

一种虚拟机的热迁移方法、设备、系统及存储介质

本申请要求于 2022 年 08 月 26 日提交中国专利局、申请号为 202211035631.8、申请名称为“一种虚拟机的热迁移方法、设备、系统及存储介质”的中国专利申请的优先权，其
5 全部内容通过引用结合在本申请中。

技术领域

本申请涉及云计算技术领域，尤其涉及一种虚拟机的热迁移方法、设备、系统及存储
10 介质。

背景技术

在公有云中，存在较多大规格虚拟机，该类虚拟机由于配置高、负载重，一直是业界
热迁移的难点，主要体现在如下两点：

- 1.内存大，脏页率特别高，通常的网络下，热迁移难以达到迭代收敛。
- 15 2.物理 CPU 资源上承载的实例多，尤其是实例已经占满整机的情况，没有空余物理
CPU 资源留给热迁移。

但是，目前业界所有的热迁移手段，都需要在虚拟机的运行环境中消耗 CPU 资源用于
热迁移工作，从而不得不使热迁移与 VCPU 共享物理 CPU 资源，这一方面会对虚拟机造
成工作抖动，另一方面，由于热迁移得不到充足的计算资源，导致迁移性能不佳。

20

发明内容

本申请的多个方面提供一种虚拟机的热迁移方法、设备、系统及存储介质，用以改善
虚拟机的热迁移对源宿主机的资源占用问题。

本申请实施例提供一种虚拟机的热迁移系统，包括：源宿主机及其对应的第一加
25 速设备、目的宿主机：

所述第一加速设备，用于在热迁移过程中，从所述源宿主机中读取脏页信息；基
于所述脏页信息查找本轮待传输脏页的源地址信息；将所述源地址信息发送至所述目
的宿主机；

所述目的宿主机，用于获取所述源地址信息；基于所述源地址信息而发起脏页读
30 取请求；利用所述第一加速设备从所述源宿主机中获取本轮待传输脏页。

本申请实施例还提供一种虚拟机的热迁移方法，适用于为源宿主机配置的第一加
速设备，所述方法包括：

在热迁移过程中，从所述源宿主机中读取脏页信息；

基于所述脏页信息查找本轮待传输脏页的源地址信息；

将所述源地址信息发送至目的宿主机，以供所述目的宿主机基于所述源地址信息而发起脏页读取请求；

从所述源宿主机中读取本轮待传输脏页；

5 将读取到的本轮待传输脏页提供给所述目的宿主机中，以响应所述脏页读取请求。

本申请实施例还提供一种虚拟机的热迁移方法，适用于为目的宿主机配置的第二加速设备，所述方法包括：

在热迁移过程中，接收为源宿主机配置的第一加速设备提供的本轮待传输脏页的源地址信息；

10 将所述源地址信息提供给所述目的宿主机；

按照所述目的宿主机基于所述源地址信息而发起的脏页读取请求，利用所述第一加速设备从所述源宿主机中读取本轮待传输脏页；

将获取到的本轮待传输脏页写入所述目的宿主机中。

15 本申请实施例还提供一种加速设备，与源宿主机连接，包括脏页迭代组件和通信组件；

所述脏页迭代组件，用于在热迁移过程中，从所述源宿主机中读取脏页信息；基于所述脏页信息查找本轮待传输脏页的源地址信息；将所述源地址信息发送至目的宿主机，以供所述目的宿主机基于所述源地址信息而发起脏页读取请求；

20 所述通信组件，用于从所述源宿主机中读取本轮待传输脏页；通过所述通信组件将读取到的本轮待传输脏页发送给所述目的宿主机中，以响应所述脏页读取请求。

本申请实施例还提供一种加速设备，与目的源宿主机连接，包括脏页迭代组件和通信组件；

所述脏页迭代组件，用于在热迁移过程中，接收为源宿主机配置的第一加速设备提供的本轮待传输脏页的源地址信息；将所述源地址信息提供给所述目的宿主机；

25 所述通信组件，用于按照所述目的宿主机基于所述源地址信息而发起的脏页读取请求，利用所述第一加速设备从所述源宿主机中读取本轮待传输脏页；将获取到的本轮待传输脏页写入所述目的宿主机中。

30 本申请实施例还提供一种存储计算机指令的计算机可读存储介质，当所述计算机指令被一个或多个处理器执行时，致使所述一个或多个处理器执行前述的虚拟机的热迁移方法。

在本申请实施例中，在源宿主机和目的宿主机上分别对应有加速设备，在源宿主机一侧，由其上的加速设备在热迁移过程中从源宿主机中读取脏页信息；基于所述脏页信息查找本轮待传输脏页的源地址信息；将所述源地址信息发送至目的宿主机侧；而在目的宿主机一侧，可获取到源地址信息；基于源地址信息而发起的脏页读取请求，并利用源宿主机侧的加速设备从源宿主机中读取本轮待传输脏页。据此，本申请实施例中，提供了一种对

35

传统热迁移架构的优化方案，通过在源宿主机侧配置加速设备，可在源宿主机侧将控制面和数据面的热迁移工作全部下沉到加速设备上执行，只需由目的宿主机和加速设备配合即可完成热迁移工作，这实现了热迁移工作对源宿主机资源的零依赖，即使在源宿主机计算资源全被占用的情况下，热迁移的性能也不受影响，且热迁移功能对源宿主机上运行中的虚拟机实现零抖动。

附图说明

此处所说明的附图用来提供对本申请的进一步理解，构成本申请的一部分，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

- 图 1a 为本申请一示例性实施例提供的一种虚拟机的热迁移系统的结构示意图；
- 图 1b 为本申请一示例性实施例提供的另一种虚拟机的热迁移系统的结构示意图；
- 图 2 为本申请一示例性实施例提供的一种加速设备的硬件实现形式的示意图；
- 图 3 为本申请另一示例性实施例提供的一种虚拟机的热迁移方法的流程示意图；
- 图 4 为本申请另一示例性实施例提供的另一种虚拟机的热迁移方法的流程示意图；
- 图 5 为本申请又一示例性实施例提供的一种加速设备的结构示意图；
- 图 6 为本申请又一示例性实施例提供的另一种加速设备的结构示意图。

具体实施方式

为使本申请的目的、技术方案和优点更加清楚，下面将结合本申请具体实施例及相应的附图对本申请技术方案进行清楚、完整地描述。显然，所描述的实施例仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，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，都属于本申请保护的范围。

目前，虚拟机的热迁移工作需要占用源宿主机上的资源，导致迁移效率不佳，还可能导致虚拟机的工作抖动。为此，本申请的一些实施例中：在源宿主机和目的宿主机上分别对应加速设备，在源宿主机一侧，由其上的加速设备在热迁移过程中从源宿主机中读取脏页信息；基于所述脏页信息查找本轮待传输脏页的源地址信息；将所述源地址信息发送至目的宿主机侧；而在目的宿主机一侧，可获取到源地址信息；基于源地址信息而发起的脏页读取请求，并利用源宿主机侧的加速设备从源宿主机中读取本轮待传输脏页。据此，本申请实施例中，提供了一种对传统热迁移架构的优化方案，通过在源宿主机侧配置加速设备，可在源宿主机侧将控制面和数据面的热迁移工作全部下沉到加速设备上执行，只需由目的宿主机和加速设备配合即可完成热迁移工作，这实现了热迁移工作对源宿主机资源的零依赖，即使在源宿主机计算资源全被占用的情况下，热迁移的性能也不受影响，且热迁移功能对源宿主机上运行中的虚拟机实现零抖动。

以下结合附图，详细说明本申请各实施例提供的技术方案。

- 图 1a 为本申请一示例性实施例提供的一种虚拟机的热迁移系统的结构示意图。如图

1a 所示，该系统包括：源宿主机及其对应的第一加速设备、目的宿主机。

本实施例提供的虚拟机热迁移系统中，继承了传统的预拷贝 pre-copy 热迁移方案，并对其进行了改进。其中，pre-copy 热迁移方案是由源宿主机通过迭代，将内存数据发送至目的宿主机，迭代第一轮发送所有内存数据，接下来每一轮发送上一轮预拷贝过程中的脏页，最后一轮是停机拷贝阶段，源宿主机被挂起，停止内存更新，将脏页整体拷贝到目的宿主机上。本实施例提供的虚拟机迁移系统对 pre-copy 热迁移方案中的脏页迭代过程和脏页传输过程均进行了改进。

本实施例中，源宿主机是指需要迁移的虚拟机所在的宿主机，目的宿主机则是需要迁移的虚拟机所将迁移到的宿主机。

10 正如前文提及的，在 pre-copy 热迁移方案中，需要进行多轮迁移，为了便于描述，本实施例中将从其中一轮迁移过程的角度来进行技术方案的阐述。应当理解的是，在各轮迁移过程中可采用同样的技术方案来实现迁移。比较特殊的是 pre-copy 热迁移方案中的第一轮迁移过程，在第一轮迁移过程中，可将脏页信息中所有的内存页都认为是脏页而无需执行脏页查找操作，在此基础上可按照本实施例提供的技术方案将源宿主机中的所有内存数据迁移到目的宿主机中。

参考图 1a，本实施例中，为源宿主机配置的第一加速设备可在热迁移过程中，从源宿主机中读取脏页信息。脏页信息可用于存储分配给虚拟机的若干内存页是否为脏页的标记。其中，第一加速设备可外接与源宿主机，当然，也可集成在源宿主机内，本实施例对双方的实际连接方式不做限定。可选地，可采用脏页位图作为脏页信息，脏页位图可以是二进制序列。应当理解的是，本实施例中脏页信息还可采用其它信息形式，而并不限于脏页位图这一种。举例来说，若为虚拟机分配了 512KB 内存，每个内存页的大小为 4KB，则该内存页的数目为 $512/4=128$ ，从而虚拟机对应一个包含 128 位的脏页位图，每个位代表一个内存页是否为脏页，若脏页位图的第一位为 1，则代表第一个内存页为脏页；若脏页位图的第二位为 0，则代表第二个内存页不为脏页。值得说明的是，本实施例中，脏页位图是源宿主机上既有的，通常可由源宿主机上的 KVM (Kernel-based Virtual Machine，基于内核的虚拟机) 进行维护。

基于此，本实施例中，可将 pre-copy 热迁移方案中的脏页迭代过程(即控制面的工作)全部下沉到第一加速设备中进行。为此，对于第一加速设备来说，可基于脏页信息查找本轮待传输脏页的源地址信息。本实施例中，第一加速设备可遍历脏页信息，也即是对脏页信息进行逐位扫描，以发现其中脏页的源地址信息。源地址信息用于表征脏页在源宿主机内存空间中的位置。这里，脏页迭代过程将完全与源宿主机解耦，该过程对源宿主机中的资源零依赖，也就是该过程完全无需占用源宿主机中的 CPU 资源。

参考图 1a，第一加速设备可将本轮待传输脏页的源地址信息发送至目的宿主机。

本实施例中，目的宿主机中可运行一定定制的热迁移程序，由于在热迁移完成之前，目的宿主机上会存在大量的空闲资源，因此，本实施例中，可使用目的宿主机上的空闲资源

来支持该热迁移程序的运行。举例来说，目的宿主机上通常会为待迁移的虚拟机预留 CPU 资源，本实施例中，热迁移程序即可使用目的宿主机上为迁移后虚拟机划定的 CPU 资源来支持该热迁移程序的运行。应当理解的是，在完成虚拟机迁移之前，为迁移后虚拟机划定的 CPU 资源是空闲的、可用的。当然，这仅是示例性的，本实施例中，还可使用目的宿主机上的其它空闲资源来支持其上热迁移程序的运行，本实施例对此不做限定。

目的宿主机可基于其上运行的定制的热迁移程序，来与第一加速设备配合，实施 pre-copy 热迁移中的脏页传输过程（即数据面的工作）。目的宿主机可在其上定制的热迁移程序的驱动下，根据获取到的源地址信息而发起脏页读取请求。由于本实施例中已经改变了传统方案中由源宿主机主动推送脏页的架构，因此，这里，目的宿主机还可将本轮待传输脏页在目的宿主机上所分配到的目的地址信息也携带在脏页读取请求中，以便读回脏页后可写入目的宿主机内存空间中的正确位置上。其中，热迁移过程中，源宿主机上为内存数据分配的地址与目的宿主机上为内存数据预留的地址通常是一一对应的，因此，目的宿主机可在接收到源地址信息的情况下，基于这种对应关系而方便快捷地查找到脏页所对应的目的地址，并携带在脏页读取请求中。

图 1b 为本申请一示例性实施例提供的另一种虚拟机的热迁移系统的结构示意图。参考图 1b，优选地，本实施例中，还可在目的宿主机一侧配置第二加速设备，在此基础上，目的宿主机也与第一加速设备和第二加速设备配合，实施 pre-copy 热迁移中的脏页传输过程（即数据面的工作）。

参考图 1b，在增配第二加速设备的基础上，第一加速设备可将待传输脏页的源地址信息发送给第二加速设备。对于第二加速设备来说，则可将收到的源地址信息提供给目的宿主机，并可触发目的宿主机生成脏页读取请求。而目的宿主机则可将脏页读取请求发送至其对应的第二加速设备，参考图 1b，对于第二加速设备来说，可按照目的宿主机发起的脏页读取请求，利用第一加速设备从源宿主机中读取本轮待传输脏页。也即是，第一加速设备和第二加速设备可相互配合，来响应目的宿主机发起的脏页读取请求。由于第一加速设备与源宿主机连接，因此，可由第一加速设备从源宿主机中读取本轮待传输脏页后，返回给第二加速设备。需要说明的是，本实施例中，第一加速设备和第二加速设备相互配合读取本轮待传输脏页的过程中，并无需限定脏页的读取顺序。

优选地，本实施例中，第一加速设备可按照直接内存访问 DMA（direct memory access，直接内存访问）的方式，从源宿主机中读取本轮待传输脏页；将读取到的本轮待传输脏页提供给第二加速设备。DMA 的方式可支持第一加速设备顺利地从源宿主机中读取本轮待传输的脏页，而无需依赖源宿主机中的 CPU 资源。这样，脏页传输过程也将完全与源宿主机中解耦，该过程对源宿主机中的资源零依赖，也就是该过程完全无需占用源宿主机中的 CPU 资源。

至此可知，本实施例中，将热迁移过程中控制面和数据面的工作全部从源宿主机中解耦出去，下沉至为源宿主机配置的第一加速设备来执行，因此，源宿主机上无需执行任何

与热迁移相关的任务,无需为此付出任何 CPU 资源,真正实现了对源宿主机资源的零依赖。

继续参考图 1a, 第一加速设备在从源宿主机中读取到本轮待传输脏页后, 可将本轮待传输脏页提供给目的宿主机。基于图 1b 所示的系统架构, 本实施例中, 第一加速设备可将本轮待传输脏页回传给第二加速设备; 第二加速设备可将获取到的本轮待传输脏页写入目的宿主机中。优选地, 本实施例中, 第二加速设备可按照直接内存访问 DMA 的方式将本轮待传输脏页写入目的宿主机中。正如前文提及的, 目的宿主机发起的脏页读取请求中携带有脏页在目的宿主机上的目的地址信息, 因此, 这里, 第二加速设备可按照脏页读取请求中的目的地址信息将接收到的本轮待传输脏页写入目的宿主机中对应的地址上。这样, 目的宿主机可在发出脏页读取请求后, 收到其对应的第二加速设备所写入的本轮待传输脏页, 从而完成本轮脏页的迁移。

可以看到, 整个脏页传输过程, 源宿主机的 CPU 完全没有参与, 只有目的宿主机 CPU 和第一加速设备 (在优选方案中还有第二加速设备) 参与来完成脏页传输。因此, 一方面, 达到了源宿主机 CPU 资源的零消耗, 源宿主机上运行中的虚拟机也不会因为 CPU 资源争抢出现业务抖动; 另一方面, 脏页传输的性能也不会受源宿主机 CPU 资源的限制, 通过目的宿主机上闲置 CPU 和两侧的加速设备来搬运脏页, 可以使脏页传输性能一直保持在硬件的峰值上, 使脏页传输的性能得到保证。

据此, 本实施例中, 在源宿主机和目的宿主机上分别对应有加速设备, 在源宿主机一侧, 由其上的加速设备在热迁移过程中从源宿主机中读取脏页信息; 基于所述脏页信息查找本轮待传输脏页的源地址信息; 将所述源地址信息发送至目的宿主机侧; 而在目的宿主机一侧, 可获取到源地址信息; 基于源地址信息而发起的脏页读取请求, 并利用源宿主机侧的加速设备从源宿主机中读取本轮待传输脏页。据此, 本申请实施例中, 提供了一种对传统热迁移架构的优化方案, 通过在源宿主机侧配置加速设备, 可在源宿主机侧将控制面和数据面的热迁移工作全部下沉到加速设备上执行, 只需由目的宿主机和加速设备配合即可完成热迁移工作, 这实现了热迁移工作对源宿主机资源的零依赖, 有效改善了热迁移工作的效率。

在上述或下述实施例中, 可采用远程直接内存访问 RDMA 的方式来实现脏页传输过程的下沉。

本实施例中, 目的宿主机可根据源地址信息以及本轮待传输脏页在目的宿主机上分配到的目的地址信息, 生成远程直接内存访问 RDMA 指令, 作为脏页读取请求; 在目的宿主机配置有第二加速设备的情况下, 目的宿主机可将远程直接内存访问指令发送给第二加速设备。相应地, 对于第二加速设备来说, 则可按照远程直接内存访问 RDMA 的方式, 利用第一加速设备从源宿主机中读取本轮待传输脏页。当然, 在目的宿主机未配置有第二加速设备的情况下, 目的宿主机也可自主按照远程直接内存访问 RDMA 的方式, 利用第一加速设备从源宿主机中读取本轮待传输脏页。图 2 为本申请一示例性实施例提供的一种加速设备的硬件实现形式的示意图。参考图 2, 在硬件实现上, 为了支持按照 RDMA 方式

来实现脏页传输过程，本实施例中，第一加速设备中可包含第一 RDMA 组件，第二加速设备中可包含第二 RDMA 组件。其中，RDMA 组件具有 RDMA 通信能力。可选地，本实施例中，可采用 FPGA 等可编程器件来模拟 RDMA 组件，也可采用 RDMA 网卡来作为 RDMA 组件，还可基于 ASIC 的具有 RDMA 能力的其它设备来作为 RDMA 组件，本实施

5 例对 RDMA 组件的硬件实现形式不做限定。尤其是 FPGA，由于其具有专用集成电路，因此，特别适配于本实施例中脏页传输过程的计算量大但逻辑简单的工作特点，FPGA 模拟出的 RDMA 组件可出色地完成脏页传输过程中的数据收发工作。当然，这里提到的这几种 RDMA 组件的硬件实现形式也仅是示例性的，本实施例并不限于此。

在此基础上，一种示例性的脏页传输方案中：对于第二 RDMA 组件来说，可将目的宿主机针对各个本轮待传输脏页发出的 RDMA 指令存入发送队列中；将各个 RDMA 指令封装为 RDMA 读报文后分别发送至第一 RDMA 组件中维护的接收队列；对于第一 RDMA 组件来说，则可从接收队列中的 RDMA 读报文中解析出相关脏页的源地址信息；将根据源地址信息从源宿主机中读取到的相关脏页封装为 RDMA 应答报文后发

10 送至第二 RDMA 组件，具体地，可以发送至第二 RDMA 组件维护的应答队列中。这样，第一 RDMA 组件可协助第二 RDMA 组件按照远程直接内存访问 RDMA 的方式从源宿主机中读取本轮待传输脏页。

可选地，考虑到 pre-copy 热迁移方案中，需要保证每轮脏页迁移的完整性和顺序性，因此，在该示例性方案中，第二 RDMA 组件中还可维护完成队列，第二 RDMA 组件可在完成队列中维护本轮待传输脏页的完成状态。对于目的宿主机来说，可在第二 RDMA 组件中预先创建本轮待传输脏页的完成队列，第二 RDMA 组件可在该队列中的任意脏页完成传输后在完成队列中填入相应脏页对应的完成标识，这样，目的宿主机可通过拉取的方式来感知完成队列中各脏页的完成状态。在完成队列中的所有脏页均获得完成标识后，目的宿主机可确认本轮待传输脏页已经全部完成传输，并可启动下一轮迁移。应当理解的是，目的宿主机及第二 RDMA 组件之间还可采用其它方式

20 来同步本轮待传输脏页是否全部完成，例如，第二 RDMA 组件可在每完成一个脏页的传输后，产生一个通知事件，并提供给目的宿主机，以供目的宿主机感知哪些脏页已完成传输，并可自主判断本轮待传输脏页是否已全部完成，等。本实施例对此不做限定。

其中，在该示例性方案中，第一 RDMA 组件和第二 RDMA 组件可按照 RDMA 标准协议来维护各自相关队列，并交互相关报文，以实现按照 RDMA 方式响应目的宿主机发起的 RDMA 指令。当然，本实施例对此并不限定，第一 RDMA 组件和第二 RDMA 组件也可按照其它自定义的传输协议来实现响应目的宿主机发起的 RDMA 请求，也即是，第一 RDMA 组件和第二 RDMA 组件之间的交互细节并不限于上述的示例性方案，在此不再详述。

30

35 另外，本实施例中，除了包含 RDMA 组件之外，第一加速设备中还可包含第一脏页

迭代组件，第二加速设备中还可包含第二脏页迭代组件。可选地，本实施例中，脏页迭代组件可采用系统级 SOC 芯片或专用宿主机进行实现。SOC 芯片，可被称为片上系统，其是一个产品，是一个有专用目标的集成电路，其中包含完整系统并有嵌入软件的全部内容，因此，特别适配于本实施例中脏页迭代过程的逻辑复杂但计算量小的工作特定，SCO 芯片可在脏页迭代过程中出色地完成脏页信息遍历工作。而对于采用专用宿主机来实现脏页迭代组件的情况下，多个源宿主机可共用同一专用宿主机，该专用宿主机可分别为不同的源宿主机提供脏页迭代功能支持，举例来说，该专用宿主机可为不同的源宿主机运行不同的脏页迭代进程，以隔离不同源宿主机对应的脏页迭代工作，每个脏页迭代进程的工作逻辑可参考前文中第一加速设备关于脏页迭代过程的描述，在此不再赘述。

10 基于此，本实施例中，可由第一脏页迭代组件承担前述的在热迁移过程中，从源宿主机中读取脏页信息，基于脏页信息查找本轮待传输脏页的源地址信息，以及将源地址信息发送至第二加速设备等操作；由第二脏页迭代组件承担将源地址信息提供给目的宿主机的操作。

值得说明的是，对于 SOC 芯片及专用宿主机等硬件形式来实现脏页迭代组件的情况下，15 可将前述的 RDMA 组件作为脏页迭代组件与源/目的宿主机之间的硬件桥，这样，第一脏页迭代组件可以第一 RDMA 组件作为硬件桥（例如，FPGA 通道）来从源宿主机中读取脏页信息，而第二脏页迭代组件和目的宿主机之间也可以第二 RDMA 组件作为硬件桥来传递本轮待传输脏页的地址信息等。当然，这仅是示例性的，本实施例并不限于此。

据此，本实施例中，加速设备可包含 RDMA 组件和脏页迭代组件，脏页迭代组件可用于下沉脏页迭代过程，而 RDMA 组件则可用于下沉脏页传输过程。

本实施例中，除了可采用前述的远程直接内存访问 RDMA 的方式来实现脏页传输过程的下沉之外，还可采用其它通信方式来实现脏页传输过程的下沉。例如，可采用 TCP 的方式来实现脏页传输过程，在 TCP 的方式下，加速设备可采用 SOC 芯片或专用宿主机来实现，在该示例性下，加速设备中不再需要配置 RDMA 组件，而是可由源宿主机对应的25 SOC 芯片或专用宿主机直接通过 DMA 等方式来从源宿主机读取脏页信息及本轮待传输脏页，而两端的加速设备之间则可通过 TCP 协议进行交互，并可由目的宿主机外接的 SOC 芯片或专用宿主机直接通过 DMA 等方式来从目的宿主机中接收脏页读取请求或向目的宿主机中写入本轮待传输脏页等。本实施例对脏页传输过程中采用的通信方式不做限定，对于不同的通信方式，可采用合适的硬件形式来实现加速设备。

30 据此，本实施例中，可采用 RDMA 等通信方式来实现虚拟机热迁移中的脏页传输过程，整个过程中，无需源宿主机参与，保证了对源宿主机的零依赖。并且可在目的宿主机单边完成脏页传输过程，实现了一种类 post-copy 的 pre-copy 热迁移效果，也即是，将原本需要在源宿主机中执行的工作，转由目的宿主机来执行，这可实现对源宿主机资源的零占用，一方面可以将热迁移导致的虚拟机抖动问题消除，另一方面，使热迁移性能脱离源35 宿主机中的 CPU 资源的束缚，从而稳定保持热迁移的高性能。

图3为本申请另一示例性实施例提供的一种虚拟机的热迁移方法的流程示意图，该方法可由加速装置执行，该加速装置可实现为软件和/或硬件的结合，该加速装置可集成在为源宿主机配置的第一加速设备中。参考图3，该方法包括：

- 步骤300、在热迁移过程中，从源宿主机中读取脏页信息；
- 5 步骤301、基于脏页信息查找本轮待传输脏页的源地址信息；
- 步骤302、将源地址信息发送至目的宿主机，以供目的宿主机基于源地址信息而发起脏页读取请求；
- 步骤303、从源宿主机中读取本轮待传输脏页；
- 步骤304、将读取到的本轮待传输脏页发送给目的宿主机以响应脏页读取请求。
- 10 在一可选实施例中，目的宿主机配置有第二加速设备，步骤302包括：
- 将源地址信息发送至为目的宿主机配置的第二加速设备，以供第二加速设备触发目的宿主机基于源地址信息而发起脏页读取请求；
- 步骤304包括：将读取到的本轮待传输脏页发送给第二加速设备，以配合第二加速设备响应脏页读取请求。
- 15 在一可选实施例中，步骤303可包括：
- 按照直接内存访问DMA的方式，从源宿主机中读取本轮待传输脏页。
- 在一可选实施例中，目的宿主机可根据源地址信息以及本轮待传输脏页在目的宿主机上分配到的目的地址信息，生成远程直接内存访问RDMA指令，作为脏页读取请求；并将远程直接内存访问指令发送给第二加速设备；
- 20 步骤将读取到的本轮待传输脏页发送给第二加速设备中，可包括：按照远程直接内存访问RDMA的方式，将读取到的本轮待传输脏页发送给第二加速设备。
- 在一可选实施例中，第一加速设备中包含第一RDMA组件，第二加速设备中包含第二RDMA组件；该方法具体包括：
- 将第二RDMA组件发送的RDMA读报文存储在接收队列中，RDMA读报文是由
- 25 第二RDMA组件将目的宿主机针对各个本轮待传输脏页发出的RDMA指令存入发送队列中并对各个RDMA指令进行封装而产生的；
- 从接收队列中的RDMA读报文中解析出相关脏页的源地址信息；
- 将根据源地址信息从源宿主机中读取到的相关脏页封装为RDMA应答报文后发送至第二RDMA组件，以协助第二RDMA组件按照远程直接内存访问RDMA的方式
- 30 从源宿主机中读取本轮待传输脏页。
- 在一可选实施例中，第一RDMA组件和第二RDMA组件采用FPGA、RDMA网卡或基于ASIC的具有RDMA能力的其它设备进行实现。
- 在一可选实施例中，第一加速设备中还包含第一脏页迭代组件，第二加速设备中还包含第二脏页迭代组件，由第一脏页迭代组件承担步骤300-302。
- 35 在一可选实施例中，第一脏页迭代组件和第二脏页迭代组件采用系统级SOC芯片

或专用宿主机进行实现。

在一可选实施例中，第一加速设备和第二加速设备采用 SOC 芯片或者专用宿主机进行实现，第一加速设备和第二加速设备之间按照 TCP 协议从源宿主机中读取本轮待传输脏页。

5 值得说明的是，上述关于虚拟机的迁移方法的各实施例中的技术细节，可参考前述的系统实施例中关于第一加速设备的相关描述，为节省篇幅，在此不再赘述，但这不应造成本申请保护范围的损失。

图 4 为本申请另一示例性实施例提供的另一种虚拟机的热迁移方法的流程示意图，该方法可由加速装置执行，该加速装置可实现为软件和/或硬件的结合，该加速装置可集成在
10 为目的宿主机配置的第二加速设备中。参考图 4，该方法包括：

步骤 400、在热迁移过程中，接收为源宿主机配置的第一加速设备提供的本轮待传输脏页的源地址信息；

步骤 401、将源地址信息提供给目的宿主机；

步骤 402、按照目的宿主机基于源地址信息而发起的脏页读取请求，利用第一加速设备从源宿主机中读取本轮待传输脏页；
15

步骤 403、将获取到的本轮待传输脏页写入目的宿主机中。

在一可选实施例中，步骤 403 包括：按照直接内存访问 DMA 的方式将本轮待传输脏页写入目的宿主机中。

在一可选实施例中，目的宿主机可根据源地址信息以及本轮待传输脏页在目的宿主机上分配到的目的地址信息，生成远程直接内存访问 RDMA 指令，作为脏页读取请求；将远程直接内存访问指令发送给第二加速设备；步骤 402 可包括：
20

按照远程直接内存访问 RDMA 的方式，利用第一加速设备从源宿主机中读取本轮待传输脏页。

在一可选实施例中，第一加速设备中包含第一 RDMA 组件，第二加速设备中包含第二 RDMA 组件；该方法具体包括：
25

将目的宿主机针对各个本轮待传输脏页发出的 RDMA 指令存入发送队列中；将各个 RDMA 指令封装为 RDMA 读报文后分别发送至第一 RDMA 组件中维护的接收队列，以供第一 RDMA 组件从接收队列中的 RDMA 读报文中解析出相关脏页的源地址信息；将根据源地址信息从源宿主机中读取到的相关脏页封装为 RDMA 应答报文后发送至
30 第二 RDMA 组件。

在一可选实施例中，该方法还包括：在完成队列中维护本轮待传输脏页的完成状态，以供目的宿主机根据完成队列中本轮待传输脏页的完成状态确定本轮待传输脏页全部完成迁移后，启动下一轮迁移。

在一可选实施例中，第一 RDMA 组件和第二 RDMA 组件采用 FPGA、RDMA 网卡或基于 ASIC 的具有 RDMA 能力的其它设备进行实现。
35

在一可选实施例中，第一加速设备中还包含第一脏页迭代组件，第二加速设备中还包含第二脏页迭代组件，由第二脏页迭代组件承担将源地址信息提供给目的宿主机的操作。

5 在一可选实施例中，第一脏页迭代组件和第二脏页迭代组件采用系统级 SOC 芯片或专用宿主机进行实现。

在一可选实施例中，第一加速设备和第二加速设备采用 SOC 芯片或者专用宿主机进行实现，第一加速设备和第二加速设备之间按照 TCP 协议从源宿主机中读取本轮待传输脏页。

10 值得说明的是，上述关于虚拟机的迁移方法的各实施例中的技术细节，可参考前述的系统实施例中关于第二加速设备的相关描述，为节省篇幅，在此不再赘述，但这不应造成本申请保护范围的损失。

另外，在上述实施例及附图中的描述的一些流程中，包含了按照特定顺序出现的多个操作，但是应该清楚了解，这些操作可以不按照其在本文中出现的顺序来执行或并行执行，操作的序号如 801、802 等，仅仅是用于区分开各个不同的操作，序号本身不代表任何的
15 执行顺序。另外，这些流程可以包括更多或更少的操作，并且这些操作可以按顺序执行或并行执行。需要说明的是，本文中的“第一”、“第二”等描述，是用于区分不同的设备、组件等，不代表先后顺序，也不限定“第一”和“第二”是不同的类型。

图 5 为本申请又一示例性实施例提供的一种加速设备的结构示意图。该加速设备可与源宿主机连接。如图 5 所示，该加速设备包括：脏页迭代组件 51 和通信组件 52；
20 脏页迭代组件 51，用于在热迁移过程中，从源宿主机中读取脏页信息；基于脏页信息查找本轮待传输脏页的源地址信息；将源地址信息发送至目的宿主机，以供目的宿主机基于源地址信息而发起脏页读取请求；

通信组件 52，用于从源宿主机中读取本轮待传输脏页；将读取到的本轮待传输脏页发送给目的宿主机中，以响应脏页读取请求。

25 在一可选实施例中，目的宿主机配置有第二加速设备，脏页迭代组件 51 在将源地址信息发送至目的宿主机的过程中，可用于：

将源地址信息发送至为目的宿主机配置的第二加速设备，以供第二加速设备触发目的宿主机基于源地址信息而发起脏页读取请求；

脏页迭代组件 51 在将读取到的本轮待传输脏页发送给目的宿主机的过程中，可用
30 于：将读取到的本轮待传输脏页发送给第二加速设备，以配合第二加速设备响应脏页读取请求。

在一可选实施例中，通信组件 52 从源宿主机中读取本轮待传输脏页过程中，可用于：

按照直接内存访问 DMA 的方式，从源宿主机中读取本轮待传输脏页。
35 在一可选实施例中，目的宿主机可根据源地址信息以及本轮待传输脏页在目的宿

主机上分配到的目的地址信息，生成远程直接内存访问 RDMA 指令，作为脏页读取请求；并将远程直接内存访问指令发送给第二加速设备；

通信组件 52 在将读取到的本轮待传输脏页发送给第二加速设备的过程中，可用于：按照远程直接内存访问 RDMA 的方式，将读取到的本轮待传输脏页发送给第二加速设备。

在一可选实施例中，通信组件 52 中可包含第一 RDMA 组件，第二加速设备中可包含第二 RDMA 组件；脏页迭代组件 51 可用于：

将第二 RDMA 组件发送的 RDMA 读报文存储在接收队列中，RDMA 读报文是由第二 RDMA 组件将目的宿主机针对各个本轮待传输脏页发出的 RDMA 指令存入发送队列中并对各个 RDMA 指令进行封装而产生的；

从接收队列中的 RDMA 读报文中解析出相关脏页的源地址信息；

将根据源地址信息从源宿主机中读取到的相关脏页封装为 RDMA 应答报文后发送至第二 RDMA 组件，以协助第二 RDMA 组件按照远程直接内存访问 RDMA 的方式从源宿主机中读取本轮待传输脏页。

在一可选实施例中，第一 RDMA 组件和第二 RDMA 组件采用 FPGA、RDMA 网卡或基于 ASIC 的具有 RDMA 能力的其它设备进行实现。

在一可选实施例中，本加速设备中的脏页迭代组件和目的宿主机连接的第二加速设备中的脏页迭代组件采用系统级 SOC 芯片或专用宿主机进行实现。

在一可选实施例中，本加速设备和第二加速设备采用 SOC 芯片或者专用宿主机进行实现，本加速设备和第二加速设备之间按照 TCP 协议从源宿主机中读取本轮待传输脏页。

进一步，该加速设备还包括：电源组件 53 等其它组件。

值得说明的是，上述关于加速设备各实施例中的技术细节，可参考前述的系统实施例中关于第一加速设备的相关描述，为节省篇幅，在此不再赘述，但这不应造成本申请保护范围的损失。

图 6 为本申请又一示例性实施例提供的另一种加速设备的结构示意图。该加速设备可与目的宿主机连接。如图 6 所示，该加速设备包括：脏页迭代组件 61 和通信组件 62；

所述脏页迭代组件 61，用于在热迁移过程中，接收为源宿主机配置的第一加速设备提供的本轮待传输脏页的源地址信息；将所述源地址信息提供给所述目的宿主机；

所述通信组件 62，用于按照所述目的宿主机基于所述源地址信息而发起的脏页读取请求，利用所述第一加速设备从所述源宿主机中读取本轮待传输脏页；将获取到的本轮待传输脏页写入所述目的宿主机中。

在一可选实施例中，通信组件 62 将获取到的本轮待传输脏页写入目的宿主机的过程中，可用于：按照直接内存访问 DMA 的方式将本轮待传输脏页写入目的宿主机中。

在一可选实施例中，目的宿主机可根据源地址信息以及本轮待传输脏页在目的宿主机上分配到的目的地址信息，生成远程直接内存访问 RDMA 指令，作为脏页读取请求；将远程直接内存访问指令发送给本加速设备；通信组件 62 在利用第一加速设备从源宿主机中读取本轮待传输脏页的过程中，可用于：

- 5 按照远程直接内存访问 RDMA 的方式，利用第一加速设备从源宿主机中读取本轮待传输脏页。

在一可选实施例中，第一加速设备中包含第一 RDMA 组件，通信组件 62 中可包含第二 RDMA 组件；该方法具体包括：

- 10 将目的宿主机针对各个本轮待传输脏页发出的 RDMA 指令存入发送队列中；将各个 RDMA 指令封装为 RDMA 读报文后分别发送至第一 RDMA 组件中维护的接收队列，以供第一 RDMA 组件从接收队列中的 RDMA 读报文中解析出相关脏页的源地址信息；将根据源地址信息从源宿主机中读取到的相关脏页封装为 RDMA 应答报文后发送至第二 RDMA 组件。

- 15 在一可选实施例中，脏页迭代组件 61 还可用于：在完成队列中维护本轮待传输脏页的完成状态，以供目的宿主机根据完成队列中本轮待传输脏页的完成状态确定本轮待传输脏页全部完成迁移后，启动下一轮迁移。

在一可选实施例中，第一 RDMA 组件和第二 RDMA 组件采用 FPGA、RDMA 网卡或基于 ASIC 的具有 RDMA 能力的其它设备进行实现。

- 20 在一可选实施例中，源宿主机上第一加速设备中的第一脏页迭代组件和本加速设备这边的脏页迭代组件 61 采用系统级 SOC 芯片或专用宿主机进行实现。

在一可选实施例中，第一加速设备和本加速设备采用 SOC 芯片或者专用宿主机进行实现，第一加速设备和本加速设备之间按照 TCP 协议从源宿主机中读取本轮待传输脏页。

- 25 进一步，该加速设备还包括：电源组件 63 等其它组件。

值得说明的是，上述关于加速设备各实施例中的技术细节，可参考前述的系统实施例中关于第二加速设备的相关描述，为节省篇幅，在此不再赘述，但这不应造成本申请保护范围的损失。

- 30 相应地，本申请实施例还提供一种存储有计算机程序的计算机可读存储介质，计算机程序被执行时能够实现上述方法实施例中可由第一加速设备或第二加速设备执行的各步骤。

- 35 本领域内的技术人员应明白，本申请的实施例可提供为方法、系统、或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

本申请是参照根据本申请实施例的方法、设备（系统）、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

在一个典型的配置中，计算设备包括一个或多个处理器（CPU）、输入/输出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器（RAM）和/或非易失性内存等形式，如只读存储器（ROM）或闪存（flash RAM）。内存是计算机可读介质的示例。

计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存（PRAM）、静态随机存取存储器（SRAM）、动态随机存取存储器（DRAM）、其他类型的随机存取存储器（RAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、快闪记忆体或其他内存技术、只读光盘只读存储器（CD-ROM）、数字多功能光盘（DVD）或其他光学存储、磁盒式磁带，磁带式磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体（transitory media），如调制的数据信号和载波。

还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

以上所述仅为本申请的实施例而已，并不用于限制本申请。对于本领域技术人员来说，

本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

权 利 要 求 书

1、一种虚拟机的热迁移系统，其特征在于，包括：源宿主机及其对应的第一加速设备、目的宿主机；

5 所述第一加速设备，用于在热迁移过程中，从所述源宿主机中读取脏页信息；基于所述脏页信息查找本轮待传输脏页的源地址信息；将所述源地址信息发送至所述目的宿主机；

所述目的宿主机，用于获取所述源地址信息；基于所述源地址信息而发起脏页读取请求；利用所述第一加速设备从所述源宿主机中获取本轮待传输脏页。

10 2、根据权利要求1所述的系统，其特征在于，所述目的宿主机对应第二加速设备，所述第一加速设备在将所述源地址信息发送至所述目的宿主机的过程中，用于：将所述源地址信息发送至为所述目的宿主机配置的第二加速设备；

所述第二加速设备，用于：

将所述源地址信息提供给所述目的宿主机；

15 按照所述脏页读取请求，利用所述第一加速设备从所述源宿主机中读取本轮待传输脏页；

将获取到的本轮待传输脏页写入所述目的宿主机中。

3、根据权利要求2所述的系统，其特征在于，所述第一加速设备还用于：按照直接内存访问 DMA 的方式，从所述源宿主机中读取本轮待传输脏页；将读取到的本轮待传输脏页提供给所述第二加速设备；

20 所述第二加速设备在将获取到的本轮待传输脏页写入所述目的宿主机中的过程中，用于：按照直接内存访问 DMA 的方式将本轮待传输脏页写入所述目的宿主机中。

4、根据权利要求2所述的系统，其特征在于，所述目的宿主机用于：根据所述源地址信息以及所述本轮待传输脏页在所述目的宿主机上分配到的目的地址信息，生成远程直接内存访问 RDMA 指令，作为所述脏页读取请求；将所述远程直接内存访问指令发送给所述第二加速设备；

所述第二加速设备在利用所述第一加速设备从所述源宿主机中读取本轮待传输脏页的过程中，用于：按照远程直接内存访问 RDMA 的方式，利用所述第一加速设备从所述源宿主机中读取本轮待传输脏页。

30 5、根据权利要求4所述的系统，其特征在于，所述第一加速设备中包含第一 RDMA 组件，所述第二加速设备中包含第二 RDMA 组件；

所述第二 RDMA 组件，用于将所述目的宿主机针对各个本轮待传输脏页发出的 RDMA 指令存入发送队列中；将各个 RDMA 指令封装为 RDMA 读报文后分别发送至所述第一 RDMA 组件中维护的接收队列；

35 所述第一 RDMA 组件，用于从接收队列中的 RDMA 读报文中解析出相关脏页的源地址信息；将根据所述源地址信息从所述源宿主机中读取到的相关脏页封装为

RDMA 应答报文后发送至所述第二 RDMA 组件，以协助所述第二 RDMA 组件按照远程直接内存访问 RDMA 的方式从所述源宿主机中读取本轮待传输脏页。

6、根据权利要求 5 所述的系统，其特征在于，所述第二 RDMA 组件，还用于在完成队列中维护本轮待传输脏页的完成状态：

5 所述目的宿主机，还用于根据所述完成队列中本轮待传输脏页的完成状态确定本轮待传输脏页全部完成迁移后，启动下一轮迁移。

7、根据权利要求 5 所述的系统，其特征在于，所述第一 RDMA 组件和所述第二 RDMA 组件采用 FPGA、RDMA 网卡或基于 ASIC 的具有 RDMA 能力的其它设备进行实现。

10 8、根据权利要求 5 所述的系统，其特征在于，所述第一加速设备中还包含第一脏页迭代组件，所述第二加速设备中还包含第二脏页迭代组件，由所述第一脏页迭代组件承担所述在热迁移过程中，从所述源宿主机中读取脏页信息，所述基于所述脏页信息查找本轮待传输脏页的源地址信息，以及所述将所述源地址信息发送至所述第二加速设备的操作：

15 由所述第二脏页迭代组件承担所述将所述源地址信息提供给所述目的宿主机的操作。

9、根据权利要求 8 所述的系统，其特征在于，所述第一脏页迭代组件和所述第二脏页迭代组件采用系统级 SOC 芯片或专用宿主机进行实现。

20 10、根据权利要求 1 所述的系统，其特征在于，所述目的宿主机上运行有热迁移程序，所述热迁移程序使用所述目的宿主机上为迁移后虚拟机划定的 CPU 资源，执行所述根据所述源地址信息而发起脏页读取请求的操作。

11、根据权利要求 2 所述的系统，其特征在于，所述第一加速设备和所述第二加速设备采用 SOC 芯片或者专用宿主机进行实现，所述第一加速设备和所述第二加速设备之间按照 TCP 协议从所述源宿主机中读取本轮待传输脏页。

25 12、一种虚拟机的热迁移方法，其特征在于，适用于源宿主机对应的第一加速设备，所述方法包括：

在热迁移过程中，从所述源宿主机中读取脏页信息；

基于所述脏页信息查找本轮待传输脏页的源地址信息；

30 将所述源地址信息发送至目的宿主机，以供所述目的宿主机基于所述源地址信息而发起脏页读取请求；

从所述源宿主机中读取本轮待传输脏页；

将读取到的本轮待传输脏页提供给所述目的宿主机中，以响应所述脏页读取请求。

13、一种虚拟机的热迁移方法，其特征在于，适用于为目的宿主机配置的第二加速设备，所述方法包括：

35 在热迁移过程中，接收为源宿主机配置的第一加速设备提供的本轮待传输脏页的

源地址信息：

将所述源地址信息提供给所述目的宿主机；

按照所述目的宿主机基于所述源地址信息而发起的脏页读取请求，利用所述第一加速设备从所述源宿主机中读取本轮待传输脏页；

5 将获取到的本轮待传输脏页写入所述目的宿主机中。

14、一种加速设备，其特征在于，与源宿主机连接，包括脏页迭代组件和通信组件：

所述脏页迭代组件，用于在热迁移过程中，通过所述通信组件从所述源宿主机中读取脏页信息；基于所述脏页信息查找本轮待传输脏页的源地址信息；将所述源地址信息发送至目的宿主机，以供所述目的宿主机基于所述源地址信息而发起脏页读取请求；

所述通信组件，用于从所述源宿主机中读取本轮待传输脏页；将读取到的本轮待传输脏页发送给所述目的宿主机中，以响应所述脏页读取请求。

15 15、一种加速设备，其特征在于，与目的源宿主机连接，包括脏页迭代组件和通信组件：

所述脏页迭代组件，用于在热迁移过程中，接收为源宿主机配置的第一加速设备提供的本轮待传输脏页的源地址信息；将所述源地址信息提供给所述目的宿主机；

20 所述通信组件，用于按照所述目的宿主机基于所述源地址信息而发起的脏页读取请求，利用所述第一加速设备从所述源宿主机中读取本轮待传输脏页；将获取到的本轮待传输脏页写入所述目的宿主机中。

25 16、一种存储计算机指令的计算机可读存储介质，其特征在于，当所述计算机指令被一个或多个处理器执行时，致使所述一个或多个处理器执行权利要求 12 或 13 任一项所述的虚拟机的热迁移方法。