(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
2012 年 9 月 20 日 (20.09.2012)

WIPO | PCT

(10) 国际公布号
WO 2012/122796 A1

(51) 国际专利分类号:
G06F 9/455 (2006.01)

(21) 国际申请号: PCT/CN2011/080573

(22) 国际申请日: 2011 年 10 月 9 日 (09.10.2011)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201110061738.5 2011 年 3 月 15 日 (15.03.2011) CN

(71) 申请人 (对除美国外的所有指定国): **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; 及

(75) 发明人/申请人 (仅对美国): **邱军 (QIU, Jun)** [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)。 **叶川 (YE, Chuan)** [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)。

(74) 代理人: **北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.)**; 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (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, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, 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, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, 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))。

(54) **Title**: METHOD FOR CREATING VIRTUAL MACHINE, VIRTUAL MACHINE MONITOR AND VIRTUAL MACHINE SYSTEM

(54) **发明名称** : 一种创建虚拟机的方法、虚拟机监控器及虚拟机系统

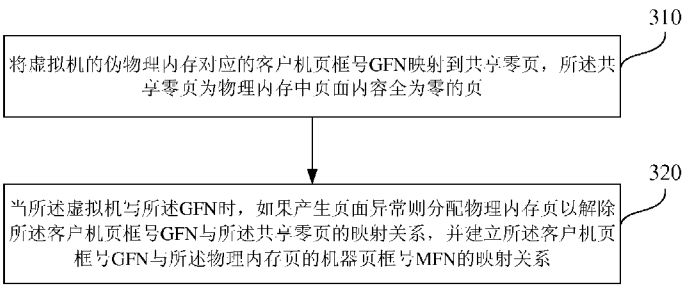

图 3 / Fig. 3

310 MAPPING GFN CORRESPONDING TO PSEUDO-PHYSICAL MEMORY OF VIRTUAL MACHINE ONTO SHARED ZERO PAGE, PAGES IN PHYSICAL MEMORY HAVING WHOLLY ZERO PAGE CONTENT

320 IF PAGE FAULT OCCURS WHEN VIRTUAL MACHINE WRITES GFN, ALLOCATION OF PHYSICAL MEMORY PAGE TO REMOVE MAPPING RELATIONSHIP BETWEEN GFN AND MFN OF SAID PHYSICAL MEMORY PAGE

(57) **Abstract**: Provided in the embodiments of the present invention are a method for creating a virtual machine, a virtual machine monitor and a virtual machine system, the method comprising: mapping a guest frame number (GFN) corresponding to a pseudo-physical memory of the virtual machine onto a shared zero page, said shared zero page being pages in the physical memory having wholly zero page content. If a page fault occurs when the virtual machine writes a GFN, a physical memory page is allocated to remove the mapping relationship between the GFN and a machine frame number (MFN) of said physical memory page. The method reduces the amount of memory used during the virtual machine start process, increases virtual machine density, and supports concurrent start-up of the number of virtual machines needed for memory overcommit.

[见续页]

WO 2012/122796 A1

(57) 摘要:

本发明实施例提供了一种创建虚拟机的方法、虚拟机监控器及虚拟机系统，该方法包括：将虚拟机的伪物理内存对应的客户机页框号 GFN 映射到一共享零页，所述共享零页为物理内存中页面内容全为零的页；当所述虚拟机写所述 GFN 时，如果产生页面异常则分配物理内存页以解除所述客户机页框号 GFN 与所述共享零页的映射关系，并建立所述客户机页框号 GFN 与所述物理内存页的机器页框号 MFN 的映射关系。该方法可以减少虚拟机启动过程中的内存使用量，提高虚拟机密度，支持并发启动 Memory Overcommitted 数量的虚拟机。

一种创建虚拟机的方法、虚拟机监控器及虚拟机系统

技术领域

本发明涉及虚拟机技术，具体地涉及一种创建虚拟机的方法、虚拟机监控器及虚拟机系统。

背景技术

虚拟化技术是一种将底层硬件设备与上层操作系统、应用程序分离的去耦合方法，其引入虚拟机监控器（Virtual Machine Monitor, VMM）层来直接管理底层硬件资源，并创建与底层硬件无关的虚拟机（Virtual Machine，VM）供上层操作系统和应用程序使用。虚拟化技术作为当前流行的云计算（Cloud Computing）平台的底层重要支撑技术之一，可以大大提高物理设备的资源使用效率。如图 1 所示，经过系统虚拟化后，一台物理机器上可以同时运行多个虚拟机 VM，物理机器上支持同时运行的虚拟机数量称为虚拟机密度(Virtual Machine Density)，虚拟机密度越大，资源利用率越高。

为提高虚拟机密度，CPU 虚拟化时采用调度的方式使得虚拟机的虚拟处理器可以共享物理 CPU，如果不考虑性能，理论上可以虚拟出任意多个虚拟 CPU。外设虚拟化通过软件模拟或连接到外部子系统，例如 SAN（存储域网络，Storage Area Network），同样可以虚拟出任意多份。但在内存虚拟化方面，将同一个物理内存页给多个虚拟机同时使用的前提是虚拟机之间具有页面内容完全相同的页，因此虚拟内存量理论上不能超过物理内存量。为使虚拟内存量大于物理内存量的限制，业界提出了 Memory Overcommitted（虚拟内存量大于物理内存）方法，该方法包括：气球驱动（Balloon Driver）、基于页内容的内存页共享（Content Based Page Sharing, CBPS）、内存页交换（Memory Page Swap）、内存页压缩（Memory Page Compression）和 Populate on Demand（PoD）等。

气球驱动（Balloon Driver）安装在客户机操作系统（Guest Operating System，GOS）内部，并诱导 GOS 释放或分配内存，气球驱动将相应的内存收回或分配给 GOS，从而实现自动伸缩内存调节。

基于页内容的内存页共享 CBPS 方法是扫描全局物理页，发现页内容相同的页就共享，从而释放冗余页，减少虚拟机的物理内存使用量。

内存页交换（Memory Page Swap）或内存页压缩方法（Memory Page Compression）选择虚拟机的若干页交换到磁盘等外部设备上，或无损压缩成 1/n 页大小，从而释放出内存给更多的虚拟机使用。

PoD 方法为每个虚拟机分配指定数量的内存页作为内存池（Memory Pool），刚启动的时候虚拟机的虚拟内存都是空的，没有对应到物理内存页，只有当虚拟内存真正被访问的时候才从内存池中拿出物理页。

发明人在实现本发明的过程中发现，现有技术至少存在以下不足：

以上技术都没有考虑虚拟机创建后并启动 GOS 这个过程当中内存使用情况，而是先将虚拟机需要的内存全部分配给虚拟机，然后再回收部分内存，这使得物理机器能并发启动虚拟机的数量受到物理内存量的限制，从而降低了虚拟机密度。

发明内容

本发明实施例提供一种创建虚拟机的方法、虚拟机监控器及虚拟机系统。

一方面，本发明实施例提供了一种创建虚拟机的方法，所述方法包括：将虚拟机的伪物理内存（Pseudo-physical Memory）对应的客户机页框号 GFN 映射到一共享零页，所述共享零页为物理内存中页面内容全为零的页；当所述虚拟机写所述 GFN 时，如果产生页面异常则分配物理内存页以解除所述客户机页框号 GFN 与所述共享零页的映射关系，并建立所述客户机页框号 GFN 与物理内存页的机器页框号 MFN 的映射关系。

另一方面，本发明实施例提供了一种虚拟机监控器，所述虚拟机监控器包括：初始化虚拟内存单元，用于将虚拟机的伪物理内存对应的客户机页框号 GFN 映射到一共享零页，所述共享零页为物理内存中页面内容全为零的页；写时拷贝单元，用于当所述虚拟机写所述 GFN 时，如果产生页面异常则分配物理内存页以解除所述客户机页框号 GFN 与所述共享零页的映射关系，并建立所述客户机页框号 GFN 与所述物理内存页的机器页框号 MFN 的映射关系。

又一方面，本发明实施例提供了一种虚拟机系统，所述系统包括：虚拟机监控器和虚拟机；其中，所述虚拟机监控器，用于将虚拟机的伪物理内存对应的客户机页框号 GFN 映射到一共享零页，所述共享零页为物理内存中页面内容全为零的页；当所述虚拟机写所述 GFN 时，如果产生页面异常则分配物理内存页以解除所述客户机页框号 GFN 与所述共享零页的映射关系，并建立所述客户机页框号 GFN 与物理内存页的机器页框号 MFN 的映射关系。

本发明实施例提供的上述技术方案，虚拟机监控器通过将虚拟机的伪物理内存映射到共享零页的方式初始化虚拟内存，使得创建虚拟机不需要消耗物理内存；当虚拟化开始使用虚拟内存时，虚拟机监控器通过写时拷贝分配物理内存给虚拟机并解除到共享零页的映射关系，从而可以减少虚拟机启动过程中的内存使用量，提高虚拟机密度。

附图说明

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

图 1 为现有技术的虚拟机架构示意图；

图 2 为本发明实施例一的系统及实施例二的装置的结构示意图；

图 3 是本发明实施例三的一种创建虚拟机的方法的流程图；

图 4 为本发明实施例四的一种创建虚拟机的方法的流程图；

图 5 为本发明实施例五的一种创建虚拟机的方法的流程图。

具体实施方式

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

实施例一：

图 2 为本发明实施例的虚拟机系统的体系架构示意图。如图 2 所示，该系统包括：虚拟机监控器和虚拟机；其中，

该虚拟机监控器，用于将虚拟机的伪物理内存对应的客户机页框号 GFN 映射到一共享零页，所述共享零页为物理内存中页面内容全为零的页；当所述虚拟机写所述 GFN 时，如果产生页面异常则分配物理内存页以解除所述客户机页框号 GFN 与所述共享零页的映射关系，并建立所述客户机页框号 GFN 与所述物理内存页的机器页框号 MFN 的映射关系。

进一步地，该虚拟机监控器，还可以用于在为所述虚拟机分配了物理内存页后，更新所述虚拟机的内存使用增量；当所述内存使用增量达到预设的阈值时，对虚拟机启动过程中分配的物理内存页进行扫描，将搜索到的内容为零的内存页释放或加入至内存池中，并将 GFN 重新映射到共享零页上，在扫描完成后将内存使用增量置零。

再请参阅图 2，以下进行更为详细的说明：

硬件层：作为虚拟化环境运行的整个硬件平台，具体可以包括处理器

CPU、内存 Memory、网卡（NIC，Network Interface Card）、外部存储设备等高速 I/O 设备和基本输入输出等低速设备。

宿主机 Host 或者虚拟机监控器 VMM：虚拟机监控器（VMM）作为一个管理层，其主要功能包括：完成硬件资源的管理、分配；为虚拟机呈现一个虚拟硬件平台；以及执行虚拟机的调度和隔离。可选地，有些 VMM 的实现里面需要一个特权虚拟机配合，两者结合组成宿主机。虚拟硬件平台对其上的虚拟机提供各种硬件资源，如虚拟处理器 VCPU、虚拟内存、虚拟磁盘、虚拟网卡等。其中虚拟内存对于 VM 而言，是一个隔离的、从零开始且具有连续性的伪物理内存空间，VMM 为每个 VM 建立一个 P2M 表以将 GFN 转换成 MFN（Machine Frame Number，机器页框号），从而使虚拟内存离散分布在物理内存中。

一个或多个虚拟机 VM：虚拟机运行宿主机为其准备的虚拟平台中。大多数时间虚拟机的执行不受宿主机的影响。

较佳地，GFN 到 MFN 的映射由 p2m 表记录。如图 2 中的 p2m 表所示，p2m 表由 p2m 表项组成，p2m 表项中记录了 mfn。把 gfn 作为索引，能找到唯一一个 p2m 表项，从而找到 mfn。假设共享零页的 mfn 等于 m1，那么在虚拟机创建阶段，host/vmm 用虚拟机伪物理内存对应的全部 GFN 去索引该虚拟机对应的 p2m 表，在每个 GFN 索引到的 p2m 项中 MFN 的位置填入 m1；假设虚拟机创建完成后，开始写内存，该内存所在的 GFN 为 g2，且产生页面异常，host/vmm 开始处理该异常，并分配一页物理内存，假设该物理内存的 MFN 等于 m2，host/vmm 将 m2 内存页清零后把 m2 填入 g2 索引 p2m 表得到的表项，从而解除 g2 到共享零页的映射；在零页扫描阶段，假设 host/vmm 发现虚拟机伪物理内存的 GFN 为 g3 对应的 MFN 内存页 m3 是零页，host/vmm 把 g3 索引 p2m 表得到的表项中填入 m1，使其重新映射到共享零页上，最后 host/vmm 回收 m3，使其成为空闲内存页。

可选地，该虚拟机监控器还可以用于在为所述虚拟机分配了物理内存页

后，更新所述虚拟机的内存使用增量；当内存使用增量达到预设的阈值时，对虚拟机启动过程中分配的物理内存页进行扫描，将搜索到的内容为零的内存页释放或加入至内存池中，并将 GFN 重新映射到共享零页上，在扫描完成后将内存使用增量置零。本发明实施例是实时扫描虚拟机启动过程中使用的内存页，而不是全局扫描，从而可以提高扫描效率低，降低扫描间隔。其中，释放表示 host/vmm 回收一页内存，使其成为空闲内存。需要说明的是，虚拟机的内存使用增量是一个计数值，用来表示距离上一次零页扫描以来，虚拟机由于写内存而解除原本映射到共享零页上的 GFN 的数量；如果这个计数值超过了阈值，那么就启动新的一次零页扫描，扫描结束后将计数值置零，即把内存使用增量置零，开始下一轮计数。也就是说内存使用增量就是零页扫描间隔内解除映射到共享零页上的 GFN 的数量。

本发明实施例的系统，可以减少虚拟机启动过程中的内存使用量，提高虚拟机密度，支持并发启动 Memory Overcommitted 数量的虚拟机，并使虚拟机启动过程当中内存使用量单调递增。

实施例二：

本发明实施例二提供了一种虚拟机监控器 VMM。请继续参阅图 2，该虚拟机监控器包括：

初始化虚拟内存单元 210，用于将虚拟机的伪物理内存对应的客户机页框号 GFN 映射到一共享零页，该共享零页为物理内存中页面内容全为零的页；

写时拷贝单元 220，用于当上述虚拟机写所述 GFN 时，如果产生页面异常则分配物理内存页以解除所述客户机页框号 GFN 与所述共享零页的映射关系，并建立所述客户机页框号 GFN 与物理内存页的机器页框号 MFN 的映射关系。

具体地，初始化虚拟内存单元 210，具体用于在虚拟机伪物理地址到机器物理地址的映射关系 P2M 表中，将所述虚拟机的伪物理内存对应的全部

GFN 都索引至所述共享零页的机器页框号 MFN。

具体地，写时拷贝单元 220，具体用于在虚拟机伪物理地址到机器物理地址的映射关系 P2M 表中，将所述客户机页框号 GFN 索引至所述物理内存页的机器页框号 MFN。

具体地，初始化虚拟内存单元 210，用于将 VM 的所有 GFN 都映射到 VMM 分配的共享零页上。该初始化虚拟内存单元 210 在创建虚拟机时被 VMM 调用。

由于 VM 的 GFN 到该共享零页的映射是只读映射，所以当 VM 写映射到共享零页的 GFN 时会产生异常，VMM 捕获该异常并调用写时拷贝 Copy on Write 单元。写时拷贝 Copy on Write 单元 220 具有以下功能：首先分配一块物理内存页并将这页清零；接着解除产生异常的 GFN 到共享零页的映射关系；最后将产生异常的 GFN 重新映射到上述分配的物理内存页对应的 MFN 上。写时拷贝单元，还用于判断内存池中是否包含零页，如是，则从所述内存池中获取零页并分配给虚拟机。

在一较佳实施例中，如图 2 所示，该虚拟机监控器还进一步包括：阈值控制单元 230 和零页扫描单元 240；该阈值控制单元 230，用于判断已分配的物理内存页的数量是否达到预设的阈值，如是，则启动该零页扫描单元；该零页扫描单元 240，用于对上述已分配的物理内存页进行扫描，并释放扫描到的的零页或者将扫描到的零页放入内存池中

具体地，阈值控制单元 230，与写时拷贝单元 220 连接，随着 VM 不断地写映射到共享零页的 GFN，VM 占用的物理内存页也在不断增加，因此阈值控制单元 230 的功能包括：判断 VM 增加的内存页数量是否超过预设的阈值，如果 VM 增加的内存页数量超过预设的阈值，则启动零页扫描单元，例如当 VM 使用的内存量超过 4096 页时启动零页扫描单元。

零页扫描单元 240，与阈值控制单元 230 连接，在 VM 启动阶段中 GOS 大量写内存操作是往内存页中写零，所以零页扫描单元 240 的功能包括：对

VM 触发写异常后调用写时拷贝单元为虚拟机分配的物理内存页进行扫描，以搜索到上述分配的物理内存页中内面内容全为零的页，将内容为全零的内存页对应的 GFN 重新映射共享零页上，并释放该 GFN 之前映射的物理内存页。

写时拷贝单元 220，还可进一步用于判断内存池中是否包含零页，如是，则从该内存池中获取零页并分配给虚拟机，如否，则从空闲内存中分配页给虚拟机。

本发明实施例的虚拟机监控器，可以减少虚拟机启动过程中的内存使用量，提高虚拟机密度，支持并发启动 Memory Overcommitted 数量的虚拟机，并使虚拟机启动过程当中内存使用量单调递增。如果虚拟机的内存使用量是单调递增的，那么可以确定某个时刻所有虚拟机使用的内存总量不会超过某个值，这样可以减少 Memory Overcommitted 失效。如果虚拟机的内存使用量与时间轴形成的是带有波峰波谷的曲线，那么无法确定到底最多能同时启动多少台虚拟机。如果虚拟机的内存使用量与时间轴形成的是与时间轴平行的直线，那么能够启动虚拟机的数量受到理内存大小的限制。

实施例三：

本发明实施例三提供了一种创建虚拟机的方法，图 3 是本发明实施例 3 的一种创建虚拟机的方法的整体流程图，如图 3 所示，该方法包括：

步骤 310、将虚拟机的伪物理内存对应的客户机页框号 GFN 映射到共享零页，所述共享零页为物理内存中页面内容全为零的页；

步骤 320、当所述虚拟机写所述 GFN 时，如果产生页面异常则分配物理内存页以解除所述客户机页框号 GFN 与所述共享零页的映射关系，并建立所述客户机页框号 GFN 与所述物理内存页的机器页框号 MFN 的映射关系。

具体地，步骤 310 中将所述虚拟机的伪物理内存对应的客户机页框号 GFN 都映射到一共享零页可以包括如下过程：在虚拟机伪物理地址到机器物理地址的映射关系 P2M 表中，将所述虚拟机的伪物理内存对应的全部 GFN

都索引至所述共享零页的机器页框号 MFN。

具体地，步骤 320 中建立所述客户机页框号 GFN 与所述物理内存页的机器页框号 MFN 的映射关系可以包括如下过程：在虚拟机伪物理地址到机器物理地址的映射关系 P2M 表中，将所述客户机页框号 MFN 索引至所述物理内存页的机器页框号 MFN。

较佳地，在步骤 320 中分配物理内存页后，图 3 所示方法还可以包括步骤：判断已分配的物理内存页的数量是否达到预设的阈值，如是，则对所述已分配的物理内存页进行扫描，并释放扫描到的零页或者将扫描到的零页放入内存池中。

较佳地，在步骤 320 中分配物理内存页的具体过程还可以包括：判断内存池中是否包含零页，如是，则从所述内存池中获取零页并分配给虚拟机，如否，则从空闲内存中分配页给虚拟机。

本发明实施例三的方法，可以减少虚拟机启动过程中的内存使用量，提高虚拟机密度，支持并发启动 Memory Overcommitted 数量的虚拟机，并使虚拟机启动过程当中内存使用量单调递增。

以下通过实施例四和实施例五，对实施例三的方法进行更为详细的说明。

实施例四：

虚拟机启动可以分为两个过程：第一，在虚拟机创建阶段，创建虚拟机并分配必要的资源，VMM 为了安全起见，分配给虚拟机的内存是经过清零的；第二，启动 Guest OS（客户操作系统，GOS）的阶段，GOS 对内存的使用包括两部分，即用于存放内核代码及数据的内存和空闲内存，GOS 为了安全起见会对空闲内存做清零工作。从前述分析可以看出，在现有 Memory Overcommitted 方法中，要么没有考虑虚拟机启动这种场景，要么没有考虑虚拟机启动过程中 Guest OS 对内存的使用。

在本发明实施例中，充分考虑了上述两种场景，首先，在创建虚拟机时

将客户机内存页全部共享映射到一张内容为零的机器物理页，即共享零页上，接着，在虚拟机写 GFN（Guest Frame Number，客户机页框号）时产生异常并用 CoW（Copy on Write，写时拷贝）机制解除共享，并记录虚拟机的内存使用增量，当内存使用增量超过预设的阈值时，启用零页扫描，搜索到内容为零的空闲页，将内容为零的页释放并将 GFN 重新映射到共享零页上，通过上述方法可减少虚拟机启动时内存使用量。

图 4 为本发明实施例四的方法流程图。结合参阅图 2 和图 4，该流程包括如下步骤：

步骤 10、创建 VM；

步骤 11、将所有 GFN 都映射到共享零页；零页是指 VM 的数据全为零的内存页；VM 中的零页是冗余的，在整个系统中，零页只需要一份就行，因此本发明实施例将 VM 中的原本是零页的 GFN 都共享映射到这一份零页上，这份给 VM 共享的零页叫共享零页。

步骤 12、启动客户操作系统 Guest OS；

具体地，在步骤 10-12 中，创建虚拟机并启动 Guest OS，在创建虚拟内存时并不给虚拟机分配任何物理内存资源，而是由初始化虚拟内存单元将 VM 的所有 GFN 都映射到 VMM 分配的共享零页上。尽管此时 VM 没有物理内存资源，但是 VM 通过 P2M 表仍然能获知全额并且内容都为零的虚拟内存，所以 Guest OS 仍能够启动。

步骤 20、VM 写 GFN；

步骤 21、判断是否产生页面异常 PageFault，如是，则转入步骤 22a，如否，则转入步骤 22b；

步骤 22a、CoW 单元申请物理内存页，更新 P2M 表，并使内存使用增量加 1；更新 P2M 表是指：解除所访问的 GFN 与共享零页的映射关系，并将所访问的 GFN 映射至申请的物理内存页；

步骤 22b、VM 访问已经申请的页；

具体地，在步骤 20、步骤 21 和步骤 22a 中，VM 开始使用虚拟内存，使用内存操作包括读内存和写内存。如果是写内存操作并且所访问的 GFN 是映射到共享零页上的，则处理器将产生页面异常 PageFault。VMM 开始处理这个异常，VMM 调用 CoW 单元解除所访问的 GFN 到共享零页的映射关系，CoW 单元再将所访问的 GFN 重新映射到一个新分配的且页内容为全零的 MFN 上，在这种情况下，VMM 还更新 VM 的内存使用增量计数，将这个计数值加 1。其中解除映射关系及重新映射是由 CoW 单元完成。在步骤 20、步骤 21 和步骤 22b 中，如果 VM 写的 GFN 不是映射到共享零页上，则不会产生 PageFault 异常，VM 正常使用内存。

步骤 30、判断 VM 的内存使用增量是否小于预设的阈值，如是，则转入步骤 20，如否，则转入步骤 40；

步骤 40、扫描零页，并将内存使用增量置零；

步骤 41、释放零页。

具体地，在 CoW 单元写成申请物理内存页、解除及重新建立映射关系之后，阈值控制单元，根据内存使用增量与预设的阈值之间的关系判断是否需要启动零页扫描单元。如果不需要启动零页扫描单元，则 VMM 从异常处理流程中退出，返回到 VM 中，VM 继续运行，如图 4 中步骤 30 和步骤 20 所示；

如果内存增量已经超过阈值，则启动零页扫描单元，阈值的大小可以由虚拟配置文件制定，也可以是 VMM 设定的默认值，例如 32MB。

在步骤 40 中，如果需要进行零页扫描，则零页扫描单元开始扫描已经分配给 VM 的物理内存，把内容为全零的页对应的 GFN 重新映射到共享零页上，并释放这些内容为零的物理页且将内存增量置为 0。扫描零页是指：找到 VM 的内存页中内容为全零的页。扫描 VM 的物理页时，可以扫描 VM 当前拥有的全部物理页，也可以只扫描增量内存。如果只扫描增量内存则需要步骤 22a 完成后记录相应的 MFN，也即 VM 新增的物理页对应的 MFN，

记录的数据结构可以是二进制位图（Bitmap）或是链表等数据结构。这一步完成后从异常处理流程中退出，返回到 VM 中，VM 继续运行。图 4 中的流程运行直到阈值控制单元单元收到启动完成通知消息，这个消息的发送者可以是 VMM 本身、特权域或者 VM 的前端驱动。

本发明实施例四的方法，可以减少虚拟机启动过程中的内存使用量，提高虚拟机密度，支持并发启动 Memory Overcommitted 数量的虚拟机，并使虚拟机启动过程当中内存使用量单调递增。

实施例五：

实施例四的流程中，零页扫描单元将零页重新映射到共享零页上后，直接释放物理内存页，而 CoW 单元每次都重新申请一块物理内存页，因此为提高性能，实施例五中将零页扫描单元发现的零页先放入一个内存池，CoW 单元申请内存页时先从该内存池中获取，如果内存已经为空再由 VMM 分配物理页。通过本发明实施例五，可以减少本发明实现流程中的反复先申请后释放的操作，从而提高效率。

本发明实施例五图 5 的具体流程与本发明实施例四图 4 类似，不同之处在于：

第一，图 4 中步骤 22a 在图 5 中对应于以下两种情况：

一种情况，如图 5 中步骤 120、步骤 121b 和步骤 123 所示，当内存池不为空的时候，VMM 在为 GFN 重新分配页时从内存池中取出一页，并把 GFN 映射到该页对应的 MFN 上，接着将内存使用增量加一。

另一种情况，当内存池为空时，执行图 5 中的步骤 120、步骤 121a、步骤 122 和步骤 123，完成与图 4 中 22a 一样的操作，即 VMM 重新分配一块物理内存页并清零，然后把 GFN 映射到该物理内存页对应的 MFN 上，最后将内存使用增量加一。

第二，图 4 中的步骤 41 变为图 5 中的步骤 141。当零页扫描单元发现内容为零的页后将该页的 GFN 映射到共享零页上，并把扫描到的零页放入内

存池中。

零页扫描单元扫描到的零页 GFN 重新映射到共享零页后，将 GFN 之前对应得 MFN 被放入了内存池，而 CoW 单元申请内存时又从内存池中取页，因此可以不要释放零页这个步骤，相应地把释放零页这个步骤换成了将零页加入内存池的步骤。从而有利于减少重复的申请内存池和释放内存。

根据图 5 流程可知，内存池的大小不会超过增量阈值，因此内存可以用数组或链表等线性表数据结构表示。这种线性表适用简单且效率高。

本发明实施例的有益效果：1、节省虚拟机启动过程中的内存使用量；2、支持并发启动 Memory Overcommitted 数量的虚拟机；这里强调的是并发数量，因为通过本发明实施例的方法能使 VM 尽量少的占用内存，并且 VM 的内存使用量是单调递增的，因此能够增加并发启动 VM 的数量 3、虚拟机启动过程当中内存使用量单调递增；4、提高了虚拟机密度；5、该方法可以应用于小型机虚拟化和聚合虚拟化等虚拟化领域。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本发明的范围。

结合本文中所公开的实施例描述的方法或算法的步骤可以用硬件、处理器执行的软件模块，或者二者的结合来实施。软件模块可以置于随机存储器（RAM）、内存、只读存储器（ROM）、电可编程 ROM、电可擦除可编程 ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

以上所述，仅为本发明较佳的具体实施方式，但本发明的保护范围并不

局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。因此本发明的保护范围应该以权利要求的保护范围为准。

权利要求书

1、一种创建虚拟机的方法，其特征在于，所述方法包括：

将虚拟机的伪物理内存对应的客户机页框号 GFN 映射到共享零页，所述共享零页为物理内存中页面内容全为零的页；

当所述虚拟机写所述 GFN 时，如果产生页面异常则分配物理内存页以解除所述客户机页框号 GFN 与所述共享零页的映射关系，并建立所述客户机页框号 GFN 与所述物理内存页的机器页框号 MFN 的映射关系。

2、根据权利要求 1 所述的方法，其特征在于，将所述虚拟机的伪物理内存对应的客户机页框号 GFN 映射到一共享零页包括：

在虚拟机伪物理地址到机器物理地址的映射关系 P2M 表中，将所述虚拟机的伪物理内存对应的全部 GFN 都索引至所述共享零页的机器页框号 MFN。

3、根据权利要求 1 所述的方法，其特征在于，建立所述客户机页框号 GFN 与所述物理内存页的机器页框号 MFN 的映射关系包括：

在虚拟机伪物理地址到机器物理地址的映射关系 P2M 表中，将所述客户机页框号 GFN 索引至所述物理内存页的机器页框号 MFN。

4、根据权利要求 1 所述的方法，其特征在于，在分配物理内存页后，所述方法还包括：

判断已分配的物理内存页的数量是否达到预设的阈值，如是，则对所述已分配的物理内存页进行扫描，并释放扫描到的零页或者将扫描到的零页放入内存池中。

5、根据权利要求 1 所述的方法，其特征在于，所述分配物理内存页包括：

判断内存池中是否包含零页，如是，则从所述内存池中获取零页并分配给虚拟机，如否，则从空闲内存中分配页给虚拟机。

6、一种虚拟机监控器，其特征在于，所述虚拟机监控器包括：

初始化虚拟内存单元，用于将虚拟机的伪物理内存对应的客户机页框号 GFN 映射到一共享零页，所述共享零页为物理内存中页面内容全为零的页；

写时拷贝单元，用于当所述虚拟机写所述 GFN 时，如果产生页面异常则分配物理内存页以解除所述客户机页框号 GFN 与所述共享零页的映射关系，并建立所述客户机页框号 GFN 与所述物理内存页的机器页框号 MFN 的映射关系。

7、根据权利要求 6 所述的虚拟机监控器，其特征在于，所述初始化虚拟内存单元，具体用于在虚拟机伪物理地址到机器物理地址的映射关系 P2M 表中，将所述虚拟机的伪物理内存对应的全部 GFN 都索引至所述共享零页的机器页框号 MFN。

8、根据权利要求 6 所述的虚拟机监控器，其特征在于，所述写时拷贝单元，具体用于在虚拟机伪物理地址到机器物理地址的映射关系 P2M 表中，将所述客户机页框号 GFN 索引至所述物理内存页的机器页框号 MFN。

9、根据权利要求 6 所述的虚拟机监控器，其特征在于，所述虚拟机监控器还包括：阈值控制单元和零页扫描单元；

所述阈值控制单元，用于判断已分配的物理内存页的数量是否达到预设的阈值，如是，则启动所述零页扫描单元；

所述零页扫描单元，用于对所述已分配的物理内存页进行扫描，并释放扫描到的零页或者将扫描到的零页放入内存池中。

10、根据权利要求 6 所述的虚拟机监控器，其特征在于，所述写时拷贝单元，还用于判断内存池中是否包含零页，如是，则从所述内存池中获取零页并分配给虚拟机，如否，则从空闲内存中分配页给虚拟机。

11、一种虚拟机系统，其特征在于，所述系统包括：虚拟机监控器和虚拟机；

所述虚拟机监控器，用于将虚拟机的伪物理内存对应的客户机页框号

GFN 映射到一共享零页，所述共享零页为物理内存中页面内容全为零的页；当所述虚拟机写所述 GFN 时，如果产生页面异常则分配物理内存页以解除所述客户机页框号 GFN 与所述共享零页的映射关系，并建立所述客户机页框号 GFN 与所述物理内存页的机器页框号 MFN 的映射关系。

12、根据权利要求 11 所述的虚拟机系统，其特征在于，所述虚拟机监控器，还用于在为所述虚拟机分配了物理内存页后，更新所述虚拟机的内存使用增量；当所述内存使用增量达到预设的阈值时，对虚拟机启动过程中分配的物理内存页进行扫描，将搜索到的内容为零的内存页释放或加入至内存池中，并将 GFN 重新映射到共享零页上，在扫描完成后将内存使用增量置零。

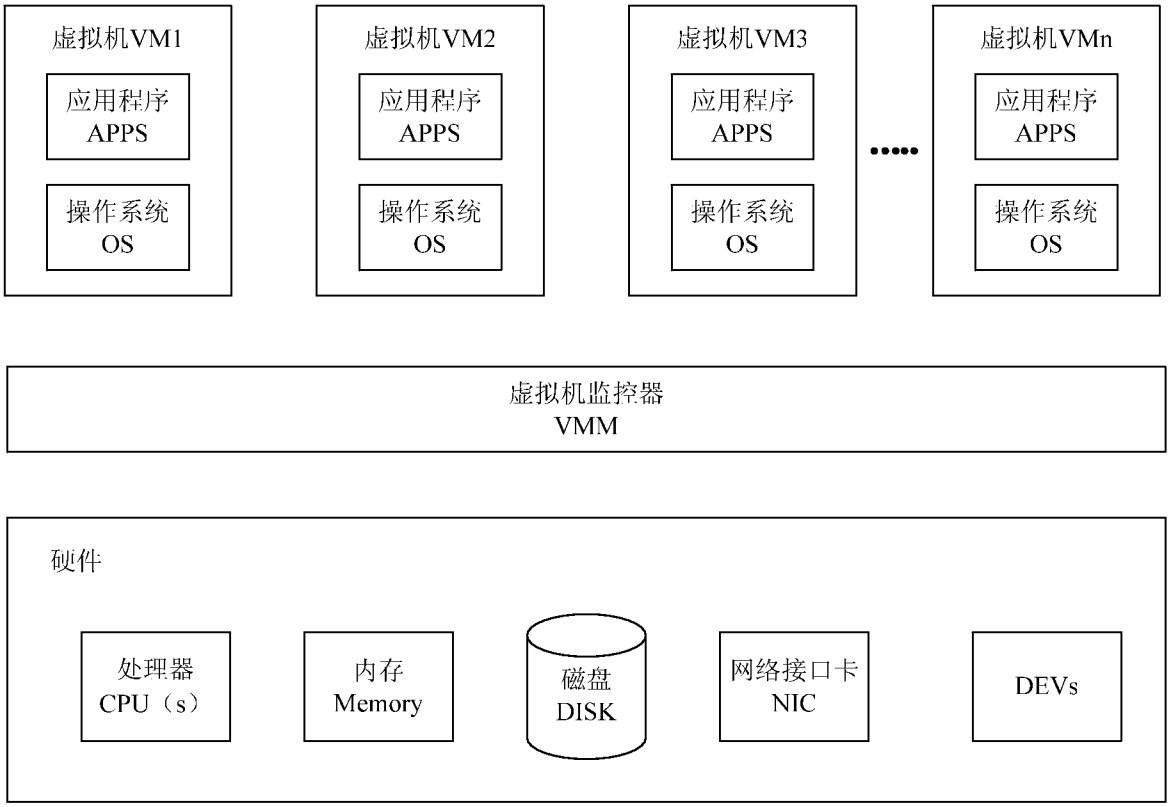
虚拟机VM1
应用程序
APPS
操作系统
OS
虚拟机VM2
应用程序
APPS
操作系统
OS
虚拟机VM3
应用程序
APPS
操作系统
OS
.....
虚拟机VMn
应用程序
APPS
操作系统
OS
虚拟机监控器
VMM
硬件
处理器
CPU（s）
内存
Memory
磁盘
DISK
网络接口卡
NIC
DEVs

图 1

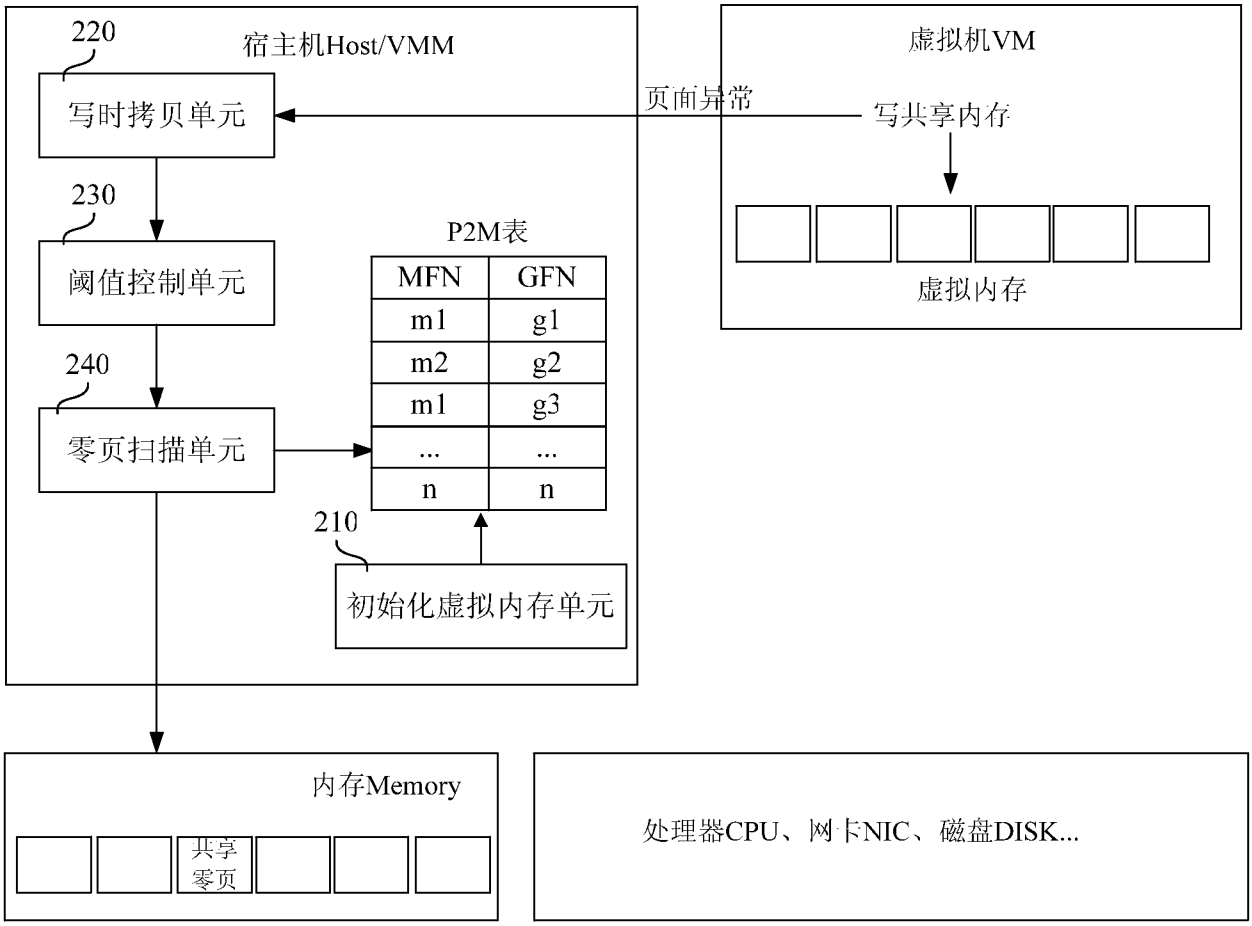
220
宿主机Host/VMM
写时拷贝单元
页面异常
虚拟机VM
写共享内存
230
P2M表
阈值控制单元
MFN
GFN
m1
g1
m2
g2
m1
g3
...
...
n
n
虚拟内存
240
零页扫描单元
210
初始化虚拟内存单元
内存Memory
共享零页
处理器CPU、网卡NIC、磁盘DISK...

图 2

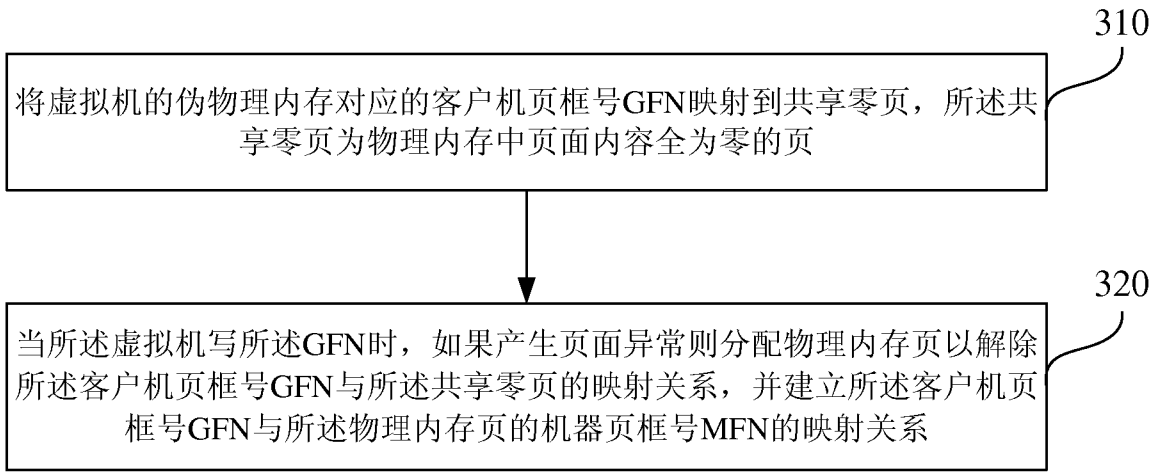
310
将虚拟机的伪物理内存对应的客户机页框号GFN映射到共享零页，所述共享零页为物理内存中页面内容全为零的页
320
当所述虚拟机写所述GFN时，如果产生页面异常则分配物理内存页以解除所述客户机页框号GFN与所述共享零页的映射关系，并建立所述客户机页框号GFN与所述物理内存页的机器页框号MFN的映射关系

图 3

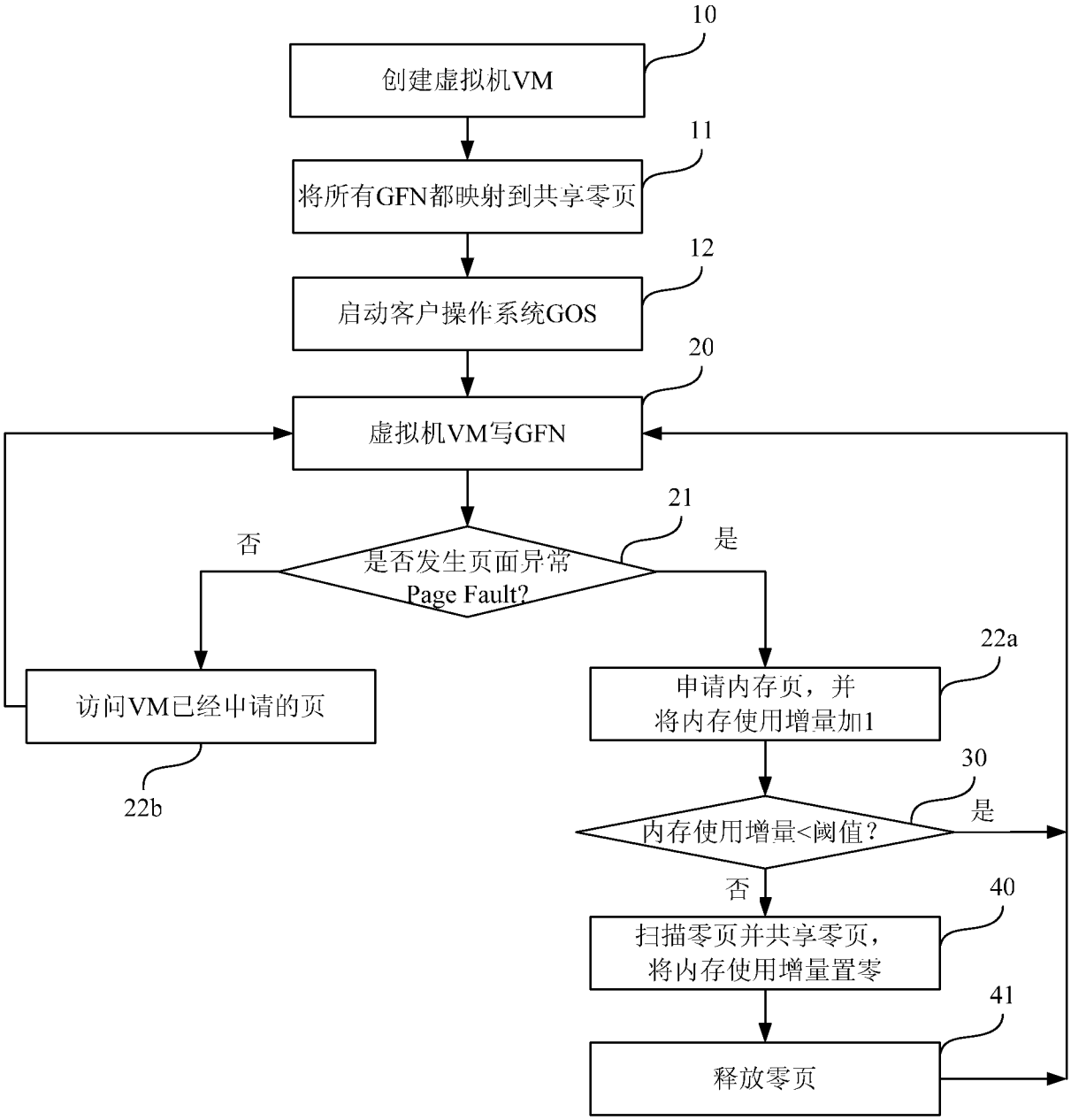
10
创建虚拟机VM
11
将所有GFN都映射到共享零页
12
启动客户操作系统GOS
20
虚拟机VM写GFN
21
是否发生页面异常
Page Fault?
否
是
访问VM已经申请的页
22b
22a
申请内存页，并
将内存使用增量加1
30
内存使用增量<阈值？
是
否
40
扫描零页并共享零页，
将内存使用增量置零
41
释放零页

图4

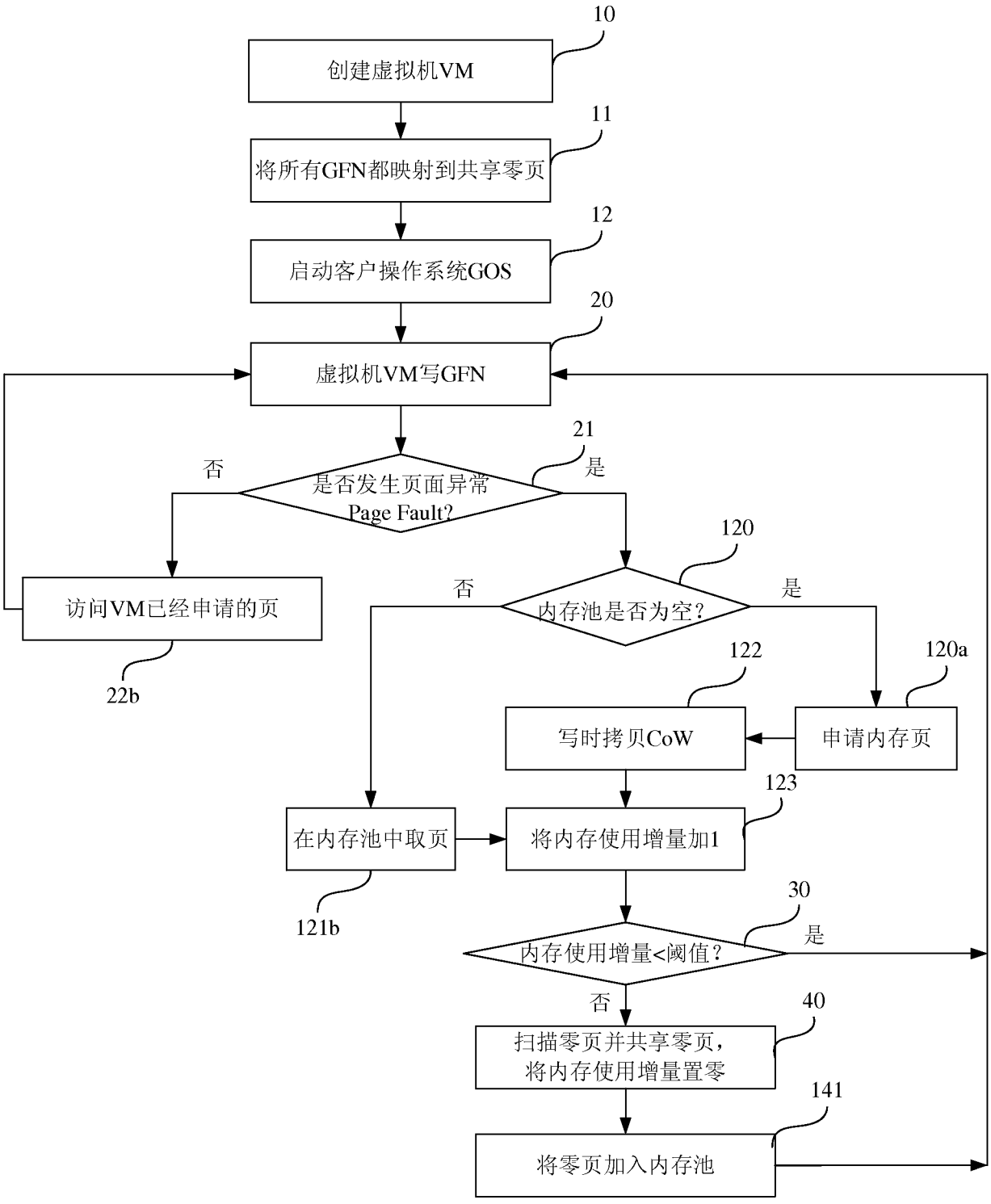
10
创建虚拟机VM
11
将所有GFN都映射到共享零页
12
启动客户操作系统GOS
20
虚拟机VM写GFN
21
是否发生页面异常
Page Fault?
否
是
访问VM已经申请的页
22b
120
内存池是否为空?
否
是
122
120a
写时拷贝CoW
申请内存页
123
在内存池中取页
将内存使用增量加1
121b
30
内存使用增量<阈值?
是
否
40
扫描零页并共享零页,
将内存使用增量置零
141
将零页加入内存池

图5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2011/080573

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)

IPC: 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)

Database: CNTXT, CNKI, CNABS, DWPI, SIPOABS

Searched words: virtual, machine, memory, allocate, mapping, page

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P, X	CN102141931A (HUAWEI TECHNOLOGIES CO., LTD.) 03 Aug. 2011 (03.08.2011) See claims 1-12	1-12
X	CN101158924A (UNIV PEKING) 09 Apr. 2008 (09.04.2008) See the description, line 3 – line 28 page 3, line 20 page 9 – line 23 page 11	1-12
A	CN101697134A (UNIV PEKING) 21 Apr. 2010 (21.04.2010) See the whole document	1-12
A	US2005235123A1 (INTEL CORP) 20 Oct. 2005 (20.10.2005) See the whole document	1-12

☐ 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
12 Nov. 2011 (12.11.2011)	29 Dec. 2011 (29.12.2011)
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10)62019451	Authorized officer YIN, Jianfeng Telephone No. (86-10)62411647

Form PCT/ISA /210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2011/080573

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN102141931A	03.08.2011	NONE	
CN101158924A	09.04.2008	CN100527098C	12.08.2009
CN101697134A	21.04.2010	NONE	
US2005235123A1	20.10.2005	US7421533B2	02.09.2008

国际检索报告

国际申请号
PCT/CN2011/080573

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)

IPC：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）)

数据库：CNTXT，CNKI，CNABS，DWPI，SIPOABS

检索词：虚拟机，内存，分配，映射，页，共享，virtual machine, memory, allocate, mapping, page

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
P，X	CN102141931A（华为技术有限公司）03. 8 月 2011（03.08.2011） 参见权利要求 1-12	1-12
X	CN101158924A（北京大学）09. 4 月 2008（09.04.2008） 参见说明书第 3 页第 3 行-第 28 行，第 9 页第 20 行-第 11 页第 23 行	1-12
A	CN101697134A（北京大学）21. 4 月 2010（21.04.2010） 参见全文	1-12
A	US2005235123A1（英特尔公司）20. 10 月 2005（20.10.2005） 参见全文	1-12

☐ 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
12. 11 月 2011（12.11.2011）	**29.12 月 2011 (29.12.2011)**
ISA/CN 的名称和邮寄地址： 中华人民共和国国家知识产权局 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号：　(86-10)62019451	受权官员 尹剑峰 电话号码：(86-10) **62411647**

国际检索报告

关于同族专利的信息

国际申请号
PCT/CN2011/080573

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN102141931A	03.08.2011	无	
CN101158924A	09.04.2008	CN100527098C	12.08.2009
CN101697134A	21.04.2010	无	
US2005235123A1	20.10.2005	US7421533B2	02.09.2008