(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115390992 A
(43)申请公布日 2022.11.25

(21)申请号 202211104740.0

(22)申请日 2022.09.09

(71)申请人 深圳威科软件科技有限公司
地址 518000 广东省深圳市南山区粤海街道滨海社区海天二路14号软件产业基地5D座7022

(72)发明人 雷浪声 王超

(74)专利代理机构 北京品源专利代理有限公司 11332
专利代理师 潘登

(51)Int.Cl.
G06F 9/455(2006.01)
G06F 9/50(2006.01)

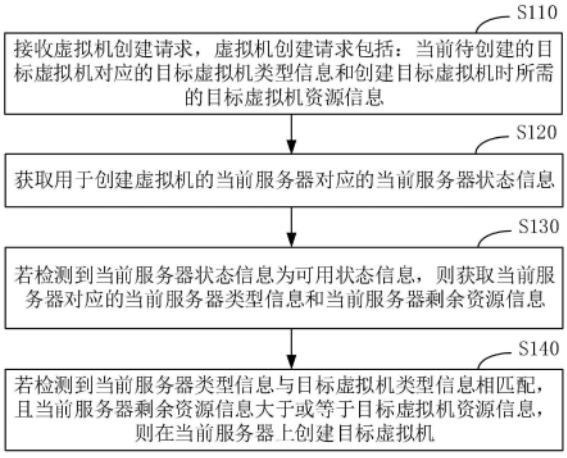
权利要求书2页 说明书11页 附图4页

(54)发明名称

一种虚拟机创建方法、装置、设备和存储介质

(57)摘要

本发明公开了一种虚拟机创建方法、装置、设备和存储介质。该方法包括:接收虚拟机创建请求,虚拟机创建请求包括:当前待创建的目标虚拟机对应的目标虚拟机类型信息和创建目标虚拟机时所需的目标虚拟机资源信息;获取用于创建虚拟机的当前服务器对应的当前服务器状态信息;若检测到当前服务器状态信息为可用状态信息,则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息;若检测到当前服务器类型信息与目标虚拟机类型信息相匹配,且当前服务器剩余资源信息大于或等于目标虚拟机资源信息,则在当前服务器上创建目标虚拟机。本发明技术方案可以保证虚拟机的成功创建,并且避免出现服务器卡顿或者奔溃的情况,提升了用户体验。



CN 115390992 A

1.一种虚拟机创建方法,其特征在于,包括:

接收虚拟机创建请求,所述虚拟机创建请求包括:当前待创建的目标虚拟机对应的目标虚拟机类型信息和创建所述目标虚拟机时所需的目标虚拟机资源信息;

获取用于创建虚拟机的当前服务器对应的当前服务器状态信息;

若检测到所述当前服务器状态信息为可用状态信息,则获取所述当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息;

若检测到所述当前服务器类型信息与所述目标虚拟机类型信息相匹配,且所述当前服务器剩余资源信息大于或等于所述目标虚拟机资源信息,则在所述当前服务器上创建所述目标虚拟机。

2.根据权利要求1所述的方法,其特征在于,在获取用于创建虚拟机的当前服务器对应的当前服务器状态信息之前,还包括:

将服务器集群中的第一个服务器确定为用于创建虚拟机的当前服务器;

所述方法还包括:

若检测到所述当前服务器状态信息为非可用状态信息、所述当前服务器类型信息与所述目标虚拟机类型信息不匹配,或者所述当前服务器剩余资源信息小于所述目标虚拟机资源信息,则基于所述服务器集群对应的服务器排列顺序,将当前服务器的下一服务器作为当前服务器,返回执行所述获取用于创建虚拟机的当前服务器对应的当前服务器状态信息的操作。

3.根据权利要求2所述的方法,其特征在于,所述方法还包括:

若所述服务器集群中的所有服务器均不满足创建所述目标虚拟机的条件,则间隔预设时长后,重新对所述服务器集群中的服务器进行遍历匹配,直到在满足条件的目标服务器中创建出所述目标虚拟机为止。

4.根据权利要求1所述的方法,其特征在于,在所述当前服务器上创建所述目标虚拟机,包括:

获取所述当前服务器对应的当前总创建虚拟机数量和当前待创建虚拟机数量;

若检测到所述当前总创建虚拟机数量小于第一数量阈值,且所述当前待创建虚拟机数量小于第二数量阈值,则在所述当前服务器上创建所述目标虚拟机。

5.根据权利要求1所述的方法,其特征在于,若检测到所述当前服务器状态信息为可用状态信息,则获取所述当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息,包括:

若检测到所述当前服务器状态信息是除了服务器重启状态、软件安装状态和软件升级状态之外的可分配资源的服务器状态,则获取所述当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息。

6.根据权利要求1所述的方法,其特征在于,所述虚拟机创建请求还包括:所述目标虚拟机对应的目标镜像文件标识信息;

在所述当前服务器上创建所述目标虚拟机,包括:

获取所述目标镜像文件标识信息对应的目标镜像文件;

通过运行所述目标镜像文件,在所述当前服务器上创建出所述目标虚拟机。

7.根据权利要求1-6任一项所述的方法,其特征在于,所述目标虚拟机资源信息包括:

所述目标虚拟机对应的CPU核心数、运行内存大小和显卡内存大小；

所述当前服务器剩余资源信息包括：所述当前服务器对应的剩余CPU核心数、剩余运行内存大小和剩余显卡内存大小。

8.一种虚拟机创建装置，其特征在于，包括：

请求接收模块，用于接收虚拟机创建请求，所述虚拟机创建请求包括：当前待创建的目标虚拟机对应的目标虚拟机类型信息和创建所述目标虚拟机时所需的目标虚拟机资源信息；

服务器状态获取模块，用于获取用于创建虚拟机的当前服务器对应的当前服务器状态信息；

服务器信息获取模块，用于若检测到所述当前服务器状态信息为可用状态信息，则获取所述当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息；

虚拟机创建模块，用于若检测到所述当前服务器类型信息与所述目标虚拟机类型信息相匹配，且所述当前服务器剩余资源信息大于或等于所述目标虚拟机资源信息，则在所述当前服务器上创建所述目标虚拟机。

9.一种电子设备，其特征在于，所述电子设备包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的计算机程序，所述计算机程序被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求1-7中任一项所述的虚拟机创建方法。

10.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有计算机指令，所述计算机指令用于使处理器执行时实现权利要求1-7中任一项所述的虚拟机创建方法。

一种虚拟机创建方法、装置、设备和存储介质

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种虚拟机创建方法、装置、设备和存储介质。

背景技术

[0002] 随着计算机技术的发展，虚拟机的使用越来越普遍。虚拟机通常是创建在服务器中并使用。目前，普遍采用模拟处理器(Quick Emulator，QEMU)或多计算机切换器(Keyboard Video Mouse，KVM)创建虚拟机。该创建方式是直接在服务器上创建虚拟机。

[0003] 然而，该种直接创建虚拟机的方式可能会存在创建失败的情况，并且即使虚拟机在服务器中成功创建，在运行虚拟机时也会出现服务器卡顿甚至崩溃的情况。一旦服务器崩溃，服务器上所有的虚拟机都会消失，需要重新创建，费时费力，同时也大大降低了用户体验。

发明内容

[0004] 本发明提供了一种虚拟机创建方法，以保证虚拟机的成功创建，并且避免出现服务器卡顿或者奔溃的情况，从而提升了用户体验。

[0005] 根据本发明的一方面，提供了一种虚拟机创建方法，该方法包括：

[0006] 接收虚拟机创建请求，虚拟机创建请求包括：当前待创建的目标虚拟机对应的目标虚拟机类型信息和创建目标虚拟机时所需的目标虚拟机资源信息；

[0007] 获取用于创建虚拟机的当前服务器对应的当前服务器状态信息；

[0008] 若检测到当前服务器状态信息为可用状态信息，则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息；

[0009] 若检测到当前服务器类型信息与目标虚拟机类型信息相匹配，且当前服务器剩余资源信息大于或等于目标虚拟机资源信息，则在当前服务器上创建目标虚拟机。

[0010] 根据本发明的另一方面，提供了一种虚拟机创建装置，该装置包括：

[0011] 请求接收模块，用于接收虚拟机创建请求，虚拟机创建请求包括：当前待创建的目标虚拟机对应的目标虚拟机类型信息和创建目标虚拟机时所需的目标虚拟机资源信息；

[0012] 服务器状态获取模块，用于获取用于创建虚拟机的当前服务器对应的当前服务器状态信息；

[0013] 服务器信息获取模块，用于若检测到当前服务器状态信息为可用状态信息，则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息；

[0014] 虚拟机创建模块，用于若检测到当前服务器类型信息与目标虚拟机类型信息相匹配，且当前服务器剩余资源信息大于或等于目标虚拟机资源信息，则在当前服务器上创建目标虚拟机。

[0015] 根据本发明的另一方面，提供了一种电子设备，该电子设备包括：

[0016] 至少一个处理器；以及

[0017] 与所述至少一个处理器通信连接的存储器;其中,

[0018] 所述存储器存储有可被所述至少一个处理器执行的计算机程序,所述计算机程序被所述至少一个处理器执行,以使所述至少一个处理器能够执行本发明任意实施例所述的虚拟机创建方法。

[0019] 根据本发明的另一方面,提供了一种计算机可读存储介质,所述计算机可读存储介质存储有计算机指令,所述计算机指令用于使处理器执行时实现本发明任意实施例所述的虚拟机创建方法。

[0020] 本发明实施例的技术方案,通过接收虚拟机创建请求,虚拟机创建请求包括:当前待创建的目标虚拟机对应的目标虚拟机类型信息和创建目标虚拟机时所需的目标虚拟机资源信息;获取用于创建虚拟机的当前服务器对应的当前服务器状态信息;若检测到当前服务器状态信息为可用状态信息,则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息;若检测到当前服务器类型信息与目标虚拟机类型信息相匹配,且当前服务器剩余资源信息大于或等于目标虚拟机资源信息,则在当前服务器上创建目标虚拟机,以实现基于服务器的实际资源情况在服务器上合理创建虚拟机,从而保证虚拟机的成功创建,并且避免出现服务器卡顿或者奔溃的情况,从而提升了用户体验。

[0021] 应当理解,本部分所描述的内容并非旨在标识本发明的实施例的关键或重要特征,也不用于限制本发明的范围。本发明的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0022] 为了更清楚地说明本发明实施例中的技术方案,下面将对实施例描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0023] 图1是根据本发明实施例一提供的一种虚拟机创建方法的流程图;

[0024] 图2是根据本发明实施例二提供的另一种虚拟机创建方法的流程图;

[0025] 图3是根据本发明实施例三提供的又一种虚拟机创建方法的流程图;

[0026] 图4是根据本发明实施例四提供的一种虚拟机创建装置的结构示意图;

[0027] 图5是实现本发明实施例的虚拟机创建方法的电子设备的结构示意图。

具体实施方式

[0028] 为了使本技术领域的人员更好地理解本发明方案,下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分的实施例,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,都应当属于本发明保护的范围。

[0029] 需要说明的是,本发明的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换,以便这里描述的本发明的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外,术语“包括”和“具有”以及他们的任何变形,意图在于覆

盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0030] 实施例一

[0031] 图1为本发明实施例一提供了一种虚拟机创建方法的流程图，本实施例可适用于在服务器中创建虚拟机的情况，该方法可以由虚拟机创建装置来执行，该虚拟机创建装置可以采用硬件和/或软件的形式实现，集成于电子设备中。如图1所示，该方法包括：

[0032] S110、接收虚拟机创建请求，虚拟机创建请求包括：当前待创建的目标虚拟机对应的目标虚拟机类型信息和创建目标虚拟机时所需的目标虚拟机资源信息。

[0033] 其中，虚拟机可以是指通过软件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系统。例如，虚拟机可以是VM、Virtual Box或Virtual PC等。目标虚拟机可以是指待创建的虚拟机。目标虚拟机的数量可以是一个或多个。虚拟机类型信息可以包括目标虚拟机对应的中央处理器(central processing unit,CPU)或图形处理器(graphics processing unit,GPU)类型。示例性地，目标虚拟机资源信息可以包括：目标虚拟机对应的CPU核心数、运行内存大小和显卡内存大小。

[0034] 具体地，在用户想要创建某个虚拟机时，可以在用户终端上触发虚拟机创建操作，并基于用户触发的虚拟机创建操作生成虚拟机创建请求，将虚拟机创建请求发送至用于对虚拟机进行创建控制的控制器，使得控制器接收到该虚拟机创建请求，并基于接收到的虚拟机创建请求确定当前待创建的目标虚拟机对应的目标虚拟机类型信息(比如CPU类型或是CPU+GPU类型)和创建目标虚拟机时所需的目标虚拟机资源信息(比如4核CPU、8G运行内存和2G显示内存)。

[0035] 例如，若软件A的教学老师需要申请200台具有软件A的虚拟机进行教学，则可以在用户终端上输入需要使用的软件A、虚拟机数量、虚拟机类型、虚拟机内存、虚拟机运存等信息，以使用户终端基于上述信息生成虚拟机创建请求，并将该虚拟机创建请求发送至控制器。

[0036] 需要说明的是，若目标虚拟机存在多个时，可以基于每个目标虚拟机对应的目标虚拟机类型信息和创建目标虚拟机时所需的目标虚拟机资源信息进行逐个创建。

[0037] S120、获取用于创建虚拟机的当前服务器对应的当前服务器状态信息。

[0038] 其中，服务器状态信息可以是指服务器是否处于可用状态的信息，可分为可用状态信息和非可用状态信息。例如，当服务器在重启阶段、安装其他软件阶段以及升级其他服务阶段等服务器不能执行创建虚拟机任务的阶段时，均为非可用状态即不可分配机器状态。当前服务器状态信息可以是指当前服务器的状态信息。

[0039] 需要说明的是，针对只有一个服务器存在的情况，控制器可以将该服务器作为当前服务器，进行当前服务器状态信息的获取。当存在多个服务器即服务器集群时，可以基于服务器集群对应的服务器创建顺序，将排序第一的服务器作为当前服务器，并获取该服务器的当前服务器状态信息。若当前服务器的状态不适合创建虚拟机，则将当前服务器的下一服务器作为当前服务器，并再次获取当前服务器对应的当前服务器状态信息。

[0040] 具体地，控制器可以获取当前服务器对应的当前服务器状态信息，并基于当前服务器状态信息确定用于创建虚拟机的当前服务器是否处于可创建虚拟机的状态。

[0041] 示例性地，在S120之前还可以包括：将服务器集群中的第一个服务器确定为用于创建虚拟机的当前服务器。

[0042] 其中，服务器可以是指一种管理计算资源的服务器。例如，服务器可以是但不限于计算机。服务器集群可以是由多个服务器组成的集群。服务器集群中的服务器按一定顺序排列。当前服务器可以是指在服务器集群中排在第一个的服务器，用于创建虚拟机。

[0043] 具体地，若存在多个服务器组成的服务器集群，则可以将服务器中排在第一个的服务器确定为用于创建虚拟机的当前服务器。

[0044] S130、若检测到当前服务器状态信息为可用状态信息，则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息。

[0045] 其中，当前服务器类型信息可以包括当前服务器对应的CPU或GPU类型。示例性地，当前服务器剩余资源信息可以包括当前服务器对应的剩余CPU核心数、剩余运行内存大小和剩余显卡内存大小。剩余CPU核心数可以是指服务器可创建的核心数。例如，若服务器使用的是8核CPU，则剩余CPU核心数可以是4核和8核。

[0046] 具体地，若检测到当前服务器状态信息为可用状态信息，表明当前服务器的当前状态为可创建虚拟机的状态，则可以获取当前服务器对应的当前服务器类型信息(比如CPU类型)和当前服务器剩余资源信息(比如4核CPU、剩余100G运行内存)。

[0047] S140、若检测到当前服务器类型信息与目标虚拟机类型信息相匹配，且当前服务器剩余资源信息大于或等于目标虚拟机资源信息，则在当前服务器上创建目标虚拟机。

[0048] 具体地，若检测到当前服务器类型信息与目标虚拟机类型信息相匹配，比如当前服务器的处理器为CPU+GPU，创建目标虚拟机的处理器也是CPU+GPU，那么两者进行类型信息匹配可以匹配成功，并将前服务器剩余资源信息中的每一项信息都与目标虚拟机资源信息中对应的信息进行匹配，若当前服务器剩余资源信息中的每一项信息都大于或等于目标虚拟机资源信息对应的信息，则资源信息匹配成功，以使类型信息与资源信息全部匹配成功的当前服务器进行目标虚拟机的创建，从而准确的分配计算服务器的可分配资源，避免超过可分配资源进行虚拟机创建而造成服务器崩溃，如创建虚拟机数量超过服务器限制而造成服务器崩溃的情况，保证服务器中虚拟机的正常运行，进而提升了用户体验。

[0049] 例如，基于上述对服务器的检测操作，可以逐一确定出软件A的教学老师需要申请200台具有软件A的虚拟机进行创建的对应服务器，并进行虚拟机的创建，避免在教学老师上课时，200台虚拟机运行出现卡顿甚至崩溃的情况，能保证老师在上课期间虚拟机的正常使用。

[0050] 示例性地，若检测到当前服务器状态信息为非可用状态信息、当前服务器类型信息与目标虚拟机类型信息不匹配，或者当前服务器剩余资源信息小于目标虚拟机资源信息，则基于服务器集群对应的服务器排列顺序，将当前服务器的下一服务器作为当前服务器，返回执行获取用于创建虚拟机的当前服务器对应的当前服务器状态信息的操作。

[0051] 具体地，若检测到当前服务器状态信息为非可用状态信息、当前服务器类型信息与目标虚拟机类型信息不匹配和当前服务器剩余资源信息小于目标虚拟机资源信息中的至少一项，则表明该当前服务器不能创建目标虚拟机，并基于服务器集群对应的服务器排列顺序，将当前服务器的下一服务器作为当前服务器，返回执行S120操作，直至检测到符合目标虚拟机的创建条件的服务器为止或是遍历服务器集群中所有服务器后中止检测。

[0052] 示例性地，若服务器集群中的所有服务器均不满足创建目标虚拟机的条件，则间隔预设时长后，重新对服务器集群中的服务器进行遍历匹配，直到在满足条件的目标服务器中创建出目标虚拟机为止。

[0053] 其中，间隔预设时长可以是指预先设置的一段间隔时长，用于在满足间隔预设时长后重启遍历服务器集群中所有服务器仍未进行创建的虚拟机的遍历匹配任务。

[0054] 具体地，若服务器集群中的所有服务器均不满足创建目标虚拟机的条件，则停止对该目标虚拟器的创建，并继续进行其他目标虚拟机的检测及创建。在该目标虚拟机的创建停止时长达到间隔预设时长后，可以基于该目标虚拟机重新对服务器集群中的服务器进行遍历匹配，直到在满足条件的目标服务器中创建出目标虚拟机为止。

[0055] 本发明实施例的技术方案，通过接收虚拟机创建请求，虚拟机创建请求包括：当前待创建的目标虚拟机对应的目标虚拟机类型信息和创建目标虚拟机时所需的目标虚拟机资源信息；获取用于创建虚拟机的当前服务器对应的当前服务器状态信息；若检测到当前服务器状态信息为可用状态信息，则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息；若检测到当前服务器类型信息与目标虚拟机类型信息相匹配，且当前服务器剩余资源信息大于或等于目标虚拟机资源信息，则在当前服务器上创建目标虚拟机，以实现基于服务器的实际资源情况在服务器上合理创建虚拟机，从而保证虚拟机的成功创建，并且避免出现服务器卡顿或者奔溃的情况，从而提升了用户体验。

[0056] 实施例二

[0057] 图2为本发明实施例二提供的一种虚拟机创建方法的流程图，本实施例在上述实施例的基础上，对在当前服务器上创建目标虚拟机的过程进行了详细描述。其中与上述各实施例相同或相应的术语的解释在此不再赘述。如图2所示，该方法包括：

[0058] S210、接收虚拟机创建请求，虚拟机创建请求包括：当前待创建的目标虚拟机对应的目标虚拟机类型信息和创建目标虚拟机时所需的目标虚拟机资源信息。

[0059] S220、获取用于创建虚拟机的当前服务器对应的当前服务器状态信息。

[0060] S230、若检测到当前服务器状态信息为可用状态信息，则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息。

[0061] 示例性地，S230可以包括：若检测到当前服务器状态信息是除了服务器重启状态、软件安装状态和软件升级状态之外的可分配资源的服务器状态，则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息。

[0062] 其中，非可用状态可以包括服务器重启状态、软件安装状态和软件升级状态。

[0063] 具体地，经主服务器检测，若检测到当前服务器状态信息是除了服务器重启状态、软件安装状态和软件升级状态之外的可分配资源的服务器状态，即检测到存在当前服务器状态为可用状态的当前服务器，则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息。

[0064] S240、若检测到当前服务器类型信息与目标虚拟机类型信息相匹配，且当前服务器剩余资源信息大于或等于目标虚拟机资源信息，则获取当前服务器对应的当前总创建虚拟机数量和当前待创建虚拟机数量。

[0065] 其中，当前总创建虚拟机数量可以包括当前已创建虚拟机数量、当前正在创建虚拟机数量和当前待创建虚拟机数量。

[0066] S250、若检测到当前总创建虚拟机数量小于第一数量阈值，且当前待创建虚拟机数量小于第二数量阈值，则在当前服务器上创建目标虚拟机。

[0067] 其中，第一数量阈值可以是指当前服务器运行不卡顿、不崩溃的情况下当前服务器允许创建虚拟机的最大数量。第二数量阈值可以是指当前服务器预设创建虚拟机队列能容纳的最大的数量。

[0068] 具体地，针对只有一个服务器存在的情况，若检测到所述目标创建虚拟机数量大于或等于第一数量阈值，或者所述当前待创建虚拟机数量大于或等于第二数量阈值，即满足其中的任意一条，表明当前无可创建虚拟机的服务器，则可以直接退出创建。当存在多个服务器即服务器集群时，可以基于服务器集群对应的服务器创建顺序，将当前服务器的下一服务器作为当前服务器，针对每个当前服务器而言，若检测到所述目标创建虚拟机数量大于或等于第一数量阈值，或者所述当前待创建虚拟机数量大于或等于第二数量阈值，即满足其中的任意一条，表明当前服务器不满足创建虚拟机的条件，则可以返回执行S220，并进行上述条件检测，直至匹配到可创建目标虚拟机的当前服务器时，或是直到所有服务器均匹配失败时，循环结束。

[0069] 需要说明的是，服务器无法做到一直创建虚拟机和运行已创建的虚拟机。若创建虚拟机和运行虚拟机的数量超过服务器的承受范围，则会出现服务器运行卡顿或崩溃的情况，因此对服务器的当前总创建虚拟机数量进行数量限制。例如，一个浏览器在开启过多页面时，会出现运行卡顿或崩溃的情况。同时，当同一时间创建过多的虚拟机时，会导致服务器CPU持续飙高形成卡顿，影响其他已创建完备的虚拟机。

[0070] 在上述技术方案的基础上，S210中“虚拟机创建请求”还包括：目标虚拟机对应的目标镜像文件标识信息；S240中“在当前服务器上创建目标虚拟机”可以包括：获取目标镜像文件标识信息对应的目标镜像文件；通过运行目标镜像文件，在当前服务器上创建出目标虚拟机。

[0071] 其中，虚拟机创建请求可以包括：当前待创建的目标虚拟机对应的目标虚拟机类型信息、创建目标虚拟机时所需的目标虚拟机资源信息以及目标虚拟机对应的目标镜像文件标识信息。镜像可以是指一种文件存储形式。镜像文件可以是指一个磁盘上的数据在另一个磁盘上存在一个完全相同的副本文件。目标镜像文件可以是指包含目标虚拟机中需要使用的软件和文件的镜像文件，用于生成目标虚拟机。镜像文件标识信息可以是指镜像文件的唯一标识信息，用于区分各个镜像文件。

[0072] 具体地，控制器可以基于虚拟机创建请求中目标虚拟机对应的目标镜像文件标识信息，获取目标镜像文件标识信息对应的目标镜像文件，并在确定创建目标虚拟机的当前服务器中运行目标镜像文件，从而在当前服务器上创建出目标虚拟机。需要说明的是，目标镜像文件在运行创建出目标虚拟机后，不会消失，可以多次运行创建对应的虚拟机，避免了镜像文件的多次生成，提高了虚拟机创建的效率，同时创建虚拟机只需要预先生成一个对应的镜像文件即可，节约了存储资源。

[0073] 需要说明的是，虚拟机是通过镜像文件创建。若需要关闭或消除基于该镜像文件创建的所有虚拟机，则只需要自动查询带有该镜像文件标识信息的虚拟机，并批量进行操作，就可以关闭或消除基于该镜像文件创建的所有虚拟机，以实现通过同一镜像创建虚拟机的批量操作，节约了时间成本和服务器存储资源。

[0074] 本发明实施例的技术方案，通过利用第一数量阈值限制当前服务器对应的当前总创建虚拟机数量，避免创建虚拟机和运行虚拟机的数量超过服务器的承受范围而导致服务器出现运行卡顿或崩溃的情况，并利用第二数量阈值限制当前待创建虚拟机数量，避免当同一时间创建过多的虚拟机时，导致服务器CPU持续飙高形成卡顿，影响其他已创建完备的虚拟机的情况，同时保证了服务器在同一时间创建虚拟机的个数不会超过服务器所能承受的最大数量，避免当同一时间创建虚拟机数量超过最大数量限制而导致服务器卡顿甚至崩溃情况，进一步提升了用户体验。

[0075] 实施例三

[0076] 图3为本发明实施例三提供的一种虚拟机创建方法的流程图，本实施例提供了一种优选实现方式。其中与上述各实施例相同或相应的术语的解释在此不再赘述。如图3所示，该方法包括：

[0077] S310、接收虚拟机创建请求。

[0078] S320、遍历所有服务器，并根据服务器排列顺序确定是否还存在未检测服务器。若是，则执行S330；若否，则执行S391。

[0079] 需要说明的是，退出创建的虚拟机可以在本次虚拟机创建请求全部执行完毕的一段时长后，基于未创建的虚拟机再次向控制器发送虚拟机创建请求。

[0080] S330、将该未检测服务器作为用于创建虚拟机的当前服务器，并获取用于当前服务器对应的当前服务器状态信息、当前服务器类型信息、当前总创建虚拟机数量、和当前待创建虚拟机数量、当前服务器剩余资源信息。

[0081] S340、检测当前服务器状态信息是否可用。若是，则执行350进行下一步检测；若否，则执行S320。

[0082] S350、检测当前服务器类型信息与目标虚拟机所需类型是否符合。若是，则执行360进行下一步检测；若否，则执行S320。

[0083] S360、检测当前总创建虚拟机数量是否达到当前服务器对应预设创建上限。

[0084] 若否，则执行370进行下一步检测；若是，则执行S320。

[0085] S370、检测当前待创建虚拟机数量是否达到当前服务器对应预设待创建上限。

[0086] 若否，则执行S380进行下一步检测；若是，则执行S320。

[0087] S380、检测当前服务器剩余资源信息是否能支撑创建目标虚拟机。若是，则执行步骤S390，若否，则执行S320。

[0088] S390、在当前服务器中创建目标虚拟机。

[0089] S391、退出创建。

[0090] 通过本发明实施例，在虚拟机创建过程中，虚拟机创建请求者无需关注虚拟机的创建，只需要提供创建虚拟机的相关信息即可，从而在简化了创建虚拟机的流程基础上，同时结合服务器的资源使用情况，以及控制创建虚拟机的频率极大减少了服务器卡顿甚至崩溃情况。

[0091] 以下是本发明实施例提供的虚拟机创建装置的实施例，该装置与上述各实施例的虚拟机创建方法属于同一个发明构思，在虚拟机创建装置的实施例中未详尽描述的细节内容，可以参考上述虚拟机创建方法的实施例。

[0092] 实施例四

[0093] 图4为本发明实施例四提供的一种虚拟机创建装置的结构示意图。如图4所示，该装置包括：请求接收模块410、服务器状态获取模块420、服务器信息获取模块430、虚拟机创建模块440。

[0094] 其中，请求接收模块410，用于接收虚拟机创建请求，虚拟机创建请求包括：当前待创建的目标虚拟机对应的目标虚拟机类型信息和创建目标虚拟机时所需的目标虚拟机资源信息；服务器状态获取模块420，用于获取用于创建虚拟机的当前服务器对应的当前服务器状态信息；服务器信息获取模块430，用于若检测到当前服务器状态信息为可用状态信息，则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息；虚拟机创建模块440，用于若检测到当前服务器类型信息与目标虚拟机类型信息相匹配，且当前服务器剩余资源信息大于或等于目标虚拟机资源信息，则在当前服务器上创建目标虚拟机。

[0095] 本发明实施例的技术方案，通过接收虚拟机创建请求，虚拟机创建请求包括：当前待创建的目标虚拟机对应的目标虚拟机类型信息和创建目标虚拟机时所需的目标虚拟机资源信息；获取用于创建虚拟机的当前服务器对应的当前服务器状态信息；若检测到当前服务器状态信息为可用状态信息，则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息；若检测到当前服务器类型信息与目标虚拟机类型信息相匹配，且当前服务器剩余资源信息大于或等于目标虚拟机资源信息，则在当前服务器上创建目标虚拟机，以实现基于服务器的实际资源情况在服务器上合理创建虚拟机，从而保证虚拟机的成功创建，并且避免出现服务器卡顿或者奔溃的情况，从而提升了用户体验。

[0096] 可选地，该装置还包括：

[0097] 当前服务器确定模块，用于在获取用于创建虚拟机的当前服务器对应的当前服务器状态信息之前，将服务器集群中的第一个服务器确定为用于创建虚拟机的当前服务器；

[0098] 当前服务器变更模块，用于若检测到当前服务器状态信息为非可用状态信息、当前服务器类型信息与目标虚拟机类型信息不匹配，或者当前服务器剩余资源信息小于目标虚拟机资源信息，则基于服务器集群对应的服务器排列顺序，将当前服务器的下一服务器作为当前服务器，返回执行获取用于创建虚拟机的当前服务器对应的当前服务器状态信息的操作。

[0099] 可选地，该装置还包括：

[0100] 重新遍历服务器模块，用于若服务器集群中的所有服务器均不满足创建目标虚拟机的条件，则间隔预设时长后，重新对服务器集群中的服务器进行遍历匹配，直到在满足条件的目标服务器中创建出目标虚拟机为止。

[0101] 可选地，虚拟机创建模块440，可以包括：

[0102] 当前服务器数量获取子模块，用于获取当前服务器对应的当前总创建虚拟机数量和当前待创建虚拟机数量；

[0103] 目标虚拟机创建子模块，用于若检测到当前总创建虚拟机数量小于第一数量阈值，且当前待创建虚拟机数量小于第二数量阈值，则在当前服务器上创建目标虚拟机。

[0104] 可选地，服务器信息获取模块430具体用于：

[0105] 若检测到当前服务器状态信息是除了服务器重启状态、软件安装状态和软件升级状态之外的可分配资源的服务器状态，则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息。

[0106] 可选地，虚拟机创建请求还包括：目标虚拟机对应的目标镜像文件标识信息；虚拟机创建模块440具体用于：获取目标镜像文件标识信息对应的目标镜像文件；通过运行目标镜像文件，在当前服务器上创建出目标虚拟机。

[0107] 可选地，目标虚拟机资源信息包括：目标虚拟机对应的CPU核心数、运行内存大小和显卡内存大小；

[0108] 当前服务器剩余资源信息包括：当前服务器对应的剩余CPU核心数、剩余运行内存大小和剩余显卡内存大小。

[0109] 本发明实施例所提供的虚拟机创建装置可执行本发明任意实施例所提供的虚拟机创建方法，具备执行方法相应的功能模块和有益效果。

[0110] 值得注意的是，上述虚拟机创建装置的实施例中，所包括的各个模块只是按照功能逻辑进行划分的，但并不局限于上述的划分，只要能够实现相应的功能即可；另外，各功能模块的具体名称也只是为了便于相互区分，并不用于限制本发明的保护范围。

[0111] 实施例五

[0112] 图5示出了可以用来实施本发明的实施例的电子设备10的结构示意图。电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备(如头盔、眼镜、手表等)和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本发明的实现。

[0113] 如图5所示，电子设备10包括至少一个处理器11，以及与至少一个处理器11通信连接的存储器，如只读存储器(ROM)12、随机访问存储器(RAM)13等，其中，存储器存储有可被至少一个处理器执行的计算机程序，处理器11可以根据存储在只读存储器(ROM)12中的计算机程序或者从存储单元18加载到随机访问存储器(RAM)13中的计算机程序，来执行各种适当的动作和处理。在RAM 13中，还可存储电子设备10操作所需的各种程序和数据。处理器11、ROM 12以及RAM 13通过总线14彼此相连。输入/输出(I/O)接口15也连接至总线14。

[0114] 电子设备10中的多个部件连接至I/O接口15，包括：输入单元16，例如键盘、鼠标等；输出单元17，例如各种类型的显示器、扬声器等；存储单元18，例如磁盘、光盘等；以及通信单元19，例如网卡、调制解调器、无线通信收发机等。通信单元19允许电子设备10通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0115] 处理器11可以是各种具有处理和计算能力的通用和/或专用处理组件。处理器11的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的处理器、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。处理器11执行上文所描述的各个方法和处理，例如虚拟机创建方法。

[0116] 在一些实施例中，虚拟机创建方法可被实现为计算机程序，其被有形地包含于计算机可读存储介质，例如存储单元18。在一些实施例中，计算机程序的部分或者全部可以经由ROM 12和/或通信单元19而被载入和/或安装到电子设备10上。当计算机程序加载到RAM 13并由处理器11执行时，可以执行上文描述的虚拟机创建方法的一个或多个步骤。备选地，在其他实施例中，处理器11可以通过其他任何适当的方式(例如，借助于固件)而被配置为

执行虚拟机创建方法。

[0117] 本文中以上描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(CPLD)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0118] 用于实施本发明的方法的计算机程序可以采用一个或多个编程语言的任何组合来编写。这些计算机程序可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器,使得计算机程序当由处理器执行时使流程图和/或框图中所规定的功能/操作被实施。计算机程序可以完全在机器上执行、部分地在机器上执行,作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0119] 在本发明的上下文中,计算机可读存储介质可以是有形的介质,其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的计算机程序。计算机可读存储介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备,或者上述内容的任何合适组合。备选地,计算机可读存储介质可以是机器可读信号介质。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0120] 为了提供与用户的交互,可以在电子设备上实施此处描述的系统和技术,该电子设备具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给电子设备。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0121] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)、区块链网络和互联网。

[0122] 计算系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器,又称为云计算服务器或云主机,是云计算服务体系中的一项主机产品,以解决了传统物理主机与VPS服务中,存在的

管理难度大，业务扩展性弱的缺陷。

[0123] 应该理解，可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本发明中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本发明的技术方案所期望的结果，本文在此不进行限制。

[0124] 上述具体实施方式，并不构成对本发明保护范围的限制。本领域技术人员应该明白的是，根据设计要求和其他因素，可以进行各种修改、组合、子组合和替代。任何在本发明的精神和原则之内所作的修改、等同替换和改进等，均应包含在本发明保护范围之内。

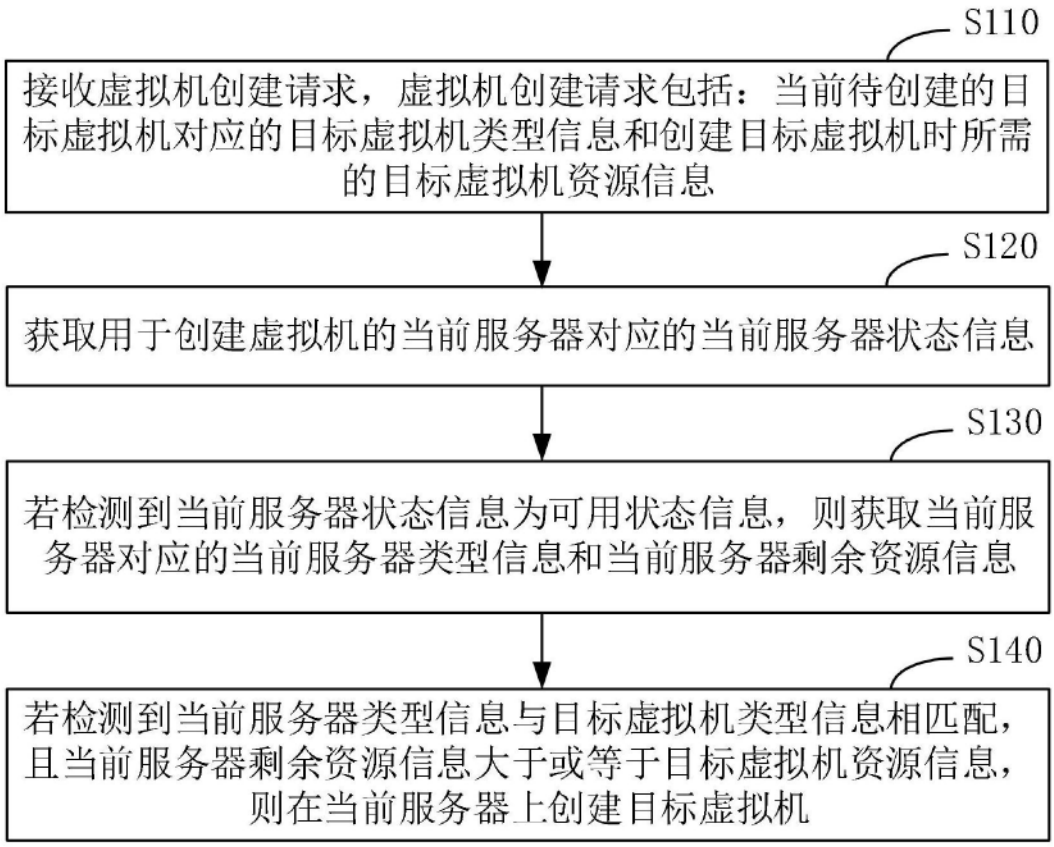

S110
接收虚拟机创建请求，虚拟机创建请求包括：当前待创建的目标虚拟机对应的目标虚拟机类型信息和创建目标虚拟机时所需的目标虚拟机资源信息
S120
获取用于创建虚拟机的当前服务器对应的当前服务器状态信息
S130
若检测到当前服务器状态信息为可用状态信息，则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息
S140
若检测到当前服务器类型信息与目标虚拟机类型信息相匹配，且当前服务器剩余资源信息大于或等于目标虚拟机资源信息，则在当前服务器上创建目标虚拟机

图1

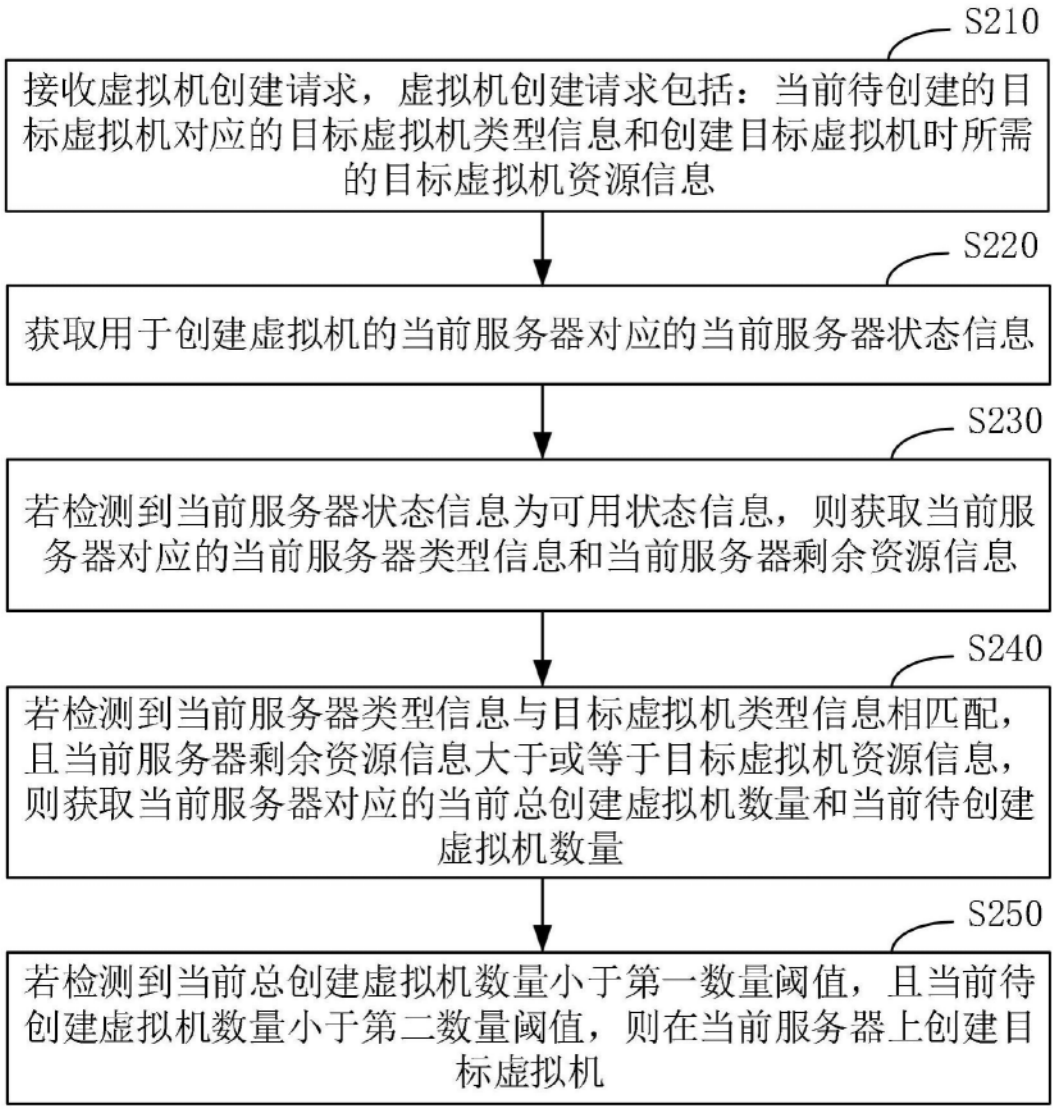

S210
接收虚拟机创建请求，虚拟机创建请求包括：当前待创建的目标虚拟机对应的目标虚拟机类型信息和创建目标虚拟机时所需的目标虚拟机资源信息
S220
获取用于创建虚拟机的当前服务器对应的当前服务器状态信息
S230
若检测到当前服务器状态信息为可用状态信息，则获取当前服务器对应的当前服务器类型信息和当前服务器剩余资源信息
S240
若检测到当前服务器类型信息与目标虚拟机类型信息相匹配，且当前服务器剩余资源信息大于或等于目标虚拟机资源信息，则获取当前服务器对应的当前总创建虚拟机数量和当前待创建虚拟机数量
S250
若检测到当前总创建虚拟机数量小于第一数量阈值，且当前待创建虚拟机数量小于第二数量阈值，则在当前服务器上创建目标虚拟机

图2

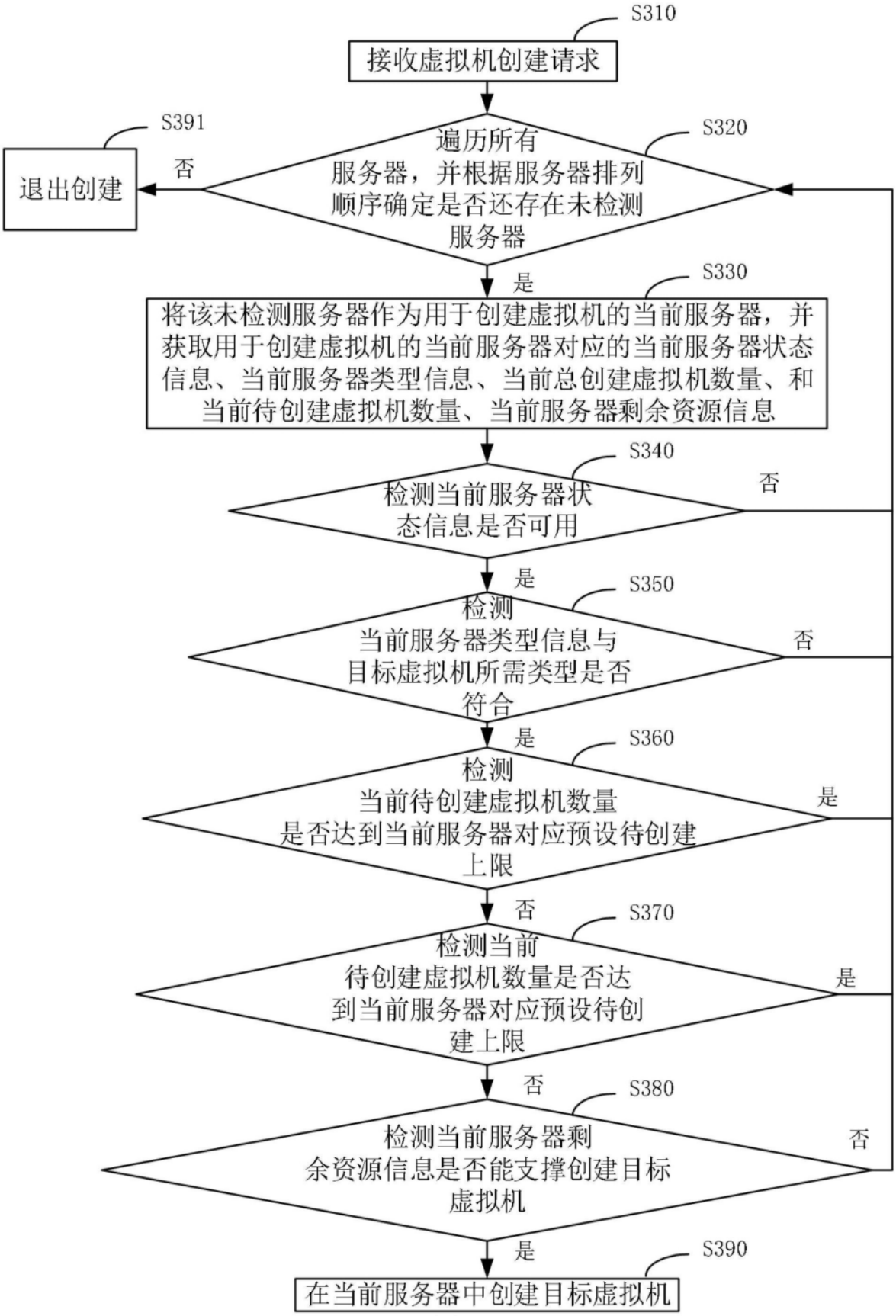

S310
接收虚拟机创建请求
S320
遍历所有服务器，并根据服务器排列顺序确定是否还存在未检测服务器
否
S391
退出创建
是
S330
将该未检测服务器作为用于创建虚拟机的当前服务器，并获取用于创建虚拟机的当前服务器对应的当前服务器状态信息、当前服务器类型信息、当前总创建虚拟机数量、和当前待创建虚拟机数量、当前服务器剩余资源信息
S340
检测当前服务器状态信息是否可用
否
是
S350
检测当前服务器类型信息与目标虚拟机所需类型是否符合
否
是
S360
检测当前待创建虚拟机数量是否达到当前服务器对应预设待创建上限
是
否
S370
检测当前待创建虚拟机数量是否达到当前服务器对应预设待创建上限
是
否
S380
检测当前服务器剩余资源信息是否能支撑创建目标虚拟机
否
是
S390
在当前服务器中创建目标虚拟机

图3

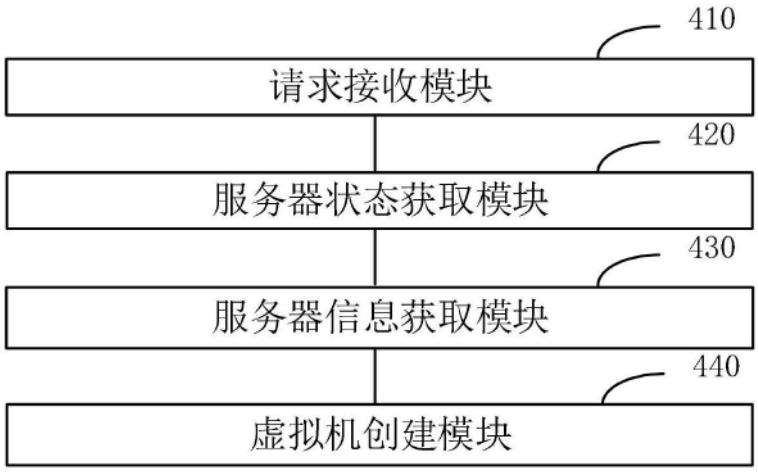

410
请求接收模块
420
服务器状态获取模块
430
服务器信息获取模块
440
虚拟机创建模块

图4

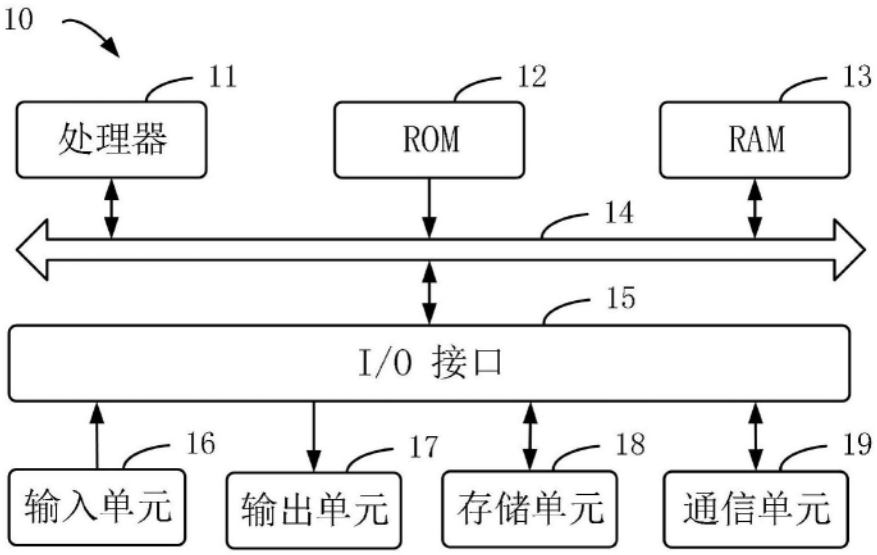

10
11
处理器
12
ROM
13
RAM
14
15
I/O 接口
16
输入单元
17
输出单元
18
存储单元
19
通信单元

图5