含用户态进程对应的虚拟机的标识，以及请求发送的内存数据的PFN；

[0151] 例如，如果有三个用户态进程，分别请求发送虚拟机1、虚拟机2和虚拟机3的部分内存数据，获取虚拟机1对应的用户态进程1触发的数据发送请求1，以及虚拟机2对应的用户态进程2触发的数据发送请求2，以及虚拟机3对应的用户态进程3触发的数据发送请求3。

[0152] 其中，数据发送请求1包含虚拟机1的标识和用户态进程1请求发送的虚拟机1的内存数据的PFN、数据发送请求2包含虚拟机2的标识和用户态进程2请求发送的虚拟机2的内存数据的PFN、数据发送请求3中包含虚拟机3的标识和用户态进程3请求发送的虚拟机3的内存数据的PFN。

[0153] 示例性的，若待迁移的虚拟机个数包含更多个时，获取每一个虚拟机分别对应的用户态进程触发的数据发送请求，每个用户态进程管理对应的虚拟机内存数据的迁移。

[0154] S504：将数据发送请求分别放入数据发送请求队列中；

[0155] 可选的，可以按照获取到数据发送请求的时间的先后顺序将数据发送请求分别放入数据发送请求队列中，例如，如图4，为本发明实施例提供的一种数据发送请求队列的示意图，如图所示，若依次获取到数据发送请求1，数据发送请求2，数据发送请求3，则，首先，将用户态进程1发送的数据发送请求1放入数据发送请求队列中，然后，将用户态进程2发送的数据发送请求2放入数据发送请求1后面的数据发送请求队列中，再然后，将用户态进程3发送的数据发送请求3放入数据发送请求2后面的数据发送请求队列中。

[0156] 可选的，可以按照获取的数据发送请求的优先级将数据发送请求分别放入数据发送请求队列中，例如，若获取到3个数据发送请求，优先级由高到低依次为数据发送请求1，数据发送请求2，数据发送请求3，则，首先，将用户态进程1发送的数据发送请求1放入数据发送请求队列中，然后，将用户态进程2发送的数据发送请求2放入数据发送请求1后面的数据发送请求队列中，再然后，将用户态进程3发送的数据发送请求3放入数据发送请求2后面的数据发送请求队列中。

[0157] 当然，若虚拟机的个数包含更多个，则可以按上述方式将每个虚拟机对应的用户态进程分别发送的数据发送请求按序放入数据发送请求队列中。

[0158] S505：按顺序提取数据发送请求队列中的数据发送请求；

[0159] 例如，如图4所示，首先从数据发送请求队列中提取数据发送请求1，再提取数据发送请求2，最后提取发送请求3。

[0160] S506：根据提取数据发送请求中虚拟机标识查询对应关系库，得到该虚拟机的对应关系；

[0161] S507：根据数据发送请求中的PFN查询该虚拟机的对应关系，查找数据发送请求请求发送的内存数据的物理内存页地址；

[0162] S508：将提取的数据发送请求对应的物理内存页地址注册至RDMA网卡；

[0163] 其中，按照提取的数据发送请求的顺序将各个数据发送请求分别对应的物理内存页地址分别注册至RDMA网卡。例如，特权虚拟机15提取数据发送请求队列中的数据发送请求1并将数据发送请求1对应的物理内存页地址注册至RDMA网卡；再提取数据发送请求队列中的数据发送请求2并将数据发送请求2对应的物理内存页地址注册至RDMA网卡；再按顺序提取数据发送请求队列中的数据发送请求3并将数据发送请求3对应的物理内存页地址注册至RDMA网卡。

[0164] S509：通过RDMA网卡依次数据发送请求队列中的数据发送请求请求发送的内存数据的物理内存页地址存储的数据至目的物理主机200；

[0165] S510：异步通知每一个用户态进程；

[0166] 例如，通过RDMA网卡发送完数据发送请求1的内存数据后，异步通知数据发送请求1对应的用户态进程1；通过RDMA网卡发送完数据发送请求2的内存数据后，通知数据发送请求2对应的用户态进程2；通过RDMA网卡发送完数据发送请求3的内存数据后，异步通知数据发送请求3对应的用户态进程3。

[0167] S511：注销已经发送的数据发送请求对应的虚拟机的内存；

[0168] S512：虚拟机16的内存数据是否全部发送至目的物理机200，若否，则执行步骤513，若是，则执行步骤514；

[0169] S513：执行步骤S503-S512。

[0170] S514：将虚拟机16的对应关系删除，并注消虚拟机16内存，将虚拟机16在源物理机100上销毁，进而虚拟机26在目的物理主机200上被启动。

[0171] 若是迁移完成，Host17将虚拟机的对应关系删除，并注消虚拟机内存。

[0172] 至此，完成虚拟机16从源物理主机100迁移至目的物理主机200的过程。

[0173] 由上可见，本发明实施例中，通过获取特权虚拟机用户态进程触发的虚拟机的迁移命令，并根据虚拟机的迁移命令中包含的虚拟机的标识建立虚拟机内存的PFN与物理内存页地址的对应关系；根据虚拟机的对应关系，可以获取虚拟机内存的PFN对应的物理内存页地址，再通过RDMA网卡发送数据之前将虚拟机的物理内存页地址注册至该RDMA网卡；再通过RDMA网卡将请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。如此，在虚拟机迁移的过程中，不需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，通过该待迁移虚拟的对应关系以及数据发送求中的PFN，可直接查询获取请求发送的内存数据对应的物理内存页地址，进而大大降低了虚拟机所在的物理机中的处理器利用率及时间开销，一定程度上解决了现有技术中由于在虚拟化平台下，虚拟机迁移的过程中，需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，导致虚拟机所

在的物理机中的处理器利用率及时间开销较大的问题。

[0174] 而且，可以将每一个虚拟机对应的用户态进程分别发送的数据发送请求通过数据发送请求队列进行统一管理，可以实现针对不同的用户态进程的数据迁移设置不同优先级，进而对不同的业务进行QoS优化，提高虚拟机迁移的性能。

[0175] 实施例四

[0176] 本发明实施例提供一种宿主机Host60，该Host60与第一虚拟机同部署于一个物理主机，在一种实现方式下，该Host 60可以包括特权虚拟机和VMM，其中特权虚拟机和VMM与第一虚拟机同部署于一个物理主机，参见图6，该Host60可以包括：

[0177] 获取单元601，用于获取发送第一虚拟机内存数据的数据发送请求，其中，所述数据发送请求包含所述第一虚拟机的标识，以及请求发送的内存数据的PFN；

[0178] 第一查询单元602，用于根据所述第一虚拟机的标识查询对应关系信息库，得到所述第一虚拟机的对应关系，其中，所述对应关系信息库包含待迁移的虚拟机的对应关系，所述第一虚拟机的对应关系为基于所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间，以及所述地址空间对应的所述第一虚拟机内存的PFN，以及所述第一虚拟机内存的PFN对应的所述第一虚拟机内存的物理内存页地址后，建立的所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址的对应关系；

[0179] 第二查询单元603，用于根据所述请求发送的内存数据的PFN查询所述第一查询单元602得到的所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址；

[0180] 驱动发送单元604，用于通过RDMA网卡将所述第二查询单元603得到的请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。

[0181] 本发明实施例提供的Host60，该Host60可获取发送第一虚拟机内存数据的数据发送请求，其中，该数据发送请求包含第一虚拟机的标识，以及请求发送的内存数据的PFN；该Host根据第一虚拟机的标识查询对应关系信息库，获取第一虚拟机的对应关系，其中，该对应关系信息库包含待迁移的虚拟机的对应关系，虚拟机的对应关系为基于该虚拟机的标识获取该虚拟机内存的地址空间，以及该地址空间对应的该虚拟机内存的PFN，以及该虚拟机内存的PFN对应的该虚拟机内存的物理内存页地址后，建立的该虚拟机内存的PFN与物理内存页地址的对应关系；所以，该Host可根据请求发送的内存数据的PFN查询第一虚拟机的对应关系，获取请求发送的内存数据的物理内存页地址；再通过RDMA网卡将请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。如此，在虚拟机迁移的过程中，不需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，通过该待迁移虚拟的对应关系以及数据发送求中的PFN，可直接查询获取请求发送的内存数据对应的物理内存页地址，进而大大降低了虚拟机所在的物理机中的处理器利用率及时间开销，一定程度上解决了现有技术中由于在虚拟化平台下，虚拟机迁移的过程中，需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，导致虚拟机所在的物理机中的处理器利用率及时间开销较大的问题。

[0182] 实施例五

[0183] 参见图7，为本发明实施例提供的另一种Host70，参见图7，该Host70包括：建立单元701，第一注册单元702，获取单元703，第一查询单元704，第二查询单元705，驱动发送单

元706。

[0184] 其中，获取单元703，第一查询单元704，第二查询单元705，驱动发送单元706的具体功能参见实施例四中所述，在此不再赘述。

[0185] 建立单元701，用于获取特权虚拟机用户态进程触发的迁移所述第一虚拟机的内存数据的迁移请求，其中，所述迁移请求中包含所述第一虚拟机的标识；根据所述迁移请求中的所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间；根据所述地址空间获取所述第一虚拟机内存的PFN；根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址；建立所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址之间的对应关系；

[0186] 第一注册单元702，用于在建立单元701根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址之后，获取单元702获取发送第一虚拟机内存数据的数据发送请求之前，将建立单元701获取的第一虚拟机内存的物理内存页地址注册至所述RDMA网卡。

[0187] 本实施例中，第一注册单元702将第一虚拟机全部内存的物理内存页地址一次注册给RDMA网卡，例如，第一虚拟机全部内存对应的物理内存页地址为0x00010000～0x0001ffff，则第一注册单元702将0x00010000～0x0001ffff一次性全部注册至RDMA网卡。

[0188] 进一步的，所述获取单元601具体用于，依序提取数据发送请求队列中的数据发送请求，其中，所述数据发送请求包含不同的待迁移的虚拟机的数据发送请求，所述数据发送请求队列中的数据发送请求按照时间先后或者优先级高低进行排序。

[0189] 本发明实施例提供的Host70，用于获取特权虚拟机用户态进程发送的虚拟机的迁移命令，并根据每个虚拟机的迁移命令中包含的虚拟机的标识建立每个虚拟机内存的PFN与物理内存页地址的对应关系；根据每个虚拟机的对应关系，可以获取每个虚拟机内存的PFN对应的物理内存页地址，并将每个虚拟机的物理内存页地址注册至该RDMA网卡；再通过RDMA网卡将请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。如此，在虚拟机迁移的过程中，不需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，通过该待迁移虚拟的对应关系以及数据发送求中的PFN，可直接查询获取请求发送的内存数据对应的物理内存页地址，进而大大降低了虚拟机所在的物理机中的处理器利用率及时间开销，一定程度上解决了现有技术中由于在虚拟化平台下，虚拟机迁移的过程中，需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，导致虚拟机所在的物理机中的处理器利用率及时间开销较大的问题。

[0190] 而且，因为在驱动发送单元706通过RDMA网卡发送数据前，第一注册单元702把每个虚拟机的物理内存页地址一次性注册至该RDMA网卡，可以保证RDMA网卡根据用户态进程发送的数据发送请求成功传输相关的数据。

[0191] 而且，获取单元703可以将分别获取的每一个虚拟机对应的用户态进程的数据发送请求通过数据发送请求队列进行统一管理，可以实现针对不同的用户态进程的数据迁移设置不同优先级，进而对不同的业务进行QoS优化，提高虚拟机迁移的性能。

[0192] 实施例六

[0193] 参见图8，为本发明实施例提供的另一种Host80，参见图8，该Host80包括：建立单元801，获取单元802，第一查询单元803，第二查询单元804，第一注册单元805，驱动发送单

元806。

[0194] 其中，建立单元801，用于获取特权虚拟机用户态进程触发的迁移所述第一虚拟机的内存数据的迁移请求，其中，所述迁移请求中包含所述第一虚拟机的标识；根据所述迁移请求中的所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间；根据所述地址空间获取所述第一虚拟机内存的PFN；根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址；建立所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址之间的对应关系；

[0195] 获取单元802，用于获取发送第一虚拟机内存数据的数据发送请求，其中，所述数据发送请求包含所述第一虚拟机的标识，以及请求发送的内存数据的PFN；

[0196] 优选的，所述获取单元802具体用于，依序提取数据发送请求队列中的数据发送请求，其中，所述数据发送请求包含不同的待迁移的虚拟机的数据发送请求，所述数据发送请求队列中的数据发送请求按照时间先后或者优先级高低进行排序。

[0197] 第一查询单元803，用于根据所述第一虚拟机的标识查询对应关系信息库，得到所述第一虚拟机的对应关系，其中，所述对应关系信息库包含待迁移的虚拟机的对应关系，所述第一虚拟机的对应关系为基于所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间，以及所述地址空间对应的所述第一虚拟机内存的PFN，以及所述第一虚拟机内存的PFN对应的所述第一虚拟机内存的物理内存页地址后，建立的所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址的对应关系；

[0198] 第二查询单元804，用于根据所述请求发送的内存数据的PFN查询所述第一查询单元803得到的所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址；

[0199] 第二注册单元805，用于将第二查询单元804得到的得到所述请求发送的内存数据的物理内存页地址注册至RDMA网卡；

[0200] 本实施例中，第二注册单元805根据每次的数据发送请求，分别就将每个数据发送请求请求发送的内存数据的物理内存页地址分批次的注册至RDMA网卡，例如，第一虚拟机全部内存对应的物理内存页地址为0x00010000～0x0001ffff，第一个数据发送请求请求发送的内存数据的物理内存页地址为0x00010000～0x000100ff，则第二注册单元805在获取第一个数据发送请求之后，通过RDMA网卡将物理内存页地址0x00010000～0x000100ff上存储的内存数据发送给目的物理主机之前，将物理内存页地址0x00010000～0x0001ffff注册至RDMA网卡；再例如，第二个数据发送请求请求发送的内存数据的物理内存页地址为0x00010100～0x000101ff，则第二注册单元805在获取第二个数据发送请求之后，通过RDMA网卡将物理内存页地址0x00010100～0x000101ff上存储的内存数据发送给目的物理主机之前，将物理内存页地址0x00010100～0x000101ff注册至RDMA网卡。

[0201] 驱动发送单元806，用于在第二注册单元805将第二查询单元804得到的得到所述请求发送的内存数据的物理内存页地址注册至RDMA网卡后，通过RDMA网卡将所述第二查询单元804得到的请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。

[0202] 本发明实施例提供的Host80，用于获取特权虚拟机用户态进程触发的虚拟机的迁移命令，并根据虚拟机的迁移命令中包含的虚拟机的标识建立虚拟机内存的PFN与物理内

存页地址的对应关系;根据虚拟机的对应关系,可以获取虚拟机内存的PFN对应的物理内存页地址,再通过RDMA网卡发送数据之前将虚拟机的物理内存页地址注册至该RDMA网卡;再通过RDMA网卡将请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。如此,在虚拟机迁移的过程中,不需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表,通过该待迁移虚拟的对应关系以及数据发送求中的PFN,可直接查询获取请求发送的内存数据对应的物理内存页地址,进而大大降低了虚拟机所在的物理机中的处理器利用率及时间开销,一定程度上解决了现有技术中由于在虚拟化平台下,虚拟机迁移的过程中,需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表,导致虚拟机所在的物理机中的处理器利用率及时间开销较大的问题。

[0203] 而且,可以将每一个虚拟机对应的用户态进程分别发送的数据发送请求通过数据发送请求队列进行统一管理,可以实现针对不同的用户态进程的数据迁移设置不同优先级,进而对不同的业务进行QoS优化,提高虚拟机迁移的性能。

[0204] 而且,获取单元703可以将分别获取的每一个虚拟机对应的用户态进程的数据发送请求通过数据发送请求队列进行统一管理,可以实现针对不同的用户态进程的数据迁移设置不同优先级,进而对不同的业务进行QoS优化,提高虚拟机迁移的性能。

[0205] 实施例七

[0206] 本发明实施例提供一种物理主机90,参见图9,该物理主机90包括硬件,其中所述硬件可以包括RDMA网卡901,可选的,所述硬件还可以包括至少一个处理器902、存储器903,用于进行该物理主机90内部各设备之间的连接的至少一个通信总线904,用于实现这些装置之间的连接和相互通信。

[0207] 其中,通信总线904可以是工业标准体系结构(Industry Standard Architecture,简称为ISA)总线、外部设备互连(Peripheral Component,简称为PCI)总线或扩展工业标准体系结构(Extended Industry Standard Architecture,简称为EISA)总线等。该总线904可以分为地址总线、数据总线、控制总线等。为便于表示,图9中仅用一条粗线表示,但并不表示仅有一根总线或一种类型的总线。

[0208] 存储器903可以包括随机存取存储器,并向处理器803提供指令和数据。

[0209] 处理器902可以是一个中央处理器(Central Processing Unit,简称为CPU),或者是特定集成电路(Application Specific Integrated Circuit,简称为ASIC),或者是被配置成实施本发明实施例的一个或多个集成电路。

[0210] RDMA网卡901可以为支持RDMA功能的各种网卡,例如,可以为无线宽带(InfiniBand)卡或以太网(RDMA over Converged Ethernet,简称RoCE)卡等。

[0211] 其中,通过读取存储器903存储的指令,处理器902用于:

[0212] 获取发送第一虚拟机内存数据的数据发送请求,其中,所述数据发送请求包含所述第一虚拟机的标识,以及请求发送的内存数据的PFN;

[0213] 根据所述第一虚拟机的标识查询对应关系信息库,得到所述第一虚拟机的对应关系,其中,所述对应关系信息库包含待迁移的虚拟机的对应关系,所述第一虚拟机的对应关系为基于所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间,以及所述地址空间对应的所述第一虚拟机内存的PFN,以及所述第一虚拟机内存的PFN对应的所述第一虚拟机内存的物理内存页地址后,建立的所述第一虚拟机内存的PFN与所述第一虚拟机内存的物

理内存页地址的对应关系；

[0214] 根据所述请求发送的内存数据的PFN查询获取的所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址；

[0215] 通过RDMA网卡901将获取的请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。

[0216] 进一步的，所述处理器902还用于：

[0217] 在获取发送第一虚拟机内存数据的数据发送请求之前，获取特权虚拟机用户态进程触发的迁移所述第一虚拟机的内存数据的迁移请求，其中，所述迁移请求中包含所述第一虚拟机的标识；

[0218] 根据所述迁移请求中的所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间；

[0219] 根据所述地址空间获取所述第一虚拟机内存的PFN；

[0220] 根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址；

[0221] 建立所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址之间的对应关系。

[0222] 进一步的，所述处理器902还用于：将第一虚拟机内存数据的物理页内存地址注册给RDMA网卡，具体可以通过以下两种方式中任意一种方式来实现：

[0223] 1、在根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址之后，获取发送第一虚拟机内存数据的数据发送请求之前，将所述第一虚拟机内存的物理内存页地址注册至所述RDMA网卡。

[0224] 在这种方式下，处理器902将第一虚拟机全部内存的物理内存页地址一次注册给RDMA网卡，例如，第一虚拟机全部内存对应的物理内存页地址为0x00010000～0x0001ffff，则Host901将0x00010000～0x0001ffff一次性全部注册至RDMA网卡。

[0225] 2、在根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址之后，通过RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机之前，将所述请求发送的内存数据的物理内存页地址注册至所述RDMA网卡。

[0226] 在这种方式下，处理器902根据每次的数据发送请求，分别就将每个数据发送请求请求发送的内存数据的物理内存页地址分批次的注册至RDMA网卡，例如，第一虚拟机全部内存对应的物理内存页地址为0x00010000～0x0001ffff，第一个数据发送请求请求发送的内存数据的物理内存页地址为0x00010000～0x000100ff，则处理器902在获取第一个数据发送请求之后，通过RDMA网卡将物理内存页地址0x00010000～0x000100ff上存储的内存数据发送给目的物理主机之前，将物理内存页地址0x00010000～0x0001ffff注册至RDMA网卡；再例如，第二个数据发送请求请求发送的内存数据的物理内存页地址为0x00010100～0x000101ff，则Host901在获取第二个数据发送请求之后，通过RDMA网卡将物理内存页地址0x00010100～0x000101ff上存储的内存数据发送给目的物理主机之前，将物理内存页地址0x00010100～0x000101ff注册至RDMA网卡。

[0227] 本发明实施例提供的物理主机90，可获取发送第一虚拟机内存数据的数据发送请求，其中，该数据发送请求包含第一虚拟机的标识，以及请求发送的内存数据的PFN；该Host

根据第一虚拟机的标识查询对应关系信息库，获取第一虚拟机的对应关系，其中，该对应关系信息库包含待迁移的虚拟机的对应关系，虚拟机的对应关系为基于该虚拟机的标识获取该虚拟机内存的地址空间，以及该地址空间对应的该虚拟机内存的PFN，以及该虚拟机内存的PFN对应的该虚拟机内存的物理内存页地址后，建立的该虚拟机内存的PFN与物理内存页地址的对应关系；所以，该Host可根据请求发送的内存数据的PFN查询第一虚拟机的对应关系，获取请求发送的内存数据的物理内存页地址；再通过RDMA网卡将请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。如此，在虚拟机迁移的过程中，不需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，通过该待迁移虚拟的对应关系以及数据发送求中的PFN，可直接查询获取请求发送的内存数据对应的物理内存页地址，进而大大降低了虚拟机所在的物理机中的处理器利用率及时间开销，一定程度上解决了现有技术中由于在虚拟化平台下，虚拟机迁移的过程中，需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，导致虚拟机所在的物理机中的处理器利用率及时间开销较大的问题。

[0228] 实施例八

[0229] 本发明实施例提供另一种物理主机100，参见图10，该物理主机100包含硬件层、运行在所述硬件层之上的Host1001、以及运行在所述Host1001之上的至少一个虚拟机VM1002，以及所述硬件层包括RDMA网卡1003，可选的，还可以包括处理器1004和内存1005；其中，Host可以包括该物理主机100上的VMM和运行于该VMM上的特权虚拟机，虚拟机1002为该物理主机100上除特权虚拟机之外的其他虚拟机，虚拟机1002包括第一虚拟机。

[0230] 其中，Host1001用于：

[0231] 获取发送第一虚拟机内存数据的数据发送请求，其中，所述数据发送请求包含所述第一虚拟机的标识，以及请求发送的内存数据的PFN；

[0232] 根据所述第一虚拟机的标识查询对应关系信息库，得到所述第一虚拟机的对应关系，其中，所述对应关系信息库包含待迁移的虚拟机的对应关系，所述第一虚拟机的对应关系为基于所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间，以及所述地址空间对应的所述第一虚拟机内存的PFN，以及所述第一虚拟机内存的PFN对应的所述第一虚拟机内存的物理内存页地址后，建立的所述第一虚拟机内存的PFN与所述第一虚拟机内存物理内存页地址的对应关系；

[0233] 根据所述请求发送的内存数据的PFN查询获取的所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址；

[0234] 通过RDMA网卡1003将获取的请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。

[0235] 进一步的，所述Host1001还用于：

[0236] 在获取发送第一虚拟机内存数据的数据发送请求之前，获取特权虚拟机用户态进程触发的迁移所述第一虚拟机的内存数据的迁移请求，其中，所述迁移请求中包含所述第一虚拟机的标识；

[0237] 根据所述迁移请求中的所述第一虚拟机的标识获取所述第一虚拟机内存的地址空间；

[0238] 根据所述地址空间获取所述第一虚拟机内存的PFN；

[0239] 根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址；

[0240] 建立所述第一虚拟机内存的PFN与所述第一虚拟机内存的物理内存页地址之间的对应关系。

[0241] 进一步的，所述Host1001还用于：

[0242] 将第一虚拟机内存数据的物理页内存地址注册给RDMA网卡，具体可以通过以下两种方式中任意一种方式来实现：

[0243] 1、在根据所述第一虚拟机内存的PFN计算所述第一虚拟机内存的物理内存页地址之后，获取发送第一虚拟机内存数据的数据发送请求之前，将所述第一虚拟机内存的物理内存页地址注册至所述RDMA网卡。

[0244] 在这种方式下，Host1001将第一虚拟机全部内存的物理内存页地址一次注册给RDMA网卡，例如，第一虚拟机全部内存对应的物理内存页地址为0x00010000～0x0001ffff，则Host901将0x00010000～0x0001ffff一次性全部注册至RDMA网卡。

[0245] 2、在根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址之后，通过RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机之前，将所述请求发送的内存数据的物理内存页地址注册至所述RDMA网卡。

[0246] 在这种方式下，Host1001根据每次的数据发送请求，分别就将每个数据发送请求请求发送的内存数据的物理内存页地址分批次的注册至RDMA网卡，例如，第一虚拟机全部内存对应的物理内存页地址为0x00010000～0x0001ffff，第一个数据发送请求请求发送的内存数据的物理内存页地址为0x00010000～0x000100ff，则Host1001在获取第一个数据发送请求之后，通过RDMA网卡将物理内存页地址0x00010000～0x000100ff上存储的内存数据发送给目的物理主机之前，将物理内存页地址0x00010000～0x0001ffff注册至RDMA网卡；再例如，第二个数据发送请求请求发送的内存数据的物理内存页地址为0x00010100～0x000101ff，则Host1001在获取第二个数据发送请求之后，通过RDMA网卡将物理内存页地址0x00010100～0x000101ff上存储的内存数据发送给目的物理主机之前，将物理内存页地址0x00010100～0x000101ff注册至RDMA网卡。

[0247] 本发明实施例的物理主机100，用于获取发送第一虚拟机内存数据的数据发送请求，其中，该数据发送请求包含第一虚拟机的标识，以及请求发送的内存数据的PFN；该Host根据第一虚拟机的标识查询对应关系信息库，获取第一虚拟机的对应关系，其中，该对应关系信息库包含待迁移的虚拟机的对应关系，虚拟机的对应关系为基于该虚拟机的标识获取该虚拟机内存的地址空间，以及该地址空间对应的该虚拟机内存的PFN，以及该虚拟机内存的PFN对应的该虚拟机内存的物理内存页地址后，建立的该虚拟机内存的PFN与物理内存页地址的对应关系；所以，可根据请求发送的内存数据的PFN查询第一虚拟机的对应关系，获取请求发送的内存数据的物理内存页地址；再通过RDMA网卡将请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。如此，在虚拟机迁移的过程中，不需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，通过该待迁移虚拟的对应关系以及数据发送求中的PFN，可直接查询获取请求发送的内存数据对应的物理内存页地址，进而大大降低了虚拟机所在的物理机中的处理器利用率及时间开销，一定程度上解决了现有技术中由于在虚拟化平台下，虚拟机迁移的过程中，需要将虚拟机的内存映射到用户态空

间并通过超级调用更新页表，导致虚拟机所在的物理机中的处理器利用率及时间开销较大的问题。

[0248] 实施例九

[0249] 本发明实施例提供一种集群系统，参见图10，该集群系统包括：多台物理主机，所述多台物理主机包括源物理主机和目的物理主机，其中每台物理主机包括硬件层、运行在所述硬件层之上的Host、以及运行在所述Host之上的至少一个虚拟机VM，以及所述硬件层包括RDMA网卡，可选的，还可以处理器和内存。其中，Host可以包括Host所在的物理主机上的VMM和运行于该VMM上的特权虚拟机。

[0250] 以及，本发明实施例的集群系统中包括的物理主机参考前述实施例介绍的物理主机，在此不再赘述。

[0251] 本发明实施例的集群系统，源物理主机中的Host可获取发送第一虚拟机内存数据的数据发送请求，其中，该数据发送请求包含第一虚拟机的标识，以及请求发送的内存数据的PFN；该Host根据第一虚拟机的标识查询对应关系信息库，获取第一虚拟机的对应关系，其中，该对应关系信息库包含待迁移的虚拟机的对应关系，虚拟机的对应关系为基于该虚拟机的标识获取该虚拟机内存的地址空间，以及该地址空间对应的该虚拟机内存的PFN，以及该虚拟机内存的PFN对应的该虚拟机内存的物理内存页地址后，建立的该虚拟机内存的PFN与物理内存页地址的对应关系；所以，该Host可根据请求发送的内存数据的PFN查询第一虚拟机的对应关系，获取请求发送的内存数据的物理内存页地址；再通过RDMA网卡将请求发送的内存数据的物理内存页地址存储的内存数据发送至目的物理主机。如此，在虚拟机迁移的过程中，不需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，通过该待迁移虚拟的对应关系以及数据发送求中的PFN，可直接查询获取请求发送的内存数据对应的物理内存页地址，进而大大降低了虚拟机所在的物理机中的处理器利用率及时间开销，一定程度上解决了现有技术中由于在虚拟化平台下，虚拟机迁移的过程中，需要将虚拟机的内存映射到用户态空间并通过超级调用更新页表，导致虚拟机所在的物理机中的处理器利用率及时间开销较大的问题。

[0252] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统，装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0253] 在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0254] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0255] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理包括，也可以两个或两个以上单元集成在一个单元中。上述集成的单

元既可以采用硬件的形式实现，也可以采用硬件加软件功能单元的形式实现。

[0256] 上述以软件功能单元的形式实现的集成的单元，可以存储在一个计算机可读取存储介质中。上述软件功能单元存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述方法的部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(Read-Only Memory，简称ROM)、随机存取存储器(Random Access Memory，简称RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0257] 本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令相关的硬件(例如处理器)来完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：只读存储器、随机存储器、磁盘或光盘等。

[0258] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

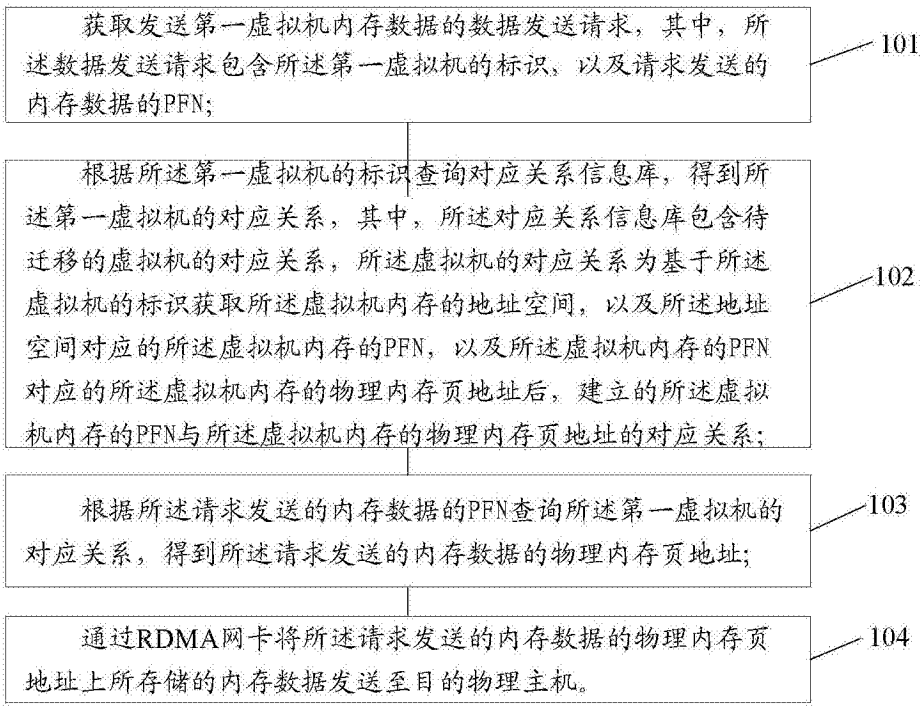

获取发送第一虚拟机内存数据的数据发送请求，其中，所述数据发送请求包含所述第一虚拟机的标识，以及请求发送的内存数据的PFN;
101
根据所述第一虚拟机的标识查询对应关系信息库，得到所述第一虚拟机的对应关系，其中，所述对应关系信息库包含待迁移的虚拟机的对应关系，所述虚拟机的对应关系为基于所述虚拟机的标识获取所述虚拟机内存的地址空间，以及所述地址空间对应的所述虚拟机内存的PFN，以及所述虚拟机内存的PFN对应的所述虚拟机内存的物理内存页地址后，建立的所述虚拟机内存的PFN与所述虚拟机内存的物理内存页地址的对应关系;
102
根据所述请求发送的内存数据的PFN查询所述第一虚拟机的对应关系，得到所述请求发送的内存数据的物理内存页地址;
103
通过RDMA网卡将所述请求发送的内存数据的物理内存页地址上所存储的内存数据发送至目的物理主机。
104

图1

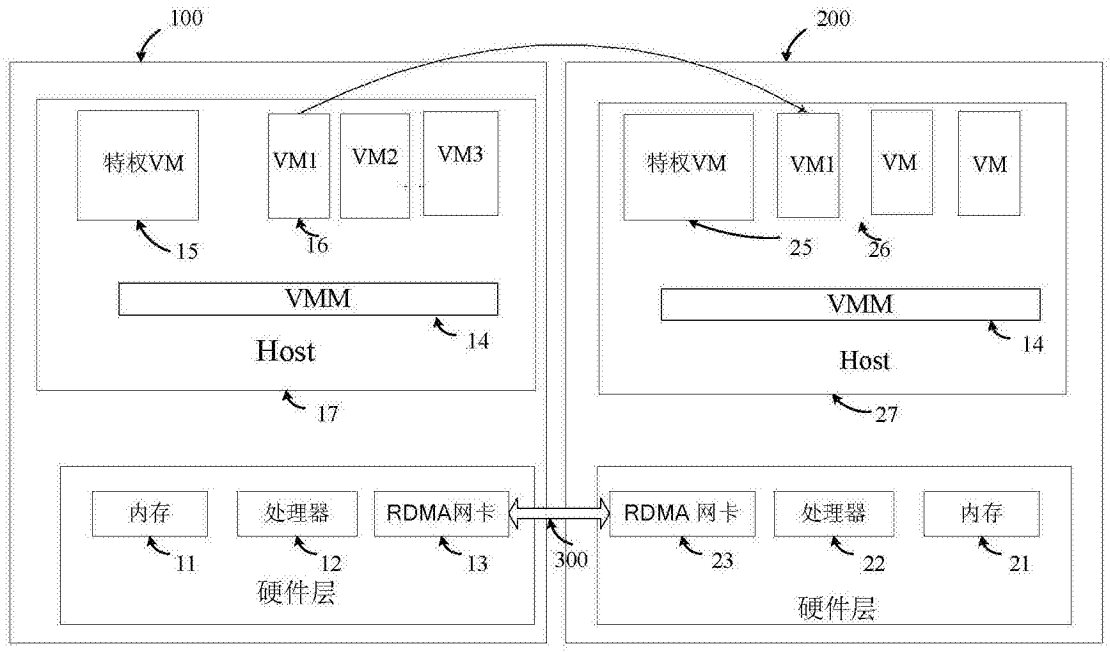

100
200
特权VM
VM1
VM2
VM3
15
16
VMM
14
Host
17
特权VM
VM1
VM
VM
25
26
VMM
14
Host
27
内存
处理器
RDMA网卡
11
12
13
300
RDMA 网卡
处理器
内存
23
22
21
硬件层
硬件层

图2

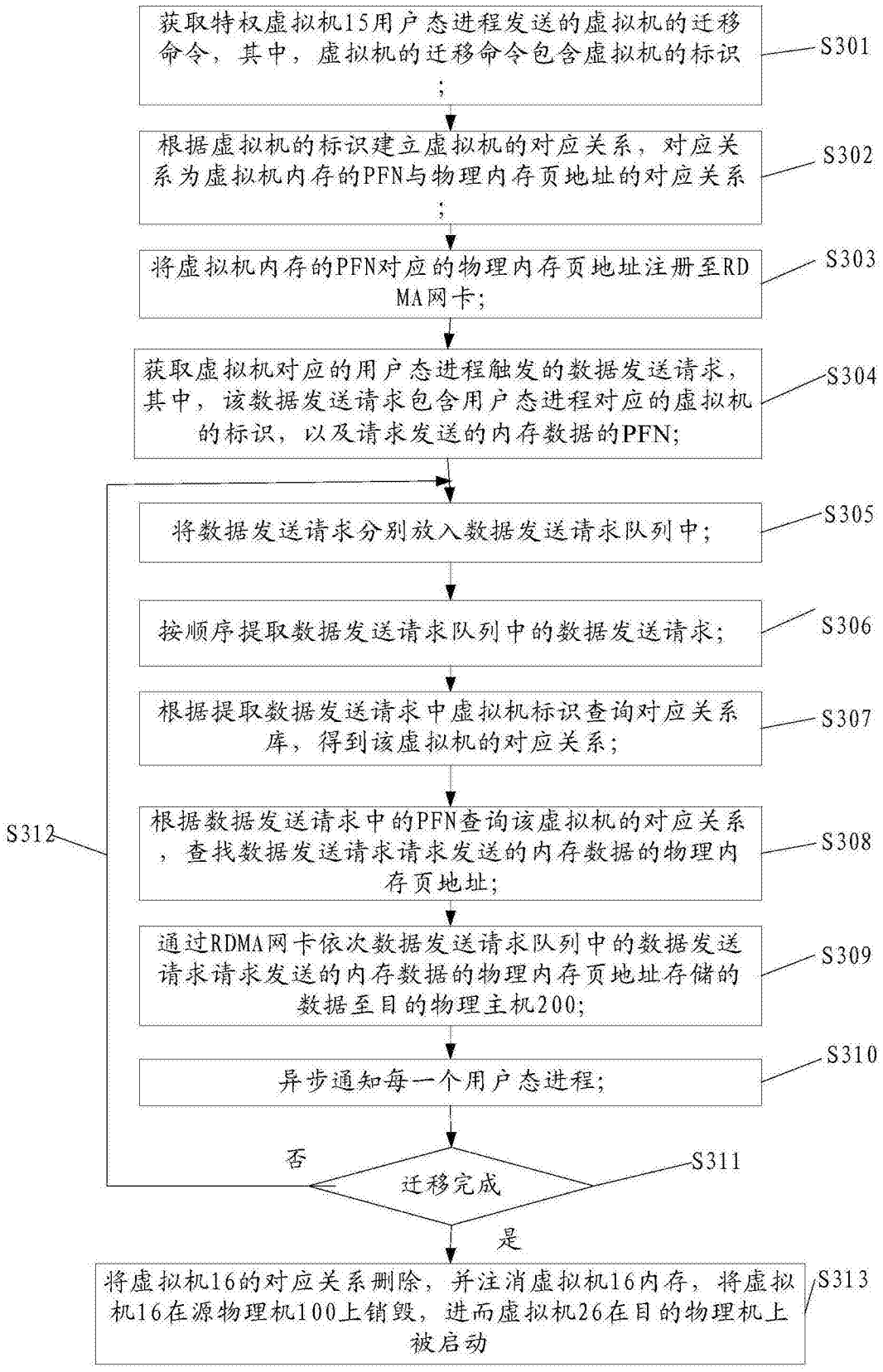

获取特权虚拟机15用户态进程发送的虚拟机的迁移命令，其中，虚拟机的迁移命令包含虚拟机的标识；
S301
根据虚拟机的标识建立虚拟机的对应关系，对应关系为虚拟机内存的PFN与物理内存页地址的对应关系；
S302
将虚拟机内存的PFN对应的物理内存页地址注册至RDMA网卡；
S303
获取虚拟机对应的用户态进程触发的数据发送请求，其中，该数据发送请求包含用户态进程对应的虚拟机的标识，以及请求发送的内存数据的PFN；
S304
将数据发送请求分别放入数据发送请求队列中；
S305
按顺序提取数据发送请求队列中的数据发送请求；
S306
根据提取数据发送请求中虚拟机标识查询对应关系库，得到该虚拟机的对应关系；
S307
根据数据发送请求中的PFN查询该虚拟机的对应关系，查找数据发送请求请求发送的内存数据的物理内存页地址；
S308
S312
通过RDMA网卡依次数据发送请求队列中的数据发送请求请求发送的内存数据的物理内存页地址存储的数据至目的物理主机200；
S309
异步通知每一个用户态进程；
S310
否
迁移完成
S311
是
将虚拟机16的对应关系删除，并注消虚拟机16内存，将虚拟机16在源物理机100上销毁，进而虚拟机26在目的物理机上被启动
S313

图3

数据发送请求1	数据发送请求2	数据发送请求3	…

图4

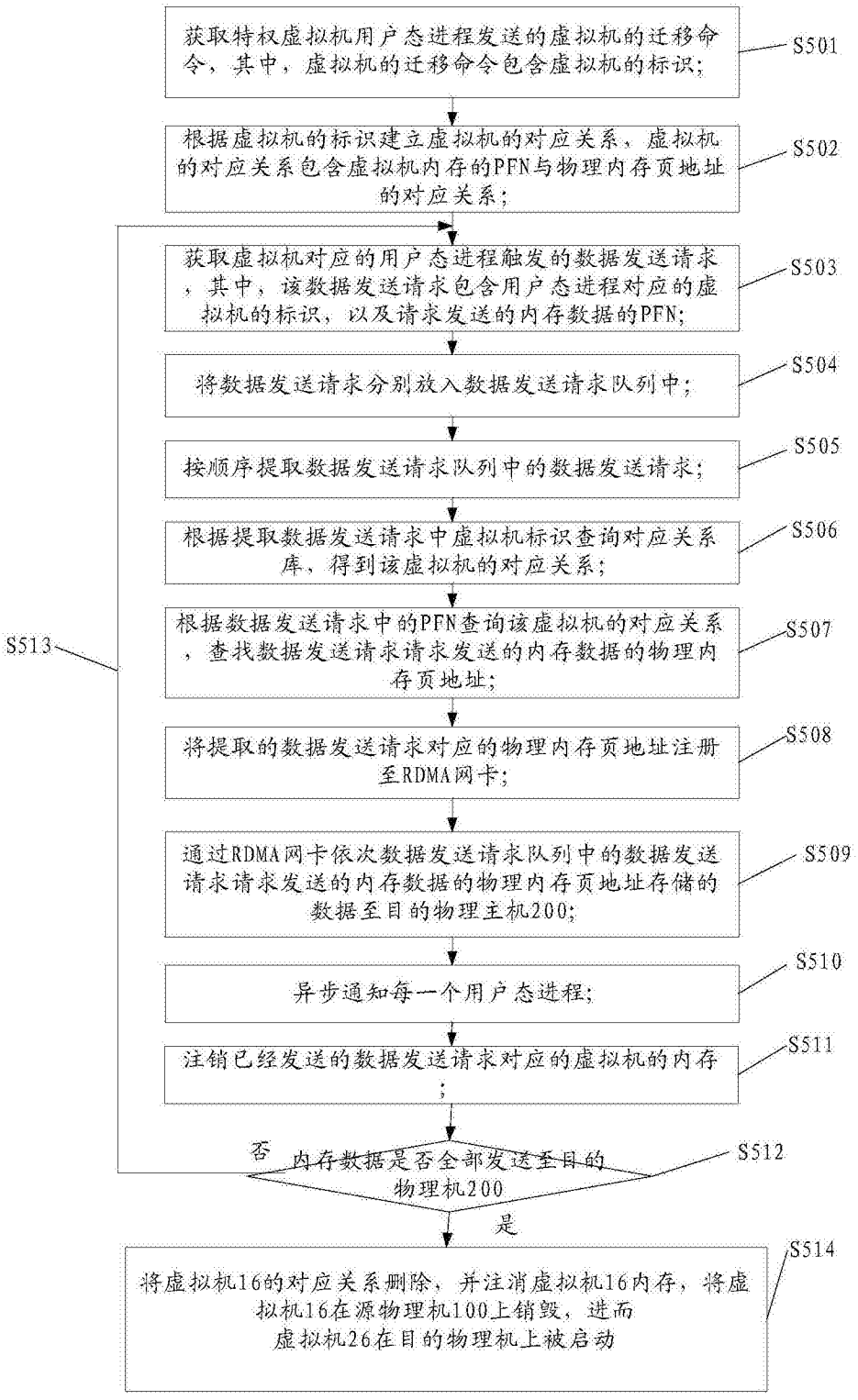


图5

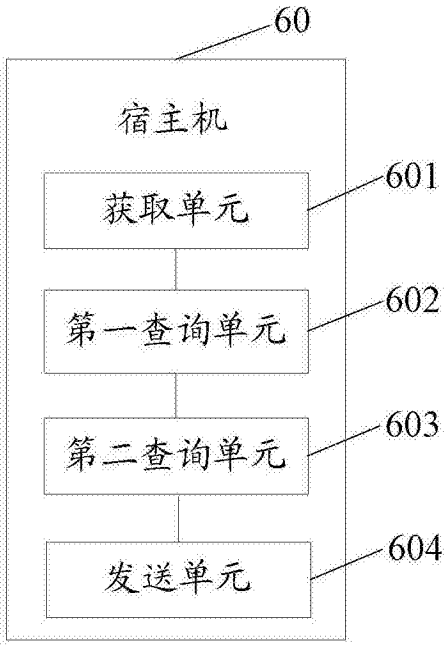

60
宿主机
获取单元
601
第一查询单元
602
第二查询单元
603
发送单元
604

图6

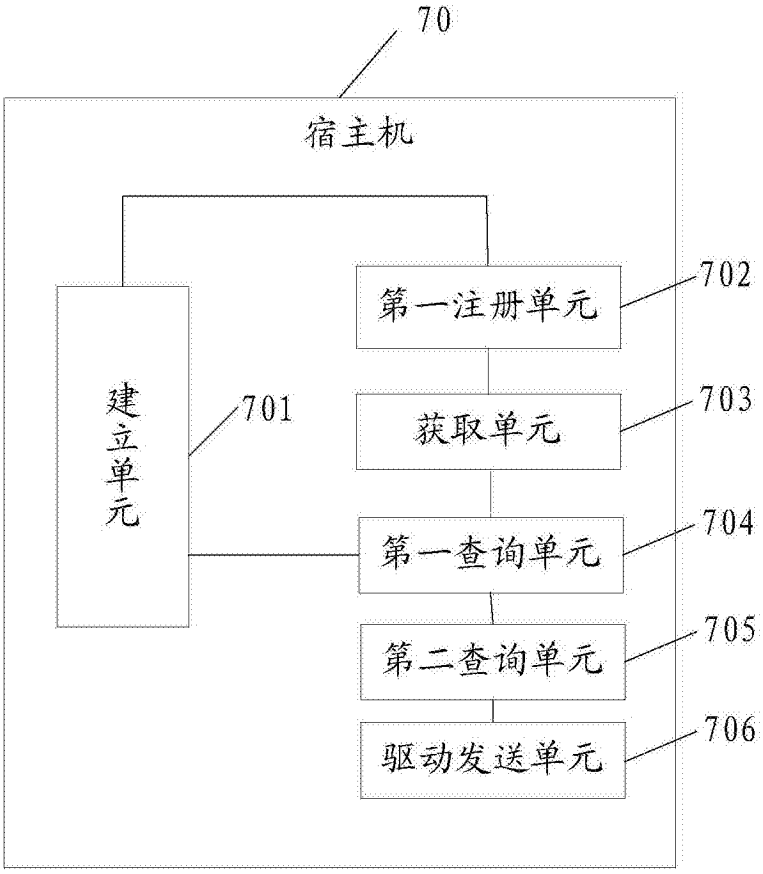

70
宿主机
建立单元
701
第一注册单元
702
获取单元
703
第一查询单元
704
第二查询单元
705
驱动发送单元
706

图7

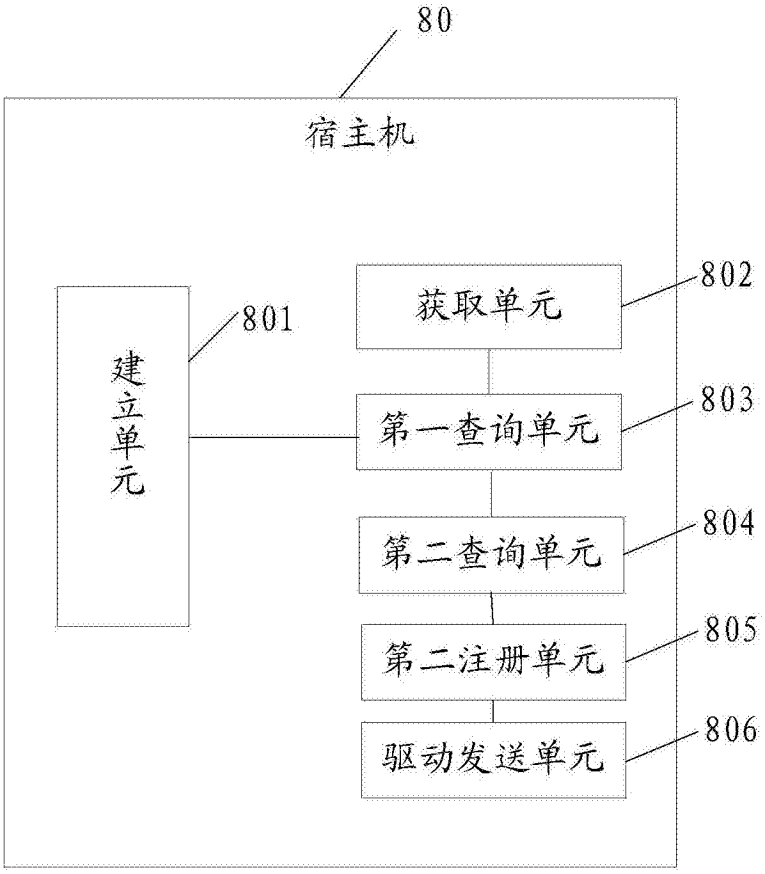

80
宿主机
建立单元
801
获取单元
802
第一查询单元
803
第二查询单元
804
第二注册单元
805
驱动发送单元
806

图8

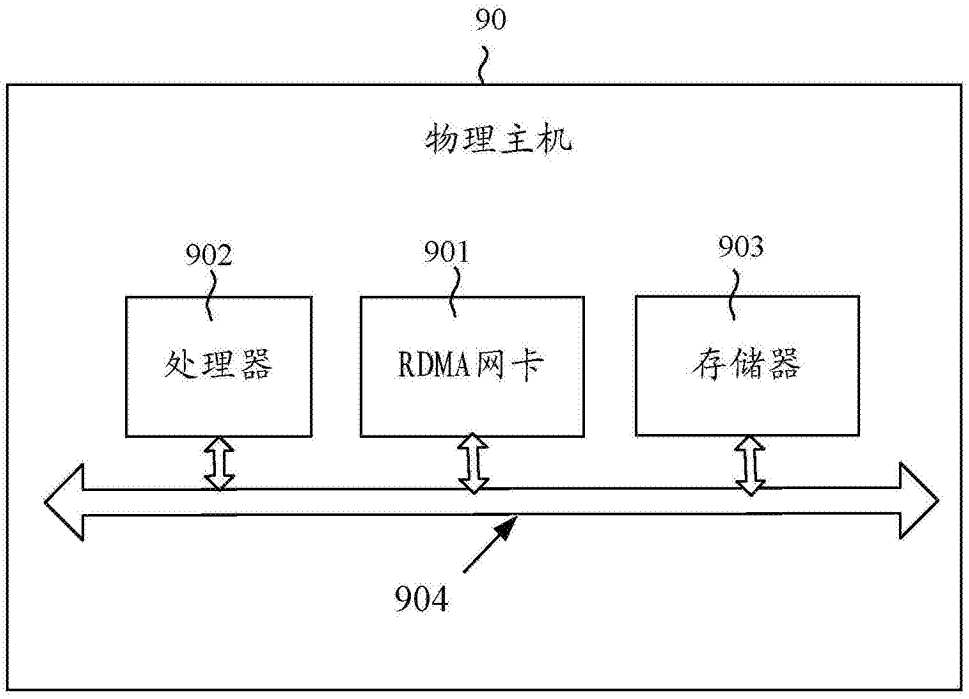

90
物理主机
902
处理器
901
RDMA网卡
903
存储器
904

图9

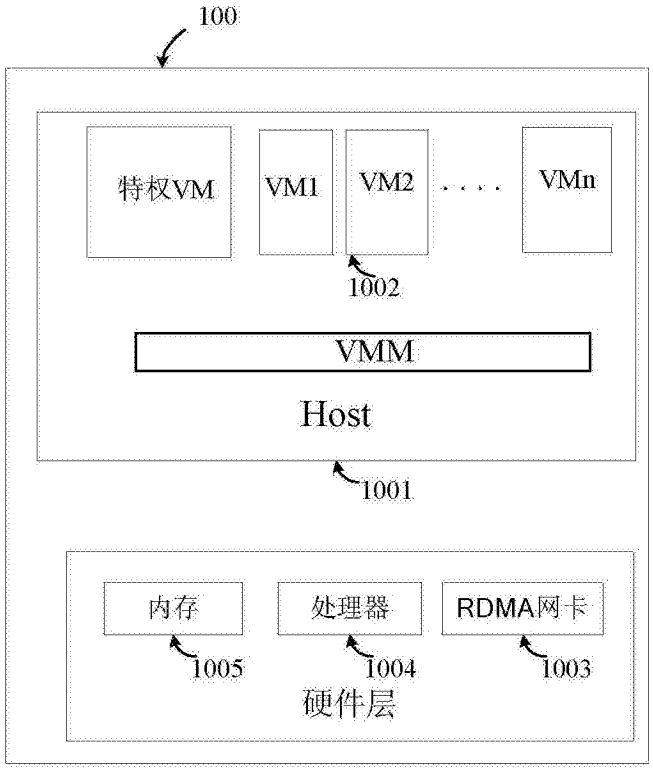

100
特权VM
VM1
VM2
. . . .
VMn
1002
VMM
Host
1001
内存
处理器
RDMA网卡
1005
1004
1003
硬件层

图10

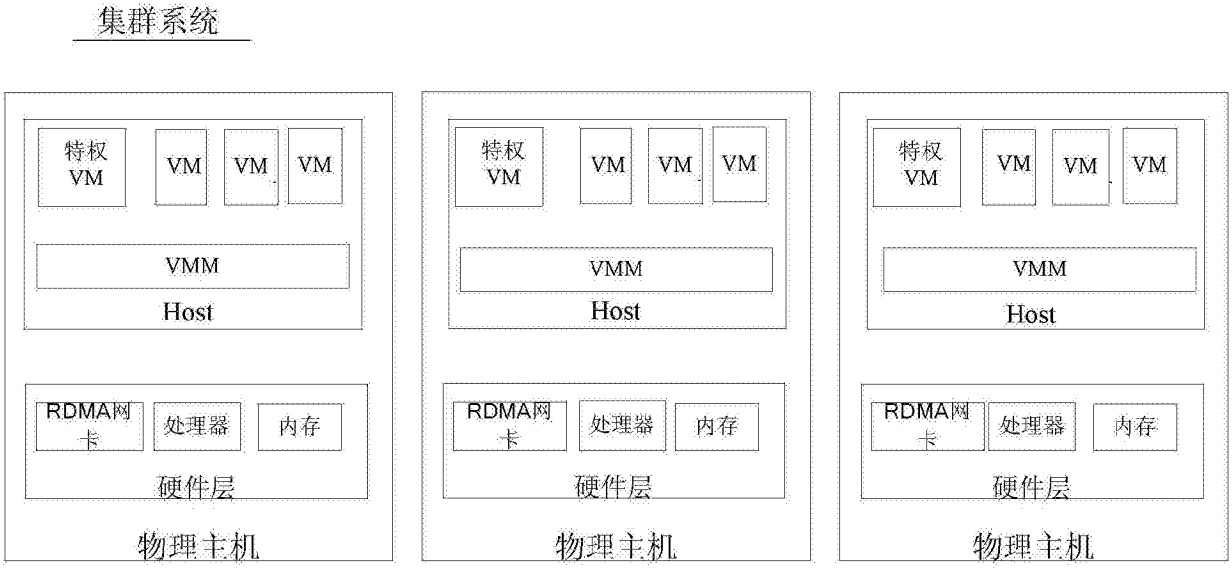

集群系统
特权VM
VM
VM
VM
VMM
Host
RDMA网卡
处理器
内存
硬件层
物理主机
特权VM
VM
VM
VM
VMM
Host
RDMA网卡
处理器
内存
硬件层
物理主机
特权VM
VM
VM
VM
VMM
Host
RDMA网卡
处理器
内存
硬件层
物理主机

图11