(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2023 年 6 月 1 日 (01.06.2023)

WIPO | PCT

(10) 国际公布号
WO 2023/093418 A1

(51) 国际专利分类号:
G06F 9/455 (2006.01)

(21) 国际申请号: PCT/CN2022/127151

(22) 国际申请日: 2022 年 10 月 24 日 (24.10.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202111426045.1 2021年11月26日 (26.11.2021) CN

(71) 申请人: 华 为 技 术 有 限 公 司 (**HUAWEI TECHNOLOGIES CO., LTD.**) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(72) 发明人: 卢 胜 文 (**LU, Shengwen**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 北京中博世达专利商标代理有限公司 (**BEIJING ZBSD PATENT & TRADEMARK AGENT LTD.**); 中国北京市海淀区交大东路31号11号楼8层, Beijing 100044 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE,

WO 2023/093418 A1

(54) **Title:** DATA MIGRATION METHOD AND APPARATUS, AND ELECTRONIC DEVICE

(54) 发明名称: 数据迁移方法、装置及电子设备

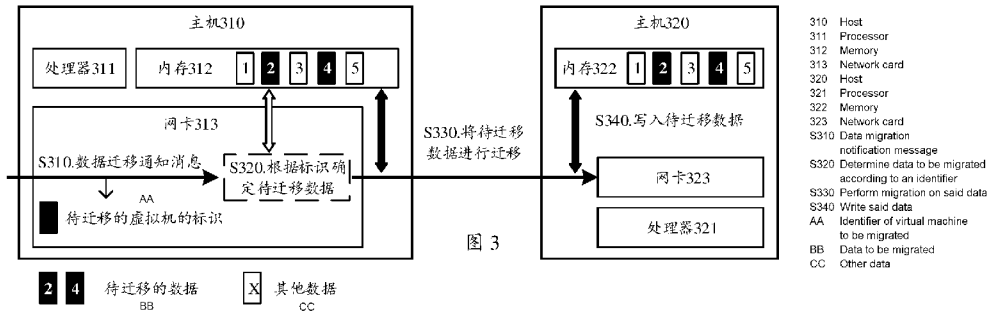

图 3

(57) **Abstract:** A data migration method and apparatus, and an electronic device, related to the technical field of computers. The data migration method comprises: first, a network card of a source host obtains a data migration notification message, the data migration notification message being used for indicating an identifier of a virtual machine to be migrated; secondly, the network card determines data to be migrated according to the identifier, wherein said data is data stored in a memory of the source host and associated with said virtual machine; and finally, the network card migrates said data to a destination host. Data to be migrated is determined by the network card according to the identifier of said virtual machine, such that the process of determining said data of the virtual machine by means of the source host is avoided, the computing resources required by the source host in the hot migration process of the virtual machine are reduced, and the capability of executing other services (such as AI, HPC, and other computing intensive and time-delay sensitive services) by the source host is improved.

(57) **摘要**: 一种数据迁移方法、装置及电子设备，涉及计算机技术领域。该数据迁移方法包括：首先，源主机的网卡获取数据迁移通知消息，该数据迁移通知消息用于指示待迁移的虚拟机的标识；其次，网卡根据标识确定待迁移数据，该待迁移数据为存储在源主机的内存中，且与待迁移的虚拟机关联的数据；最后，网卡将待迁移数据迁移至目的主机。待迁移数据是由网卡依据待迁移的虚拟机的标识确定的，避免了源主机对虚拟机的待迁移数据进行确定的过程，减少了源主机在虚拟机热迁移过程中所需的计算资源，提高了源主机执行其他业务(如AI、HPC等计算密集型和时延敏感型业务)的能力。

[见续页]

PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) **指定国**(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

数据迁移方法、装置及电子设备

本申请要求于2021年11月26日提交国家知识产权局、申请号为202111426045.1、申请名称为“数据迁移方法、装置及电子设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及计算机技术领域，尤其涉及一种数据迁移方法、装置及电子设备。

背景技术

虚拟机（virtual machine，VM）指通过虚拟化技术将物理的计算资源、存储资源和网络资源进行虚拟获得的虚拟设备。运行虚拟机所在物理设备称为宿主机。由于宿主机存在硬件故障等问题，为了保证虚拟机的正常运行，通常采用热迁移技术实现将VM从源宿主机（也可以称为源主机）迁移到目的宿主机（也可以称为目的主机），来实现硬件的双机容错或者负载均衡功能。为了保证业务不中断，热迁移技术需包括VM的内存数据的迁移，这就涉及内存中脏页数据的处理，而脏页是指内存中数据发生修改的存储空间。通常地，由源主机的处理器执行热迁移处理，包括识别脏页、搬迁脏页中数据、以及指示网卡向目的主机发送上述脏页中数据。上述过程完全依赖于源主机的处理器的处理能力和输入/输出（input/output，I/O）带宽，对于部署人工智能（artificial intelligent，AI）、高性能计算（high performance computing，HPC）等计算密集型和时延敏感型业务而言，现有方案无法满足业务性能需求，因此，如何提供一种更高效的数据迁移方法成为亟待解决的技术问题。

发明内容

本申请提供一种数据迁移方法、装置及电子设备，解决了源主机在虚拟机的数据迁移过程中，业务性能降低的问题。

第一方面，提供了一种数据迁移方法，该数据迁移方法可应用于源主机，或者支持实现该数据迁移方法的物理设备，例如该物理设备包括芯片系统。例如，源主机包括内存和网卡，该数据迁移方法由源主机的网卡执行，该数据迁移方法包括：首先，源主机的网卡获取第一数据迁移通知消息，该第一数据迁移通知消息用于指示待迁移的虚拟机的标识。其次，源主机的网卡根据标识确定待迁移数据，该待迁移数据为存储在源主机的内存中，且与前述待迁移的虚拟机关联的数据。最后，源主机的网卡将待迁移数据迁移至目的主机。

在本申请中待迁移数据是由网卡依据待迁移的虚拟机的标识确定的，避免了源主机对虚拟机的待迁移数据进行确定的过程，减少了源主机在虚拟机热迁移过程中所需的计算资源，提高了源主机执行其他业务（如AI、HPC等计算密集型和时延敏感型业务）的能力。

在一种可选的实现方式中，网卡根据标识确定待迁移数据，包括：网卡根据标识确定源主机的内存中与待迁移的虚拟机关联的内存页集合，并将内存页集合中存储的数据作为待迁移数据。其中，内存页集合包括一个或多个内存页。由网卡依据待迁移的虚拟机的标识确定待迁移数据，避免了由源主机的处理器确定待迁移数据导致源主机的性能下降，提高了源主机处理其他业务的能力，减少了源主机的卡顿。

在另一种可选的实现方式中，上述实施例提供的待迁移数据包括脏页数据，脏页数据

为一个或多个内存页中，数据发生修改的内存页所存储的数据。

在另一种可选的实现方式中，网卡根据标识确定待迁移数据包括的脏页数据，包括：网卡查询网卡中保存的脏页标记信息，确定与标识关联的脏页集合；继而，网卡将该脏页集合中存储的数据作为脏页数据。其中，脏页集合包括一个或多个脏页，该脏页为一个或多个内存页中数据发生修改的内存页，前述的脏页标记信息用于指示脏页的内存地址。

由网卡来对源主机的内存中与待迁移的虚拟机关联的内存页中的脏页进行标记，即内存中的标脏功能由源主机的处理器卸载到了网卡，避免了处理器标记内存中脏页的过程，减少了处理器的资源消耗，进而，避免了由于处理器管理虚拟机的热迁移过程导致的源主机的其他计算业务受到影响。

在另一种可选的实现方式中，脏页标记信息包括第一脏页表和第二脏页表中至少一个。其中，第一脏页表用于标记脏页为标脏状态，该标脏状态为源主机对脏页中存储的数据进行修改的状态；第二脏页表用于标记脏页为数据迁移状态，该数据迁移状态为网卡对脏页中存储的数据进行迁移的状态。

在另一种可选的实现方式中，网卡将待迁移数据迁移至目的主机，包括：第一，网卡向目的主机发送待迁移数据的页面信息，页面信息用于指示待迁移数据在内存中的内存地址和偏移量。第二，网卡接收目的主机基于页面信息反馈的迁移消息，迁移消息用于指示目的主机中与待迁移的虚拟机相应的接收队列（receive queue，RQ）。第三，网卡向迁移消息指示的 RQ 发送待迁移数据。在目的主机中网卡中存在多个 RQ，一个 RQ 对应一个虚拟机的情况下，由目的主机中网卡依据源主机中网卡发送的页面信息，为源主机中待迁移的虚拟机分配迁移消息，并由源主机中网卡向该迁移消息指示的 RQ 迁移待迁移数据，避免了待迁移的虚拟机（如 VM1）的待迁移数据被发送到其他虚拟机（如 VM2）对应的 RQ，提高了虚拟机的数据迁移准确性。

在另一种可选的实现方式中，网卡中发送队列（send queue，SQ）维护有包含脏页的内存地址和偏移量的 SG 信息，前述的页面信息包括脏页对应的 SG 信息。

在另一种可选的实现方式中，网卡向迁移消息指示的 RQ 发送待迁移数据，包括：网卡从 SQ 中获取脏页对应的 SG 信息，并向迁移消息指示的 RQ 发送 SG 信息指示的数据。在虚拟机热迁移过程中，将源主机的数据迁移功能卸载到网卡中，避免了源主机中的处理器执行多次虚拟机的数据拷贝操作，降低了源主机的计算资源消耗，提高了源主机处理其他计算业务的效率。

在另一种可选的实现方式中，源主机通过传输控制协议/网络之间互连协议（transmission control protocol/internet protocol，TCP/IP）和目的主机建立有控制连接和数据连接，其中，该控制连接用于传输页面信息和迁移消息，该数据连接用于传输待迁移数据。在本实施例中，不同的传输通道用于传输不同的信息或数据，避免了页面信息由数据连接传输，导致该页面信息无法被接收侧（目的主机）的网卡进行处理，以及虚拟机的热迁移出现问题，提高了虚拟机的数据迁移稳定性。

在另一种可选的实现方式中，源主机通过 TCP/IP 和目的主机建立有单一连接，单一连接用于传输页面信息、迁移消息和待迁移数据。

在另一种可选的实现方式中，迁移消息是目的主机为待迁移的虚拟机分配的消息处理标识（identifier，ID）。由于源主机和目的主机之间的单一连接可以传输不同 VM 的数据，

上述的消息处理ID还可以用于区分单一连接中数据所属的VM，避免单一连接中多个VM的数据发生传输错误，如将VM（1）的数据错误的识别为VM（2）的数据，进而，提高了VM热迁移的准确性。

在另一种可选的实现方式中，源主机通过远程直接内存访问（remote direct memory access，RDMA）网络和目的主机进行通信，网卡中存储有内存保护表（memory protect table，MPT），MPT表用于指示内存的主机物理地址（host physical address，HPA）和待迁移的虚拟机的客户机物理地址（guest physical address，GPA）的对应关系，且MPT表中包含有待迁移的虚拟机所使用的物理功能（physical function，PF）信息。网卡向迁移消息指示的RQ发送待迁移数据，包括：首先，网卡依据迁移信息确定待迁移的虚拟机的PF信息。其次，网卡根据待迁移的虚拟机的PF信息和GPA查询MPT表，确定页面信息对应的HPA。最后，网卡向目的主机发送待迁移的虚拟机的HPA中存储的待迁移数据。

在另一种可选的实现方式中，数据迁移方法还包括：网卡获取第二数据迁移通知消息，该第二数据迁移通知消息用于指示待接收的虚拟机的标识。网卡将其他主机发送的待接收数据迁移至源主机的内存中，待接收数据为存储在其他主机的内存中，且与待接收的虚拟机关联的数据。在本实施例中，源主机可以作为发送端将待迁移的虚拟机的待迁移数据发送到目的主机，源主机还可以作为接收端接收其他主机发送的虚拟机的数据，从而实现源主机的多个虚拟机的迁移过程中的收发功能，提高了源主机的数据迁移性能。

第二方面，本申请提供了一种数据迁移装置，该数据迁移装置应用于源主机的网卡，该数据迁移装置包括用于执行第一方面或第一方面任一种可能实现方式中的数据迁移方法的各个模块。

示例的，数据迁移装置包括：通信单元、数据识别单元和迁移单元。其中，通信单元用于获取第一数据迁移通知消息，该第一数据迁移通知消息用于指示待迁移的虚拟机的标识。数据识别单元用于根据标识确定待迁移数据，该待迁移数据为存储在源主机的内存中，且与待迁移的虚拟机关联的数据。迁移单元用于将前述的待迁移数据迁移至目的主机。

有益效果可以参见第一方面中任一实现方式的描述，此处不再赘述。所述数据迁移装置具有实现上述第一方面中任一实现方式的方法实例中行为的功能。所述功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。所述硬件或软件包括一个或多个与上述功能相对应的模块。

第三方面，本申请提供了一种电子设备，该电子设备包括：接口电路和控制电路。接口电路用于接收来自电子设备之外的其它设备的信号并传输至控制电路，或将来自控制电路的信号发送给电子设备之外的其它设备，控制电路通过逻辑电路或执行代码指令用于实现第一方面中任一实现方式的方法。有益效果可以参见第一方面中任一实现方式的描述，此处不再赘述。

在一种可能的示例中，该电子设备可以是指网卡。

在另一种可能的示例中，该电子设备也可以是指网卡包含的处理器。

在又一种可能的示例中，该电子设备还可以是指包含有网卡的专用处理设备，该专用处理设备可以实现第一方面中任一实现方式的方法。

值得注意的是，上述三种示例仅为本实施例提供的可能的实现方式，不应理解为对本申请的限定。

第四方面，本申请提供一种计算机可读存储介质，存储介质中存储有计算机程序或指令，当计算机程序或指令被主机或网卡执行时，实现第一方面和第一方面中任一种可能实现方式中的方法。

第五方面，本申请提供一种计算机程序产品，该计算程序产品包括指令，当计算机程序产品在主机或网卡上运行时，使得主机或网卡执行该指令，以实现第一方面和第一方面中任一种可能实现方式中的方法。

第六方面，本申请提供一种芯片，包括存储器和处理器，存储器用于存储计算机指令，处理器用于从存储器中调用并运行该计算机指令，以执行上述第一方面及其第一方面任意可能的实现方式中的方法。

本申请在上述各方面提供的实现方式的基础上，还可以进行进一步组合以提供更多实现方式。

附图说明

图 1 为本申请提供的一种通信系统的应用场景图；

图 2 为本申请提供的一种主机的结构示意图；

图 3 为本申请提供的一种数据迁移方法的流程示意图一；

图 4 为本申请提供的一种数据迁移方法的流程示意图二；

图 5 为本申请提供的一种数据迁移的示意图；

图 6 为本申请提供的一种数据迁移装置的结构示意图。

具体实施方式

本申请提供一种数据迁移方法：首先，源主机的网卡获取数据迁移通知消息，该数据迁移通知消息用于指示待迁移的虚拟机的标识；其次，网卡根据标识确定待迁移数据，该待迁移数据为存储在源主机的内存中，且与待迁移的虚拟机关联的数据；最后，网卡将待迁移数据迁移至目的主机。在本实施例中，待迁移数据是由网卡依据待迁移的虚拟机的标识确定的，避免了源主机对虚拟机的待迁移数据进行确定的过程，减少了源主机在虚拟机热迁移过程中所需的计算资源，提高了源主机执行其他业务（如 AI、HPC 等计算密集型和时延敏感型业务）的能力。

下面结合附图详细介绍本申请提供的数据迁移方法。

图 1 为本申请提供的一种通信系统的应用场景图，该通信系统包括计算机集群（computer cluster）110 和客户端 120，计算机集群 110 可以通过网络 130 与客户端 120 进行通信，网络 130 可以是因特网，或其他网络（如以太网）。网络 130 可以包括一个或多个网络设备，如网络设备可以是路由器或交换机等。

客户端 120 可以是运行有应用程序的计算机，该运行有应用程序的计算机可以是物理机，也可以是虚拟机。例如，若该运行有应用程序的计算机为物理计算设备，该物理计算设备可以是主机或终端（Terminal）。其中，终端也可以称为终端设备、用户设备（user equipment，UE）、移动台（mobile station，MS）、移动终端（mobile terminal，MT）等。终端可以是手机、平板电脑、笔记本电脑、桌面电脑、台式计算机、虚拟现实（virtual reality，VR）终端设备、增强现实（augmented reality，AR）终端设备、工业控制中的无线终端、无人驾驶中的无线终端、远程手术（remote medical surgery）中的无线终端、智能电网（smart

grid）中的无线终端、运输安全（transportation safety）中的无线终端、智慧城市（smart city）中的无线终端、智慧家庭（smart home）中的无线终端等等。本申请的实施例对客户端120所采用的具体技术和具体设备形态不做限定。在一种可能的实现方式中，该客户端120可以是运行在计算机集群110中任意一个或多个主机上的软件模块。

计算机集群110是指通过局域网或互联网连接计算机集合，通常用于执行大型任务（也可以称为作业（job）），这里的作业通常是需要较多计算资源并行处理的大型作业，本实施例不限定作业的性质和数量。一个作业可能包含多个计算任务，这些任务可以分配给多个计算资源执行。大多数任务是并发或并行执行的，而有一些任务则需要依赖于其他任务所产生的数据。计算机集群中每台计算设备使用相同的硬件和相同的操作系统；也可以根据业务需求，在计算机集群的主机中采用不同的硬件和不同的操作系统。由于采用计算机集群部署的任务可并发度执行，可以提升总体性能。

如图1所示，计算机集群110包括多个主机，例如主机111~主机114，各个主机可以用于提供计算资源。就一台主机来说，它可以包含多个处理器或处理器核，每个处理器或者处理核可以是一个计算资源，因此一台物理主机可以提供多个计算资源。例如，物理主机可以是指服务器。

计算机集群110可以处理多种类型的作业。本实施例对任务的数量，以及可以并行执行的任务的数据都没有予以限制。示例的，上述的作业是指虚拟机的热迁移或者主机之间的主从备份等，如数据备份。

在图1中，作业可以从客户端120通过网络130向主机111提交给计算机集群110。在作业从主机111提交给计算机集群110的情况下，主机111可以用于管理计算机集群110中所有的主机，以完成该作业所包括的一个或多个任务，如调度其他主机的计算资源或存储资源等。在另一种可能的实现方式中，作业提交的位置也可以是计算机集群110中的其他主机，本实施例不限定提交作业的位置。

如图1所示，计算机集群110中可以运行有一个或多个虚拟机。虚拟机是指通过虚拟化技术将物理的计算资源、存储资源和网络资源进行虚拟获得的虚拟设备。

在一种可能的示例中，一个主机上运行有一个或多个VM，如主机111中运行有2个VM，主机114中运行有1个VM。

在另一种可能的示例中，一个VM运行在多个主机上，如一个VM利用主机111的处理资源和主机114的存储资源等。

图1仅为本实施例提供的示例，不应理解为对本申请的限定，本申请以一个VM运行在一个主机为例进行说明。

值得注意的是，图1只是示意图，该通信系统中还可以包括其他设备，在图1中未画出，本申请的实施例对该系统中包括的主机（计算设备）、客户端的数量和类型不做限定。示例的，计算机集群110还可以包括更多或更少的计算设备，如计算机集群110包括两个计算设备，一个计算设备用于实现上述主机111和主机112的功能，另一个计算设备用于实现上述主机113和主机114的功能。

图2为本申请提供的一种主机的结构示意图，示例性的，图1中的任一个主机可通过图2所示的主机200来实现，该主机200包括基板管理控制器（baseboard management controller，BMC）210、处理器220、内存230、硬盘240和网卡250。

基板管理控制器 210，可以对设备进行固件升级，对设备的运行状态进行管理以及排除故障等。处理器 220 可通过外围器件互联（Peripheral Component Interconnect express，PCIe）总线、通用串行总线（Universal Serial Bus，USB），或者集成电路总线（Inter-Integrated Circuit，I2C）等总线访问基板管理控制器 210。基板管理控制器 210 还可以和至少一个传感器相连。通过传感器获取计算机设备的状态数据，其中状态数据包括：温度数据，电流数据、电压数据等等。在本申请中不对状态数据的类型做具体限制。基板管理控制器 210 通过 PCIe 总线或者其他类型的总线和处理器 220 通信，例如，将获取到的状态数据，传递给处理器 220 进行处理。基板管理控制器 210 也可以对存储器中的程序代码进行维护，包括升级或恢复等等。基板管理控制器 210 还可以对主机 200 内的电源电路或时钟电路进行控制等。总之，基板管理控制器 210 可以通过以上方式实现对主机 200 的管理。然而，基板管理控制器 210 只是一个可选设备。在一些实施方式中，处理器 220 可以直接和传感器通信，从而对计算机设备直接进行管理和维护。

值得说明的是，主机中器件的连接方式除了采用上述 PCIe 总线、USB 总线、I2C 总线外，还可以通过扩展工业标准结构（extended industry standard architecture，EISA）总线、统一总线（unified bus，Ubus 或 UB）、计算机快速链接（compute express link，CXL）、缓存一致互联协议（cache coherent interconnect for accelerators，CCIX）等。总线还可以分为地址总线、数据总线、控制总线等。

处理器 220 通过双倍速率（double data rate，DDR）总线和内存 230 相连。这里，不同的内存 230 可能采用不同的数据总线与处理器 220 通信，因此 DDR 总线也可以替换为其他类型的数据总线，本申请实施例不对总线类型进行限定。

另外，主机 200 还包括各种输入输出（input/output，I/O）设备，处理器 220 可以通过 PCIe 总线访问这些 I/O 设备。

处理器（processor）220 是主机 200 的运算核心和控制核心。处理器 220 中可以包括一个或多个处理器核（core）221。处理器 220 可以是一块超大规模的集成电路。在处理器 220 中安装有操作系统和其他软件程序，从而处理器 220 能够实现对内存 230 及各种 PCIe 设备的访问。可以理解的是，在本发明实施例中，处理器 220 可以是中央处理器（central processing unit，CPU），可以是其他特定集成电路（application specific integrated circuit，ASIC）。处理器 220 还可以是其他通用处理器、数字信号处理器（digital signal processing，DSP）、现场可编程门阵列（field programmable gate array，FPGA）或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。实际应用中，主机 200 也可以包括多个处理器。

内存 230，也称为主存（main memory）。内存 230 通常用来存放操作系统中各种正在运行的软件、输入和输出数据以及与外存交换的信息等。为了提高处理器 220 的访问速度，内存 230 需要具备访问速度快的优点。在传统的计算机系统架构中，通常采用动态随机存取存储器（dynamic random access memory，DRAM）作为内存 230。处理器 220 能够通过内存控制器高速访问内存 230，对内存 230 中的任意一个存储单元进行读操作和写操作。除了 DRAM 之外，内存 230 还可以是其他随机存取存储器，例如静态随机存取存储器（static random access memory，SRAM）等。本实施例不对内存 230 的数量和类型进行限定。此外，可对内存 230 进行配置使其具有保电功能。保电功能是指系统发生掉电又重新上电时，存

储器中存储的数据也不会丢失。具有保电功能的内存 230 被称为非易失性存储器。

示例的，内存 230 包括多个内存页（page），内存页是内存 230 的数据 I/O 操作的最小单位，内存页也称为数据读写的原子单位。每个内存页对应内存 230 的一段存储地址空间，如一个内存页可以用于存储 4 千字节（kilo bytes，KB）的数据，则该内存页对应 4KB 的存储地址空间。另外，一个内存页也可以对应更大或更小的存储地址空间，如 2KB 或 8KB 等。

在一些可能的情形中，如虚拟机热迁移过程中，若内存页中所存储的虚拟机数据在一段时间内被修改，则该内存页可以被称为该虚拟机的热迁移过程的脏页，该修改后的内存页（脏页）中所存储的数据被称为该虚拟机的脏页数据。

I/O 设备是指可以进行数据传输的硬件，也可以理解为与 I/O 接口对接的设备。常见的 I/O 设备有网卡、打印机、键盘、鼠标等，如 I/O 设备可以是图 2 所示出的网卡 250。所有的外存也可以作为 I/O 设备，如硬盘、软盘、光盘等。处理器 220 可通过 PCIe 总线访问各个 I/O 设备。需要说明的是，PCIe 总线只是其中的一个示例，可以被替换为其他总线，例如统一（Unified Bus，UB 或 Ubus）总线或计算机快速链接（compute express link，CXL）等。

如图 2 所示，网卡 250 包括处理器 251、存储器 252 和通信接口 253。在一些可能的示例中，包含处理单元和网络适配器（network interface card，NIC）的网卡也被称为智能网卡(intelligent NIC，iNIC)。

处理器 251 是指具有处理能力的处理器，例如数据处理单元（data processing unit，DPU）。DPU 具有 CPU 的通用性和可编程性，但更具有专用性，可以在网络数据包，存储请求或分析请求上高效运行。DPU 通过较大程度的并行性（需要处理大量请求）与 CPU 区别开来。可选的，这里的 DPU 也可以替换成图形处理单元（graphics processing unit，GPU）、嵌入式神经网络处理器（neural-network processing units，NPU）等处理芯片。

存储器 252 可以是指与处理器 251 直接交换数据的内部存储器，它可以随时读写数据，而且速度很快，作为操作系统或其他正在运行中的程序的临时数据存储器。存储器 252 包括至少两种存储器，例如存储器 252 既可以是随机存取存储器，也可以是 ROM。举例来说，随机存取存储器是 DRAM，或者存储级存储器（storage class memory，SCM）。DRAM 是一种半导体存储器，与大部分 RAM 一样，属于一种易失性存储器（volatile memory）设备。SCM 是一种同时结合传统储存装置与存储器特性的复合型储存技术，存储级存储器能够提供比硬盘更快速的读写速度，但存取速度上比 DRAM 慢，在成本上也比 DRAM 更为便宜。然而，DRAM 和 SCM 在本实施例中只是示例性的说明，存储器 252 还可以包括其他随机存取存储器，例如 SRAM 等。而对于只读存储器，举例来说，可以是 PROM、EPROM 等。另外，存储器 252 还可以是双列直插式存储器模块或双线存储器模块（dual in-line memory module，DIMM），即由 DRAM 组成的模块，还可以是固态硬盘（solid state disk，SSD）。实际应用中，网卡 250 中可配置多个存储器 252，以及不同类型的存储器 252。本实施例不对存储器 252 的数量和类型进行限定。此外，可对存储器 252 进行配置使其具有保电功能。保电功能是指系统发生掉电又重新上电时，存储器 252 中存储的数据也不会丢失。具有保电功能的存储器被称为非易失性存储器。

在一种可能的情形中，存储器 252 中存储有软件程序，处理器 251 运行存储器 252 中

的软件程序可实现 VM 迁移的管理，例如，将内存 230 存储的 VM 数据迁移到其他设备。

通信接口 253 用于实现的主机 200 与其他设备进行通信的网络接口卡，例如，网络适配器 253 可以实现从并行到串行的数据转换、数据包的装配和拆装、网络存取控制、数据缓存和网络信号中的一种或多种功能。

如图 2 所示，主机 200 上可以运行有一个或多个 VM，如图 2 所示出的虚拟机 200A~虚拟机 200C。VM 所需的计算资源来源于主机 200 本地的处理器 220 和内存 230，而 VM 所需的存储资源既可以来源于主机 200 本地的硬盘 240，也可以来自其他主机中的硬盘。例如，虚拟机 200A 包括虚拟处理器 200A1、虚拟内存 200A2 和虚拟网卡 200A3，虚拟处理器 200A1 所需的计算资源由处理器 200 提供、虚拟内存 200A2 所需的存储资源由内存 230（或硬盘 240）提供，虚拟网卡 200A3 所需的网络资源由网卡 250 提供。此外，VM 中可运行各种应用程序，用户可通过 VM 中的应用程序触发读/写数据请求等业务。

示例的，VM 中虚拟内存 200A2 的存储空间可由内存 230 包括的内存页来提供。通常，VM 的热迁移可以在保证业务不中断的情况下，将 VM 从一台主机迁移到另外一台主机，而保证业务不中断的关键技术即为内存数据的迁移，该内存数据是指存储在内存 230 包括的内存页的数据。若该内存页中所存储的数据在一段时间内被修改，则该内存页可以被称为该 VM 热迁移过程中的脏页，因此，该修改后的内存页（脏页）中所存储的数据被称为 VM 的脏页数据。

这里以源主机是图 1 中主机 111、目的主机是主机 112 为例进行说明，内存的迁移可为以下几个阶段。

1、迭代预拷贝阶段：VM 迁移开始时，依然在源主机上运行；为了保证 VM 业务不中断，待迁移虚拟机的数据会被同时写入源主机和目的主机的内存中。

2、停机拷贝阶段：VM 在源主机上的运行中断，且 VM 在源主机的内存页中所存储的数据被传输到目的主机的内存上。

3、脏页拷贝阶段：VM 依然在源主机上的运行，源主机监控并记录下迁移过程中所有已被传输的内存页的任何修改（即内存页中的脏页），并在 VM 所使用的所有内存页都传输完成后，将传输脏页所存储的脏页数据。

另外，源主机估计迁移过程中的数据传输速度，当剩余的内存数据量能够在一个可以设定的时间周期（如 30 毫秒）内传输完成时，关闭源主机上的 VM，再将 VM 剩余的脏页数据传输到目的主机上。

4、虚拟机恢复阶段：在目的主机启动 VM，VM 的整个迁移过程完成。

在一种可能的情形中，如果源主机和目的主机共享存储系统，则源主机只需要通过网络发送 VM 的执行状态、内存中的内容、虚机设备的状态到目的主机上。否则，还需要将 VM 的磁盘存储发到目的主机上。共享存储系统指的是源和目的虚机的镜像文件目录是在一个共享的存储上的。

值得注意的是，本实施例提供的内存迁移过程是在计算机集群 110 内部的多个主机之间实现的，在一种可选的实现方式中，内存的迁移也可以是在计算机集群 110 中主机与计算机集群 110 外部的其他设备之间的数据迁移，本申请对此不予限定。

下面将结合附图对本申请实施例的实施方式进行详细描述。

图 3 为本申请提供的一种数据迁移方法的流程示意图一，该数据迁移方法可应用于图

1 所示出的通信系统，示例的，主机 310 可以实现图 1 中主机 111 的功能，主机 320 可以实现图 1 中主机 112 的功能。该数据迁移方法可以由网卡 313 执行，网卡 313 所在主机 310 的硬件实现可以参考图 2 所示出的主机 200，该网卡 313 也可以具有图 2 中网卡 253 的功能，主机 320 和主机 310 相似，此处不再赘述。

为了便于描述，将图 3 所示的主机 310 称为源主机，或称发送端主机、第一主机、源节点等，内存 312 存储有一个或多个 VM 的待迁移数据，该待迁移数据包括 VM 在拷贝前的运行数据，以及热迁移过程中脏页存储的数据等。脏页是指内存 312 中数据发生修改的内存页（page）。

示例的，网卡 313 中可以运行有虚拟机热迁移管理程序，该虚拟机热迁移管理程序所需的计算资源由网卡 313 所包括的处理器和存储器提供，具体的，该虚拟机热迁移管理程序可以对主机 310 中内存 312 存储的脏页数据进行管理，如读取、写入或擦除等。在虚拟机热迁移过程中，将处理器 311 的数据迁移功能卸载到网卡 313 中，避免了主机 310 中的处理器 311 执行多次虚拟机的数据拷贝操作，降低了处理器 311 的资源消耗，提高了主机 310 处理其他计算业务的效率。

为了便于描述，将图 3 所示中主机 320 称为目的主机，或称接收端主机、第二主机、目的节点等，内存 322 用于存储一个或多个 VM 的待迁移数据，该待迁移数据包括 VM 在拷贝前的运行数据，以及热迁移过程中脏页存储的数据等。脏页是指在 VM 的热迁移过程中，内存 322 存储的数据发生修改的内存页（page）。示例的，为实现虚拟机的热迁移管理过程，网卡 323 中也可以运行有上述的虚拟机热迁移管理程序。

值得注意的是，源主机和目的主机中可以运行有虚拟机管理软件，该虚拟机管理软件用于管理源主机和目的主机上的虚拟机。示例的，前述的虚拟机热迁移管理程序可以是虚拟机管理软件的一部分，也可以是虚拟机管理软件为实现虚拟机热迁移而在网卡中启动的一个线程。

网卡接收和发送数据通常采用消息队列的方式，消息队列包括一组队列对（queue pair，QP），QP 包括发送队列和接收队列，如网卡 313 用于发送数据的消息队列是发送队列（SQ），网卡 323 中用于接收数据的消息队列是接收队列（RQ）。消息队列是多个主机之间进行通信所采用的连接方式，例如，多个主机之间可以利用 TCP/IP 协议建立多条连接，每条连接都有一个接收队列和发送队列，该接收队列和发送队列用于传输该连接的数据。

示例的，主机 310 和主机 320 基于硬直通（single root I/O virtualization，SR-IOV）技术来实现硬件网络资源（如网卡）的虚拟化，SR-IOV 技术是“虚拟通道”的一个技术实现。示例的，SR-IOV 技术将一个实现物理功能（physical function，PF）的 PCIe 设备进行虚拟化，获得一个或多个实现虚拟功能（virtual function，VF）的 PCIe 设备，每个可以实现 VF 的 PCIe 设备（下文简称为 VF）被直接分配到一个虚拟机，主机还为每个 VF 提供独立的内存空间、中断（号）和直接内存访问（direct memory access，DMA）流。

在 SR-IOV 技术中，SR-IOV 设备的 PF 驱动程序用于管理具有 SR-IOV 功能的设备的物理功能，物理功能支持 SR-IOV 规范定义的 SR-IOV 功能的 PCI 功能，物理功能是全面的 PCIe 功能，可以像其他任何物理 PCIe 设备一样发现、管理和处理。物理功能可用于配置和控制虚拟 PCIe 设备。

VF 是支持 SR-IOV 的物理网卡所虚拟出的一个虚拟网卡或实例，VF 会以一个独立网

卡的形式呈现给虚拟机，每个 VF 具有独享的 PCI 配置区域，并且可能与其他 VF 共享同一个物理资源（如公用一个物理网口），VF 具有轻量级的 PCIe 功能，可以与 PF 以及该 PF 关联的其他 VF 共享一个或多个物理资源（如物理网卡）。

例如，对于虚拟机的一个 QP 来说，在源主机中，一个 VF 对应一个或多个 SQ；在目的主机中，一个 VF 对应一个或多个 RQ。

消息队列中存储有工作队列元素（work queue element，WQE），WQE 中存储有指向发送队列、或接收队列数据的地址和长度的信息，数据的长度可以由数据的地址和偏移量来确定，WQE 中指示了数据的地址和偏移量的信息又被称为散集（scatter/gather，SG）信息。如果一组数据包括多段的数据，一段数据的 SG 信息包括该数据的地址和偏移量，则 WQE 中关于该组数据的多个 SG 信息也可称为 SG 的链，或称散集列表（scatter/gather list，SGL）。

在第一种可行的实现方式中，上述 SG 信息所包括的地址是指 VM 数据的客户机物理地址（guest physical address，GPA），客户机是指运行在源主机上的虚拟机（VM），该 GPA 是指 VM 数据在内存（例如，内存 312）中所存储的内存地址。在一种可能的情形中，GPA 可以是指中间的物理地址（intermediate phyical address，IPA）。VM 可以基于该 GPA 访问相应的数据，但硬件设备（如主机的处理器或网卡）需要使用主机物理地址（host physical address，HPA）才能对该数据进行访问，因此，该 GPA 需经过 GPA→主机虚拟地址（host virtual address，HVA），以及 HVA→HPA 的两级地址转换，以实现硬件设备对该 SG 信息指示的数据的访问。

在第二种可行的实现方式中，上述 SG 信息所包括的地址是指 VM 数据的 HPA，硬件设备（如主机的处理器或网卡）可以基于该 GPA 访问 VM 数据。

值得注意的是，上述的 SG 信息仅为本实施例提供的一种示例，不应理解为对本申请的限定。

请继续参见图 3，内存 312 包括多个内存页（page），如序号为 1~5 的 page，分别记为 page1、page2、page3、page4、page5，该多个 page 中具有 2 个与待迁移的虚拟机关联的内存页，如内存 312 中 page2 和 page4。例如，在 VM 热迁移的过程中，网卡 313 可以将内存 312 中 page2 和 page4 存储的数据发送到其他主机，如主机 320。

如图 3 所示，本实施例提供的数据迁移方法包括以下步骤。

S310，网卡 313 获取数据迁移通知消息。

该数据迁移通知消息用于指示待迁移的虚拟机的标识，如图 3 中黑色填充的小块。

示例的，若主机 310 中运行有多个虚拟机，该数据迁移通知消息可以用于指示该多个虚拟机中待迁移的虚拟机，避免无需迁移的虚拟机数据被迁移到其他主机，降低数据迁移过程中多个虚拟机的数据发生紊乱的概率。

可选的，数据迁移通知消息是虚拟机管理软件触发虚拟机的热迁移操作生成的。例如，虚拟机管理软件管理了多个虚拟机，在主机 310（或客户端）触发热迁移操作后，由虚拟机管理软件生成一个数据迁移通知消息，该数据迁移通知消息可以包含待迁移的虚拟机的标识，例如，该标识是虚拟机的序号，或者虚拟机在虚拟局域网（virtual local area network，VLAN）中的标签等。

虚拟机的热迁移操作的触发条件可以是：用户操作和主机的资源使用情况。该资源使用情况可以是例如主机待释放的资源量等信息，具体的，如主机的计算资源、存储资源和

网络带宽等。

S320，网卡 313 根据数据迁移通知消息指示的标识确定待迁移数据。

该待迁移数据为存储在源主机（主机 310）的内存（内存 312）中，且与待迁移的虚拟机关联的数据。

可选的，网卡 313 可以根据数据迁移通知消息指示的标识确定内存 312 中与待迁移的虚拟机关联的内存页集合，并将内存页集合中存储的数据作为待迁移数据。示例的，该内存页集合包括一个或多个内存页。

如图 3 所示，内存 312 中与待迁移的虚拟机关联的内存页为 page2 和 page4，该 page2 和 page4 中存储的数据为 VM 的待迁移数据。

由网卡 313 依据待迁移的虚拟机的标识确定待迁移数据，避免了由处理器 311 确定待迁移数据导致主机 310 的性能下降，提高了主机 310 处理其他业务的能力，减少了主机 310 的卡顿。

在一种可能的情形中，该待迁移数据包括脏页数据，脏页数据为一个或多个内存页中，数据发生修改的内存页所存储的数据。在内存 312 中与待迁移的虚拟机关联的内存页集合中，数据发生修改的内存页为脏页，如图 3 所示出的 page2 为脏页，则该 page2 中所存储的待迁移的虚拟机的数据为脏页数据。

针对于待迁移数据包括的脏页数据，下面提供一种可能的具体实现方式：主机 310 查询网卡中保存的脏页标记信息，确定与标识关联的脏页集合，并将脏页集合中存储的数据作为脏页数据。

其中，脏页集合包括一个或多个脏页，如图 3 所示出的 page2。

前述的脏页标记信息用于指示脏页的内存地址。在一种可选的示例中，脏页标记信息包括第一脏页表和第二脏页表中至少一个。例如，网卡 313 中可以保存有一张或多张脏页表，该多张脏页表在同一个时间可以用于实现不同的功能。

第一脏页表用于标记脏页为标脏状态，该标脏状态为源主机（如主机 310 中的处理器 311）对脏页中存储的数据进行修改的状态。

第二脏页表用于标记脏页为数据迁移状态，该数据迁移状态为网卡（如网卡 313）对脏页中存储的数据进行迁移的状态。

在数据迁移过程中，脏页标记信息可以仅包括一个脏页表，该脏页表用于记录脏页的迁移。例如，当源主机对脏页进行数据访问时，网卡 313 中仅需维护有第一脏页表，记该脏页为标脏状态；又如，当网卡 313 对脏页进行数据访问时，网卡 313 中仅需维护有第二脏页表，记该脏页为数据迁移状态。

另外，由于一个脏页表无法被多个程序或硬件设备同时访问，因此，网卡 313 中也可以设置有 2 个脏页表来标记内存页的迁移状态，如上述的第一脏页表和第二脏页表，在同一个时间节点，该 2 个脏页表分别用于实现不同的功能。

在网卡 313 确定与待迁移的虚拟机关联的内存页集合后，下面提供一种可行的实现方式，来对上述的脏页标记信息包括的脏页表进行说明。

对于内存 312 中一个内存页（page）而言，网卡 313 可以用脏页表来标记该 page 的状态。例如，一张脏页表表示了与待迁移的虚拟机关联的所有内存页（page）中的脏页的迁移状态，如每个脏页的迁移状态用一个位元（Bit）来表示。

示例的，该脏页表中可以设置有一个或多个状态标志，例如，脏页表的结构可以如下表 1 所示。

表 1

	M	S
情况 1	0	1
情况 2	1	0

其中，M 和 S 是用于标记 page 的迁移状态的状态标志。

状态标志 M 用于指示主机 310 对该 page 的访问信息，例如，M=1 是指主机 310 对该 page 进行了数据访问，M=0 是指主机 310 未对该 page 进行数据访问。

状态标志 S 用于指示网卡 313 对该 page 的访问信息，例如，S=1 是指网卡 313 对该 page 进行了数据访问(如将该 page 的数据迁移到其他主机)，S=0 是指网卡 313 未对该 page 进行数据访问。

如表 1 所示，上述状态标志对该 page 的状态标记过程存在以下两种可能的情况。

情况 1：M=1，S=0，则该 page 为标脏状态，标脏状态为源主机（如主机 310 中的处理器 311）对脏页中存储的数据进行修改的状态，如上述的第一脏页表。

情况 2：M=0，S=1，则该 page 为数据迁移状态，数据迁移状态为网卡（如网卡 313）对脏页中存储的数据进行迁移的状态，如上述的第二脏页表。

由于一个 page 无法在同一个时间节点为多个硬件设备提供数据访问服务，因此，脏页表不存在“M=1，S=1”的情况。另外，由于“M=0，S=0”指示该 page 不是脏页，且处理器 311 和网卡 313 均未对脏页表指示的 page 进行数据访问，而脏页表无需记录非脏页的迁移状态，因此，脏页表也不存在“M=0，S=0”的情况。

值得注意的是，表 1 仅为本实施例提供的脏页表的示例，脏页表中的状态标志（如 M 和 S）所代表的含义可根据不同的使用场景和需求进行调整，本申请对此不予限定。

在数据迁移过程中，为了实现对脏页标记信息的管理，虚拟机管理软件中可以运行标脏程序和迁移程序。标脏程序和迁移程序可以是虚拟机热迁移管理程序中一个软件模块，也可以是虚拟机管理软件触发的一个独立软件单元，本申请对此不予限定。

在一种可能的示例中，该标脏程序和迁移程序运行在网卡 313。

在另一种可能的示例中，该标脏程序和迁移程序运行在处理器 311。

这里以标脏程序和迁移程序运行在网卡 313 为例进行说明。

例如，标脏程序用于管理第一脏页表，如在内存页中存储的数据发生修改后，标脏程序将该内存页在第一脏页表中的状态标志位记为“M=1，S=0”。

又如，迁移程序用于管理第二脏页表，如在内存页中发生修改后的数据被迁移到其他主机时，迁移程序将该内存页在第二脏页表中的状态标志位记为“M=0，S=1”。在内存页中发生修改后的数据被迁移到其他主机后，迁移程序将该内存页在第二脏页表中的状态标志位记为“M=0，S=1”。

在本实施例中，由网卡 313 来对内存 312 中与待迁移的虚拟机关联的内存页中的脏页进行标记，即内存 312 中的标脏功能由处理器 311 卸载到了网卡 313，避免了处理器 311 标记内存 312 中脏页的过程，减少了处理器 311 的资源消耗，进而，避免了由于处理器 311 管理虚拟机的热迁移过程导致的主机 310 的其他计算业务受到影响。

在网卡313中，针对内存312的一个page，网卡313可以利用两张脏页表来标记该page：若该page为脏页，且处理器311正在对该page存储的数据进行修改，则网卡313利用脏页表1（第一脏页表）对该page进行标记，记为M=1，S=0，如上述的情况1；若网卡313需要发送该脏页中存储的数据，则网卡313利用脏页表2（第二脏页表）对该脏页进行标记，记为M=0，S=1，如上述的情况2。在网卡313中设置有2个脏页表来标记一个脏页的迁移状态的情况下，由于标脏程序可以对第一脏页表进行修改，迁移程序可以对第二脏页表进行修改，避免了标脏程序和迁移程序同时修改一个脏页表，造成数据迁移错误。

关于网卡313利用脏页标记信息来识别内存页集合中的脏页的过程，这里给出一种可能的具体示例，该脏页扫描过程包括以下可能的步骤：

步骤1、网卡313设置两个脏页表：脏页表1和脏页表2。这两个脏页表中每个脏页表均有两个标志位（如上述表1示出的M和S），并设置脏页表1中的标志位全为0，脏页表2中的标志位全为1，其中，“1”表示脏页，“0”表示非脏页，脏页表2中全部为“1”表示脏页表2中记录的任一个page均为脏页。

步骤2、网卡313对标志位全为0的脏页表1进行初始化，将脏页表1中所有page记为“M=1，S=0”，网卡313对标志全为1的脏页表2进行初始化，将脏页表2中所有page记为“M=0，S=1”。网卡313对脏页表1和脏页表2进行初始化，使得脏页表1中记录的所有脏页的状态记为标脏状态，将脏页表1的管理交由上述的标脏程序；脏页表2中记录的所有脏页的状态记为数据迁移状态，将脏页表2的管理交由上述的迁移程序。由不同的程序来对2个脏页表分别进行管理，避免一张脏页表被不同的程序进行修改，导致数据迁移发生错误。

步骤3、网卡313对主机310中内存312的多个page进行扫描，并将多个page中的脏页记为“M=1，S=0”。在网卡313确定内存312中被处理器311进行了数据修改后的内存页后，网卡313将这些内存页记为脏页，以便网卡313启动脏页的数据迁移操作。

步骤4、网卡313扫描M=0，S=1的脏页表2。在完成全部的扫描后，将脏页表2中记录的任一个脏页对应的状态标志位全部清0（记为M=1，S=0），并将脏页表1中与脏页表2中相应的page置1（记为M=0，S=1）。

如此，网卡313可以扫描脏页表1（或脏页表2）中“M=0，S=1”的page，并在网卡313将该page中存储的数据发往其他主机后，将脏页表1中该page的标志记为“M=1，S=0”，从而实现内存312中脏页的标脏和数据迁移（或推送）过程。

值得注意的是，上述实施例是以一个page是脏页为例进行说明的，在VM的热迁移过程涉及多个脏页的情况下，网卡313中保存的脏页表还记录有该脏页的地址信息，如该地址信息是指在主机310中脏页的HPA，或在VM中脏页对应的GPA等。

请继续参见图3，本实施例提供的数据迁移方法还包括以下步骤。

S330，网卡313将待迁移数据迁移至主机320。

在本实施例中，待迁移数据是由网卡依据待迁移的虚拟机的标识确定的，避免了源主机对虚拟机的待迁移数据进行确定的过程，减少了源主机在虚拟机热迁移过程中所需的计算资源，提高了源主机执行其他业务（如AI、HPC等计算密集型和时延敏感型业务）的能力。

针对于主机320接收到待迁移数据后，请继续参见图3，本实施例提供的数据迁移方

法还包括以下步骤 S340。

S340，网卡 323 将待迁移数据迁移至主机 320 的内存 322 中。

由目的主机的网卡来将待迁移数据写入目的主机的内存中，避免了目的主机中，待迁移数据从网卡到内存的拷贝过程，减少了数据迁移时延，提高了数据迁移的效率。另外，由于目的主机的处理器无需对虚拟机的待迁移数据进行拷贝，减少了虚拟机热迁移过程中，目的主机的计算资源和存储资源消耗，提高了目的主机执行其他业务的能力。

另外，主机 310 也可以作为目的主机来接收其他主机发送的虚拟机的数据，如主机 310 接收到另一个数据迁移通知消息，该数据迁移通知消息用于指示主机 310 待接收的虚拟机的标识；主机 310 将其他主机发送的待接收数据迁移至内存 312 中，该待接收数据为存储在其他主机的内存中，且与待接收的虚拟机关联的数据。关于主机 310 接收其他主机迁移的虚拟机的数据过程，可以参考上述实施例中主机 320 接收数据的内容，此处不予赘述。

在本实施例中，源主机可以作为发送端将待迁移的虚拟机的待迁移数据发送到目的主机，源主机还可以作为接收端接收其他主机发送的虚拟机的数据，从而实现源主机的多个虚拟机的迁移过程中的收发功能，提高了源主机的数据迁移性能。

在一种可选的实现方式中，针对于源主机将待迁移数据迁移至目的主机的过程，这里提供一种可行的具体实现方式，如图 4 所示，图 4 为本申请提供的一种数据迁移方法的流程示意图二，图 4 示出了上述 S330 的一种可能的实现方式，上述的 S330 包括以下步骤。

S410，网卡 313 向主机 320 发送待迁移数据的页面信息。

其中，该页面信息用于指示待迁移数据在内存 312 中的内存地址和偏移量。

示例的，网卡 313 中发送队列 SQ 维护有包含脏页的内存地址和偏移量的 SG 信息，前述的页面信息包括脏页对应的 SG 信息。在一些情形中，该页面信息可以称为待迁移数据的描述信息，如该描述信息是指用于描述脏页中存储的业务数据的元数据。

S420，网卡 323 向网卡 313 发送基于页面信息确定的迁移消息。

其中，该迁移消息用于指示目的主机中与待迁移的虚拟机相应的接收队列（RQ）。

示例的，该迁移消息是网卡 323 依据页面信息为主机 310 中待迁移的虚拟机分配的 RQ 标识，或者 RQ 序号等。

在一种可能的情形中，网卡 313 和网卡 323 获取到数据迁移通知消息后，为该待迁移的虚拟机分配了一个或多个 RQ；该迁移消息是在网卡 323 收到待迁移数据的页面信息后，从该一个或多个 RQ 中选择的一个 RQ 的序号（或标识）。

在另一种可能的情形中，该迁移消息是网卡 323 收到待迁移数据的页面信息后，网卡 323 与网卡 313 建立了一个用于传输待迁移数据的数据连接（或传输通道），该数据连接中的数据传输是通过 QP 实现的，该迁移消息是指该 QP 所包括的接收队列（RQ）的序号（或标识）。

以上两种可能的情形仅为本实施例提供的迁移消息的生成方式的示例，不应理解为对本申请的限定。

S430，网卡 313 向迁移消息指示的 RQ 发送待迁移数据。

在网卡 323 中存在多个 RQ，一个 RQ 对应一个虚拟机的情况下，由网卡 323 依据网卡 313 发送的页面信息，为主机 310 中待迁移的虚拟机分配迁移消息，并由网卡 313 向该迁移消息指示的 RQ 迁移待迁移数据，避免了待迁移的虚拟机（如 VM1）的待迁移数据被

发送到其他虚拟机（如 VM2）对应的 RQ，提高了虚拟机的数据迁移准确性。

针对于上述的页面信息和待迁移数据的关系，这里提供一种可能的实现方式，如图 5 所示，图 5 为本申请提供的一种数据迁移的示意图，网卡 313 和网卡 323 通过 QP 的方式建立有传输通道，这里的 QP 包括网卡 313 中的发送队列和网卡 323 中的接收队列，发送队列位于该传输通道的发送端，接收队列位于该传输通道的接收端。

如图 5 所示，网卡 313 包括的发送队列（SQ）维护的 SG 信息和内存 312 中与待迁移的虚拟机关联的内存页具有第一关系。

示例的，该第一关系是由网卡 313 在依据待迁移的虚拟机的标识确定待迁移数据后，网卡 313 利用内存 312 中与标识关联的内存页的内存地址构建的。

在一种情形中，该第一关系是指内存 312 中内存页存储的业务数据与 SG 信息的对应关系。如图 5 所示出网卡 313 的 page2-SG1，以及 page4-SG2。

在另一种情形中，该第一关系是指内存页的内存地址与 SG 信息的对应关系，例如，page2 的内存地址为 001，page4 的内存地址为 011，该对应关系包括 001-SG1，以及 011-SG2。

值得注意的是，以上两种情形是本实施例提供的示例，不应理解为对本申请的限定，如在另一种可能的情形中，该第一关系是指与待迁移的虚拟机关联的脏页的内存地址（或脏页数据）与 SG 信息的对应关系。

如图 5 所示，当多个脏页的内存地址不连续时，则在 SQ 中，每个脏页对应一个 SG 信息。如 SG1 包括“001 4KB”，表示内存 312 中地址为“001”的 page2 为脏页，且该脏页中存储的业务数据的数据长度为 4KB；SG2 包括“011 4KB”，表示内存 312 中地址为“011”的 page4 为脏页，且该脏页中存储的业务数据的数据长度为 4KB。

值得注意的是，基于图 5 给出的示例是以内存 312 中数据读写的基本单位为 page，单个 page 的存储空间是 4KB 为例进行说明的，但在一个 VM 的热迁移过程中，可能会涉及到多个脏页，且这多个脏页的内存地址是连续的，则该多个脏页在 SQ 中可以仅对应一个 SG 信息。例如，2 个脏页的地址分别为“001”和“010”，且每个脏页中存储的业务数据的数据长度均为 4KB，则在网卡 313 将该 2 个脏页存储的业务数据映射到 SQ 时，该 2 个脏页对应的 SG 信息可以为“001 8KB”。

在 VM 的数据迁移过程中，由网卡 313 将内存 312 中的脏页映射到 SQ 中的 SG 信息，该脏页中存储的业务数据无需拷贝到网卡 313 中的存储器，在主机 310 中，避免了业务数据从内存 312 到网卡 313 的数据拷贝，减少了 VM 的数据迁移时间，提高了 VM 的热迁移效率。

如图 5 所示，网卡 323 包括的接收队列（RQ）维护的 SG 信息和内存 322 中的内存页具有第二关系。

例如，该第二关系是指内存页存储的业务数据与 SG 信息的对应关系。如图 5 所示出网卡 323 中的 page2-SG1，以及 page4-SG2。

又如，该第二关系是指内存 322 中内存页的内存地址与 SG 信息的对应关系，如 page2 的内存地址为 001，page4 的内存地址为 011，该第二关系包括 001-SG1，以及 011-SG2。

在另一种可能的情形中，该第二关系是指与待迁移的虚拟机关联的脏页的内存地址（或脏页数据）与 SG 信息的对应关系。

示例的，在网卡 313 和网卡 323 构建待迁移数据的传输通道后，网卡 323 依据页面信

息包含的内存地址构建上述的第二关系。

在本文的下述内容中，在源主机中，将待迁移的虚拟机称为源 VM；在目的主机中，将待迁移的虚拟机称为目的 VM。

为避免待迁移的数据在源主机和目的主机中的存储空间发生变化，针对于一组数据在源主机的内存地址（如源 VM GPA）和在目的主机的内存地址（如目的 VM GPA）应是一致的。在一种可行的示例中，源 VM GPA 与目的 VM GPA 的一致是由虚拟机管理软件实现的，如在网卡 313 构建了上述的第一关系后，网卡 323 利用虚拟机管理软件和前述的第一关系来构建第二关系，使得源 VM GPA 与目的 VM GPA 保持一致，从而避免虚拟机迁移后，虚拟机中指示数据的内存地址发生变化，提高虚拟机迁移的准确性。

为避免网卡 313 从内存 312 中拷贝数据（如脏页数据）后，再将该拷贝的脏页数据发送到主机 320，导致数据迁移的时延较大，虚拟机的卡顿明显，针对于上述的 S430，结合图 5 所示出的 SQ、RQ 和 SG 信息，这里提供一种可选的实现方式：网卡 313 从 SQ 中获取脏页对应的 SG 信息；网卡向迁移消息指示的 RQ 发送 SG 信息指示的数据。

在本实施例中，脏页对应的 SG 信息包含的内存地址，可以是上述示例提供的 HPA 或者 GPA。

例如，若 SG 信息包含的内存地址是指 HPA，则网卡 313 基于该 HPA 指示的存储地址空间，将该脏页中存储的脏页数据迁移到主机 320 中。

又如，若 SG 信息包含的内存地址是指 GPA，则网卡 313 基于该 GPA 进行地址转换，获得脏页的 HPA，进而将该 HPA 指示的存储地址空间中存储的脏页数据迁移到主机 320 中。

网卡 313 基于 GPA 进行地址转换可以是通过 IO 内存管理单元（memory management unit，MMU）实现的，具体的，该 IOMMU 可以基于待迁移的虚拟机在主机 310 所采用的 VF 信息，对该 GPA 进行地址转换，得到脏页的 HPA。该 VF 信息用于指示待迁移的虚拟机采用的虚拟 PCIe 设备（如网卡或内存）的标识。

由于主机为每个 VF 提供了独立的内存空间，且在源主机中，一个 VF 对应一个或多个 SQ，网卡 313 基于 VF 信息对 SQ 中 SG 信息包括 GPA 进行地址转换，获得脏页的 HPA，使得网卡 313 可以基于该 HPA 将脏页数据迁移到主机 320，实现了待迁移的虚拟机的数据迁移，避免了脏页数据从内存 312 到网卡 313 的拷贝过程，减少了数据迁移的时间，提高了虚拟机的迁移效率。

在本实施例中，由网卡 313 将内存 312 中的脏页映射到 SQ 中的 SG 信息，该脏页中存储的业务数据无需拷贝到网卡 313 中的存储器，在主机 310 中，避免了业务数据从内存 312 到网卡 313 的数据拷贝，减少了 VM 的热迁移时间，提高了 VM 的热迁移效率。

另外，上述的实施例是以消息队列对（QP）中 SQ 和 RQ 是 1:1 为例进行说明的，在另一些可能的情形中，SQ 和 RQ 的数量比例也可以是 N：M，N 和 M 均为正整数，且 N≠M。例如，SQ 和 RQ 的数量比例为 2:1，如在网卡 313 中，page2 对应的 SG1 信息存储在 SQ（1）、page4 对应的 SG2 信息存储在 SQ（2）；在网卡 323 中，SG1 信息和 SG2 信息保存在一个 RQ，在待迁移数据迁移过程中，网卡 313 基于上述的 SG1 信息和 SG2 信息，将 page2 和 page4 中存储的数据迁移到 RQ 映射的存储地址空间。

在本申请提供的上述实施例中，待迁移数据的迁移可以是通过 TCP/IP 协议或 RDMA

网络等实现的。

在 TCP/IP 场景中，主机 310（源主机）通过 TCP/IP 协议和主机 320（目的主机）建立有 2 个传输通道：控制连接和数据连接。其中，控制连接用于传输页面信息和迁移消息，数据连接用于传输待迁移数据。

在本实施例中，不同的传输通道用于传输不同的信息或数据，避免了页面信息由数据连接传输，导致该页面信息无法被接收侧的网卡 323 进行处理，以及虚拟机的热迁移出现问题，提高了虚拟机的数据迁移稳定性。

另外，在 TCP/IP 场景中，主机 310（源主机）通过 TCP/IP 协议和主机 320（目的主机）也可以仅建立有 1 个传输通道：单一连接。该单一连接可用于传输上述的页面信息、迁移消息和待迁移数据。

这里给出一种可能的具体示例，来说明 TCP/IP 场景中，源主机和目的数据进行数据迁移的过程。

步骤 1、网卡 313 将本次要复制的脏页的页面信息（SG 信息）通过单一连接发送给网卡 323。

其中，发送队列和接收队列中 SG 信息包含的地址可以是 GPA 或者 HPA，只需两者（源 VM 和目的 VM）对应的内存空间在 GPA 层面是完全一致的，能够实现源主机的内存到目的主机的内存的完全复制即可。

如果是 SG 信息包含的地址是 GPA，则需要由 IOMMU 实现 GPA→HPA 的地址转换。

如果是 SG 信息包含的地址是 HPA，则不需要 IOMMU，进行地址转换。由网卡 313 指向直接指定主机 320 归属的用于实现 PF 的 PCIe 设备，直接上送给主机 320 即可，这时由于网卡 313 基于单一连接传输的数据都是发往主机 320，即在该单一连接中，实现 PF 的 PCIe 设备的信息（简称 PF 信息）相同，网卡 313 可以将 PF 配置在 TCP 的连接上下文中。该 PF 信息用于指示待迁移的虚拟机采用的物理 PCIe 设备的标识。

在本示例中，网卡 313 将 PF 信息配置在 TCP 的连接上下文中，避免了一个虚拟机的多组数据在源主机和目的主机的迁移过程中，每组数据都需要配置 PF 信息，减少了数据迁移的总时间，提高了虚拟机的数据迁移效率。

值得注意的是，PF 和 VF 是 PCIe 接口中虚拟化设备的概念，不应理解为对本申请的限定。如果是其它接口，只需使用对应的标识来标记虚拟设备即可。

步骤 2、目的主机基于网卡 323 接收到的页面信息分配一个消息处理 ID（如上述的迁移消息），一个消息处理 ID 对应网卡 323 中的一个接收队列。

在一种可能的示例中，该消息处理 ID 是由网卡 323 分配的，以减少目的主机的负载，提高目的主机处理其他业务的计算能力。

另外，网卡 323 还可以利用热迁移管理程序根据接收到的脏页的页面信息（SG 信息），设置好接收队列。如接收队列中对应的接收内存块的 SG 顺序与发送队列中对应的内存块的 SG 顺序完全一致，其对应的内存空间在 VM 的 GPA 这个层面是完全一致的，接收队列中放置的就是前述的页面信息（SG 信息或 SG 信息块）。

步骤 3、网卡 313 将本次要发送的脏页对应的内存块（脏页数据），加上一个携带前述消息处理 ID 的消息头后，按与 SG 信息相同的内存组织顺序以 SG 方式放入 TCP 的 SQ 中，网卡 313 通知硬件（如通信接口），从 SQ 指定的内存去取数据发往目的主机的 RQ。

步骤4、网卡323接收到携带有消息头（消息处理ID）的数据后，依据消息处理ID，获取对应RQ中存储的SG信息，并将数据写入SG信息对应的目的主机内存。

在一种可选的实现方式中，网卡323向网卡313发送迁移消息的过程中，网卡323还可以向网卡313通知内存322的缓存信息，示例的，该缓存信息包括主机320中用于进行虚拟机热迁移的缓存（buffer）空间的可用余量，如10百万字节（mega byte，MB），在业务数据的传输过程中，网卡313发送的业务数据不超过主机320通知的buffer空间的可用余量，避免业务数据的数据量过大，导致网卡323无法快速的将buffer空间中的数据写入内存322中。

另外，由于源主机和目的主机之间的单一连接可以传输不同VM的数据，上述的PF信息/VF信息还可以用于区分单一连接中数据所属的VM，避免单一连接中多个VM的数据发生传输错误，如将VM（1）的数据错误的识别为VM（2）的数据，进而，提高了VM热迁移的准确性。

VM的热迁移管理过程由网卡实现，不需要消耗源主机的处理器所具有的计算资源；且VM的业务数据由网卡从源主机的内存直接复制到目的主机的内存，减少了业务数据的数据拷贝次数和数据拷贝时间，从而降低了VM的热迁移时延。

由于单一连接中配置了待迁移的虚拟机采用的PF信息，网卡313还可以在一个单位（如10个脏页）的业务数据迁移到目的主机（主机320），且网卡313接收到目的主机反馈的响应，如该响应指示网卡323已经完成该一个单位的业务数据的写操作之后，网卡313发起下一个单位的业务数据迁移过程。

综上，由于VM的热迁移过程由主机中的处理器卸载到网卡，因此，减少了主机中处理器的处理资源消耗。

其次，VM的热迁移过程中，由于源主机的网卡中仅保存脏页的元数据（如SG信息），该网卡中不需要保存脏页的业务数据，且元数据的数据量小于业务数据的数据量，这减少了网卡中存储资源的消耗。

而且，由于网卡可以利用脏页的元数据，将内存的脏页中存储的业务数据发往目的主机，这避免了业务数据从内存-网卡的拷贝过程，减少了数据拷贝时间，从而提高了VM热迁移的时延，提高了VM热迁移的效率。

在本申请的上述实施例中，虚拟机热迁移过程是基于TCP协议来实现的，但是一些可能的情况中，虚拟机热迁移过程也可以是基于RDMA网络实现的。

RDMA是一种绕过远程主机操作系统内核访问其内存中数据的技术，由于不经过操作系统，不仅节省了大量CPU资源，同样也提高了系统吞吐量、降低了系统的网络通信延迟，尤其适合在大规模并行计算机集群中有广泛应用。RDMA有几大特点，（1）数据通过网络与远程机器间进行数据传输；（2）没有操作系统内核的参与，有关发送传输的所有内容都卸载到智能网卡上；（3）在用户空间虚拟内存与智能网卡直接进行数据传输不涉及操作系统内核，没有额外的数据移动和复制。

目前，大致有三类RDMA网络，分别是无限带宽（Infiniband，IB）、承载融合以太网上的RDMA（RDMA over Converged Ethernet，RoCE）、互联网广域RDMA协议（internet wide area RDMA protocol，iWARP）。其中，Infiniband是一种专为RDMA设计的网络，从硬件上保证可靠传输，需要支持该技术的网卡和交换机。而RoCE和iWARP都是基于以

太网的 RDMA 技术，只需要配置特殊的网卡即可。从性能上，Infiniband 网络最好，但网卡和交换机是价格也很高，而 RoCE 和 iWARP 仅需使用特殊的网卡就可以了，价格也相对便宜很多。

下面给出一种可能的示例，对 RDMA 场景中虚拟机的数据迁移方法进行说明。

当上述的网卡 313 和网卡 323 基于 RDMA 网络来实现本申请提供的数据迁移方法时，网卡 313 和网卡 323 也可以称为智能网卡，这里给出一种该智能网卡可能的硬件和软件实现：智能网卡包括 CPU 和网络适配器（network interface card，NIC），该 CPU 上运行有热迁移管理程序。

在 RDMA 场景中，网卡 313 将待迁移 VM（或称源 VM）的全部内存空间 GPA 与 HPA 注册成一个存储（memory）区，生成内存保护表（memory protect table，MPT），以及内存翻译表（memory translation table，MTT），其中，MPT 表中增加访问主机的 PF 信息，得到本地的源本地密钥（source local key，S_LKey）和源远程密钥（source remote key，S_RKey）。其中，该 MPT 表用于指示内存的 HPA 和待迁移的虚拟机的 GPA 的对应关系。

同样的，网卡 323 将 VM 的全部内存空间 GPA 与 HPA 注册成一个 memory 区，生成 MPT 表和 MTT 表，其中，MPT 表中增加访问主机的 PF 信息，得到本地的目的本地密钥（destination local key，D_LKey）和目的远程密钥（destination remote key，D_RKey）。

由 RDMA 场景中的 MPT 表来实现上述 TCP/IP 场景中页面信息和迁移消息的功能，以实现本申请提供的数据迁移方法。

示例的，网卡 313 向迁移消息指示的 RQ 发送待迁移数据，可以包括以下过程：首先，网卡 313 依据迁移信息确定待迁移的虚拟机的 PF 信息；其次，根据待迁移的虚拟机的 PF 信息和 GPA 查询 MPT 表，确定页面信息对应的 HPA；最后，网卡 313 向目的主机（主机 320）发送待迁移的虚拟机的 HPA 中存储的待迁移数据。

例如，网卡 313 中运行的热迁移管理程序获得内存 312 中脏页的描述信息（元数据），以 GPA 地址确定网卡 323 中注册的 D_RKey 为 R-KEY，逐批发起“RDMA write”将数据写往目的主机（主机 320）中的内存 322。

值得注意的是，RDMA 的传输模式有双边操作也有单边操作。send/receive 属于双边操作，即需要远端的应用感知参与才能完成收发。read 和 write 是单边操作，只需要本端明确信息的源地址和目的地址，远端应用不必感知此次通信，数据的读或存都通过远端的网卡完成，再由远端网卡封装成消息返回到本端。

例如，在本实施例提供的数据迁移方法中，send/receive 可以用于传输脏页的描述信息，read 和 write 可以用于传输脏页中存储的业务数据。

这里以 RDMA 写操作为例对本实施例的数据迁移方法进行说明，RDMA 写操作用于请求端（如网卡 313）将数据写入响应端（如主机 320）的存储空间。

在允许网卡 313 进行 RDMA 写操作之前，主机 320 首先分配一个存储空间供主机 320 的 QP（或 QP 组）访问。主机 320 的通道适配器将一个密钥与此存储空间的虚拟地址相关联。主机 320 将该存储空间的虚拟地址、长度和密钥发送给可以访问该内存区域的网卡 313。示例性的，可以通过前文所述的发送操作来将上述信息发送给网卡 313。存储空间的虚拟地址和密钥可以用于确定脏页的 HPA。

网卡 313 可以通过发送 RDMA write 消息来发起 RDMA 写操作，该消息中包括待写至

主机 320 的数据、主机 320 的存储空间的虚拟地址、数据的长度和密钥。数据的长度可以在 0 字节到 231 字节之间，与发送操作类似的，如果数据的长度大于路径最大传输单元(path maximum transmission unit，PMTU)，将按照 PMTU 大小分段为多个报文，主机 320 再将这些报文重新组合得到数据。对于可靠连接，如果数据是短消息(即不必分段为多个报文)，主机 320 针对每个报文向网卡 313 发送确认报文；如果数据是长消息(即分段为多个报文)，主机 320 可以针对每个报文向网卡 313 发送确认报文，或者，针对同一数据的连续多个报文向网卡 313 发送一个确认报文，或者，针对报文的尾包向网卡 313 发送确认报文；另外，无论数据是短消息还是长消息，主机 320 可以针对之前接收的多个报文发送一个确认报文，例如，一个报文序列号（packet sequence numbers，PSN）为 X 的 RDMA write 消息的确认报文可以用于确认该 RDMA write 消息之前的 PSN 小于 X 的消息已被主机 320 成功接收。

可以理解的是，为了实现上述实施例中功能，主机和网卡包括了执行各个功能相应的硬件结构和/或软件模块。本领域技术人员应该很容易意识到，结合本申请中所公开的实施例描述的各示例的单元及方法步骤，本申请能够以硬件或硬件和计算机软件相结合的形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行，取决于技术方案的特定应用场景和设计约束条件。

上文中结合图 1 至图 5，详细描述了根据本申请所提供的数据迁移的方法，下面将结合图 6，描述根据本申请所提供的数据迁移装置。

图 6 为本申请提供的一种数据迁移装置的结构示意图，数据迁移装置 600 可以用于实现上述方法实施例中主机或网卡的功能，因此也能实现上述方法实施例所具备的有益效果。在本申请的实施例中，该数据迁移装置 600 可以是如图 2 所示的网卡 250，也可以是图 3~图 5 中所示出的网卡 313 或网卡 323，还可以是应用于网卡的模块（如芯片）。

如图 6 所示，数据迁移装置 600 包括通信单元 610、数据识别单元 620、迁移单元 630 和存储单元 640。该数据迁移装置 600 用于实现上述图 3~图 5 中所示的方法实施例中网卡的功能。在一种可能的示例中，该数据迁移装置 600 用于实现上述数据迁移方法的具体过程包括以下内容 1~3。

1、通信单元 610 用于获取数据迁移通知消息。该第一数据迁移通知消息用于指示待迁移的虚拟机的标识。

2、数据识别单元 620 用于根据标识确定待迁移数据，该待迁移数据为存储在源主机的内存中，且与待迁移的虚拟机关联的数据。

示例的，数据识别单元 620 可以根据前述的标识确定源主机的内存中与待迁移的虚拟机关联的内存页集合，并将该内存页集合中存储的数据作为待迁移数据。

另外，如图 6 所示，数据迁移装置 600 包括的存储单元 640 用于存储脏页标记信息，数据识别单元 620 可以根据数据迁移通知消息中待迁移的虚拟机的标识，和该脏页标记信息，确定源主机内存中与待迁移的虚拟机关联的脏页，以确定该关联的脏页中存储的脏页数据。关于脏页标记信息更详细的内容可以参考上述方法实施例中表 1 的相关阐述，此处不予赘述。

3、迁移单元 630 用于将前述的待迁移数据迁移至目的主机。

待迁移数据是由数据识别单元依据待迁移的虚拟机的标识确定的，避免了源主机对虚拟机的待迁移数据进行确定的过程，减少了源主机在虚拟机热迁移过程中所需的计算资

源，提高了源主机执行其他业务（如 AI、HPC 等计算密集型和时延敏感型业务）的能力。

当数据迁移装置 600 用于实现图 3 所示的方法实施例中主机 310 的功能时：通信单元 610 用于执行 S310；数据识别单元 620 用于执行 S320；迁移单元 630 用于执行 S330。

当数据迁移装置 600 用于实现图 3 所示的方法实施例中主机 320 的功能时：通信单元 610 用于执行 S330；迁移单元 630 用于执行 S340。

当数据迁移装置 600 用于实现图 4 所示的方法实施例中主机 310 的功能时：通信单元 610 用于执行 S410；迁移单元 630 用于执行 S430。

当数据迁移装置 600 用于实现图 4 所示的方法实施例中主机 320 的功能时：通信单元 610 用于执行 S410~S430。

另外，当数据迁移装置 600 部署在接收侧主机（如虚拟机迁移过程中的目的主机）时，通信单元 610 可以用于接收其他主机发送的虚拟机数据，该迁移单元 630 可以将虚拟机数据迁移到接收队列中 SG 信息映射的内存地址空间，以避免数据迁移装置 600 在接收侧主机对该虚拟机数据进行多次拷贝，减少目的主机的计算资源和存储资源消耗，提高虚拟机的热迁移效率，以及目的主机执行其他计算业务的处理能力。

应理解的是，本发明本申请实施例的数据迁移装置 600 可以通过 CPU 实现，也可以通过 ASIC 实现，或可编程逻辑器件（programmable logic device，PLD）实现，上述 PLD 可以是复杂程序逻辑器件（complex programmable logical device，CPLD）、FPGA、通用阵列逻辑（generic array logic，GAL）或其任意组合。数据迁移装置 600 通过软件实现图 3 至图 5 中任一所示的数据迁移方法时，数据迁移装置 600 及其各个模块也可以为软件模块。

有关上述数据迁移装置 600 更详细的描述可以直接参考上述图 3~图 5 所示的实施例中相关描述直接得到，这里不加赘述。

示例的，当数据迁移装置 600 通过硬件实现时，该硬件可以是一种电子设备，如上述的网卡，或应用在网卡的处理器或芯片等，如该电子设备包括接口电路和控制电路。

接口电路用于接收来自电子设备之外的其它设备的信号并传输至控制电路，或将来自控制电路的信号发送给电子设备之外的其它设备。

控制电路通过逻辑电路或执行代码指令用于实现上述实施例中任一种可能实现方式的方法。有益效果可以参见上述任一实施例中的描述，此处不再赘述。

应理解，根据本申请实施例的网卡可对应于申请实施例中的数据迁移装置 600，并可以对应于执行根据本申请实施例的方法图 3~图 5 中的相应主体，并且数据迁移装置 600 中的各个模块的上述和其它操作和/或功能分别为了实现图 3 至图 5 中的各个方法的相应流程，为了简洁，在此不再赘述。

可以理解的是，本申请的实施例中的处理器可以是 CPU、NPU 或 GPU，还可以是其它通用处理器、DSP、ASIC、FPGA 或者其它可编程逻辑器件、晶体管逻辑器件，硬件部件或者其任意组合。通用处理器可以是微处理器，也可以是任何常规的处理器。

本申请的实施例中的方法步骤可以通过硬件的方式来实现，也可以由处理器执行软件指令的方式来实现。软件指令可以由相应的软件模块组成，软件模块可以被存放于随机存取存储器（random access memory，RAM）、闪存、只读存储器（read-only memory，ROM）、可编程只读存储器（programmable ROM，PROM）、可擦除可编程只读存储器（erasable PROM，EPROM）、电可擦除可编程只读存储器（electrically EPROM，EEPROM）、寄存器、

硬盘、移动硬盘、CD-ROM 或者本领域熟知的任何其它形式的存储介质中。一种示例性的存储介质耦合至处理器，从而使处理器能够从该存储介质读取信息，且可向该存储介质写入信息。当然，存储介质也可以是处理器的组成部分。处理器和存储介质可以位于 ASIC 中。另外，该 ASIC 可以位于网络设备或终端设备中。当然，处理器和存储介质也可以作为分立组件存在于网络设备或终端设备中。

在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机程序或指令。在计算机上加载和执行所述计算机程序或指令时，全部或部分地执行本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、网络设备、用户设备或者其它可编程装置。所述计算机程序或指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，所述计算机程序或指令可以从一个网站站点、计算机、服务器或数据中心通过有线或无线方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是集成一个或多个可用介质的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质，例如，软盘、硬盘、磁带；也可以是光介质，例如，数字视频光盘（digital video disc，DVD）；还可以是半导体介质，例如，固态硬盘（solid state drive，SSD）。

在本申请的各个实施例中，如果没有特殊说明以及逻辑冲突，不同的实施例之间的术语和/或描述具有一致性、且可以相互引用，不同的实施例中的技术特征根据其内在的逻辑关系可以组合形成新的实施例。在本申请的实施例中涉及的各种数字编号仅为描述方便进行的区分，并不用来限制本申请的实施例的范围。上述各过程的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定。

权 利 要 求 书

1.一种数据迁移方法，其特征在于，所述方法由源主机的网卡执行，所述方法包括：

获取第一数据迁移通知消息，所述第一数据迁移通知消息用于指示待迁移的虚拟机的标识；

根据所述标识确定待迁移数据，所述待迁移数据为存储在所述源主机的内存中，且与所述待迁移的虚拟机关联的数据；

将所述待迁移数据迁移至目的主机。

2.根据权利要求 1 所述的方法，其特征在于，

根据所述标识确定待迁移数据，包括：

根据所述标识确定所述源主机的内存中与所述待迁移的虚拟机关联的内存页集合，所述内存页集合包括一个或多个内存页；

将所述内存页集合中存储的数据作为所述待迁移数据。

3.根据权利要求 2 所述的方法，其特征在于，所述待迁移数据包括脏页数据，所述脏页数据为所述一个或多个内存页中，数据发生修改的内存页所存储的数据。

4.根据权利要求 3 所述的方法，其特征在于，

根据所述标识确定所述待迁移数据包括的脏页数据，包括：

查询所述网卡中保存的脏页标记信息，确定与所述标识关联的脏页集合；所述脏页集合包括一个或多个脏页，所述脏页为所述一个或多个内存页中数据发生修改的内存页，所述脏页标记信息用于指示所述脏页的内存地址；

将所述脏页集合中存储的数据作为所述脏页数据。

5.根据权利要求 4 所述的方法，其特征在于，

所述脏页标记信息包括第一脏页表和第二脏页表中至少一个；

所述第一脏页表用于标记所述脏页为标脏状态，所述标脏状态为所述源主机对所述脏页中存储的数据进行修改的状态；

所述第二脏页表用于标记所述脏页为数据迁移状态，所述数据迁移状态为所述网卡对所述脏页中存储的数据进行迁移的状态。

6.根据权利要求 3-5 中任一项所述的方法，其特征在于，

将所述待迁移数据迁移至目的主机，包括：

向目的主机发送所述待迁移数据的页面信息，所述页面信息用于指示所述待迁移数据在所述内存中的内存地址和偏移量；

接收所述目的主机基于所述页面信息反馈的迁移消息，所述迁移消息用于指示所述目的主机中与所述待迁移的虚拟机相应的接收队列 RQ；

向所述迁移消息指示的 RQ 发送所述待迁移数据。

7.根据权利要求 6 所述的方法，其特征在于，所述网卡中发送队列 SQ 维护有包含所述脏页的内存地址和偏移量的 SG 信息，所述页面信息包括所述脏页对应的 SG 信息。

8.根据权利要求 7 所述的方法，其特征在于，

向所述迁移消息指示的 RQ 发送所述待迁移数据，包括：

从所述 SQ 中获取所述脏页对应的 SG 信息；

向所述迁移消息指示的 RQ 发送所述 SG 信息指示的数据。

9.根据权利要求 6 所述的方法，其特征在于，所述源主机通过传输控制协议/网络之间互连协议 TCP/IP 和所述目的主机建立有控制连接和数据连接，所述控制连接用于传输所述页面信息和所述迁移消息，所述数据连接用于传输所述待迁移数据。

10.根据权利要求 6 所述的方法，其特征在于，所述源主机通过 TCP/IP 和所述目的主机建立有单一连接，所述单一连接用于传输所述页面信息、所述迁移消息和所述待迁移数据。

11.根据权利要求 10 所述的方法，其特征在于，

所述迁移消息是所述目的主机为所述待迁移的虚拟机分配的消息处理标识 ID。

12.根据权利要求 6 所述的方法，其特征在于，所述源主机通过远程直接内存访问 RDMA 网络和所述目的主机进行通信，所述网卡中存储有内存保护 MPT 表，所述 MPT 表用于指示所述内存的主机物理地址 HPA 和所述待迁移的虚拟机的客户机物理地址 GPA 的对应关系，且所述 MPT 表中包含有所述待迁移的虚拟机所使用的物理功能 PF 信息；

向所述迁移消息指示的 RQ 发送所述待迁移数据，包括：

依据所述迁移信息确定所述待迁移的虚拟机的 PF 信息；

根据所述待迁移的虚拟机的 PF 信息和 GPA 查询所述 MPT 表，确定所述页面信息对应的 HPA；

向所述目的主机发送所述待迁移的虚拟机的 HPA 中存储的待迁移数据。

13.根据权利要求 1 所述的方法，其特征在于，所述方法还包括：

获取第二数据迁移通知消息，所述第二数据迁移通知消息用于指示待接收的虚拟机的标识；

将其他主机发送的待接收数据迁移至所述源主机的内存中，所述待接收数据为存储在所述其他主机的内存中，且与所述待接收的虚拟机关联的数据。

14.一种数据迁移装置，其特征在于，所述数据迁移装置应用于源主机的网卡，所述数据迁移装置包括：

通信单元，用于获取第一数据迁移通知消息，所述第一数据迁移通知消息用于指示待迁移的虚拟机的标识；

数据识别单元，用于根据所述标识确定待迁移数据，所述待迁移数据为存储在所述源主机的内存中，且与所述待迁移的虚拟机关联的数据；

迁移单元，用于将所述待迁移数据迁移至目的主机。

15.一种电子设备，其特征在于，包括：接口电路和控制电路；

所述接口电路用于接收来自所述电子设备之外的其它设备的信号并传输至所述控制电路，或将来自所述控制电路的信号发送给所述电子设备之外的其它设备，所述控制电路通过逻辑电路或执行代码指令用于实现如权利要求 1 至 13 中任一项所述的方法。

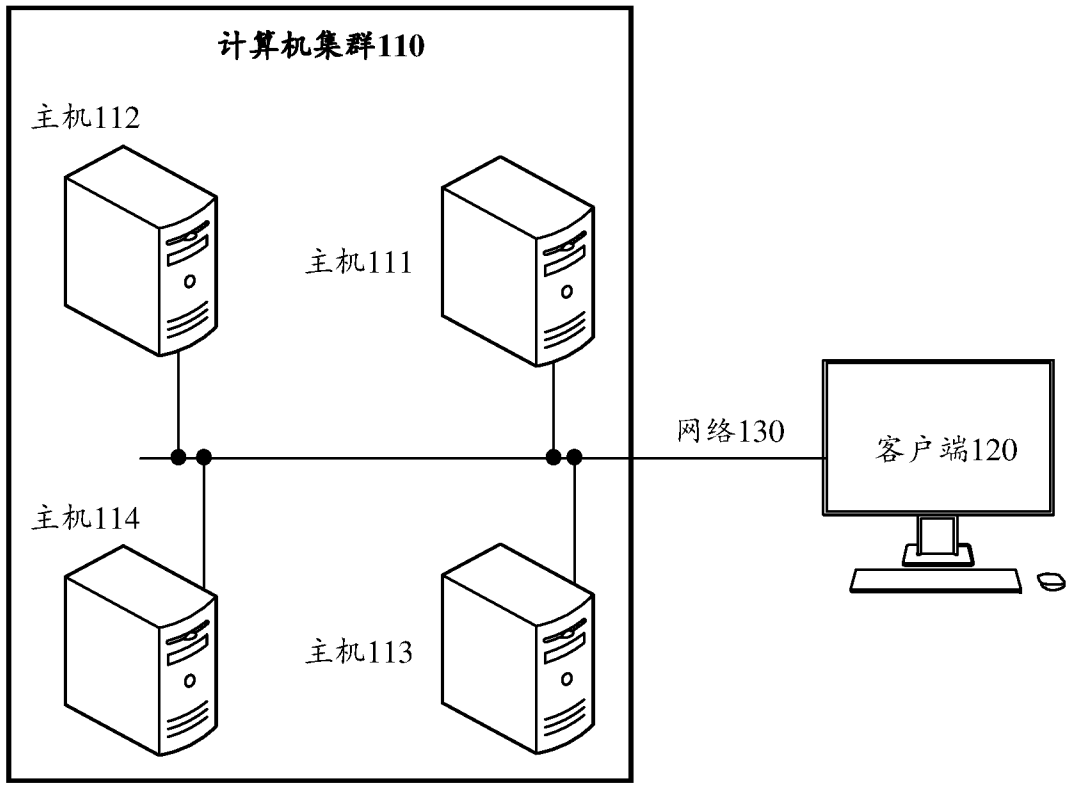
计算机集群110
主机112
主机111
主机114
主机113
网络130
客户端120

图 1

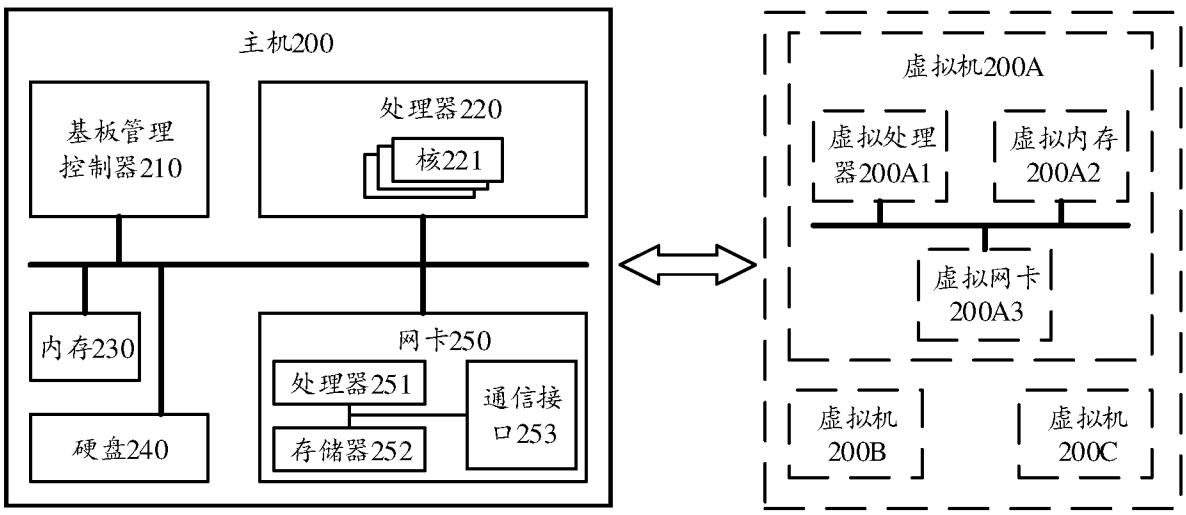
主机200
基板管理控制器210
处理器220
核221
内存230
硬盘240
网卡250
处理器251
存储器252
通信接口253
虚拟机200A
虚拟处理器200A1
虚拟内存200A2
虚拟网卡200A3
虚拟机200B
虚拟机200C

图 2

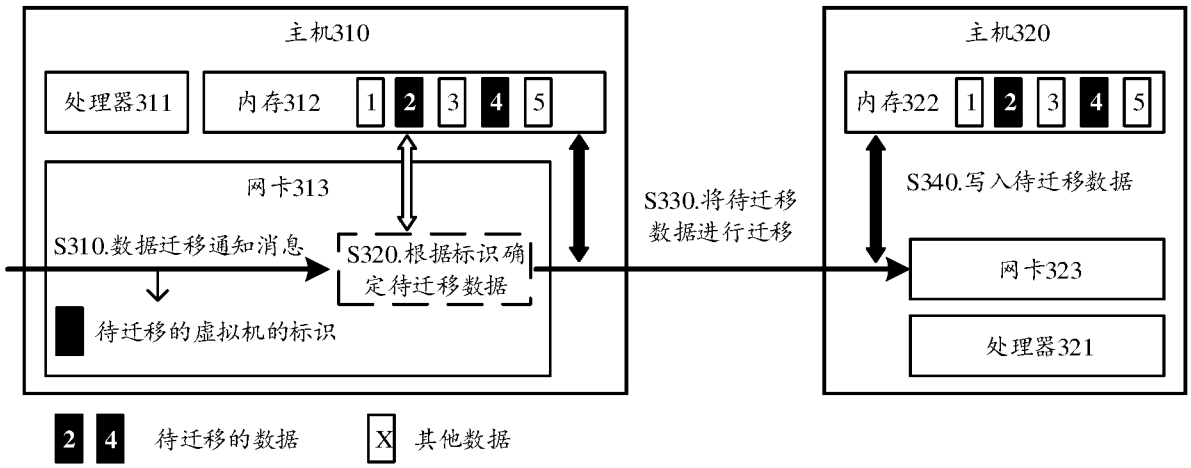
主机310
处理器311
内存312
1
2
3
4
5
网卡313
S310.数据迁移通知消息
S320.根据标识确定待迁移数据
待迁移的虚拟机的标识
S330.将待迁移数据进行迁移
主机320
内存322
S340.写入待迁移数据
网卡323
处理器321
待迁移的数据
其他数据

图 3

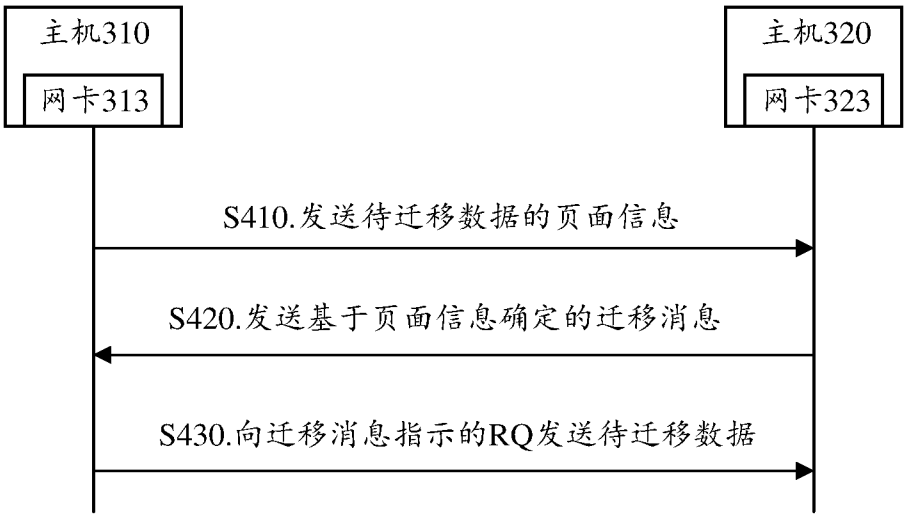
主机310
网卡313
主机320
网卡323
S410.发送待迁移数据的页面信息
S420.发送基于页面信息确定的迁移消息
S430.向迁移消息指示的RQ发送待迁移数据

图 4

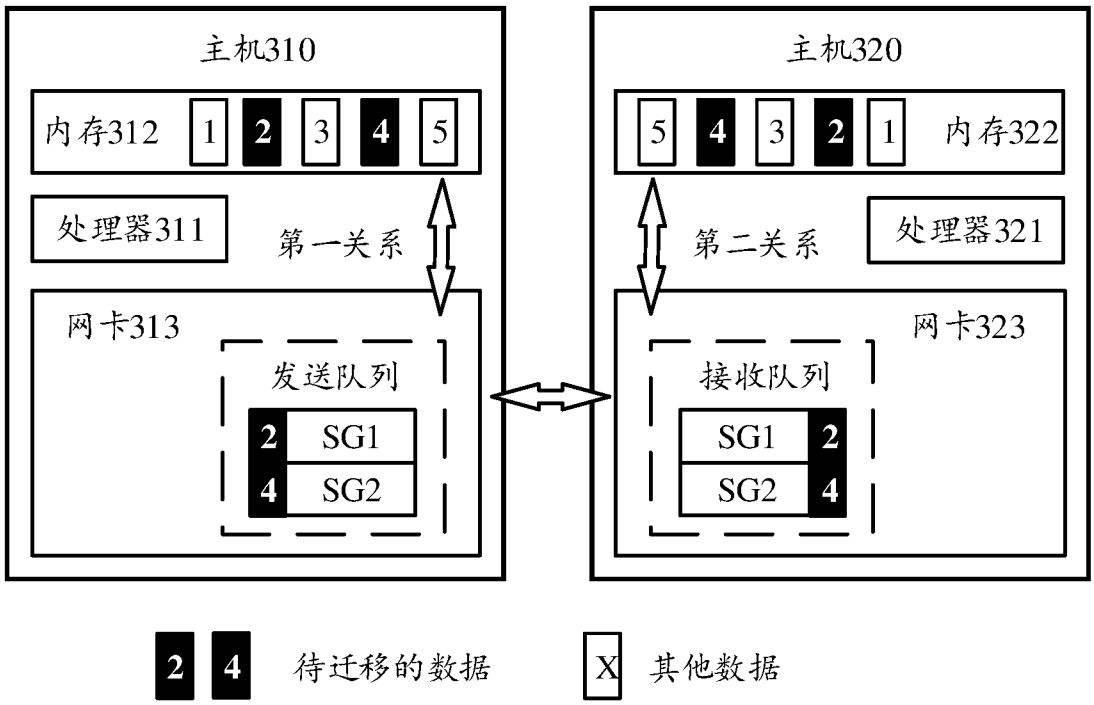
主机310
内存312
1
2
3
4
5
处理器311
第一关系
网卡313
发送队列
2
SG1
4
SG2
主机320
5
4
3
2
1
内存322
第二关系
处理器321
网卡323
接收队列
SG1
2
SG2
4
2
4
待迁移的数据
X
其他数据

图 5

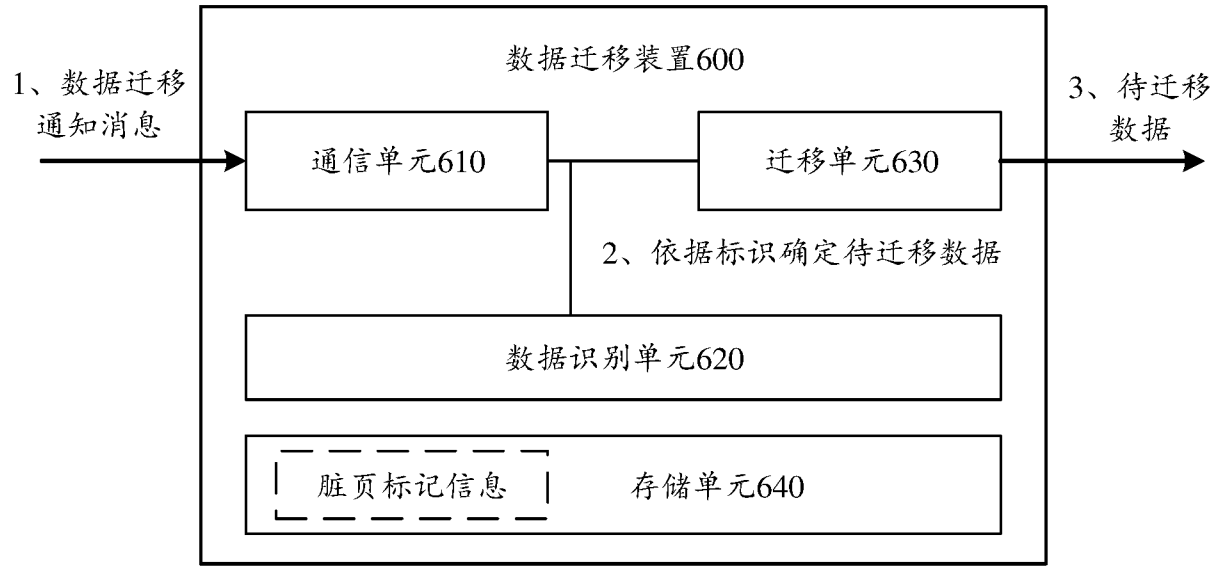
数据迁移装置600
1、数据迁移通知消息
通信单元610
迁移单元630
3、待迁移数据
2、依据标识确定待迁移数据
数据识别单元620
脏页标记信息
存储单元640

图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/127151

A. CLASSIFICATION OF SUBJECT MATTER

G06F 9/455(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT; CNKI; ENTXTC; ENTXT; DWPI: 华为, 虚拟机, 迁移, 网卡, 智能网卡, 队列, 性能, 处理器, 资源, virtual machine, VM, migrat+, transfer+, network, card, NIC, smart, intelligent, queue, performance, processor, resource

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 111736945 A (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD. et al.) 02 October 2020 (2020-10-02) description, paragraphs 25-103, and figures 1-5	1-15
Y	CN 111666036 A (HUAWEI TECHNOLOGIES CO., LTD.) 15 September 2020 (2020-09-15) description, paragraphs 4-5 and 41-105	1-15
Y	CN 103618809 A (HUAWEI TECHNOLOGIES CO., LTD.) 05 March 2014 (2014-03-05) description, paragraphs 77-88	6-12, 15
A	CN 103530167 A (HUAWEI TECHNOLOGIES CO., LTD.) 22 January 2014 (2014-01-22) entire document	1-15
A	CN 109918172 A (FIBERHOME TELECOMMUNICATION TECHNOLOGIES CO., LTD.) 21 June 2019 (2019-06-21) entire document	1-15
A	US 2016139944 A1 (FREESCALE SEMICONDUCTOR INC. et al.) 19 May 2016 (2016-05-19) entire document	1-15

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
02 January 2023	**11 January 2023**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/127151

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 111736945 A	02 October 2020	None	
CN 111666036 A	15 September 2020	CN 111666036 B	22 April 2022
		WO 2020177567 A1	10 September 2020
CN 103618809 A	05 March 2014	None	
CN 103530167 A	22 January 2014	WO 2015043147 A1	02 April 2015
		EP 2879053 A1	03 June 2015
		US 2015095443 A1	02 April 2015
		US 9854036 B2	26 December 2017
		EP 2879053 A4	14 October 2015
		EP 2879053 B1	16 November 2016
CN 109918172 A	21 June 2019	None	
US 2016139944 A1	19 May 2016	US 9811367 B2	07 November 2017

国际检索报告

国际申请号
PCT/CN2022/127151

A. 主题的分类

G06F 9/455(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNTXT;CNKI;ENTXTC;ENTXT;DWPI:华为，虚拟机，迁移，网卡，智能网卡，队列，性能，处理器，资源，virtual machine, VM, migrat+, transfer+, network, card, NIC, smart, intelligent, queue, performance, processor, resource

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
Y	CN 111736945 A (北京京东尚科信息技术有限公司等) 2020年10月2日 (2020-10-02) 说明书第25-103段、图1-5	1-15
Y	CN 111666036 A (华为技术有限公司) 2020年9月15日 (2020-09-15) 说明书第4-5、41-105段	1-15
Y	CN 103618809 A (华为技术有限公司) 2014年3月5日 (2014-03-05) 说明书第77-88段	6-12，15
A	CN 103530167 A (华为技术有限公司) 2014年1月22日 (2014-01-22) 全文	1-15
A	CN 109918172 A (烽火通信科技股份有限公司) 2019年6月21日 (2019-06-21) 全文	1-15
A	US 2016139944 A1 (FREESCALE SEMICONDUCTOR INC等) 2016年5月19日 (2016-05-19) 全文	1-15

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2023年1月2日	2023年1月11日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 王越 电话号码 62089109

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2022/127151

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	111736945	A	2020年10月2日	无			
CN	111666036	A	2020年9月15日	CN	111666036	B	2022年4月22日
				WO	2020177567	A1	2020年9月10日
CN	103618809	A	2014年3月5日	无			
CN	103530167	A	2014年1月22日	WO	2015043147	A1	2015年4月2日
				EP	2879053	A1	2015年6月3日
				US	2015095443	A1	2015年4月2日
				US	9854036	B2	2017年12月26日
				EP	2879053	A4	2015年10月14日
				EP	2879053	B1	2016年11月16日
CN	109918172	A	2019年6月21日	无			
US	2016139944	A1	2016年5月19日	US	9811367	B2	2017年11月7日