

(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
2024 年 3 月 28 日 (28.03.2024)



(10) 国际公布号
WO 2024/060710 A1

(51) 国际专利分类号:
G06F 9/455 (2018.01)

(21) 国际申请号: PCT/CN2023/100492

(22) 国际申请日: 2023 年 6 月 15 日 (15.06.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202211142315.0 2022年9月20日 (20.09.2022) 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: 钟刊 (ZHONG, Kan); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。崔文林 (CUI, Wenlin); 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: 北京同达信恒知识产权代理有限公司 (TDIP & PARTNERS); 中国北京市西城区裕民路18号北环中心A座2002, Beijing 100029 (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,

(54) Title: PAGE SWAP-IN METHOD AND APPARATUS

(54) 发明名称: 一种页面换入方法以及装置

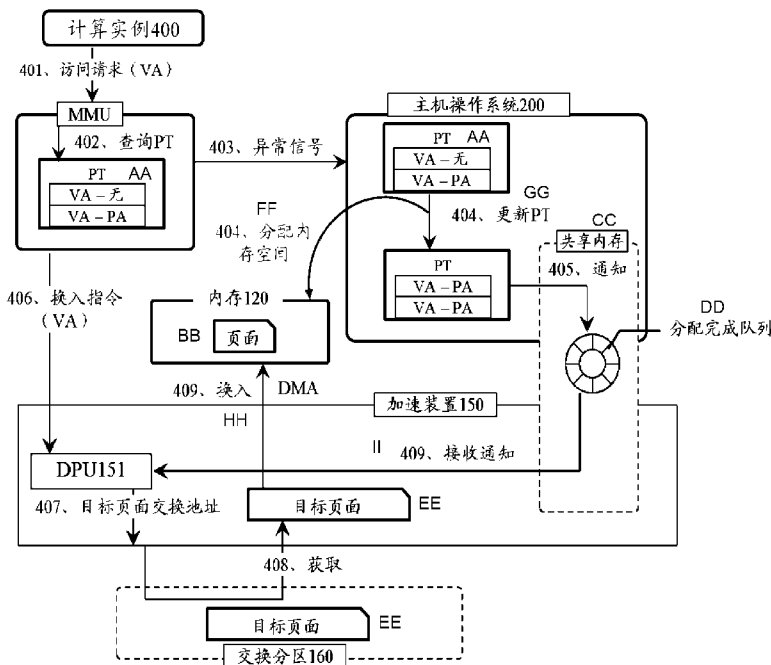


图 4

(57) Abstract: A page swap-in method and apparatus. In the present application, a DPU in a host replaces a processor in the host to swap a page from a storage device into a memory of the host. The DPU in the host receives a swap-in instruction, the swap-in instruction being used for instructing to swap a target page into the memory of the host. After receiving the swap-in instruction, the DPU acquires the target page from the storage device, swaps the target page into the memory, and notifies the host (such as a processor or MMU in the host) that the target page has been swapped into the memory. A page swap-in operation can be executed by the DPU, and the

MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, 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, SC, 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)。

根据细则4.17的声明:

— 关于申请人有权要求在先申请的优先权(细则4.17(iii))

本国际公布:

— 包括国际检索报告(条约第21条(3))。

processor in the host no longer needs to swap the target page from the storage device into the memory of the host, so that the computing power of the processor is effectively released, and the occupation of the page swap-in operation on the processor is reduced, thereby ensuring the execution efficiency of service work in the host.

(57) 摘要: 一种页面换入方法以及装置, 本申请中, 主机中的DPU代替主机中的处理器将页面从存储设备换入至主机的内存中。主机中的DPU接收换入指令, 换入指令用于指示将目标页面换入至主机的内存。DPU在接收到该换入指令后, 从存储设备获取目标页面, 将目标页面换入至内存, 通知主机(如主机中的处理器或MMU)目标页面已换入到内存中。页面的换入操作可以由DPU执行, 主机中的处理器不再需要将目标页面从存储设备换入到主机的内存中, 有效释放处理器的算力, 减少对页面换入操作对处理器的占用, 进而保证了主机中业务类工作的执行效率。

一种页面换入方法以及装置

相关申请的交叉引用

本申请要求在2022年09月20日提交中国专利局、申请号为202211142315.0、申请名称为“一种页面换入方法以及装置”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

本申请涉及通信技术领域，尤其涉及一种页面换入方法以及装置。

背景技术

为了对该主机的内存进行扩展，主机中的处理器可以将主机的内存中不活跃的页面换出到交换分区中，以此释放主机的内存的空间，从而实现对主机的内存扩展，该交换分区被认为是一种虚拟内存，该交换分区通常会位于该主机的硬盘中。

当主机中的处理器需要调用这些换出至交换分区中的页面时，处理器会从交换分区中获取页面，将该页面换入到主机的内存。

但页面从交换分区换入到主机的内存会占用处理器资源，使得处理器无法提供较多的算力支持主机中业务类的工作。

发明内容

本申请提供一种页面换入方法以及装置，用以减少页面换入操作对处理器的占用。

第一方面，本申请提供了一种页面换入方法，在该方法中与主机通信系统总线连接的加速装置可以代替主机中的处理器将页面从存储设备换入至主机的内存中。该方法中，主机中的加速装置接收换入指令，换入指令用于指示将目标页面换入至主机的内存。加速装置在接收到该换入指令后，从存储设备获取目标页面，将目标页面换入至内存，通知主机（如主机中的处理器或 MMU）目标页面已换入到内存中。

通过上述方法，页面的换入操作可以由加速装置执行，主机中的处理器不再需要将目标页面从存储设备换入到主机的内存中，有效释放处理器的算力，减少对页面换入对处理器的占用，进而保证了主机中业务类工作的执行效率。

在一种可能的实施方式中，加速装置可以从主机中的处理器或主机中 MMU 接收该换入指令。例如当 MMU 发现 PT 中缺页时，可以向加速装置发送换入指令，加速装置将目标页面换入到主机的内存后，加速装置通知 MMU 目标页面已换入到内存中。又例如，当处理器确定内存中有足够的内存空间时，可以向加速装置发送换入指令，加速装置将目标页面换入到主机的内存后，加速装置通知处理器目标页面已换入到内存中。

通过上述方法，主机中的处理器以及 MMU 仅需发送换入指令，即可通过加速装置将目标页面换入到主机的内存中，无需处理器的参与，页面换入的方式更加简单、快捷。

在一种可能的实施方式中，加速装置在从存储设备获取目标页面时，先获取目标页面的虚拟地址与目标页面的交换地址的对应关系，目标页面的交换地址为目标页面在存储设备的地址。加速装置在获取该对应关系后，根据对应关系确定目标页面的交换地址。再根据目标页面的交换地址从存储设备获取目标页面。加速装置可以与存储设备进行通信，向存储设备

发送携带有目标页面的交换地址的数据请求，以请求获取该目标页面，存储设备在接收到该数据请求后，可以向该加速装置反馈该目标页面。

通过上述方法，加速装置获取对应关系后，能够利用该对应关系获取目标页面的交换地址，以便与存储设备进行通信请求获取该目标页面。

5 在一种可能的实施方式中，存储设备能够提供存储空间作为该主机的虚拟内存，该存储设备可以是部署在主机之外，与主机通过网络连接的存储设备。存储设备也可以是主机的本地存储设备。存储设备还可以是部署在云端，如该存储设备可以是私有云或公有云中为该主机预先分配的存储空间。

通过上述方法，存储设备的具体形态有多种，有效地扩展了页面换入方法所适用的场景。

10 在一种可能的实施方式中，对应关系可以是加速装置将目标页面从内存换出至存储设备时、记录的。对应关系也可以是加速装置从主机的处理器获取的，该对应关系可以是处理器将目标页面从内存换出至存储设备时、记录的。

通过上述方法，加速装置获取该对应关系的方式有很多种，适用于不同的场景。

15 在一种可能的实施方式中，主机还包括处理器，加速装置通过系统总线与处理器连接，加速装置将目标页面换入至内存时，可以接收处理器发送的通知，通知用于指示处理器在内存中为目标页面配置的物理地址；加速装置在获取该物理地址后，可以通过 DMA 将目标页面写入到物理地址。

通过上述方法，加速装置获取物理地址后可以利用 DMA 将目标页面写入到内存中，绕过了处理器，进一步减少了对处理器的占用。

20 在一种可能的实施方式中，加速装置接收换入指令之前，加速装置还可以执行目标页面的换出操作。例如，加速装置接收主机的处理器发送的换出指令，换出指令用于指示将主机内存中的目标页面换出至存储设备，换出指令包括目标页面的虚拟地址以及目标页面；加速装置在接收到该换出指令后，将目标页面换出至存储设备，记录对应关系。

25 通过上述方法，加速装置承担了页面换入以及页面换出的主要操作，能够减少对处理器的占用，保证主机的处理器能够提供足够的算力以支持业务类的工作。

在一种可能的实施方式中，加速装置还具备数据解压缩功能，加速装置从存储设备获取压缩后的目标页面时，加速装置对压缩后的目标页面解压，获得目标页面，之后，再将目标页面换入至主机的内存中。

30 通过上述方法，加速装置可自行对压缩后的目标页面进行解压，无需处理器参与，进一步释放了处理器的算力。

在一种可能的实施方式中，加速装置将目标页面换出至存储设备时，为了能够降低该目标页面所占用的存储空间，加速装置将目标页面压缩后，换出至存储设备。

35 第二方面，本申请还提供了一种页面换入装置，该页面换入装置具有实现上述第一方面的方法实例中行为的功能，有益效果可以参见第一方面的描述此处不再赘述。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。硬件或软件包括一个或多个与上述功能相对应的单元。在一个可能的设计中，装置的结构中包括接收模块（用于接收换入指令或换出指令）、以及换出模块（用于将目标页面换出至存储设备），可选的，还包括换入模块（用于将目标页面从存储设备换入至主机的内存），这些模块可以执行上述第一方面方法示例中 DPU 的相应功能，具体参见方法示例中的详细描述，此处不做赘述。

40 第三方面，本申请还提供了一种加速装置，该加速装置包括 DPU、GPU、NPU、或 TPU 等具有处理功能的处理器。前述列举的处理器类型仅是举例，本申请并不限定加速装置中

所包括的处理器的具体类型，下面仅是以加速装置所包括的处理器为 DPU 为例进行说明。可选的，还包括供电电路以及存储器。供电电路用于对 DPU 进行供电。

在一种可能的实施方式中，该 DPU 具有实现上述第一方面以及第一方面的各个可能的实现方式中的方法实例中 DPU 的行为的功能，有益效果可以参见第一方面的描述此处不再赘述。

在另一种可能的实施方式中，存储器中存储计算机程序指令，DPU 与存储器耦合，DPU 可调用该存储器中存储的计算机执行指令，执行上述第一方面以及第一方面的各个可能的实现方式中 DPU 所执行的方法。

第四方面，本申请实施例还提供了一种计算设备，该计算设备可以为第一方面中提及的主机，有益效果可以参见第一方面的描述此处不再赘述，该计算设备包括加速装置和内存。

加速装置接收换入指令，换入指令用于指示将目标页面换入至计算设备的内存，加速装置在接收到该换入指令后，加速装置从存储设备获取目标页面；将目标页面换入至内存，通知计算设备目标页面已换入到内存中。

在一种可能的实施方式中，加速装置从计算设备的处理器或 MMU 接收换入指令。

在一种可能的实施方式中，加速装置从存储设备获取目标页面时，先获取目标页面的虚拟地址与目标页面的交换地址的对应关系，目标页面的交换地址为目标页面在存储设备的地址。加速装置根据对应关系确定目标页面的交换地址。加速装置与存储设备进行通信，根据目标页面的交换地址从存储设备获取目标页面。

在一种可能的实施方式中，存储设备为部署在计算设备之外，与计算设备通过网络连接的存储设备，或为计算设备的本地存储设备，或为公有云或私有云中为计算设备分配的存储空间。

在一种可能的实施方式中，对应关系是加速装置将目标页面从内存换出至存储设备时、记录的。对应关系是也可以是加速装置从计算设备的处理器获取的，该对应关系可以是处理器将目标页面从内存换出至存储设备时、记录的。

在一种可能的实施方式中，计算设备还包括处理器，加速装置通过系统总线与处理器连接。

加速装置将目标页面换入至内存时，处理器可以向加速装置发送通知，该通知用于指示处理器在内存中为目标页面配置的物理地址。加速装置在接收到该通知后，通过 DMA 将目标页面写入到物理地址。

在一种可能的实施方式中，计算设备的处理器向加速装置发送的换出指令，换出指令用于指示将计算设备内存中的目标页面换出至存储设备，换出指令包括目标页面的虚拟地址以及目标页面。

加速装置接收换出指令后，将目标页面换出至存储设备，记录对应关系。

在一种可能的实施方式中，加速装置从存储设备获取压缩后的目标页面，对压缩后的目标页面解压，获得目标页面，将目标页面换入至内存中。

在一种可能的实施方式中，加速装置将目标页面换出至存储设备时，将目标页面压缩后，换出至存储设备。

第五方面，本申请提了一种数据处理系统，该数据处理系统包括主机和存储设备，存储设备通过内部总线或网络与主机通信，主机设置有加速装置。

主机向加速装置发送换出指令，换出指令中包含数据。加速装置在从换入指令获取该数据后，对数据进行压缩，将压缩后的数据保存在加速装置的内存中，或者保存在主机的内存中。加速装置通过对数据进行压缩，使得数据在加速装置在内存中或主机内存中占用的空间

减少，提升加速装置在内存中或主机内存的利用率。

在一种可能的实施方式中，压缩后的数据保存在加速装置的内存中，加速装置可以在加速装置的内存中空闲的内存空间小于第一空闲阈值、或压缩后的数据的访问频率低于阈值的情况下，将压缩后的数据迁移至存储设备，这样可以减少对加速装置中内存的占用。

在一种可能的实施方式中，压缩后的数据保存在主机的内存中，加速装置可以在主机的内存中空闲的内存空间小于第二空闲阈值、或压缩后的数据的访问频率低于阈值的情况下，将压缩后的数据迁移至加速装置的内存中，这样可以减少对主机中内存的占用。

在一种可能的实施方式中，主机向加速装置发送换入指令，换入指令指示获取数据。加速装置在接收到该换入指令后，获取压缩后的数据，对压缩后的数据进行解压，在解压后，将解压后的数据保存在主机的内存中。

在一种可能的实施方式中，若压缩后的数据已迁移至存储设备，加速装置从存储设备获取压缩后的数据，若压缩后的数据保存在加速装置中的内存或主机的内存中，加速装置可以加速装置中的内存或主机的内存获取该压缩后的数据。

第六方面，本申请还提供一种计算机可读存储介质，计算机可读存储介质中存储有指令，当其在计算机上运行时，使得计算机执行上述第一方面以及第一方面的各个可能的实现方式中的方法。

第七方面，本申请还提供一种包含指令的计算机程序产品，当其在计算机上运行时，使得计算机执行上述第一方面以及第一方面的各个可能的实现方式中的方法。

第八方面，本申请还提供一种计算机芯片，芯片与存储器相连，芯片用于读取并执行存储器中存储的软件程序，执行上述第一方面以及第一方面的各个可能的实现方式中的方法。

附图说明

图 1 为本申请提供的一种主机的虚拟化架构示意图；

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页面换出方法示意图；

图 4 为本申请提供的一种页面换入方法示意图；

图 5 为本申请提供的一种页面换入装置的结构示意图。

具体实施方式

在本申请实施例提供的一种页面换入方法以及装置介绍之前，对本申请涉及的一些概念进行说明：

(1)、虚拟化技术、计算实例、容器 (container)、虚拟机 (virtual machine, VM)。

虚拟化(virtualization)是一种资源管理技术，通过虚拟化技术将主机的各种物理资源，如处理器、内存及接口等，予以抽象、转换后呈现出来。虚拟化是一种从逻辑角度出发的资源配置方法，是物理资源的逻辑抽象。

主机借助虚拟化技术能够在主机上形成具备独立的运行环境的软件模块，在本申请实施例中将在主机上形成的具备独立的运行环境的软件模块称为计算实例。该计算实例可以为虚拟机，也可以为容器。

虚拟机是通过虚拟机化技术模拟的具有完整硬件系统功能的、运行在一个完全隔离环境中的“完整计算机”。实体计算机中能够完成的工作在虚拟机中都能够实现。虚拟机具备处理器 (该处理器也被称为虚拟处理器)、内存、硬盘等组件，虚拟机的处理器、内存、硬盘等组

件均是由主机的处理器、内存、硬盘等组件虚拟而成的。虚拟机上安装有操作系统，虚拟机上的操作系统与主机自身的操作系统相互独立。为了区分这两种不同的操作系统，主机的上的操作系统通常被称为主机操作系统（host operating system, hostOS），虚拟机上的操作系统通常被称为客户操作系统（guest operating system, guestOS）。

容器是通过虚拟化技术模拟的一种独立的运行环境，容器类似于一个轻量级的沙盒，实现对容器外部的软件以及硬件进行屏蔽，容器是在操作系统层面上实现虚拟化，直接复用主机的操作系统。

对主机来说，计算实例会被看做是一种特殊的“进程”。该“进程”会执行计算任务、还会占用主机的处理器、内存、硬盘等资源。

举例来说，主机操作系统会为计算实例配置专属于该计算实例的内存空间，计算实例会占用该内存空间，以用于支持计算实例所需执行的计算任务。

（2）、内存管理单元（memory management unit, MMU）。

MMU 也被称为分页内存管理单元（paged memory management unit, PMMU）。主机中，MMU 是一个位于处理器的内核和连接高速缓存以及内存的总线之间的硬件装置，MMU 通常也被认为是处理器的一部分，这种情况下 MMU 所执行的操作可以认为是处理器所执行的操作。在本申请实施例中，仅是以 MMU 是独立的硬件装置为例进行说明。MMU 主要用于处理处理器发起的针对内存的访问请求，MMU 具备地址翻译功能，能够实现内存的虚拟地址与内存的物理地址之间的转换。例如，MMU 能够解析处理器发出的访问请求，将该访问请求中携带的内存的虚拟地址转换为内存的物理地址，以将该访问请求中携带的数据写入到该内存的物理地址上（访问请求用于请求在该内存的虚拟地址上写入数据）；或者从该内存的物理地址上读取数据，将数据反馈给处理器（访问请求用于请求读取该内存的虚拟地址上的数据）。除此之外，MMU 还具备内存保护功能、以及针对处理器的缓存的控制功能。

MMU 可以以页为粒度对内存的空间进行管理（也即页式内存管理）。在以页为粒度对内存的空间的管理中，将内存空间划分为大小固定的块，一块即为一个页面。MMU 可以以页面为单元分配、管理、以及保护内存。

为了实现虚拟地址到物理地址的转换，MMU 需要获取地址翻译表，该地址翻译表称为页表（page table, PT），页表中包括一个或多个页表项（page table entry, PTE）。每个页表项对应一个页面，记录了该页面的虚拟地址与该页面的物理地址的对应关系以及一些控制信息。该地址翻译表是由主机操作系统维护并保存在主机的内存中的，也就是说，主机操作系统能够对该 PT 执行如增加页表项、删除页表项、修改页表项等的一些更新操作，并将更新后的页表保存在主机的内存中。MMU 在需要进行地址翻译时，从主机的内存中调取该 PT。

主机上运行的不同进程所使用的页面的虚拟地址可能相同，为了避免不同进程所使用的虚拟地址之间混淆。主机中，主机操作系统为主机上的每个进程（这里的进程也包括计算实例）维护对应的 PT。

以计算实例为例，该计算实例对应的 PT 中记录了该计算实例所占用主机的内存空间的虚拟地址与物理地址之间的映射关系。如表 1 所示，为本申请实施例提供的一种 PT 的抽象表格，表 1 中将 PT 抽象为一个表格，该表格的每一行指示了一个 PTE。每一行记录了一个页面的虚拟地址（virtual address, VA）、和该页面的物理地址（physical address, PA），该 PTE 中还包括一个标志位。该标志位用于标识该 PTE 是否被访问。在本申请实施例中以标识位为 1，表征该 PTE 被访问，PTE 被访问说明该 PTE 中的物理地址所指示的内存空间被访问；标

识位为 0，表征为该 PTE 未被访问，PTE 被访问说明该 PTE 中的物理地址所指示的内存空间未被访问。MMU 每读取一次 PTE，会将该 PTE 中的标志位置 1，若标志位置 1 之后的特定时间内 PTE 未再次被读取，该 PTE 的标志位会自行翻转变为 0。

表 1

	虚拟地址	物理地址	标志位
PTE	V1	P1	1
PTE	V2	P2	0
			

当计算实例需要将数据写入到虚拟地址 V1，该计算实例会向 MMU 发送访问请求 1（从硬件交互的角度来看，该访问请求 1 是由处理器发送给 MMU 的），该访问请求 1 用于请求在虚拟地址 V1 写入数据，该访问请求 1 还携带了页面的虚拟地址 V1、以及需要写入的数据。MMU 在接收到该访问请求 1 后，找到与该计算实例对应的 PT，在该 PT 中查找是否存在该虚拟地址 V1 对应的物理地址。若在 PT 中未查找到记录了虚拟地址 V1 的 PTE，MMU 会触发缺页中断（page fault）流程，向主机操作系统发出用于指示缺页中断的异常信号，主机操作系统在内存中为计算实例分配一段空闲的内存空间（也即空闲页面），该内存空间的物理地址 P1 作为该虚拟地址 V1 对应的物理地址。主机操作系统更新 PT，也就是说在该 PT 中增加 PTE，该 PTE 记录该物理地址 P1 与该虚拟地址 V1 的对应关系。主机操作系统将更新后的 PT 保存在主机的内存中，通知 MMU 继续处理。MMU 获取更新后的 PT，查找记录了虚拟地址 V1 的 PTE，确定该虚拟地址 V1 对应的物理地址 P1，将该 PTE 中的标志位置 1。MMU 将该数据写入到该物理地址上 P1。

当计算实例需要从虚拟地址 V1 读取数据，该计算实例会向 MMU 发送访问请求 2（从硬件交互的角度看，该访问请求 2 是由处理器发送给 MMU 的），该访问请求 2 用于请求读取虚拟地址 V1 的数据，该访问请求 2 还携带了页面的虚拟地址 V1。MMU 在接收到该访问请求 2 后，找到与该计算实例对应的 PT，在该 PT 中查找是否存在与该虚拟地址 V1 对应的物理地址。若在 PT 中查找到了记录了虚拟地址 V1 的 PTE，MMU 从与该虚拟地址对应的物理地址上读取数据，将数据反馈给该计算实例，并将标志位置 1。若在 PT 中未查找到记录了虚拟地址 V1 的 PTE，说明该数据未存储在主机的内存中。这种情况存在两种可能的情况：情况一、在该主机中存在交换机制，部署有交换模块的场景下，该 MMU 可以触发缺页中断流程，通过交换模块获取该虚拟地址 V1 上的数据，MMU 通过交换模块获取该虚拟地址 V1 上的数据的过程可以参见下述关于交换机制的相关描述，此处不再赘述。情况二、在该主机中不存在交换机制，MMU 会触发缺页中断流程，向主机操作系统发出用于指示缺页中断的异常信号，交由主机操作系统继续处理。

（3）、交换（swap）机制、交换分区（swap area）、交换模块。

交换机制指的是内存交换，交换机制中提出了物理内存以及虚拟内存这两个概念，物理内存为该主机的实际内存，在本申请实施例中可以称为主机的内存。该虚拟内存即为交换分区，通常位于主机的本地存储设备中，主机的本地存储设备指与主机通过系统总线连接的存储设备，如主机的硬盘等。在本申请实施例该虚拟内存也可以位于远端存储设备中，远端存储设备是指与主机通过网络连接的存储设备，如远端存储系统中的存储节点等，该虚拟

内存也可以部署在私有云或公有云中，也即也可以位于云数据中心中，例如，该虚拟内存是云数据中心中为该主机预先分配的存储空间，该存储空间为分布在一个或多个设备上。也即该虚拟内存可以部署在云端，虚拟内存所在的设备位于云端。交换机制中，允许从主机的存储设备（这里的存储设备包括主机的本地存储设备、远端存储设备以及云中的设备）中开辟出一部分存储空间，将该部分存储空间虚拟为主机的内存空间，作为虚拟内存。当主机的内存空间不足时，在物理内存和虚拟内存之间进行数据交换。

为了方便说明交换机制的实施方式，这里将实施该交换机制的执行方称为交换模块，该交换模块实质上是主机上运行一种软件模块，是主机操作系统中的一部分。

交换机制的实施方式如下：

交换模块在主机的物理内存中空闲空间较少（如物理内存中空闲的内存空间大小低于预设的阈值），或者交换模块预期后续可能存在物理内存中空闲内存空间不足的情况，交换模块可以将物理内存中一些页面换出到虚拟内存中，将这些页面中的数据保存到虚拟内存中，以释放物理内存的一些内存空间，这个过程即为页面的换出（swap out），这些换出的页面可以称为换出页面。在这种过程中，由于换出页面不再保存到物理内存中，PT 中关于换出页面的 PTE（也即记录了换出页面的虚拟地址与换出页面的物理地址之前的对应关系的 PTE）也会被删除。而交换模块需要记录该换出页面的虚拟地址与该换出页面在虚拟内存中的地址的对应关系。

交换模块在主机需要调用这些页面（如 MMU 接收到主机上的进程发起的访问请求 2，该访问请求携带了虚拟地址为换出页面的虚拟地址），或者主机的物理内存存在足够多的空闲空间（如物理内存中空闲的内存空间大小大于预设的阈值）的情况下，交换模块可以将换出页面从虚拟内存换入至物理内存中，这个过程即为页面的换入（swap in），这些换入的页面可以称为换入页面。在这种过程中，由于换入页面保存到了物理内存中，主机操作系统需要在物理内存中为换入页面分配新的物理地址（该物理地址称为换入页面的物理地址），MMU 使用的 PT 中会增加关于换入页面的 PTE（也即记录了换入页面的虚拟地址与换入页面的物理地址之前的对应关系的 PTE）。而交换模块也会删除之前记录该换入页面的虚拟地址与该换入页面在虚拟内存中的地址的对应关系。

下面结合 MMU 以及交换模块的工作流程，对常见的针对计算实例发起的用于读取页面 A 的访问请求的处理流程进行进一步说明，由于交换模块通常被认为是主机操作系统的一部分，这该处理流程的说明中，不再区分交换模块以及主机操作系统，统一称为主机操作系统。

计算实例发起访问请求，该访问请求用于请求读取页面 A 中的数据，该访问请求携带了页面 A 的虚拟地址 V1。MMU 获取该访问请求，从内存中获取该计算实例对应的 PT。MMU 从 PT 查找记录了该虚拟地址 V1 的 PTE。

若 MMU 在该 PT 找到了记录了该虚拟地址 V1 的 PTE，MMU 确定该虚拟地址 V1 对应的物理地址 P1，将该物理地址 P1 发送给内存，将内存返回的、该物理地址 P1 上的数据反馈给计算实例。

若 MMU 在该 PT 未找到了记录了该虚拟地址 V1 的 PTE，MMU 会触发 page fault 流程，向主机操作系统发出缺页中断的异常信号，主机操作系统会暂停该计算实例的运行。主机操作系统查找在将页面 A 换出时、记录的该虚拟地址 V1 与页面 A 中的数据在虚拟内存中的地址 X1 的对应关系，根据该地址 X1 将该页面 A 中的数据从虚拟内存换入到主机的内存中、并为该页面 A 重新分配一段空闲的内存空间。主机操作系统为该页面重新分配的空闲的内存空间的物理地址 P2，该物理地址 P2 作为该虚拟地址 V1 对应的物理地址，更新 PT。也就是

说, 主机操作系统在该 PT 中增加 PTE, 该 PTE 记录该物理地址 P2 与该虚拟地址 V1 的对应关系。主机操作系统 PT 将更新后的 PT 保存在主机的内存中, 通知 MMU 继续处理, 并启动该计算实例的运行。MMU 获取更新后的 PT, 查找记录了虚拟地址 V1 的 PTE, 确定该虚拟地址 V1 对应的物理地址 P2, 将该 PTE 中的标志位置 1。MMU 将该物理地址 P2 发送至主机的内存, 将从主机的内存反馈的数据反馈至计算实例。

在主机应用虚拟化技术、部署计算实例的场景下, 计算实例能够占用主机的内存空间(也即占用内存中的一些页面), 但计算实例(尤其是虚拟机)对所占用的内存空间的使用情况, 主机操作系统是无法感知的。

在页面换出时, 主机操作系统会决定内存中的哪一个或哪几个页面会被换出。常见的方式是, 主机操作系统会利用最近最少使用(least recently used, LRU)算法, 确定内存中页面最近最少使用的一个或多个页面, 将所确定的该一个或多个页面换出到虚拟内存中。

但这种方式中, 由于主机操作系统无法获取计算实例(如虚拟机)对该计算实例所占用的页面的使用情况, 如主机操作系统无法获取计算实例所占用的各个页面是否为计算实例的内核占用的页面、或计算实例中应用所占用的页面, 主机操作系统也无法获知计算实例对所占用的页面的访问频率。鉴于此, 导致主机操作系统会将计算实例一些较为重要的页面作为换出页面, 换出到虚拟内存中, 当计算实例需要调用这些页面时, MMU 会触发缺页中断流程, 缺页中断流程伴随着页面的换入操作, 增加时延, 导致计算实例的卡顿, 影响计算实例的工作效率。

为此本申请实施例提出了一种页面换出方法, 该方法中, 计算实例可以向主机操作系统发送页面标签, 该页面标签指示了主机的内存中计算实例所占用的各个页面的重要程度。主机操作系统在需要将该内存中的页面换出至内存之外、部署有交换分区的存储设备时, 主机操作系统根据该页面标签确定主机的内存中需要换出的目标页面, 将所确定的目标页面换出至该存储设备。在本申请实施例中, 由于计算实例能够将所占用的页面的重要程度告知主机操作系统, 使得主机操作系统能够避免将计算实例的占用的重要程度较高的页面换出到存储设备。计算实例在需要调用重要程度较高的页面时, MMU 不会触发缺页中断流程, 计算实例也不会被暂停运行。

另外, 为了减少主机操作系统的压力, 主机操作系统在将页面换出至该存储设备时, 主机操作系统向主机的加速装置发送换出指令, 该换出指令用于指示加速装置将页面换出至该存储设备, 该换出指令携带了页面的虚拟地址以及该页面。加速装置获取该页面, 将该页面存储至该存储设备, 并记录该页面的虚拟地址与该页面在该存储设备中的地址的对应关系。也即是说, 加速装置可以代替主机操作系统(也即处理器)将页面从主机的内存中交换至该存储设备中, 能够减少对处理器的占用, 释放处理器的算力。

在将该内存中的页面换出至该存储设备后, 可能会存在需要将该存储设备中的页面换入至主机的内存的情况。本申请实施例中提出了一种页面换入方法, 在本申请实施例中, 主机的加速装置可以代替主机中的处理器(也可以理解为主机操作系统)将页面从该存储设备换入至主机的内存。主机中的 MMU 向该加速装置发送换入指令, 该换入指令用于指示将页面换入至内存中, 该换入指令中可以携带有页面的虚拟地址, 加速装置在接收到该换入指令后, 可以根据该换入指令将页面从存储设备换入至主机的内存中, 并通知 MMU 页面已换入至主机的内存中。例如, 加速装置可以根据已记录的对应关系确定该页面在该存储设备中的地址, 根据该页面在该存储设备中的地址从该存储设备获取该页面, 将该页面写入到主机的内存中。

这样可以进一步减少对处理器的占用。

在本申请实施例中，页面换出至内存之外的存储设备是指将页面换出至交换分区，其中，该存储设备是部署有交换分区的存储设备。也就是说，将页面换出至内存之外的存储设备以及将页面换出至交换分区的两种表述方式所表达的实质是相同的，在下文中，统一为将页面换出至交换分区。但应需理解，将页面换出至交换分区是指将页面交换到存储设备中，将页面存储在该存储设备中。类似的，将页面从存储设备换入至主机的内存是指将交换分区中的页面换入至主机的内存中，页面从存储设备换入至主机的内存以及页面从交换分区中换入至主机的内存的两种表述方式所表达的实质是相同的，在下文中，统一为页面从交换分区换入至主机的内存。也就是说，为了表述上更加明确，在下文中涉及到页面的换入以及换出的过程中用交换分区指代该交换分区所在的存储设备。

如图 1 所示，为本申请实施例提供的一种主机的虚拟化架构示意图。将虚拟化技术应用于主机 10，形成了如图 1 所示的虚拟化架构，主机 10 的虚拟化架构包括底层硬件 100、主机操作系统 200、计算实例管理单元 300、以及至少一个计算实例 400。主机 10 为一种计算设备。该主机 10 可以为服务器、移动终端、平板电脑等计算设备。

底层硬件 100 是指该主机 10 中的一些硬件组件，如处理器 110、内存 120、输入输出(input output, I/O) 接口 130 等。底层硬件 100 可以理解为主机 10 上的物理资源，是虚拟化技术中需要虚拟的对象。

以底层硬件 100 为基础，该主机 10 运行了主机操作系统 200、计算实例管理单元 300、计算实例 400 等软件模块。

主机操作系统 200 运行在处理器 110 上，用于实现主机 10 的基础功能；基础功能包括但不限于：针对底层硬件 100 的管理功能、针对与主机 10 连接的输入/输出设备（如显示器、键盘、鼠标等）的控制功能、对主机 10 中进程的管理。本申请实施例中主要涉及针对底层硬件 100 的管理功能，以主机操作系统 200 对主机 10 的内存 120 的管理为例，主机操作系统 200 能够监控内存 120 的占用情况，分配内存空间；主机操作系统 200 还能够检测内存 120 中页面的访问频率；主机操作系统 200 能够实现主机 10 的内存 120 与交换分区 160 之间的页面交换。

在本申请实施例，主机操作系统 200 能够从计算实例 400 获取计算实例 400 的页面标签，该页面标签指示了主机 10 的内存 120 中计算实例 400 所占用的页面的重要程度。主机操作系统 200 还能够获取计算实例 400 所占用的各个页面的页面访问频率。主机操作系统 200 在需要将主机 10 的内存 120 的页面换出到交换分区 160 时，主机操作系统 200 可以根据页面标签、页面访问频率以及用户配置的页面交换参数中的部分或全部从主机 10 的内页中确定目标页面。主机操作系统 200 将所确定的目标页面换出至交换分区 160。主机操作系统 200 也可以将被换出至交换分区 160 的页面换入至主机 10 的内存 120 中。

从逻辑功能上，该主机操作系统 200 包括实例页面识别模块 210、交换模块 220。

实例页面识别模块 210 能够获取来自计算实例 400 的页面标签，还能够根据 MMU 对该计算实例 400 对应的 PT 的访问情况获得计算实例 400 占用的各个页面的页面访问频率。

实例页面识别模块 210 还能够面向用户提供参数配置接口，允许用户配置页面交换参数。页面交换参数指示了目标页面的筛选条件。该页面交换参数包括但不限于：允许换出页面的计算实例 400 的标识、页面扫描频率、冷页面的判断阈值、以及计算实例 400 允许占用的最大内存空间。

在需要将页面从主机 10 的内存 120 换出至交换分区 160 时，实例页面识别模块 210 根据

计算实例 400 的页面标签、页面访问频率以及页面交换参数中的部分或全部从计算实例 400 所占用的页面中筛选目标页面，该目标页面为该计算实例 400 所占用的页面中允许换出的页面。实例页面识别模块 210 向交换模块 220 发送目标页面的虚拟地址。

交换模块 220 用于实现主机 10 的内存 120 与交换分区 160 之间的页面交换。也即交换模块 220 能够将主机 10 的内存 120 中的页面换出至交换分区 160、以及将交换分区 160 中的页面换入至主机 10 的内存 120。

交换模块 220 在需要将主机 10 的内存 120 中的页面交换至交换分区 160 时，交换模块 220 将该目标页面换出到交换分区 160。交换模块 220 还可以更新该计算实例 400 对应的 PT，从中删除与该目标页面相关的 PTE。

交换模块 220 将该目标页面换出到交换分区 160 时，可以向将该目标页面发送到交换分区 160，也即交换模块 220 自身完成目标页面的换出。交换模块 220 也可以指示该主机 10 的加速装置 150 执行目标页面的换出。例如，交换模块 220 向该主机 10 的加速装置 150 发送换出指令，告知加速装置 150 将目标页面换出至交换分区 160，该换出指令携带了目标页面的虚拟地址。

交换模块 220 在需要交换分区 160 中目标页面换入至主机 10 的内存 120 时，交换模块 220 可以直接从交换分区 160 获取目标页面，将获取的目标页面换入至主机 10 的内存 120 中。交换模块 220 也可以指示该主机 10 的加速装置 150 执行页面的换入。例如，交换模块 220 向该主机 10 的加速装置 150 发送换入指令，告知加速装置 150 将目标页面换入至主机 10 的内存 120，该换入指令携带了目标页面的虚拟地址。

本申请实施例中对模块的划分是示意性的，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。在本申请的实施例中的各功能模块可以集成在一个处理模块中，也可以是各个模块单独物理存在，也可以两个或两个以上模块集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。

计算实例管理单元 300 用于管理计算实例 400，该计算实例管理单元 300 能够将底层硬件 100 进行虚拟化，以向计算实例 400 提供运行环境。计算实例管理单元 300 向计算实例 400 提供运行环境的方式有很多种，例如，计算实例管理单元 300 能够通过软件的方式为该计算实例 400 模拟运行环境。又例如，计算实例 400 可以将底层硬件 100（如内存 120）直通给计算实例 400。

当计算实例 400 为虚拟机时，该计算实例管理单元 300 包括 QEMU（quick emulator）以及虚拟机管理器（virtual machine monitor, VMM）。当计算实例 400 为容器时，该计算实例管理单元 300 可以为容器引擎（container engine）。

在本申请实施例中，计算实例管理单元 300 可以将计算实例 400 的页面标签传输给主机操作系统 200。

在一些可能的场景中，计算实例管理单元 300 可以内置在主机操作系统 200 中。这种场景下，主机操作系统 200 具备了计算实例管理单元 300 的功能。也就是说，主机操作系统 200 可以直接从计算实例 400 获取该计算实例 400 的页面标签。

计算实例 400 具备独立的运行环境，计算实例 400 占用了主机 10 的物理资源，能够基于所占用的物理资源执行各种任务，实现相关业务。计算实例 400 之间相互独立，彼此不会产生影响。在本申请实施例中该计算实例 400 可以为虚拟机，也可以为容器，还可以是其他借助主机 10 的物理资源的虚拟化形成的模块。

在本申请实施例中，计算实例 400 能够评价主机 10 的内存 120 中该计算实例 400 所占用

的页面的重要程度,生成该计算实例 400 的页面标签。计算实例 400 还可以将该计算实例 400 的页面信息发送给主机操作系统 200。

下面对主机 10 包括的硬件组件进行说明,如图 2 所示,为本申请实施例提供的一种主机 10 的结构示意图,主机 10 包括 I/O 接口 130、处理器 110、存储器 120、以及加速装置 150。I/O 接口 130、处理器 110、存储器 120、以及加速装置 150 之间可通过系统总线连接,该系统总线可以为快捷外围部件互连标准(peripheral component interconnect express, PCIe)总线,也可以为计算快速互联(compute express link, CXL)、通用串行总线(universal serial bus, USB)协议或其他协议的总线。

图 2 示例性的展示了其中一种连接方式,图 2 中,加速装置 150 可以直接插在主机 10 的主板上的卡槽中,通过 PCIe 总线 140 与处理器 110 交换数据。

I/O 接口 130 用于与位于主机 10 外部的设备通信。例如,通过 I/O 接口 130 接收主机 10 之外的设备发送的数据或通过 I/O 接口 130 向主机 10 之外的设备发送数据。

处理器 110 是主机 10 的运算核心和控制核心,它可以是中央处理器(central processing unit, CPU),也可以是其他特定的集成电路。处理器 110 还可以是其他通用处理器、数字信号处理器(digital signal processing, DSP)、专用集成电路(application specific integrated circuit, ASIC)、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。

存储器 120 通常用来存放主机操作系统 200 相关的计算机程序指令、计算实例管理单元 300 以及计算实例 400 相关的计算机程序指令、以及主机操作系统 200、计算实例管理单元 300、计算实例 400 在运行过程中所产生的数据。存储器 120 具备访问速度快的优点。存储器 120 通常采用动态随机存取存储器(dynamic random access memory, DRAM)。除了 DRAM 之外,存储器 120 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)、存储级存储器(storage class memory, SCM)等。另外,存储器 120 也可以是只读存储器(read only memory, ROM)。而对于只读存储器,举例来说,可以是可编程只读存储器(programmable read only memory, PROM)、可抹除可编程只读存储器(erasable programmable read only memory, EPROM)等。存储器 120 还可以为双列直插式存储器模块或双线存储器模块(dual in-line memory module, DIMM)、闪存介质(FLASH)、硬盘驱动器(hard disk drive, HDD)或固态驱动器(solid state disk, SSD)等。

处理器 110 通过双倍速率(double data rate, DDR)总线或者其他类型的总线和存储器 120 相连。存储器 120 理解为主机 10 的内存 120(internal memory),内存 120 又称为主存(main memory)。

处理器 110 可以通过调用该内存 120 中的计算机程序指令在主机 10 中形成主机操作系统 200、计算实例管理单元 300 以及计算实例 400 等软件模块。处理器 110 作为主机 10 的运算核心以及控制核心能够支持如图 3 所示的实施例中主机操作系统 200 以及计算实例 400 所执行的步骤。

加速装置 150 与主机 10 连接,该加速装置 150 可以作为该主机 10 的外接设备;加速装置 150 也可以部署在该主机 10 内部,如加速装置 150 位于在主机 10 的主板或背板上。图 2 为加速装置 150 部署在主机 10 内部的一种示意图。

该加速装置 150 可以作为主机 10 附带的具有数据处理功能的模块,承担该主机 10 的部分功能。也就是说,主机 10 的部分功能卸载到该加速装置 150 上,由该加速装置 150 代替主

机 10 (如主机 10 中的处理器 110) 处理数据, 执行部分的任务, 以减轻主机 10 中处理器 110 的压力, 释放该处理器 110 的算力。在本申请实施例中该加速装置 150 能够实现对交换分区 160 的数据访问, 这里的数据访问包括将页面从主机 10 的内存 120 换出至交换分区 160、以及将页面从交换分区 160 换入至主机 10 的内存 120。

例如, 加速装置 150 可以接收主机操作系统 200 发送的换出指令, 该换出指令中携带了页面的虚拟地址以及该页面 (该页面可以为图 3 所示实施例中的目标页面), 可选的, 该换出指令还包括进程的标识 (该进程可以为计算实例 400), 加速装置 150 在接收到该换出指令后, 将该页面发送至交换分区 160 所在的存储设备, 以将该页面存储在该交换分区 160, 并保存页面的虚拟地址 (可选的, 还包括进程的标识)、与页面在交换分区 160 的地址之间的对应关系。

又例如, 加速装置 150 可以接收 MMU 发送的换入指令, 该换入指令指示将页面 (该页面可以为图 4 所示实施例中的目标页面) 换入至主机 10 的内存 120 中, 加速装置 150 根据该换入指令, 将页面从交换分区 160 中换入至主机 10 的内存 120 中。

加速装置 150 接收的换入指令中携带了页面的虚拟地址, 可选的, 该换出指令还包括进程的标识 (该进程可以为计算实例 400), 加速装置 150 在接收到该换入出指令后, 根据该页面的虚拟地址确定与页面在交换分区 160 的地址, 根据页面在交换分区 160 的地址从交换分区 160 获取页面, 将该页面写入到内存 120 中。

此外, 该加速装置 150 还具备数据解压以及数据压缩功能。例如加速装置 150 可以对先对需要换出至交换分区 160 的页面 (也即页面中的数据) 进行压缩, 将压缩后的页面换入至交换分区 160。又例如, 加速装置 150 可以对从交换分区 160 中获取的压缩后的页面进行解压, 将页面换入至主机 10 的内存 120 中。

在本申请实施例中, 该交换分区 160 位于内存 120 之外的存储设备 20, 该存储设备 20 可以为位于该主机 10 之外的远端存储设备 30。该存储设备 20 也可以为该主机 10 的本地存储设备 170。该存储设备 20 也可以位于云端, 如位于公有云或私有云。该交换分区 160 为云端预先为该主机 10 分配的存储空间。

所谓“远端存储设备 30”是指与主机 10 通过网络连接、位于主机 10 之外的存储设备。远端存储设备 30 能够存储数据, 本申请实施例并不限远端存储设备 30 的类型, 该远端存储设备 30 的存储介质可以为易失性存储器 (volatile memory), 例如 RAM、DRAM、SCM、SRAM。该远端存储设备 30 的存储介质也可以为非易失性存储器 (non-volatile memory), 例如 ROM、快闪存储器、HDD、SSD、SCM 等。

所谓“本地存储设备 170”是指主机 10 中、用于持久化存储的存储设备, 本地存储设备 170 通过系统总线连接到主机 10 上。该本地存储设备 170 可以为 ROM、快闪存储器、HDD、SSD 等非易失性存储器。

所谓“位于云端的存储设备”, 是指该存储设备 20 部署在云数据中心中。该存储设备 20 可以理解为云端为该主机分配的存储空间。该存储设备 20 可以为该云数据中心的某一个设备, 也即该设备能够提供存储空间以作为交换分区 160。该存储设备 20 也可以为该云数据中心的多个设备, 也即该多个设备共同提供存储空间以作为交换分区 160。

在本申请实施例中并不限定加速装置 150 对交换分区 160 的访问方式。例如, 当该交换分区 160 位于远端存储设备 30, 该远端存储设备 30 的存储介质为 DRAM、SCM 等易失性存储器, 加速装置 150 可以通过远程直接内存 120 访问 (remote direct memory access, RDMA) 的方式访问远端存储设备 30。又例如, 当交换分区 160 位于远端存储设备 30 且该远端存储

设备 30 的存储介质为 HDD、SSD 等非易失性存储器、或者交换分区 160 位于本地存储设备 170，加速装置 150 可以基于互联网小型计算机系统接口 (internet small computer system interface, iSCSI)、光纤通道 (fibre channel, FC)、或以太网光纤通道 (fibre channel over ethernet, FCoE) 等网络协议访问交换分区 160。

加速装置 150 可以面向主机 10 中的处理器 110 或 MMU 提供键值对 (key value, KV) 接口，该 KV 接口允许加速装置 150 与处理器 110 之间，或加速装置 150 与 MMU 之间以 KV 结构的数据结构进行数据交互。其中，键值对中的键 (key, K) 可以为页面的虚拟地址、或者页面的虚拟地址加进程的标识的组合。键值对中的值 (value, V) 可以为页面本身。

举例来说，MMU 向加速装置 150 发送的换入指令，即可表现为 KV 的形式，加速装置 150 在该 KV 接口接收到包括页面的虚拟地址 (或者页面的虚拟地址加进程的标识的组合) 为键，页面为值的数据时，加速装置 150 由此可认为需要将该页面换出至交换分区 160。

主机 10 的处理器 110 (也可以理解为主机操作系统 200) 向加速装置 150 发送的换出指令同样可以表示为 KV 的形式，加速装置 150 在该 KV 接口接收到包括页面的虚拟地址 (或者页面的虚拟地址加进程的标识的组合) 为键，值为空白或无效数据时，加速装置 150 由此可认为需要将该页面从交换分区 160 换入至主机 10 的内存 120。

加速装置 150 包括处理器，该处理器可以为数据处理单元 (data process unit, DPU) 151 图像处理器 (graphics processing unit, GPU)、张量处理器 (tensor processing unit, TPU) 或神经网络处理器 (neural network processing unit, NPU) 等具有数据处理功能的处理器，在图 2 中仅是以加速装置 150 包括的处理器为 DPU151 为例进行说明。可选的，该加速装置 150 还包括存储器 152、供电电路等，DPU151 与存储器 152 通过系统总线连接，该系统总线可以为基于 PCIe 的线路，也可以为 CXL、USB 协议或其他协议的总线。

DPU151 为加速装置 150 的主要运算单元，是加速装置 150 的核心单元，DPU151 承担了加速装置 150 的主要功能。在本申请实施例中，DPU151 能够对交换分区 160 进行数据访问，还能够进行数据压缩以及解压。

下面结合附图 3 以及图 4 对本申请实施例提供的一种页面换出方法以及页面换入方法进行说明，如图 3 所示的本申请实施例提供的一种页面换出方法，如图 4 所示的本申请实施例提供的一种页面换入方法。

需要说明的是，图 3 所示的页面换出方法以及图 4 的页面换入方法是独立存在的。图 3 所示的页面换出方法以及图 4 的页面换入方法可以结合使用。也就是说，在需要将主机 10 的内存 120 中的页面换出至交换分区 160 (也即是说需要将主机 10 的内存 120 中的页面换出至存储设备 20) 时，可以采用图 3 所示的页面换出方法。若后续需要将交换分区 160 中的页面换入至主机 10 的内存 120 (也即是说需要将存储设备 20 中的页面换入至主机 10 的内存 120 中)，可以采用图 4 所示的页面换入方法。图 3 所示的页面换出方法以及图 4 的页面换入方法也可以独立使用。也就是说，当采用图 3 所示的页面换出方法进行页面换出时，也可以采用其他方式进行页面换入，其他页面换入的方式可以参见前述关于交换机制的描述。在采用图 4 所示的页面换入方法进行页面换入之前，也可以采用其他方式，进行页面换出，其他页面换出的方式可以参见前述关于交换机制的描述。

如图 3 所示，为本申请实施例提供的一种页面换出方法，该方法包括：

步骤 301: 计算实例 400 对主机 10 的内存 120 中该计算实例 400 所占用的页面进行分类，生成该计算实例 400 的页面标签。该计算实例 400 的页面标签指示了主机 10 的内存 120 中计

算实例 400 所占用的各个页面的重要程度。

计算实例 400 可以按照页面的重要程度对该计算实例 400 所占用的页面进行分类。计算实例 400 可以从占用页面的对象为角度，对该计算实例 400 所占用的页面进行分类。例如，计算实例 400 可以将被该计算实例 400 的操作系统（如 guestOS）所占用的页面归为一类，认为该类页面具备较高的重要程度；计算实例 400 可以将被该计算实例 400 上的应用所占用的页面归为一类，认为该类页面具备较低的重要程度。计算实例 400 可以从页面的调用频率（也可以称为访问频率），对该计算实例 400 所占用的页面进行分类，该计算实例 400 可以将该计算实例 400 经常调用的页面（如调用频率大于第一频率阈值的页面）归为一类，认为该类页面具备较高的重要程度；该计算实例 400 可以将该计算实例 400 不经常调用的页面（如调用频率小于第一频率阈值的页面）归为一类，认为该类页面具备较低的重要程度。

本申请实施例并不限定该计算实例 400 对该计算实例 400 所占用的页面进行分类的方式以及页面的重要程度的评价标准。

计算实例 400 对主机 10 的内存 120 中该计算实例 400 所占用的页面进行分类后，可以按照各类页面的重要程度为每一类页面设置一个重要程度数值，计算实例 400 的页面标签页面的重要程度数值。可选的，该计算实例 400 的页面标签还包括页面的虚拟地址（或页面的虚拟地址与计算实例 400 的标识的组合）。

计算实例 400 对主机 10 的内存 120 中该计算实例 400 所占用的页面进行分类后，也可以直接将页面所属的类别作为该页面的重要程度，计算实例 400 的页面标签包括页面的所属的类别。可选的，该计算实例 400 的页面标签还包括页面的虚拟地址（或页面的虚拟地址与计算实例 400 的标识的组合）。

例如，从占用页面的对象的角度，当计算实例 400 为虚拟机时，该虚拟机所占用的页面的类别可以包括被该虚拟机的操作系统（如 guestOS）所占用的页面、该虚拟机上的应用（application, APP）所占用的页面、虚拟机管理页面。被该计算实例 400 的操作系统（如 guestOS）所占用的页面的类别为可以标识为内核（kernel），该虚拟机上的应用所占用的页面可以标识为 APP。

其中，虚拟机的操作系统（如 guestOS）所占用的页面是在存储有与虚拟机操作系统相关数据或计算机程序指令的页面。虚拟机管理页面是指存储有虚拟机管理数据（如虚拟机的属性信息、虚拟机的标识等）的页面，虚拟机上的应用所占用的页面是存储有与应用相关数据或计算机程序指令的页面。

类似的，当计算实例 400 为容器时，该容器所占用的页面的类别可以包括该容器上的应用所占用的页面、容器管理页面。

其中，容器的操作系统（如 guestOS）所占用的页面是在存储有与容器操作系统相关数据或计算机程序指令的页面。容器管理页面是指存储有容器管理数据（如容器的属性信息、容器标识等）的页面，容器上的应用所占用的页面是存储有与应用相关数据或计算机程序指令的页面。

又例如，从页面的访问情况，该计算实例所占用的页面的类别可以包括冷页面、热页面。冷页面是页面的访问程度较低的页面（如页面的访问频率低于某一值），热页面是页面访问程度较高的页面（如页面的访问频率大于某一值）。

步骤 302：计算实例 400 向主机操作系统 200 发送该计算实例 400 的页面标签。主机操作系统 200 中可以由实例页面识别模块 210 接收到页面标签。

计算实例 400 可以直接将该页面标签传递给主机操作系统 200，也可以通过计算实例管

理单元 300 将该页面标签传递给主机操作系统 200。

需要说明的是，该主机操作系统 200 所接收到的页面标签中页面的虚拟地址应为该主机操作系统 200 能够识别的虚拟地址。

举例来说，当计算实例 400 为虚拟机时，该计算实例 400 在访问页面时，所采用的页面虚拟地址可以是 guestOS 为该页面所配置的客户机虚拟地址（guest virtual address, GVA）。MMU 获取的该计算实例 400 对应的 PT 包括两种页表，一个页表为客户端页表（guest page table, GPT），其中记录了客户端虚拟地址与客户端物理地址（guest physical address, GPA）之间的对应关系。另一个页表为扩展页表（extended page table, EPT），其中记录了客户端物理地址与主机 10 物理地址（host physical address, HPA）之间的对应关系，MMU 可以根据该 GPV 查询 GPT 以及 EPT，确定 HPA。而对于主机操作系统 200 所能识别的页面的虚拟地址为主机虚拟地址（host virtual address, HVA）。计算实例管理单元 300 能够将 GPA 转换为 HVA，计算实例 400 最初生成的页面标签中携带的页面的虚拟地址可以为 GPA（计算实例 400 内部保存有 GPT，计算实例 400 内部在生成页面标签时，可以自行将页面的 GVA 转换为 GPA），计算实例 400 在经过计算实例管理单元 300 传输该页面标签时，计算实例管理单元 300 可以将该页面标签中的 GPA 转换为 HVA，更新页面标签，将更新后的页面标签发送给主机操作系统 200。这样主机操作系统 200 所接收的页面标签中页面的虚拟地址即为该主机操作系统 200 能够识别的 HVA。

本申请实施例中计算实例 400 可以周期性或实时评估对该计算实例 400 所占用的页面进行分类，以确定该计算实例 400 所占用的页面的重要程度，进而生成该页面标签。计算实例 400 可以将该页面标签周期性的发送给主机操作系统 200。计算实例 400 也可以在页面标签发生变化（或计算实例 400 所占用的页面的重要程度发生变化）的情况下，向主机操作系统 200 发送变化后的页面标签或重要程度变化后生成的页面标签。

步骤 303：主机操作系统 200 获取该计算实例 400 所占的各个页面的页面访问频率。该步骤可以由实例页面识别模块 210 执行。

在前面关于 MMU 的描述可知，MMU 每查询到 PT 中的一个 PTE，会将该 PTE 中的标志位置 1，若该 PTE 在之后一段时间内未被访问，标志位将恢复为 0。

主机操作系统 200 可以周期性的扫描该计算实例 400 对应的 PT 中各个 PTE 的标志位，以确定该各个页面的页面访问频率。针对该计算实例 400 对应的 PT 中任一 PTE，主机操作系统 200 到达一个监控时间点时查看 PTE 的标志位，若主机操作系统 200 发现 PTE 的标志位为 1 时，记录该 PTE 对应的页面的访问次数加一，主机操作系统 200 并将该标志位置 0。若主机操作系统 200 发现 PTE 的标志位为 0 时，认为该 PTE 对应的页面未被访问，该 PTE 对应的页面的访问次数维持不变。在一段时间内重发执行上述操作，主机操作系统 200 能够获得该 PTE 对应的页面在该段时间内的访问次数，进而确定出该页面的页面访问频率。

步骤 304：主机操作系统 200 在需要将页面换出至交换分区 160 时，从计算实例 400 所占用的页面中筛选目标页面。该目标页面为该计算实例 400 所占用的页面中允许换出的页面。该步骤可以由实例页面识别模块 210 执行。

主机操作系统 200 需要将页面换出至交换分区 160 的情况有很多种，本申请实施例并不限定主机操作系统 200 需要将页面交换至交换分区 160 的具体场景。例如，主机操作系统 200 在接收到 MMU 发出的、用于指示缺页中断的异常信号时，若确定主机 10 的内存 120 中无空闲的内存空间或空闲的内存空间小于预设的空闲阈值时，可以确定需要将主机 10 的内存 120 中的页面换出至交换分区 160。又例如，主机操作系统 200 也可以预测未来可能存在主机 10

的内存空间不足的情况，如主机操作系统 200 检查到需要创建新的计算实例 400，又如，计算实例 400 中的 guest OS 需要进行版本更新。主机操作系统 200 可以确定需要将主机 10 的内存 120 中的页面换出至交换分区 160。

主机操作系统 200 还能够面向用户提供参数配置接口，允许用户配置页面交换参数。页面交换参数指示了目标页面的筛选条件。该页面交换参数包括但不限于：允许换出页面的计算实例 400 的标识、页面扫描频率、冷页面的判断阈值、以及计算实例 400 允许占用的最大内存空间。

由于计算实例 400 所占的页面的换入换出总是在一些情况下影响计算实例 400 的工作效率，并不是所有计算实例 400 均允许所占用的页面被换出，总是存在一些不允许换出页面的计算实例 400。例如，在主机 10 中一些承担重要业务的计算实例 400，或者该主机 10 中由重要用户所租用的计算实例 400。用户在主机操作系统 200 为用户提供的参数配置接口中配置允许换出页面的计算实例 400 的标识，或者不允许换出页面的计算实例 400 的标识。若用户未配置允许换出页面的计算实例 400 的标识或者主机操作系统 200 未向用户提供配置该允许换出页面的计算实例 400 的标识的接口，主机操作系统 200 可以认为主机 10 所有计算实例 400 均允许换出页面。

在步骤 303 中提及主机操作系统 200 能够周期性的扫描该计算实例 400 对应的 PT 中各个 PTE 的标志位，扫描该计算实例 400 对应的 PT 中各个 PTE 的频率即为页面扫描频率。过大的页面扫描频率会导致主机操作系统 200 需要频繁的扫描计算实例 400 的对应的 PT，会增大主机操作系统 200 的负担。而过小的扫描频率会导致主机操作系统 200 不能准确的监控到 PTE 的标志位置 1 的次数，使得最终获得的页面的访问频率的准确性较差。主机操作系统 200 允许用户根据自身经验或者需求配置该页面扫描频率。若用户未配置页面扫描频率或者主机操作系统 200 未向用户提供配置该页面扫描频率的接口，主机操作系统 200 可以以预设的页面扫描频率扫描该计算实例 400 对应的 PT 中各个 PTE 的标志位。

页面的访问频率小于某一阈值的页面通常被称为冷页面，该阈值即为冷页面的判断阈值。冷页面的访问频率较低，通常能够作为需要换出的页面换出到交换分区 160。主机操作系统 200 可以根据默认的阈值判断页面是否属于冷页面，也可以向用户提供配置该阈值的接口，由用户配置该阈值。

由于计算实例 400 会占用主机 10 内存 120 中的一些内存 120 空间，若计算实例 400 占用的内存 120 空间较大会影响其他计算实例 400 或主机操作系统 200 自身的运行。若计算实例 400 占用的内存 120 空间较小，该计算实例 400 的工作效率也会受到影响。主机操作系统 200 可以预设有计算实例 400 允许占用的最大内存 120 空间，主机操作系统 200 为该计算实例 400 所分配的内存 120 空间不允许超过该最大内存 120 空间。当然，也可以由用户配置该计算实例 400 允许占用的最大内存空间，向用户提供配置该计算实例 400 允许占用的最大内存空间的接口，在用户配置了该计算实例 400 允许占用的最大内存 120 空间后，主机操作系统 200 为该计算实例 400 所分配的内存 120 空间不允许超过用户配置的最大内存 120 空间。

应需理解的是，用户配置页面交换参数的时机应早于步骤 304，如主机操作系统 200 可以在计算实例 400 创建时或计算实例 400 运行过程中，向用户显示该页面交换参数的配置界面，以使得用户能够在该页面交换参数的配置界面中，完成页面交换参数的配置。

主机操作系统 200 能够获取该计算实例 400 的页面标签、页面访问频率以及页面交换参数这三种信息。主机操作系统 200 能够根据这三种信息中的一种或多种，对计算实例 400 所

占用的页面进行排序，从中选出目标页面。

主机操作系统 200 能够根据这三种信息中的一种或多种，对计算实例 400 所占用的页面进行排序，从中选出目标页面的方式有很多种，本申请并不限定。

例如，主机操作系统 200 能够根据页面标签指示的该计算实例 400 所占用的各个页面的重要程度对计算实例 400 所占用的各个页面进行排序，重要程度较高的页面排在靠前的位置。主机操作系统 200 将排序位置在最后的 N 个页面作为目标页面，N 为预设的正整数。以步骤 301 所列举的几种类别为例，目标页面包括计算实例 400 上应用所占用的页面，也可以包括计算实例 400 的冷页面。

又例如，主机操作系统 200 能够根据页面访问频率指示的该计算实例 400 所占用的各个页面的访问频率、按照访问频率由大到小顺序对计算实例 400 所占用的各个页面进行排序。主机操作系统 200 将排序位置在最后的 M 个页面作为目标页面，M 为预设的正整数。

又例如，主机操作系统 200 能够根据页面标签指示的该计算实例 400 所占用的各个页面的重要程度对计算实例 400 所占用的各个页面进行排序，重要程度较高的页面排在靠前的位置，对于重要程度相同的页面，主机操作系统 200 能够根据页面访问频率中该重要程度下的各个页面的访问频率、按照访问频率由大到小顺序进行排序，主机操作系统 200 将排序位置在最后的 P 个页面作为目标页面，P 为预设的正整数。

又例如，主机操作系统 200 能够根据页面标签以及页面访问频率对计算实例 400 所占用的各个页面进行排序后，根据页面交换参数中的冷页面的判断阈值，将排序位置在最后的 Q 个冷页面作为目标页面，Q 为预设的正整数。

又例如，主机操作系统 200 能够根据页面标签以及页面访问频率对计算实例 400 所占用的各个页面进行排序后，根据页面交换参数中的计算实例 400 允许占用的最大内存空间，将排序位置在最后的 H 个页面作为目标页面，H 为正整数，其中，计算实例 400 所占用的页面中除该 H 个页面外的剩余页面的内存空间的大小等于计算实例 400 允许占用的最大内存空间，或计算实例 400 所占用的页面中除该 H 个页面外的剩余页面的内存空间的大小小于计算实例 400 允许占用的最大内存 120 空间，这两者的差值处于预设范围。

在实际应用中，主机 10 的内存 120 中的页面并非均是由计算实例 400 所占用的，主机 10 的内存 120 中的页面还存在被主机操作系统 200 所占用的页面，以及该主机 10 上的应用所占用的页面。主机操作系统 200 也可以从主机 10 的内存 120 中除计算实例 400 所占用的页面之外的页面选出页面，作为需要被换出的页面，在下面步骤 305 中仅是以计算实例 400 所占用的页面作为目标页面为例进行说明，但对于主机操作系统 200 从主机 10 的内存 120 中除计算实例 400 所占用的页面之外的页面选出的需要换出的页面，主机操作系统 200 也可以采用与步骤 305 类似的方式将页面换出到交换分区 160。

步骤 305：主机操作系统 200（如主机操作系统 200 中的交换模块 220）将目标页面从主机 10 的内存 120 中换出至交换分区 160。主机操作系统 200 更新该计算实例 400 对应的 PT，删除该计算实例 400 对应的 PT 中记录该目标页面的虚拟地址的 PTE。

主机操作系统 200 在执行该步骤 305 时，主机操作系统 200 可以直接将主机 10 的内存 120 中的目标页面发送到交换分区 160。主机操作系统 200 可以向交换分区 160 所在的存储设备 20 发送第一数据请求，该第一数据请求用于请求将该目标页面存储在该存储设备 20 中，该第一数据请求中携带了该目标页面，以及主机操作系统 200 设置的该目标页面在交换分区 160 中的地址（为方便说明，该目标页面在交换分区 160 中的地址或该目标页面在存储设备 20 中的地址称为目标页面的交换地址），主机操作系统 200 可以记录该目标页面的虚拟地址

与该目标页面的交换地址之间的对应关系。

应需理解的是,由于主机 10 上不同进程所使用的虚拟地址可能相同,为了区分不同进程。主机操作系统 200 记录的对应关系可以为该目标页面的虚拟地址加计算实例 400 的标识的组合、与该目标页面的交换地址之间的对应关系。下文出现的该目标页面的虚拟地址与该目标页面的交换地址之间的对应关系均具备该指定意义,但为了方便说明,仍采用该目标页面的虚拟地址与该目标页面的交换地址之间的对应关系的表述。也就是说,该对应关系可以为该目标页面的虚拟地址与该目标页面的交换地址之间的对应关系(如在主机 10 上不同进程或计算实例 400 所使用的虚拟地址完全不同的场景下),该对应关系可以为该目标页面的虚拟地址加计算实例 400 的标识的组合、与该目标页面的交换地址之间的对应关系(如在主机 10 上不同进程或计算实例 400 所使用的虚拟地址可能相同的场景下)。

主机操作系统 200 可以为换出的页面维护页面交换表,页面交换表包括多个表项,每个表项对应一个被换出至交换分区 160 的页面,该表项记录了该页面的虚拟地址与页面的交换地址(页面的交换地址即为该页面在交换分区 160 的地址)的对应关系。主机操作系统 200 在该目标页面换出至交换分区 160 时,主机操作系统 200 在页面交换表中增加与该目标页面对应的表项,用于记录目标页面的虚拟地址与该目标页面的交换地址之间的对应关系。

该存储设备 20 在接收到该第一数据请求后,为该目标页面分配存储位置,将该目标页面保存在该存储位置中,并记录该目标页面的交换地址与该存储位置之间的对应关系。

主机操作系统 200 在执行该步骤 305 时,也可以通过主机 10 的加速装置 150 将目标页面从主机 10 的内存 120 中换出至交换分区 160。例如,主机操作系统 200 向该主机 10 的加速装置 150 发送换出指令,该换出指令中指示加速装置 150 将目标页面换出至交换分区 160,该换出指令携带了目标页面的虚拟地址以及该目标页面,可选的,还换出指令中还携带了该计算实例 400 的标识。

加速装置 150 在接收到该换出指令后,向交换分区 160 所在的存储设备 20 发送第二数据请求,该第二数据请求用于请求将该目标页面存储在该存储设备 20 中,该第二数据请求中携带了该目标页面,以及加速装置 150 所配置的该目标页面的交换地址,加速装置 150 记录该目标页面的虚拟地址与该目标页面的交换地址之间的对应关系。

类似的,加速装置 150 也可以为换出的页面维护页面交换表,页面交换表包括多个表项,每个表项对应一个被换出至交换分区 160 的页面,该表项记录了该页面的虚拟地址与页面的交换地址的对应关系。加速装置 150 在将该目标页面换出至交换分区 160 时,加速装置 150 在页面交换表中增加与该目标页面对应的表项,用于记录目标页面的虚拟地址与该目标页面的交换地址之间的对应关系。

该存储设备 20 在接收到该第二数据请求后,为该目标页面分配存储位置,将该目标页面保存在该存储位置中,并记录该目标页面的交换地址与该存储位置之间的对应关系。

在由加速装置 150 将页面从主机 10 的内存 120 换出至交换分区 160 时,加速装置 150 可以直接将该目标页面换出至该交换分区 160。加速装置 150 也可以先对该目标页面进行压缩,将压缩后的目标页面换出至该交换分区 160。

至此,页面的换出流程结束,当计算实例 400 需要调用该目标页面时,会触发页面的换入流程,具体可以参见图 4。

如图 4 所示,为本申请实施例提供的一种页面换入方法,在图 4 中以计算实例 400 调用目标页面的场景为例进行说明。需要说明的是,在该主机 10 的其他进程调用被换出至交换分区 160 的页面的场景,如图 4 所示的实施例也同样适用,将计算实例 400 理解为主机 10 的其

他进程，目标页面理解为该其他进程被换出至交换分区 160 的页面。该方法包括：

步骤 401：计算实例 400 发起访问请求，该访问请求中携带了目标页面的虚拟地址。该访问请求用于请求获取该目标页面中的数据。

步骤 402：MMU 获取该访问请求，在该计算实例 400 对应的 PT 中查询相关的 PTE。

MMU 在获取该访问请求后，可以查询该计算实例 400 对应的 PT，由于该目标页面之前已被换出至交换分区 160，PT 并不存在该目标页面对应的 PTE（也即记录了该目标页面的虚拟地址的 PTE），MMU 会发现该 PT 中缺少页表项。

若 MMU 在计算实例 400 对应的 PT 中未查询到相关的 PTE，MMU 会触发如下两个流程。一个流程为主机操作系统 200 的页面重映射流程（参见步骤 403~步骤 405），在该流程中主机操作系统 200 需要为该目标页面中的数据重新配置内存空间，也即为该目标页面中的数据在主机 10 的内存 120 中找寻空白页面（也即空闲页面）。另一个流程为加速装置 150 的页面换入流程（参见步骤 406~步骤 409），在该流程中，加速装置 150 从交换分区 160 中获取目标页面，并将目标页面写入到主机 10 的内存 120 中。这两个流程可以同步进行。

流程一、主机操作系统 200 的页面重映射流程（参见步骤 403~步骤 405）。

步骤 403：MMU 向主机操作系统 200 发送用于指示缺页中断的异常信号。

步骤 404：主机操作系统 200 为该目标页面分配内存空间，更新该计算实例 400 对应的 PT。

主机操作系统 200 在主机 10 的内存 120 为该目标页面分配内存空间，内存空间的物理地址即为该目标页面新的物理地址。

主机操作系统 200 在该计算实例 400 对应的 PT 中增加 PTE，该 PTE 记录该目标页面的虚拟地址与该目标页面新的物理地址之间的对应关系。

步骤 405：主机操作系统 200 通知加速装置 150 页面分配完成，也即主机操作系统 200 告知加速装置 150 已在该主机 10 的内存 120 中为该目标页面重新分配的存储空间，并告知该加速装置 150 该目标页面新的物理地址。

主机操作系统 200 通知加速装置 150 页面分配完成的方式有很多种。例如，主机操作系统 200 通过与 DPU151 连接的系统总线发送通知消息，该通知消息指示已在该主机 10 的内存 120 中为该目标页面重新分配存储空间，该通知消息还携带了该目标页面新的物理地址。

又例如，主机操作系统 200 与加速装置 150 之间可以设置共享内存，该共享内存可以位于主机 10 的内存 120 中，也可以位于该加速装置 150 的存储器 152 中。共享内存 120 是指主机操作系统 200 与加速装置 150 均能够使用的存储空间。在该共享内存中保存了由主机操作系统 200 与加速装置 150 共同维护的分配完成队列（completion queue）。主机操作系统 200 在该主机 10 的内存 120 中为该目标页面重新分配的内存空间，更新了该计算实例 400 对应的 PT 后，可在该完成队列中加入一个完成指示，该完成指示用于指示页面分配完成，该完成指示还携带有该目标页面新的物理地址。

流程二、加速装置 150 的页面换入流程（参见步骤 406~步骤 409）。

步骤 406：MMU 向加速装置 150 发送换入指令，该换入指令携带了目标页面的虚拟地址。可选的，该换入指令还携带了计算实例 400 的标识。

步骤 407：加速装置 150 根据该目标页面的虚拟地址确定该目标页面的交换地址。

加速装置 150 可以获取目标页面的虚拟地址以及目标页面的交换地址的对应关系，确定与该目标页面的虚拟地址对应的交换地址。

其中，目标页面的虚拟地址以及目标页面的交换地址的对应关系可以是主机操作系统 200

记录的，如图 3 所示的实施例的步骤 305 中主机操作系统 200 记录的。加速装置 150 可以从主机操作系统 200 获取该页面的虚拟地址以及页面的交换地址的对应关系。例如，加速装置 150 可以从主机操作系统 200 获取主机操作系统 200 维护的页面交换表。

目标页面的虚拟地址以及目标页面的交换地址的对应关系也可以是加速装置 150 记录的，如图 3 所示的实施例的步骤 305 中加速装置 150 记录的。例如，加速装置 150 从页面交换表查询目标页面的虚拟地址以及目标页面的交换地址的对应关系。

步骤 408: 加速装置 150 根据该目标页面的交换地址从该交换分区 160 获取该目标页面。

加速装置 150 向交换分区 160 所在的存储设备 20 发送第三数据请求，该第三数据请求用于请求从存储设备 20 获取该目标页面，该第三数据请求中携带了该目标页面的交换地址。

该存储设备 20 在接收到该第三数据请求后，根据该目标页面的交换地址确定存储该目标页面的存储位置，从存储位置读取该目标页面，将该目标页面反馈给加速装置 150。

步骤 409: 加速装置 150 在接收到主机操作系统 200 的通知后，将该目标页面写入到该主机 10 的内存 120 中，通知 MMU 页面换入完成。

加速装置 150 可以接收该主机操作系统 200 的通知消息，确定该主机操作系统 200 已经为该目标页面重新分配的内存空间，加速装置 150 可以根据该通知消息中携带的目标页面新的物理地址、通过直接内存访问 (direct memory access, DMA)，将该目标页面写入到该新的物理地址处，加速装置 150 在将该目标页面写入到该新的物理地址后，通知 MMU 页面换入完成。

加速装置 150 也可以从共享内存的分配完成队列中查看是否有完成指示，若该分配完成队列中有完成指示，加速装置 150 将该完成指示取出，获取该目标页面新的物理地址。加速装置 150 可以根据该目标页面新的物理地址、通过 DMA，将该目标页面写入到该新的物理地址处，加速装置 150 在将该目标页面写入到该新的物理地址后，通知 MMU 页面换入完成。若该分配完成队列中不存在完成指示，加速装置 150 可以暂缓将该目标页面写入到该主机 10 的内存 120 中，直至检测到该分配完成队列中出现完成指示。

在由加速装置 150 将页面从交换分区 160 换出至主机 10 的内存 120 时，当加速装置 150 从交换分区 160 获取的目标页面可以是该目标页面本身，也即该目标页面未被压缩，这种情况下加速装置 150 可以直接将该目标页面换入至主机 10 的内存 120 中。当加速装置 150 从交换分区 160 获取的目标页面是压缩后的目标页面，该目标页面的压缩可是在该目标页面被交换到交换分区 160 之前由加速装置 150 执行的，也可以是目标页面在被换出至交换分区 160 后，由交换分区 160 所在的存储设备 20 执行的。这种情况下加速装置 150 可以先对该压缩后的目标页面进行解压，获得目标页面，将目标页面换入至主机 10 的内存 120 中。

至此，目标页面从交换分区 160 换入到主机 10 的内存 120 中。

MMU 在接收到加速装置 150 的通知后，获取更新后的、该计算实例 400 对应的 PT，查找记录该目标页面的虚拟地址的 PTE，从主机 10 的内存 120 中获取目标页面，将该目标页面反馈给计算实例 400。

MMU 从更新后的 PT 中查找记录该目标页面的虚拟地址的 PTE，确定该目标页面新的物理地址，从该新的物理地址上读取该目标页面，将该目标页面反馈给计算实例 400。

需要说明的是，在本申请实施例中目标页面的换出以及目标页面的换出，针对的是该目标页面中的数据，为了方便表述，在本申请实施例中均是以目标页面指代该目标页面中的数据。

基于与方法实施例同一发明构思，本申请实施例还提供了一种页面换入装置，该页面换

入装置用于执行上述如图3或图4所示的方法实施例中DPU151或加速装置150执行的方法，相关特征可参见上述方法实施例，此处不再赘述。如图5所示，页面换入装置500包括接收模块501以及换入模块502。可选的，还可以包括换出模块503。

接收模块501，用于接收换入指令，换入指令用于指示将目标页面换入至主机的内存。

换入模块502，用于从存储设备获取目标页面；将目标页面换入至内存，通知主机目标页面已换入到内存中。

在一种可能的实施方式中，接收模块501可以接收主机中的MMU或处理器发送的换入指令。相应的，换入模块502知主机的MMU或处理器目标页面已换入到内存中。

在一种可能的实施方式中，换入模块502在从存储设备获取目标页面时，获取目标页面的虚拟地址与目标页面的交换地址的对应关系，目标页面的交换地址为目标页面在存储设备的地址。换入模块502根据对应关系确定目标页面的交换地址，根据目标页面的交换地址从存储设备获取目标页面。

在一种可能的实施方式中，对应关系可以是页面换入装置将目标页面从内存换出至存储设备时、记录的。对应关系也可以是换入模块502从主机的处理器获取的。

在一种可能的实施方式中，换入模块502在将目标页面换入至内存时，可以接收主机的处理器发送的通知，通知用于指示处理器在内存中为目标页面配置的物理地址。换入模块502在获取该物理地址后，通过DMA将目标页面写入到物理地址。

一种可能的实施方式中，接收模块501接收主机的处理器发送的换出指令，换出指令用于指示将主机内存中的目标页面换出至存储设备，换出指令包括目标页面的虚拟地址以及目标页面。

换出模块503将目标页面换出至存储设备，记录对应关系。

在一种可能的实施方式中，换入模块502在从存储设备获取目标页面时，若从存储设备获取了压缩后的目标页面；换入模块502可以对压缩后的目标页面解压，获得目标页面，再将该目标页面换入到内存中。

在一种可能的实施方式中，换出模块503在将目标页面换出至存储设备时，可以将目标页面压缩后，换出至存储设备。

需要说明的是，本申请实施例中对模块的划分是示意性的，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。在本申请的实施例中的各功能模块可以集成在一个处理模块中，也可以是各个模块单独物理存在，也可以两个或两个以上模块集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。

上述实施例，可以全部或部分地通过软件、硬件、固件或其他任意组合来实现。当使用软件实现时，上述实施例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指令。在计算机上加载或执行所述计算机程序指令时，全部或部分地产生按照本发明实施例所述的流程或功能。所述计算机可以为通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机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（例如同轴电缆、光纤、数字用户线（DSL））或无线（例如红外、无线、微波等）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包含一个或多个可用介质集合的服务器、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）、光介质（例如，DVD）、或者半导体介质。半导体介质

可以是固态硬盘（solid state drive, SSD）。

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。

显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变形在内。

权 利 要 求

1.一种页面换入出方法，其特征在于，所述方法包括：

加速装置接收主机发送的换入指令，所述换入指令用于指示将目标页面换入至所述主机的内存，所述加速装置通过系统总线与所述主机连接；

所述加速装置从存储设备获取所述目标页面；

所述加速装置将所述目标页面换入至所述内存，通知所述主机所述目标页面已换入到所述内存中。

2.如权利要求 1 所述的方法，其特征在于，所述主机中的加速装置接收换入指令，包括：

所述加速装置接收所述主机中的内存管理单元 MMU 发送的换入指令。

3.如权利要求 1 或 2 所述的方法，其特征在于，所述加速装置从存储设备获取所述目标页面，包括：

所述加速装置获取所述目标页面的虚拟地址与所述目标页面的交换地址的对应关系，所述目标页面的交换地址为所述目标页面在所述存储设备的地址；

所述加速装置根据所述对应关系确定所述目标页面的交换地址；

所述加速装置根据所述目标页面的交换地址从所述存储设备获取所述目标页面。

4.如权利要求 3 所述的方法，其特征在于，所述对应关系是所述加速装置将所述目标页面从所述内存换出至所述存储设备时、记录的。

5.如权利要求 1-4 任一项所述的方法，其特征在于，所述主机还包括处理器，所述加速装置通过系统总线与所述处理器连接，所述加速装置将所述目标页面换入至所述内存，包括：

所述加速装置接收所述处理器发送的通知，所述通知用于指示所述处理器在所述内存中为所述目标页面配置的物理地址；

所述加速装置通过直接内存访问 DMA 将所述目标页面写入到所述物理地址。

6.如权利要求 5 所述的方法，其特征在于，所述加速装置接收换入指令之前，还包括：

所述加速装置接收所述主机的处理器发送的换出指令，所述换出指令用于指示将所述主机内存中的所述目标页面换出至所述存储设备，所述换出指令包括所述目标页面的虚拟地址以及所述目标页面；

所述加速装置将所述目标页面换出至所述存储设备，记录所述对应关系。

7.如权利要求 1-6 任一项所述的方法，其特征在于，所述加速装置从存储设备获取所述目标页面，包括：

所述加速装置从所述存储设备获取压缩后的所述目标页面；

所述加速装置对压缩后的所述目标页面解压，获得所述目标页面。

8.如权利要求 6 所述的方法，其特征在于，所述加速装置将所述目标页面换出至所述存储设备，包括：

所述加速装置将所述目标页面压缩后，换出至所述存储设备。

9.如权利要求 3 所述的方法，其特征在于，所述对应关系是所述加速装置从所述主机的处理器获取的，所述加速装置通过系统总线与所述处理器连接。

10.如权利要求 1-9 任一项所述的方法，其特征在于，所述存储设备为部署在所述主机之外，与所述主机通过网络连接的存储设备，或为所述主机的本地存储设备，或为公有云或私有云中为所述主机分配的存储空间。

11.一种页面换入装置，其特征在于，所述装置位于主机中，所述装置包括：

接收模块，用于接收换入指令，所述换入指令用于指示将目标页面换入至所述主机的内存；

换入模块，用于从存储设备获取所述目标页面；将所述目标页面换入至所述内存，通知所述主机所述目标页面已换入到所述内存中。

5 12.如权利要求 11 所述的装置，其特征在于，所述接收模块，用于：

接收所述主机中的内存管理模块 MMU 发送的换入指令。

13.如权利要求 11 或 12 所述的装置，其特征在于，所述换入模块在从所述存储设备获取所述目标页面，用于：

10 获取所述目标页面的虚拟地址与所述目标页面的交换地址的对应关系，所述目标页面的交换地址为所述目标页面在所述存储设备的地址；

根据所述对应关系确定所述目标页面的交换地址；

根据所述目标页面的交换地址从所述存储设备获取所述目标页面。

14.如权利要求 13 所述的装置，其特征在于，所述对应关系是所述换入模块将所述目标页面从所述内存换出至所述存储设备时、记录的。

15 15.如权利要求 11-14 任一项所述的装置，其特征在于，所述换入模块在将所述目标页面换入至所述内存，用于：

接收所述主机的处理器发送的通知，所述通知用于指示所述处理器在所述内存中为所述目标页面配置的物理地址；

通过直接内存访问 DMA 将所述目标页面写入到所述物理地址。

20 16.如权利要求 15 所述的装置，其特征在于，所述装置还包括换出模块，

所述接收模块，还用于：接收所述主机的处理器发送的换出指令，所述换出指令用于指示将所述主机内存中的所述目标页面换出至所述存储设备，所述换出指令包括所述目标页面的虚拟地址以及所述目标页面；

所述换出模块，用于将所述目标页面换出至所述存储设备，记录所述对应关系。

25 17.如权利要求 11-16 任一项所述的装置，其特征在于，所述换入模块在从所述存储设备获取所述目标页面，用于：

从所述存储设备获取压缩后的所述目标页面；

对压缩后的所述目标页面解压，获得所述目标页面。

30 18.如权利要求 16 所述的装置，其特征在于，所述换出模块在将所述目标页面换出至所述存储设备，用于：

将所述目标页面压缩后，换出至所述存储设备。

19.如权利要求 13 所述的装置，其特征在于，所述对应关系是所述换入模块从所述主机的处理器获取的。

35 20.如权利要求 11-19 任一项所述的装置，其特征在于，所述存储设备为部署在所述主机之外，与所述主机通过网络连接的存储设备，或为所述主机的本地存储设备，或为公有云或私有云中为所述主机分配的存储空间。

21.一种加速装置，其特征在于，所述加速装置包括数据处理单元 DPU 和供电电路，所述供电电路用于对所述 DPU 供电，所述 DPU 用于执行如权利要求 1-10 任一项所述的方法。

40 22.一种计算机存储介质，其特征在于，所述计算机可读存储介质存储有计算机可执行指令，所述计算机可执行指令用于使计算机执行权利要求 1-10 任一项所述的方法。

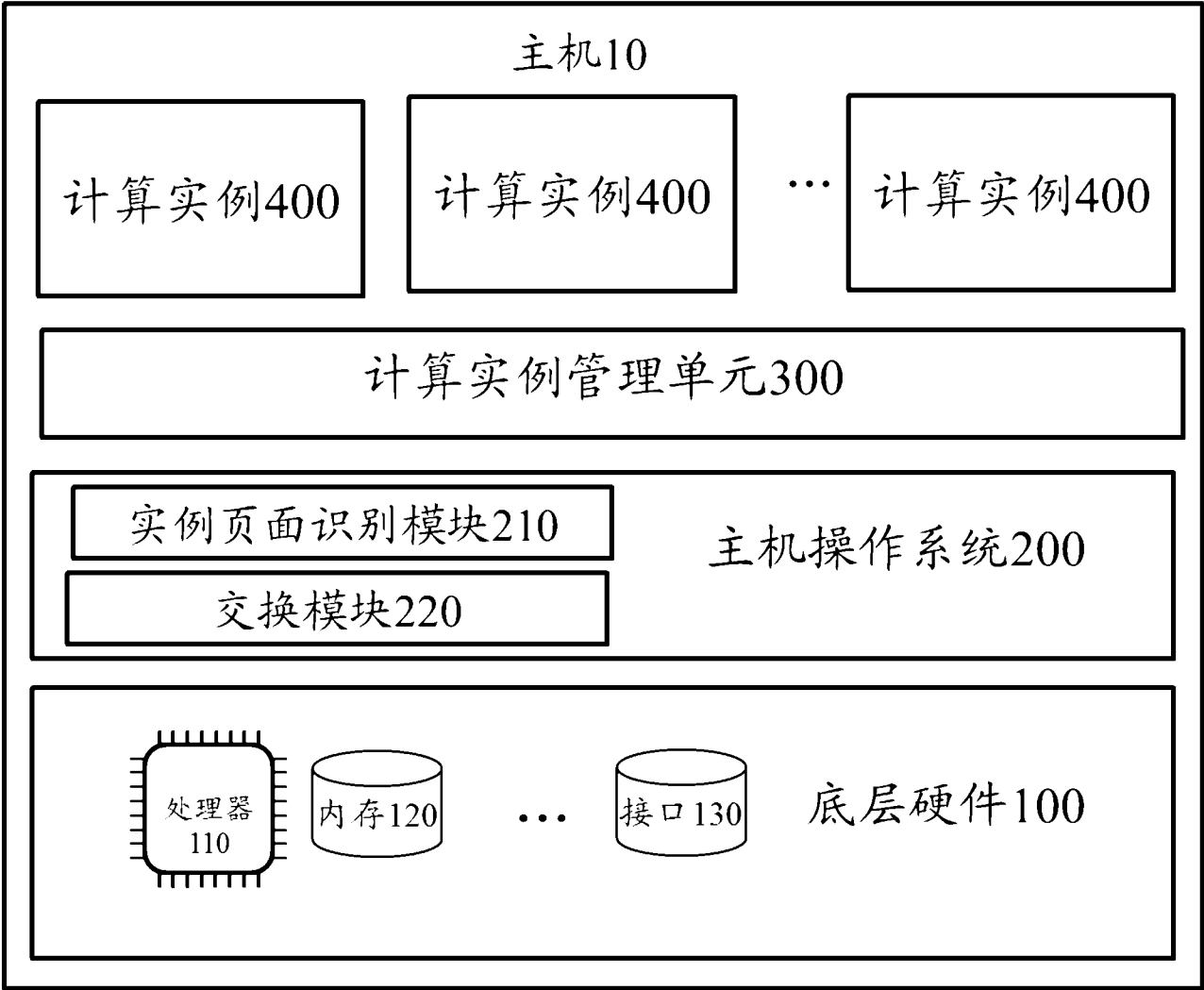


图 1

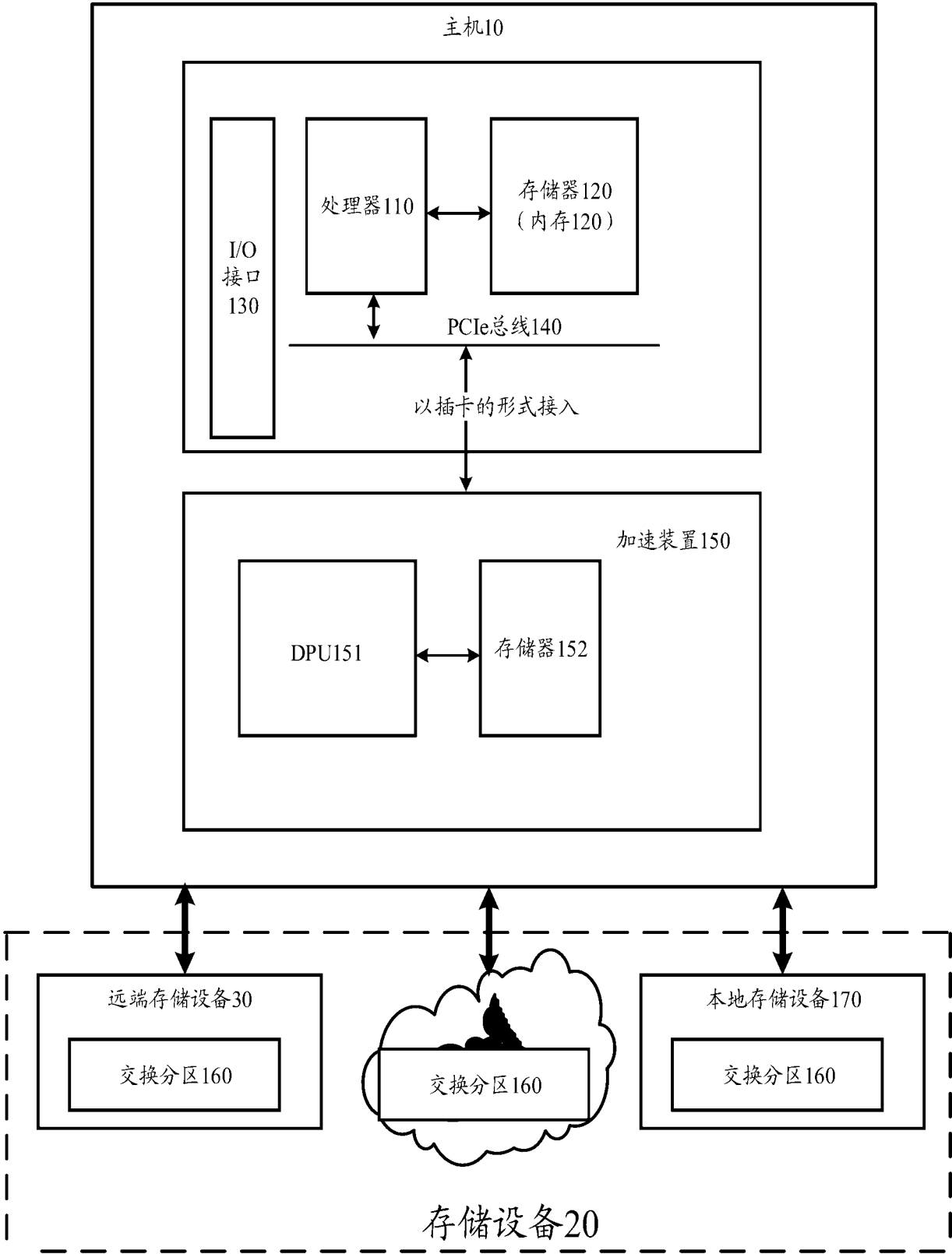


图 2

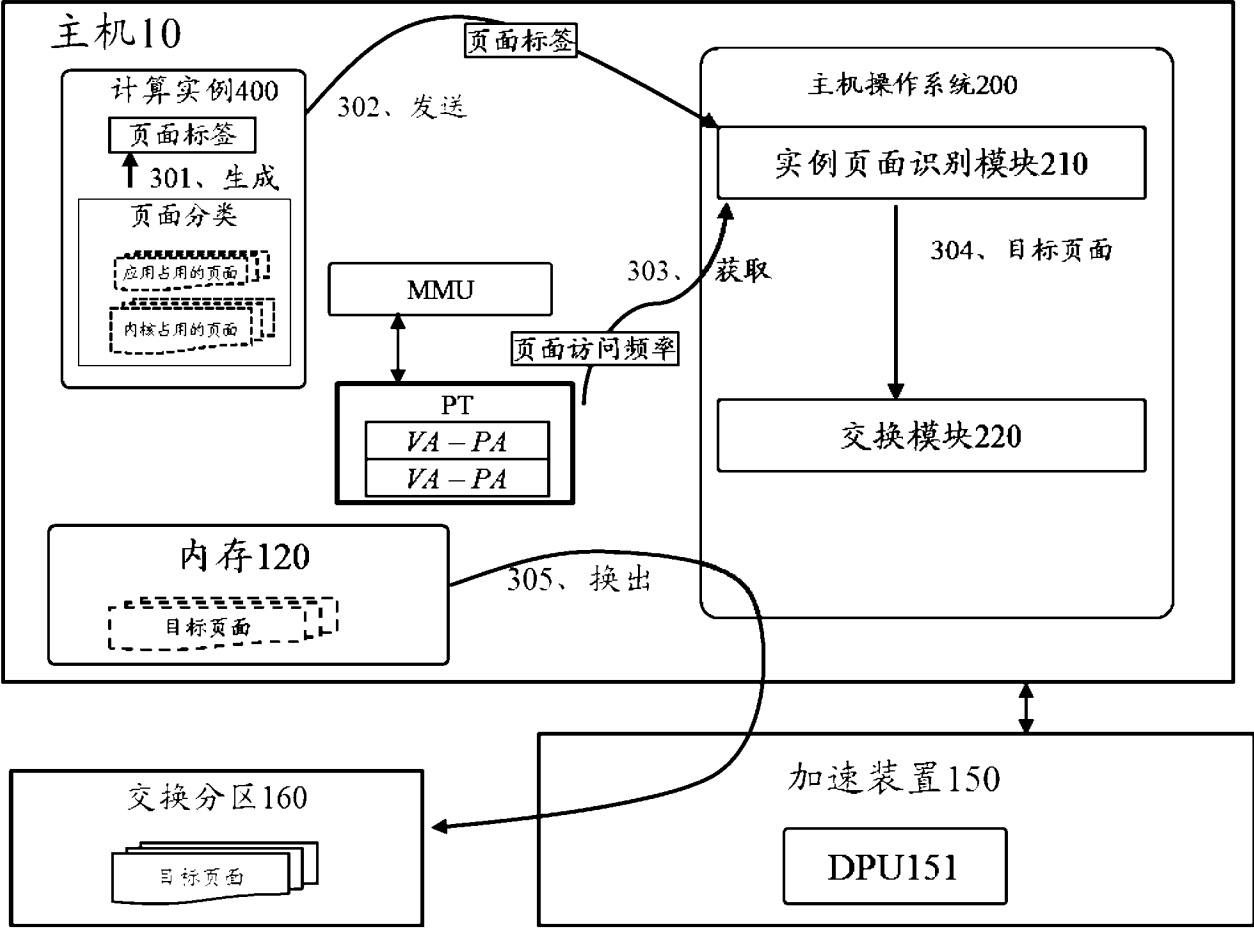


图 3

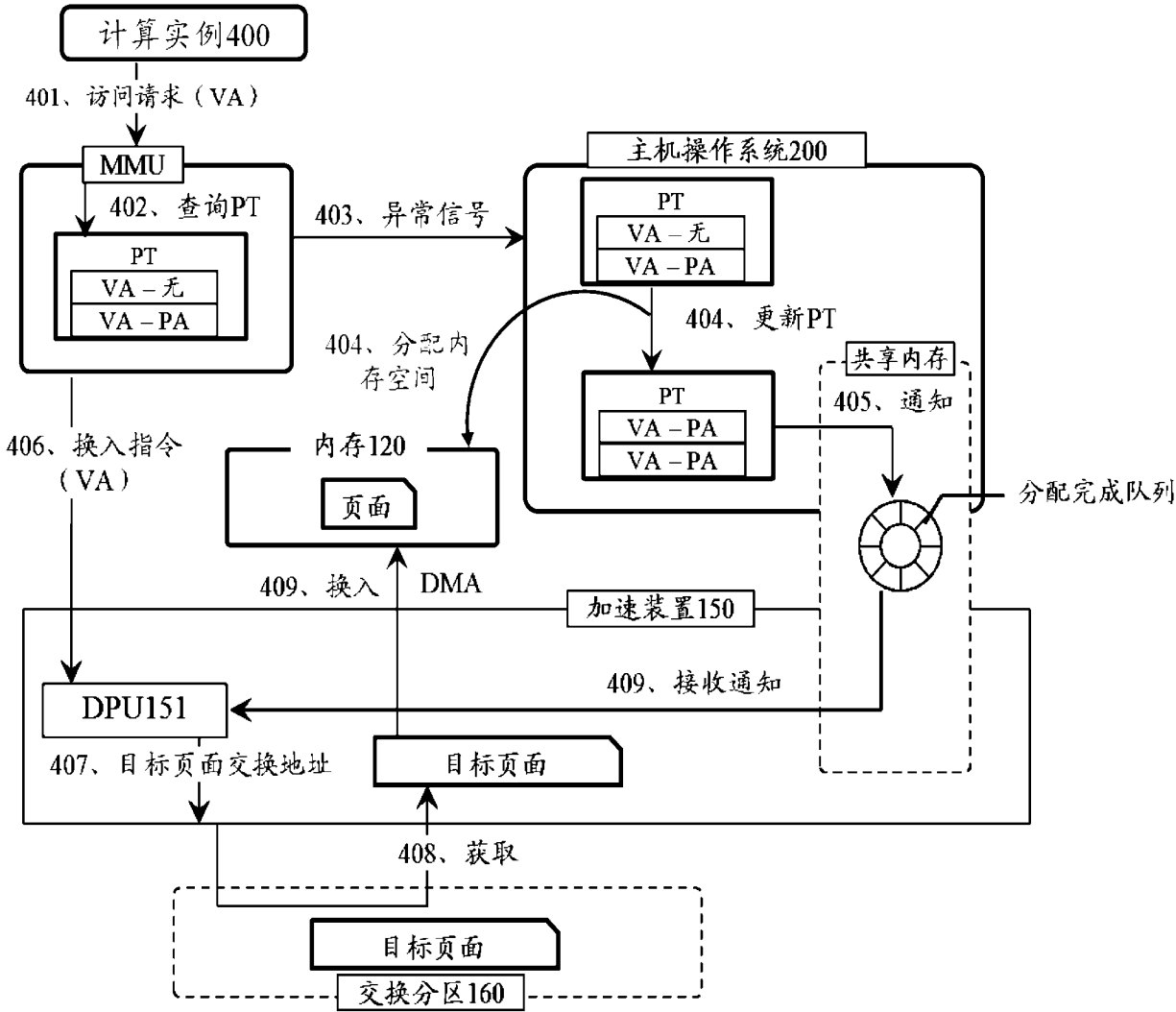


图 4

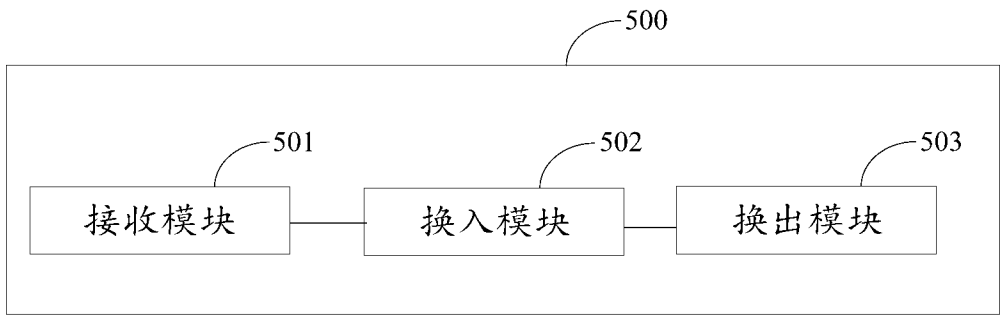


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/100492

A. CLASSIFICATION OF SUBJECT MATTER

G06F9/455(2018.01)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, G06T

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)

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 页面, 换入, 换出, 加速, 存储设备, 目标页, 内存, 虚拟地址, 物理地址, 直接内存访问, 压缩, 解压, page?, swap-in, swap-out, accelerat+, storage, target, memory, virtual w address, physical w address, direct w memory w access, DMA, compress+, decompress+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 109670345 A (CHENGDU HIGON INTEGRATED CIRCUIT DESIGN CO., LTD.) 23 April 2019 (2019-04-23) description, paragraphs 0080-0166	1-22
X	CN 111967065 A (HAIGUANG INFORMATION TECHNOLOGY CO., LTD.) 20 November 2020 (2020-11-20) description, paragraphs 0033-0103	1-22
A	CN 113590509 A (HUAWEI TECHNOLOGIES CO., LTD.) 02 November 2021 (2021-11-02) entire document	1-22
A	WO 2022121866 A1 (FOURTH PARADIGM BEIJING TECHNOLOGY CO., LTD.) 16 June 2022 (2022-06-16) entire document	1-22



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
 “D” document cited by the applicant in the international application
 “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

01 August 2023

Date of mailing of the international search report

07 August 2023

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,
Beijing 100088

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2023/100492

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109670345	A	23 April 2019	None			
CN	111967065	A	20 November 2020	None			
CN	113590509	A	02 November 2021	WO	2021218502	A1	04 November 2021
				EP	4134829	A1	15 February 2023
WO	2022121866	A1	16 June 2022	CN	112598565	A	02 April 2021

国际检索报告

国际申请号

PCT/CN2023/100492

A. 主题的分类

G06F9/455 (2018.01) 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, G06T

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))

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 页面, 换入, 换出, 加速, 存储设备, 目标页, 内存, 虚拟地址, 物理地址, 直接内存访问, 压缩, 解压, page?, swap-in, swap-out, accelerat+, storage, target, memory, virtual w address, physical w address, direct w memory w access, DMA, compress+, decompress+

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
X	CN 109670345 A (成都海光集成电路设计有限公司) 2019年4月23日 (2019 - 04 - 23) 说明书第0080-0166段	1-22
X	CN 111967065 A (海光信息技术有限公司) 2020年11月20日 (2020 - 11 - 20) 说明书第0033-0103段	1-22
A	CN 113590509 A (华为技术有限公司) 2021年11月2日 (2021 - 11 - 02) 全文	1-22
A	WO 2022121866 A1 (第四范式 (北京) 技术有限公司) 2022年6月16日 (2022 - 06 - 16) 全文	1-22

☐ 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

“D” 申请人在国际申请中引证的文件

“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

2023年8月1日

国际检索报告邮寄日期

2023年8月7日

ISA/CN的名称和邮寄地址

中国国家知识产权局

中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

霍玉明

电话号码 (+86) 010-53961327

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2023/100492

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	109670345	A	2019年4月23日	无			
CN	111967065	A	2020年11月20日	无			
CN	113590509	A	2021年11月2日	W0	2021218502	A1	2021年11月4日
				EP	4134829	A1	2023年2月15日
W0	2022121866	A1	2022年6月16日	CN	112598565	A	2021年4月2日