(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
2016 年 10 月 20 日 (20.10.2016)

WIPO | PCT

(10) 国际公布号
WO 2016/165304 A1

(51) 国际专利分类号:
H04L 29/08 (2006.01)

(21) 国际申请号: PCT/CN2015/092667

(22) 国际申请日: 2015 年 10 月 23 日 (23.10.2015)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201510180650.3 2015 年 4 月 16 日 (16.04.2015) 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: **潘方敏 (PAN, Fangmin)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **周通 (ZHOU, Tong)**; 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: **深圳市深佳知识产权代理事务所（普通合伙） (SHENPAT INTELLECTUAL PROPERTY AGENCY)**; 中国广东省深圳市国贸大厦 15 楼西座 1521 室, Guangdong 518014 (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, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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, 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, 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, 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)。

[见续页]

(54) **Title**: METHOD FOR MANAGING INSTANCE NODE AND MANAGEMENT DEVICE

(54) **发明名称** : 一种实例节点管理的方法及管理设备

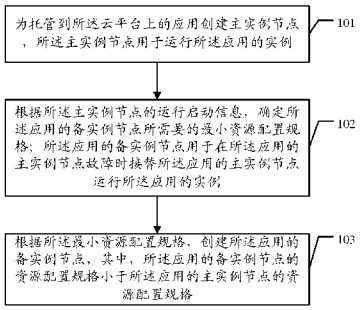

图 7

101 Creating a primary instance node for an application managed by a cloud platform, wherein the primary instance node is used to run an instance of the application

102 According to running and starting information about the primary instance node, determining a minimum resource configuration specification needed by a standby instance node of the application, wherein the standby instance node of the application is used to replace the primary instance node of the application to run the instance of the application when a failure occurs on the primary instance node of the application

103 According to the minimum resource configuration specification, creating the standby instance node of the application, wherein a resource configuration specification of the standby instance node of the application is less than a resource configuration specification of the primary instance node of the application

(57) **Abstract**: Disclosed is a method for managing an instance node, comprising: creating a primary instance node for an application managed by a cloud platform, wherein the primary instance node is used to run an instance of the application; according to running and starting information about the primary instance node, determining a minimum resource configuration specification needed by a standby instance node of the application, wherein the standby instance node of the application is used to replace the primary instance node of the application to run the instance of the application when a failure occurs on the primary instance node of the application; and according to the minimum resource configuration specification, creating the standby instance node of the application, wherein a resource configuration specification of the standby instance node of the application is less than a resource configuration specification of the primary instance node of the application. The method for managing the instance node provided in an embodiment of the present invention can, on the basis of ensuring that an application is highly available when being managed by the cloud platform, reduce the occupation for a standby resource and improve the management capability and scale of the application on the cloud platform.

(57) **摘要**:

[见续页]

WO 2016/165304 A1

根据细则 4.17 的声明:

— 关于申请人有权申请并被授予专利(细则 4.17(ii))

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

本发明公开了一种实例节点管理的方法，包括：为托管到所述云平台上的应用创建主实例节点，所述主实例节点用于运行所述应用的实例；根据所述主实例节点的运行启动信息，确定所述应用的备实例节点所需要的最小资源配置规格；所述应用的备实例节点用于在所述应用的主实例节点故障时接替所述应用的主实例节点运行所述应用的实例；根据所述最小资源配置规格，创建所述应用的备实例节点，其中，所述应用的备实例节点的资源配置规格小于所述应用的主实例节点的资源配置规格。本发明实施例提供的实例节点管理的方法，可以在保证应用托管到云平台上高可用的基础上，可以减少备用资源的占用，提高云平台上应用的托管能力和规模。

一种实例节点管理的方法及管理设备

本申请要求于 2015 年 4 月 16 日提交中国专利局、申请号为 201510180650.3、发明名称为“一种实例节点管理的方法及管理设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本发明涉及云计算技术领域，具体涉及一种实例节点管理的方法及管理设备。

背景技术

在云计算时代，大量的应用和服务都被托管在云平台上。云平台除了要保证自身的高可用，还要对托管在其上的应用和服务提供高可用的保证。现有的方案通常是通过部署主主实例节点和主备实例节点来实现应用和服务的高可用的。针对主主实例节点的高可用方案，多个应用实例之间是并列的关系，所有的实例节点都能正常接收到外部的请求信息并进行处理。在其中某个实例节点出现故障的情况下，其所承担的负载，将被分担到其他的实例节点上去。主备实例节点方案是业界最主流的高可靠方案，在备节实例点上备份主实例节点的应用实例，在主实例节点正常时，外部请求全部访问主实例节点，只有在主实例节点出现故障不能正常运行时，外部请求才切换到备实例节点上去。

由此可见，主备实例节点方案，备实例节点占用了与主实例节点完全相同的资源，却一直处于待命的状态，造成了物理资源的浪费，而主主实例节点方案下，两个主实例节点的资源利用率都无法达到 50%，否则出现单点故障时，全部的压力集中到另一个存活的主实例节点上，压力过大将导致业务无法正常，这样也就造成了更多的节点资源的使用浪费。

发明内容

本发明实施例提供的一种实例节点管理的方法，在保证应用托管到云平台上高可用的基础上，可以减少备用资源的占用，提高云平台上应用的托管能力和规模。本发明实施例还提供了相应的管理设备。

本发明第一方面提供一种实例节点管理的方法，所述方法应用于云平台的管理设备，所述方法包括：

为托管到所述云平台上的应用创建主实例节点，所述主实例节点用于运行所述应用的实例；

根据所述主实例节点的运行启动信息，确定所述应用的备实例节点所需要的最小资源配置规格；所述应用的备实例节点用于在所述应用的主实例节点故障时接替所述应用的主实例节点运行所述应用的实例；

根据所述最小资源配置规格，创建所述应用的备实例节点，其中，所述应用的备实例节点的资源配置规格小于所述应用的主实例节点的资源配置规格。

结合第一方面，在第一种可能的实现方式中，所述根据所述最小资源配置规格，创建所述应用的备实例节点之后，所述方法还包括：

当所述主实例节点故障时，将所述备实例节点的资源配置规格调整到与所述主实例节点的资源配置规格相同，并将所述备实例节点设置为工作实例节点，以接替所述应用的主实例节点运行所述应用的实例。

结合第一方面第一种可能的实现方式，在第二种可能的实现方式中，所述将所述工作实例节点的资源配置规格调整到与所述应用的主实例节点的资源配置规格相同之后，所述方法还包括：

根据所述最小资源配置规格，再次为所述应用创建备实例节点。

结合第一方面第一种可能的实现方式，在第三种可能的实现方式中，所述将所述工作实例节点的资源配置规格调整到与所述应用的主实例节点的资源配置规格相同之后，所述方法还包括：

修复发生故障的所述主实例节点，以得到修复后的主实例节点；

将所述修复后的主实例节点的资源配置规格调整到与所述最小资源配置规格相同，并将所述修复后的主实例节点设置为所述应用的备实例节点，以便于在所述工作实例节点故障时接替所述工作实例节点运行所述应用的实例。

结合第一方面、第一方面第一种至第三种中任意一种可能的实现方式，在第四种可能的实现方式中，所述确定所述应用的备实例节点的最小资源配置规格之后，所述方法还包括：

获取用户设备对所述应用的访问量;

所述根据所述最小资源配置规格，创建所述应用的备实例节点，包括:

当所述应用的访问量超过预置阈值时，根据所述最小资源配置规格，为所述应用创建至少两个备实例节点，其中，创建的每个备实例节点的资源配置规格与所述最小资源配置规格相同。

结合第一方面、第一方面第一种至第三种中任意一种可能的实现方式，在第五种可能的实现方式中，所述根据所述最小资源配置规格，创建所述应用的备实例节点之后，所述方法还包括:

监测已创建的所述应用的备实例节点;

当监测到所述应用的备实例节点发生故障时，创建与发生故障的备实例节点数量相同的备实例节点。

结合第一方面、第一方面第一种至第三种中任意一种可能的实现方式，在第六种可能的实现方式中，所述根据所述最小资源配置规格，创建所述应用的备实例节点之后，所述方法还包括:

检测运行所述应用的实例所需的资源量;

根据所述所需的资源量，调整已创建的所述应用的备实例节点的资源配置规格。

本发明第二方面提供一种云平台的管理设备，所述管理设备包括:

第一创建模块，用于为托管到所述云平台上的应用创建主实例节点，所述主实例节点用于运行所述应用的实例;

确定模块，用于根据所述第一创建模块创建的所述主实例节点的运行启动信息，确定所述应用的备实例节点所需要的最小资源配置规格;所述应用的备实例节点用于在所述应用的主实例节点故障时接替所述应用的主实例节点运行所述应用的实例;

第二创建模块，用于根据所述确定模块确定的所述最小资源配置规格，创建所述应用的备实例节点，其中，所述应用的备实例节点的资源配置规格小于所述应用的主实例节点的资源配置规格。

结合第二方面，在第一种可能的实现方式中，所述管理设备还包括:

第一调整模块，用于当所述主实例节点故障时，将所述第二创建模块创建的所述备实例节点的资源配置规格调整到与所述主实例节点的资源配置规格相同，并将所述备实例节点设置为工作实例节点，以接替所述应用的主实例节点运行所述应用的实例。

结合第二方面第一种可能的实现方式，在第二种可能的实现方式中，

所述第二创建模块，还用于根据所述最小资源配置规格，再次为所述应用创建备实例节点。

结合第二方面第一种可能的实现方式，在第三种可能的实现方式中，

所述管理设备还包括：修复模块，

所述修复模块，用于修复发生故障的所述主实例节点，以得到修复后的主实例节点；

所述第一调整模块，还用于将所述修复后的主实例节点的资源配置规格调整到所述应用的最小资源配置规格，并将所述修复后的主实例节点设置为所述应用的备实例节点，以便于在所述工作实例节点故障时接替所述工作实例节点运行所述应用的实例。

结合第二方面、第二方面第一种至第三种中任意一种可能的实现方式，在第四种可能的实现方式中，所述管理设备还包括获取模块，

所述获取模块，用于获取用户设备对所述应用的访问量；

所述第二创建模块，具体用于当所述获取模块获取的所述应用的访问量超过预置阈值时，根据所述确定模块确定的所述最小资源配置规格，为所述应用创建至少两个备实例节点，其中，创建的每个备实例节点的资源配置规格与所述最小资源配置规格相同。

结合第二方面、第二方面第一种至第三种中任意一种可能的实现方式，在第五种可能的实现方式中，所述管理设备还包括监测模块，

所述监测模块，用于监测所述第二创建模块已创建的所述应用的备实例节点；

所述第二创建模块，还用于当所述监测模块监测到所述应用的备实例节点发生故障时，创建与发生故障的备实例节点数量相同的备实例节点。

结合第二方面、第二方面第一种至第三种中任意一种可能的实现方式，在第六种可能的实现方式中，所述管理设备还包括检测模块和第二调整模块，

所述检测模块，用于检测运行所述应用的实例所需的资源量；

所述第二调整模块，用于根据所述检测模块检测到的所述所需的资源量，调整已创建的所述应用的备实例节点的资源配置规格。

本发明实施例采用为托管到所述云平台上的应用创建主实例节点，所述主实例节点用于运行所述应用的实例；根据所述主实例节点的运行启动信息，确定所述应用的备实例节点所需要的最小资源配置规格；所述应用的备实例节点用于在所述应用的主实例节点故障时接替所述应用的主实例节点运行所述应用的实例；根据所述最小资源配置规格，创建所述应用的备实例节点，其中，所述应用的备实例节点的资源配置规格小于所述应用的主实例节点的资源配置规格。与现有技术中为了保证云平台上应用的高可用性，备实例节点需要占用与主实例节点相同的资源相比，本发明实施例提供的实例节点管理的方法，可以在保证应用托管到云平台上高可用的基础上，可以减少备用资源的占用，提高云平台上应用的托管能力和规模。

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的前提下，还可以根据这些附图获得其他的附图。

图1是本发明实施例中实例节点管理的方法的一实施例示意图；

图2是本发明实施例中实例节点管理的方法的另一实施例示意图；

图3是本发明实施例中实例节点管理的方法的另一实施例示意图；

图4是本发明实施例中实例节点管理的方法的另一实施例示意图；

图5是本发明实施例中实例节点管理的方法的另一实施例示意图；

图6是本发明实施例中实例节点管理的方法的另一实施例示意图；

图7是本发明实施例中实例节点管理的方法的另一实施例示意图；

图8是本发明实施例中管理设备的一实施例示意图；

图9是本发明实施例中管理设备的另一实施例示意图；

图10是本发明实施例中管理设备的另一实施例示意图；

图11是本发明实施例中管理设备的另一实施例示意图；

图12是本发明实施例中管理设备的另一实施例示意图；

图13是本发明实施例中管理设备的另一实施例示意图；

图14是本发明实施例中管理设备的另一实施例示意图；

图15是本发明实施例中管理设备的另一实施例示意图；

图 16 是本发明实施例中管理设备的另一实施例示意图。

具体实施方式

本发明实施例提供一种实例节点管理的方法，在保证应用托管到云平台上高可用的基础上，可以减少备用资源的占用，提高云平台上应用的托管能力和规模。本发明实施例还提供了相应的管理设备。以下分别进行详细说明。

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。

云平台可以托管大量的应用，为了保证每个应用高可用的性能，在云平台的运行环境中，需要为每个应用创建主实例节点和备实例节点，主实例节点用于运行应用的实例，备实例节点用于在所述应用的主实例节点故障时接替所述应用的主实例节点运行所述应用的实例。云平台上的管理设备会对主实例节点和备实例节点进行统一的管理。主实例节点和备实例节点实际上都是由管理设备为应用所分配的的虚拟机或者比虚拟机粒度更小的功能节点，主实例节点和备实例节点均可以利用云平台上的资源。云平台上的资源是指由众多物理主机所提供的硬件资源，包括但不限于中央处理器（Central Processing Unit，CPU）的核数，内存，硬盘大小，网络带宽等。

本发明实施例中，在用户将应用托管到云平台上后，为用户托管到所述云平台上的应用创建运行所述应用的实例的主实例节点，根据所述主实例节点的运行启动信息，确定所述应用的备实例节点的最小资源配置规格；根据所述最

小资源配置规格，创建所述应用的备实例节点，以使所述备实例节点的资源配置规格与所述最小资源配置规格相同，且所述应用的备实例节点的资源配置规格小于所述应用的主实例节点的资源配置规格，所述应用的备实例节点用于在所述应用的主实例节点故障时接替所述应用的主实例节点运行所述应用的实例，其中，实例节点的资源配置规格用于指示为该实例节点分配的资源量，比如为实例节点分配的CPU核数，内存，硬盘大小，网络带宽等等；最小资源配置规格用于指示启动或运行该应用的实例所需要的最小的资源量。应用的实例可以理解为应用的可执行版本的拷贝或镜像，云平台可以为一个应用创建一个或多个实例，应用的每一个实例具体可以是一个独立的进程或线程。

其中，用户可以是应用运营商，主实例节点的运行启动信息可以指示运行所述应用的实例所需要的基本资源，例如：主实例节点开始运行时的CPU的核数，内存，硬盘大小，网络带宽。备实例节点的最小资源配置规格可以是指备实例节点用于运行所述应用所需的CPU的核数，内存，硬盘大小，网络带宽等规格信息。

如图1所示，在主实例节点正常时，由主实例节点为用户设备的访问提供服务，在主实例节点出现故障时，管理设备将用户设备的访问切换到备实例节点，并将所述应用的备实例节点设置为对用户设备的访问提供服务的工作实例节点；将所述工作实例节点的资源配置规格调整到与所述应用的主实例节点的资源配置规格相同。当然，除了调整到与所述应用的主实例节点的资源配置规格相同外，也可以是调整到比所述应用的主实例节点的资源配置规格略大或略小。

本发明实施例中的管理设备可以包括图2中的云控制器、备节点管理子系统和健康检查系统。如图2所示：

步骤1、应用的部署者将云应用通过云控制器部署到云平台上。

步骤2、云控制器根据该应用的规格创建运行该应用的的主实例节点，该主实例节点具有初始资源配置规格，该初始资源配置规格通常为系统默认值或者开发者/维护者根据经验设置的，且该初始资源配置规格能够保证主实例节点有充足的资源可以使用，以承载应用的实例。

步骤3、在主实例节点构建完成后，健康检查系统识别到新增云应用的主实例节点。

步骤4、健康检查系统通知备节点管理子系统进行最小资源配置规格的计算。

步骤5、备节点管理子系统根据已经启动并正常运行的主实例节点的运行信息计算备实例节点的最小资源配置规格。运行信息可以指示正常运行所述应用的实例实际占用的资源，最小资源配置规格指示了启动或运行该应用的实例所需的最小的资源量，该最小资源配置规格通常远远小于主实例节点的初始资源配置规格。

步骤6、云控制器根据得到的最小资源配置规格，再创建一个备实例节点，其中，创建的备实例节点的资源配置规格与该最小资源配置规格。

图3为主实例节点故障时的处理示例流程，健康检查系统检测到主实例节点故障时，通知到备节点管理子系统，由备节点管理子系统去通知云控制器启用和扩容最小化的备实例节点作为常规实例节点，以接替所述应用的主实例节点运行所述应用的实例，并再根据最小资源配置规格创建出一个备实例节点。这里的常规实例也可以称为工作实例节点，是指当前运行所述应用的实例的实例节点。

如图4所示，本发明实施例提供的实例节点管理的方法的另一实施例中，云平台包括了以下几个部分：云控制器、健康检查系统、备节点管理子系统，路由分发系统和应用运行环境。其中，云控制器、健康检查系统、备节点管理子系统和路由分发系统都可以理解为是管理设备的组成部分，管理设备具有云控制器、健康检查系统、备节点管理子系统和路由分发系统的功能。

各部分的功能包括：

备节点管理子系统：在云平台上，备节点管理系统主要负责的是计算备实例节点的最小资源配置规格，以及在主实例节点故障时，通知云控制器启动并扩容备实例节点，以及在备实例节点启用成常规工作实例节点后，后续再次根据最小资源配置规格创建备实例节点。

备实例节点扩容成常规工作实例节点是指将所述备实例节点的资源配置

规格调整到与所述应用的主实例节点的资源配置规格相同。需要说明的是，本发明实施例所说的“相同”，并不是是指数值上绝对的相同，在实际应用中，允许存在一定的误差，比如，所述备实例节点比所述主实例节点的资源配置规格略大或略小，或者实质相同。

如图5所示，备节点管理子系统主要包括以下三个模块：最小化规格计算模块、应用备节点实例管理模块和故障处理配置通知模块。

其中，最小化规格计算模块，主要是在应用初次部署在云平台上，其主实例节点已经启动成功后，从主实例节点初始化过程中获取所需要资源的最小的配置规格。规格包括且不限于，CPU核数，内存，硬盘大小，网络带宽等。这个规格将通知给云控制器，并在云控制器上记录下来，作为后续启动新的备实例节点的规格依据。

应用备节点实例管理模块，主要用于判断和检查应用的备实例节点的情况。备实例节点的个数和运行状态，由该模块负责管理。当应用持续压力较大时，系统可以判断增加最小化的备实例节点的个数，以支撑突发故障的情况下，可以快速启用多个备实例，保证运行无障碍。

故障处理配置通知模块，是在健康检查系统检测到有实例故障时的处理机制。这里包括两种实例的故障，一是常规实例节点故障，二是最小化规格的备实例节点故障。常规实例故障时，该故障处理配置通知模块通知云控制器启用一个备实例节点，并将之快速扩容到常规实例节点规格；此时最小化的备实例节点个数相应的减少了一个，即通知云控制器，新建一个最小化的备实例节点，最小化的备实例节点出现故障时也是与常规实例节点故障进行相同的操作。

云控制器：实例节点的管理和路由分发的配置。所有云应用的实例节点，都是由云控制器分发到指定的应用运行环境中，实例的访问方式，由云控制器将策略发送到路由分发系统中去，作为云平台上的通用功能系统，本发明不作详细描述。

路由分发系统：主要控制将外部的访问请求分发到对应的应用的运行实例节点中去，接收实例节点的处理结果并反馈给发送访问请求的用户设备，作为云平台上的通用功能系统，本发明不作详细描述。

健康检查系统：负责检查所有应用的实例节点的健康状态，并通知备节点管理系统进行故障恢复，以及最小化的备实例节点的创建等，作为云平台上的通用功能系统，本发明不作详细描述。

应用运行环境：在云平台上的应用运行环境，主要是虚拟机、容器等载体，并提供了应用运行所需要的操作系统、软件栈、计算/存储/网络等资源，以供给应用实例正常运行，作为云平台上的通用功能系统，本发明不作详细描述。

本发明实施例中对实例节点的管理实现流程，分为初次部署和故障出现两种情况，可以参阅图1至图5部分的相应描述进行理解，本处不做过多赘述。

可见，本发明实施例中，通过上述步骤，可以保证云应用拥有足够的主备保障能力，而且备实例节点的规格能远小于常规的实例节点，有效的减少了因为高可用需要而造成的资源浪费情况。

如图6所示，本发明实施例中对实例节点的管理的另一实施例与图4对应的实施例基本相同，只是在结构上做了一部分的简化。备实例管理节点，在实现上，可以包含在云控制器上，以云控制器的一个插件/组件等方式，与云控制器部署在一起。

在本实施例上，系统间通信先通知到云控制器，再由云控制器将相关的处理交由备节点管理插件。备节点管理节点在计算/配置完成后，可以直接通知路由/应用运行环境等，进行路由分发管理配置，以及备实例节点创建/扩容等实现。

在处理流程上，所有云控制器与备节点管理系统之间的交互过程，全部在云控制器内部实现。其他流程一样。

进一步的，还可以在备节点管理系统/插件上进行进一步的简化。大部分云应用的实现上，只要存在1个备实例节点，就能够满足基本的高可用需求。因此，可以去除应用备节点实例管理模块，只留一部分故障处理和最小化规格计算的功能即可，

初次创建云应用的过程，与上述实施例相同，在故障出现时，不需要判断检测备节点资源池，直接将云应用中的备实例节点转化成常规节点，并再创建一个备实例节点。

可选地，本发明实施例提供的实例节点管的方法的另一实施例中，还可以是初次创建云应用的过程相同，区别在故障出现时，备实例节点转化成常规节点后，并非再次创建一个备实例节点，而是由云控制器将出故障的常规实例进行恢复，并在恢复后，由健康检查系统通知备节点管理系统，将被启用的备实例节点回收，恢复初始状态。也可以是在故障修复后，将修复后的常规实例节点的资源配置规格调整到所述应用的最小资源配置规格，成为所述应用的备实例节点。

可选地，本发明实施例提供的实例节点管的方法的另一实施例中，还可以是在备节点最小规格的计算和管理上，增加了动态调整的方法。在常规实例运行过程中，根据统计得到的应用整体运行状态，动态计算最小规格，并实时对当前环境下的最小规格备份节点实例进行规格刷新。

可选地，还可以是在创建最小规格实例的方法上，考虑到有些应用可能会在初始化启动过程中使用大量的临时性资源，而在初始化完成后，所占用消耗的资源量反而较低，因此创建最小化的备实例节点的过程，首先是创建一个完整规格的实例，待启动完成后，再将其规格缩小到计算所得的最小化规格。以降低资源消耗。

在故障发生恢复时，也是首先恢复一个常规实例节点，再将其作为最小规格实例节点使用，缩小到最小规格。

另外，本发明实施例中，面向云平台上运行环境可以不同，如容器，虚拟机等时，最小规格备实例节点的实现方式上有所区别。在应用容器技术时，上述的各种实施例方法都能快速实现；在使用的是虚拟机技术时，由于针对虚拟机的纵向伸缩扩容技术的不成熟，这里采用的是纵向+横向伸缩的方法，即满足纵向伸缩的云环境上，直接使用上述实施例的方法，而在不满足纵向伸缩的云环境上，采用横向伸缩的机制，备实例节点的创建与上述相同，但是在使用上，通过建立多个备实例节点来满足短暂的可靠性要求，并在常规实例节点恢复了，再将临时生成的备实例节点释放，以保证资源的有效利用。

与现有技术中为了保证云平台上应用的高可用性，备实例节点需要占用与主实例节点相同的资源相比，本发明实施例提供的实例节点管理的方法，可以

在保证应用托管到云平台上高可用的基础上，可以减少备用资源的占用，提高云平台上应用的托管能力和规模。

需要说明的是，本发明实施例的方案不仅可以适用于云平台，还可以适用于除云平台以外的需要提高应用可靠性的多种场景。

参阅图7，本发明实施例提供的实例节点管理的方法的一实施例包括:

101、为托管到所述云平台上的应用创建主实例节点，所述主实例节点用于运行所述应用的实例。

本发明实施例提供的实例节点的管理方法应用于云平台的管理设备。

102、根据所述主实例节点的运行启动信息，确定所述应用的备实例节点所需要的最小资源配置规格; 所述应用的备实例节点用于在所述应用的主实例节点故障时接替所述应用的主实例节点运行所述应用的实例。

其中，主实例节点的运行启动信息可以是运行起所述应用所需要的基本资源，例如: 主实例节点开始运行时的CPU的核数，内存，硬盘大小，网络带宽。备实例节点的最小资源配置规格可以是指备实例节点的CPU的核数，内存，硬盘大小，网络带宽等规格信息。

103、根据所述最小资源配置规格，创建所述应用的备实例节点，其中，所述应用的备实例节点的资源配置规格小于所述应用的主实例节点的资源配置规格。

本发明实施例采用为托管到所述云平台上的应用创建主实例节点，所述主实例节点用于运行所述应用的实例; 根据所述主实例节点的运行启动信息，确定所述应用的备实例节点所需要的最小资源配置规格; 所述应用的备实例节点用于在所述应用的主实例节点故障时接替所述应用的主实例节点运行所述应用的实例; 根据所述最小资源配置规格，创建所述应用的备实例节点，其中，所述应用的备实例节点的资源配置规格小于所述应用的主实例节点的资源配置规格。与现有技术中为了保证云平台上应用的高可用性，备实例节点需要占用与主实例节点相同的资源相比，本发明实施例提供的实例节点管理的方法，可以在保证应用托管到云平台上高可用的基础上，可以减少备用资源的占用，提高云平台上应用的托管能力和规模。

可选地，在上述图7对应的实施例的基础上，本发明实施例提供的实例节点管理的方法的第一个可选实施例中，所述根据所述最小资源配置规格，创建所述应用的备实例节点之后，所述方法还可以包括:

当所述主实例节点故障时，将所述备实例节点的资源配置规格调整到与所述主实例节点的资源配置规格相同，并将所述备实例节点设置为工作实例节点，以接替所述应用的主实例节点运行所述应用的实例。

具体可以是:

接收所述应用的主实例节点故障的切换指令;

根据所述切换指令，将所述应用的备实例节点设置为对用户设备访问提供服务的工作实例节点。

将所述工作实例节点的资源配置规格调整到与所述应用的主实例节点的资源配置规格相同。

本发明实施例中，在主实例节点正常时，备实例节点的资源占用率很低，在主实例节点故障时，迅速扩容来接替主实例节点的工作，从而满足应用的访问需求，提高云平台上应用的托管能力和规模。

可选地，在上述实例节点管理的方法的第一个可选实施例的基础上，本发明实施例提供的实例节点管理的方法的第二个可选实施例中，所述将所述工作实例节点的资源配置规格调整到与所述应用的主实例节点的资源配置规格相同之后，所述方法还可以包括:

根据所述最小资源配置规格，再次为所述应用创建备实例节点。

本发明实施例中，当最小化的备实例节点数量减少时，及时的动态补充，避免了下次工作实例节点出现故障时，导致无法响应应用的访问请求，从而提高了可靠性，提高了用户体验。

可选地，在上述实例节点管理的方法的第一个可选实施例的基础上，本发明实施例提供的实例节点管理的方法的第三个可选实施例中，所述将所述工作实例节点的资源配置规格调整到与所述应用的主实例节点的资源配置规格相同之后，所述方法还可以包括:

修复发生故障的所述主实例节点，以得到修复后的主实例节点;

将所述修复后的主实例节点的资源配置规格调整到与所述最小资源配置规格相同，并将所述修复后的主实例节点设置为所述应用的备实例节点，以便于在所述工作实例节点故障时接替所述工作实例节点运行所述应用的实例。

可选地，也可以是：

修复发生故障的所述主实例节点；

将对所述用户设备访问提供服务的工作实例节点切换到所述主实例节点；

将所述工作实例节点的资源配置规格调整到所述应用的最小资源配置规格。

本发明实施例中，还可以及时修复发生故障的主实例节点，这样进一步提高了云平台中资源的利用率。

可选地，在上述实例节点管理的方法的任一实施例的基础上，本发明实施例提供的实例节点管理的方法的第四个可选实施例中，所述确定所述应用的备实例节点的最小资源配置规格之后，所述方法还可以包括：

获取用户设备对所述应用的访问量；

所述根据所述最小资源配置规格，创建所述应用的备实例节点，包括：

当所述应用的访问量超过预置阈值时，根据所述最小资源配置规格，为所述应用创建至少两个备实例节点，其中，创建的每个备实例节点的资源配置规格与所述最小资源配置规格相同。

本发明实施例中，可以根据应用的访问量创建相应数量的最小化的备实例节点，这样，可以预防应用压力过大时，导致突发的意外故障，从而进一步提高了云平台的高可用性。

可选地，在上述实例节点管理的方法的任一实施例的基础上，本发明实施例提供的实例节点管理的方法的第五个可选实施例中，所述根据所述最小资源配置规格，创建所述应用的备实例节点之后，所述方法还可以包括：

监测已创建的所述应用的备实例节点；

当监测到所述应用的备实例节点发生故障时，创建与发生故障的备实例节点数量相同的备实例节点。

本发明实施例中，还可以及时监测已创建的备实例节点中的故障实例节

点，及时创建新的可用备实例节点，以避免主实例节点出现故障时，没有可用的备实例节点接替主实例节点的工作，从而进一步提高了云平台的高可用性。

可选地，在上述实例节点管理的方法的任一实施例的基础上，本发明实施例提供的实例节点管理的方法的第六个可选实施例中，所述根据所述最小资源配置规格，创建所述应用的备实例节点之后，所述方法还可以包括:

检测运行所述应用的实例所需的资源量;

根据所述所需的资源量，调整已创建的所述应用的备实例节点的资源配置规格。

本发明实施例中，针对已创建的备实例节点，还可以根据运行应用所需要的资源量，对齐资源配置规格进行动态调整，如果应用的需求量低时，还可以进一步节省资源，如果应用的需求量高时，还可以提高主实例节点故障时的接替效率。

图7对应的实施例以及其相应的可选实施例，可以参阅图1至图6部分的相关描述进行理解，本处不做过多赘述。

需要说明的是下述在描述云平台的管理设备20的过程中所涉及的第一创建模块201、第二创建模块203、第一调整模块204、修复模块205、获取模块206、检测模块208、第二调整模块209的功能与图1-图4实施例中的所描述的云控制器的功能相同，确定模块202、监测模块207所执行的功能与图1-图4实施例中的所描述的备节点管理子系统的功能相同。其中，确定模块202可以为图5中的最小化规格计算模块，监测模块207可以为应用备节点实例管理模块和故障处理配置通知模块。

参阅图8，本发明实施例提供的云平台的管理设备20的一实施例包括:

第一创建模块201，用于为托管到所述云平台上的应用创建主实例节点，所述主实例节点用于运行所述应用的实例;

确定模块202，用于根据所述第一创建模块201创建的所述主实例节点的运行启动信息，确定所述应用的备实例节点所需要的最小资源配置规格;所述应用的备实例节点用于在所述应用的主实例节点故障时接替所述应用的主实例节点运行所述应用的实例;

第二创建模块203，用于根据所述确定模块202确定的所述最小资源配置规格，创建所述应用的备实例节点，其中，所述应用的备实例节点的资源配置规格小于所述应用的主实例节点的资源配置规格。

与现有技术中为了保证云平台上应用的高可用性，备实例节点需要占用与主实例节点相同的资源相比，本发明实施例提供的实例节点的管理设备，可以在保证应用托管到云平台上高可用的基础上，可以减少备用资源的占用，提高云平台上应用的托管能力和规模。

可选地，在上述图8对应的实施例的基础上，参阅图9，本发明实施例提供的云平台的管理设备20的第一个可选实施例中，所述管理设备20还包括：

第一调整模块204，用于当所述主实例节点故障时，将所述第二创建模块203创建的所述应用的备实例节点的资源配置规格调整到与所述应用的主实例节点的资源配置规格相同，并将所述备实例节点设置为工作实例节点，以接替所述应用的主实例节点运行所述应用的实例。

本发明实施例中，在主实例节点正常时，备实例节点的资源占用率很低，在主实例节点故障时，迅速扩容来接替主实例节点的工作，从而满足应用的访问需求，提高云平台上应用的托管能力和规模。

可选地，在上述图9对应的实施例的基础上，本发明实施例提供的云平台的管理设备20的第二个可选实施例中，

所述第二创建模块203，还用于根据所述最小资源配置规格，再次为所述应用创建备实例节点。

本发明实施例中，当最小化的备实例节点数量减少时，及时的动态补充，避免了下次工作实例节点出现故障时，导致无法响应应用的访问请求，从而提高了可靠性，提高了用户体验。

可选地，在上述图9对应的实施例的基础上，参阅图10，本发明实施例提供的云平台的管理设备20的第三个可选实施例中，所述管理设备20还包括：修复模块205，

所述修复模块205，用于修复发生故障的所述主实例节点，以得到修复后的主实例节点；

所述第一调整模块204，还用于将所述修复模块205修复后的主实例节点的资源配置规格调整到所述应用的最小资源配置规格，并将所述修复后的主实例节点设置为所述应用的备实例节点，以便于在所述工作实例节点故障时接替所述工作实例节点运行所述应用的实例。

本发明实施例中，还可以及时修复发生故障的主实例节点，这样进一步提高了云平台中资源的利用率。

可选地，在上述管理设备任一实施例的基础上，参阅图11，本发明实施例提供的云平台的管理设备20的第四个可选实施例中，所述管理设备20还包括获取模块206，

所述获取模块206，用于获取用户设备对所述应用的访问量；

所述第二创建模块203，具体用于当所述获取模块206获取的所述应用的访问量超过预置阈值时，根据所述确定模块确定的所述最小资源配置规格，为所述应用创建至少两个备实例节点，其中，创建的每个备实例节点的资源配置规格与所述最小资源配置规格相同。

本发明实施例中，可以根据应用的访问量创建相应数量的最小化的备实例节点，这样，可以预防应用压力过大时，导致突发的意外故障，从而进一步提高了云平台的高可用性。

可选地，在上述管理设备任一实施例的基础上，参阅图12，本发明实施例提供的云平台的管理设备20的第五个可选实施例中，所述管理设备20还包括监测模块207，

所述监测模块207，用于监测所述第二创建模块203已创建的所述应用的备实例节点；

所述第二创建模块203，还用于当所述监测模块211监测到所述应用的备实例节点发生故障时，创建与发生故障的备实例节点数量相同的备实例节点。

本发明实施例中，还可以及时监测已创建的备实例节点中的故障实例节点，及时创建新的可用备实例节点，以避免主实例节点出现故障时，没有可用的备实例节点接替主实例节点的工作，从而进一步提高了云平台的高可用性。

可选地，在上述管理设备任一实施例的基础上，参阅图13，本发明实施例

提供的云平台的管理设备20的第六个可选实施例中，所述管理设备20还包括检测模块212和第二调整模块213，

所述检测模块208，用于检测运行所述应用的实例所需的资源量；

所述第二调整模块209，用于根据所述检测模块212检测到的所述所需的资源量，调整已创建的所述应用的备实例节点的资源配置规格。

本发明实施例中，针对已创建的备实例节点，还可以根据应用的运行状态，对齐资源配置规格进行动态调整，如果应用的需求量低时，还可以进一步节省资源，如果应用的需求量高时，还可以提高主实例节点故障时的接替效率。

图8至图13所描述的管理设备可以参阅图1至图7中的描述进行理解，本处不做过多赘述。

在上述管理设备的多个实施例中，应当理解的是，在一种实现方式下，接收模块可以是由输入/输出I/O设备（比如网卡）来实现，确定模块、第一创建模块、第二创建模块、第一调整模块、修复模块、获取模块、监测模块、检测模块和第二调整模块可以由处理器执行存储器中的程序或指令来实现的（换言之，即由处理器以及与所述处理器耦合的存储器中的特殊指令相互配合来实现）；在另一种实现方式下，接收模块可以是由输入/输出I/O设备（比如网卡）来实现，确定模块、第一创建模块、第二创建模块、第一调整模块、修复模块、获取模块、监测模块、检测模块和第二调整模块也可以分别通过专有电路来实现，具体实现方式参见现有技术，这里不再赘述；在再一种实现方式下，接收模块可以是由输入/输出I/O设备（比如网卡）来实现，确定模块、第一创建模块、第二创建模块、第一调整模块、修复模块、获取模块、监测模块、检测模块和第二调整模块也可以通过现场可编程门阵列（FPGA，Field－Programmable Gate Array）来实现，具体实现方式参见现有技术，这里不再赘述，本发明包括但不限于前述实现方式，应当理解的是，只要按照本发明的思想实现的方案，都落入本发明实施例所保护的范围。

本实施例提供了一种管理设备的硬件结构，参见图14所示，一种管理设备的硬件结构可以包括：

收发器件、软件器件以及硬件器件三部分；

收发器件为用于完成包收发的硬件电路;

硬件器件也可称“硬件处理模块”，或者更简单的，也可简称为“硬件”，硬件器件主要包括基于FPGA、ASIC之类专用硬件电路（也会配合其他配套器件，如存储器）来实现某些特定功能的硬件电路，其处理速度相比通用处理器往往要快很多，但功能一经定制，便很难更改，因此，实现起来并不灵活，通常用来处理一些固定的功能。需要说明的是，硬件器件在实际应用中，也可以包括MCU（微处理器，如单片机）、或者CPU等处理器，但这些处理器的主要功能并不是完成大数据的处理，而主要用于进行一些控制，在这种应用场景下，由这些器件搭配的系统为硬件器件。

软件器件（或者也简单“软件”）主要包括通用的处理器（例如CPU）及其一些配套的器件（如内存、硬盘等存储设备），可以通过编程来让处理器具备相应的处理功能，用软件来实现时，可以根据业务需求灵活配置，但往往速度相比硬件器件来说要慢。软件处理完后，可以通过硬件器件将处理完的数据通过收发器件进行发送，也可以通过一个与收发器件相连的接口向收发器件发送处理完的数据。

本实施例中，收发器件用于进行上述实施例中切换指令的接收，软件器件或硬件器件用于APP注册、服务质量控制等。

硬件器件及软件器件的其他功能在前述实施例中已经详细论述，这里不再赘述。

下面结合附图就接收模块可以是由输入/输出I/O设备（比如网卡）来实现，确定模块、第一创建模块、第二创建模块、第一调整模块、修复模块、获取模块、监测模块、检测模块和第二调整模块可以是可以由处理器执行存储器中的程序或指令来实现的技术方案来做详细的介绍:

图15是本发明实施例提供的管理设备20的结构示意图。所述管理设备20包括处理器210、存储器250和输入/输出I/O设备230，存储器250可以包括只读存储器和随机存取存储器，并向处理器210提供操作指令和数据。存储器250的一部分还可以包括非易失性随机存取存储器（NVRAM）。

在一些实施方式中，存储器250存储了如下的元素，可执行模块或者数据

结构，或者他们的子集，或者他们的扩展集:

在本发明实施例中，通过调用存储器250存储的操作指令（该操作指令可存储在操作系统中），

为托管到所述云平台上的应用创建主实例节点，所述主实例节点用于运行所述应用的实例;

根据所述主实例节点的运行启动信息，确定所述应用的备实例节点所需要的最小资源配置规格;所述应用的备实例节点用于在所述应用的主实例节点故障时接替所述应用的主实例节点运行所述应用的实例;

根据所述最小资源配置规格，创建所述应用的备实例节点，其中，所述应用的备实例节点的资源配置规格小于所述应用的主实例节点的资源配置规格。

可见，与现有技术中为了保证云平台上应用的高可用性，备实例节点需要占用与主实例节点相同的资源相比，本发明实施例提供的管理设备，可以在保证应用托管到云平台上高可用的基础上，可以减少备用资源的占用，提高云平台上应用的托管能力和规模。

处理器210控制管理设备20的操作，处理器210还可以称为CPU（Central Processing Unit，中央处理单元）。存储器250可以包括只读存储器和随机存取存储器，并向处理器210提供指令和数据。存储器250的一部分还可以包括非易失性随机存取存储器（NVRAM）。具体的应用中管理设备20的各个组件通过总线系统220耦合在一起，其中总线系统220除包括数据总线之外，还可以包括电源总线、控制总线和状态信号总线等。但是为了清楚说明起见，在图中将各种总线都标为总线系统220。

上述本发明实施例揭示的方法可以应用于处理器210中，或者由处理器210实现。处理器210可能是一种集成电路芯片，具有信号的处理能力。在实现过程中，上述方法的各步骤可以通过处理器210中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器210可以是通用处理器、数字信号处理器（DSP）、专用集成电路（ASIC）、现成可编程门阵列（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。可以实现或者执行本发明实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理

器或者该处理器也可以是任何常规的处理器等。结合本发明实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成，或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器，闪存、只读存储器，可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器250，处理器210读取存储器250中的信息，结合其硬件完成上述方法的步骤。

可选地，处理器210还用于当所述主实例节点故障时，将所述备实例节点的资源配置规格调整到与所述主实例节点的资源配置规格相同，并将所述备实例节点设置为工作实例节点，以接替所述应用的主实例节点运行所述应用的实例。

本发明实施例中，在主实例节点正常时，备实例节点的资源占用率很低，在主实例节点故障时，迅速扩容来接替主实例节点的工作，从而满足应用的访问需求，提高云平台上应用的托管能力和规模。

可选地，处理器210还用于根据所述最小资源配置规格，再次为所述应用创建备实例节点。

本发明实施例中，当最小化的备实例节点数量减少时，及时的动态补充，避免了下次工作实例节点出现故障时，导致无法响应应用的访问请求，从而提高了可靠性，提高了用户体验。

可选地，处理器210还用于修复发生故障的所述主实例节点，以得到修复后的主实例节点；将所述修复后的主实例节点的资源配置规格调整到与所述最小资源配置规格相同，并将所述修复后的主实例节点设置为所述应用的备实例节点，以便于在所述工作实例节点故障时接替所述工作实例节点运行所述应用的实例。

本发明实施例中，还可以及时修复发生故障的主实例节点，这样进一步提高了云平台中资源的利用率。

可选地，处理器210还用于获取用户设备对所述应用的访问量；当所述应用的访问量超过预置阈值时，根据所述最小资源配置规格，为所述应用创建至少两个备实例节点，其中，创建的每个备实例节点的资源配置规格与所述最小

资源配置规格相同。

本发明实施例中，可以根据应用的访问量创建相应数量的最小化的备实例节点，这样，可以预防应用压力过大时，导致突发的意外故障，从而进一步提高了云平台的高可用性。

可选地，处理器210还用于监测已创建的所述应用的备实例节点；当监测到所述应用的备实例节点发生故障时，创建与发生故障的备实例节点数量相同的备实例节点。

本发明实施例中，还可以及时监测已创建的备实例节点中的故障实例节点，及时创建新的可用备实例节点，以避免主实例节点出现故障时，没有可用的备实例节点接替主实例节点的工作，从而进一步提高了云平台的高可用性。

可选地，处理器210还用于检测运行所述应用的实例所需的资源量；根据所述所需的资源量，调整已创建的所述应用的备实例节点的资源配置规格。

本发明实施例中，针对已创建的备实例节点，还可以根据应用的运行状态，对齐资源配置规格进行动态调整，如果应用的需求量低时，还可以进一步节省资源，如果应用的需求量高时，还可以提高主实例节点故障时的接替效率。

需要说明的是，本发明实施例提供的云平台的管理设备，具体可以为云计算系统中的一台云主机，该云主机可以为运行在物理机上的虚拟机。如图16所示，物理机1200包括硬件层100，运行在硬件层100之上的VMM（Virtual Machine Monitor，虚拟机监视器）110，以及运行在VMM 110之上的宿主机Host 1201和若干虚拟机（VM，Virtual Machine），其中，硬件层包括但不限于：I/O设备、CPU和memory。本发明实施例提供的云平台的管理设备具体可以为物理机1200中的一台虚拟机，比如VM 1202，VM 1202上运行有一个或多个云应用，其中，每一个云应用都用于实现相应的业务功能，比如数据库应用、地图应用等等，这些云应用可以由开发者开发然后部署到云计算系统中。此外VM1202还运行有可以执行程序，VM 1202通过运行该可执行程序，并在程序运行的过程中通过宿主机Host 1201来调用硬件层100的硬件资源，以实现云平台的管理设备的确定模块、第一创建模块、第二创建模块、第一调整模块、修复模块、获取模块、监测模块、检测模块和第二调整模块的功能，具体而言，

确定模块、第一创建模块、第二创建模块、第一调整模块、修复模块、获取模块、监测模块、检测模块和第二调整模块可以以软件模块或函数的形式被包含在上述可执行程序中，比如该可执行程序可以包括：确定模块、第一创建模块、第二创建模块、第一调整模块、修复模块、获取模块、监测模块、检测模块和第二调整模块，VM1202通过调用硬件层100中的CPU、Memory等资源，以运行该可执行程序，从而实现确定模块、第一创建模块、第二创建模块、第一调整模块、修复模块、获取模块、监测模块、检测模块和第二调整模块的功能。

图15对应的实施例以及其他可选实施例可以参阅图1-图13部分描述进行理解，本处不做过多赘述。

本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令相关的硬件（例如处理器）来完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：ROM、RAM、磁盘或光盘等。

以上对本发明实施例所提供的实例节点管理的方法、管理设备进行了详细介绍，本文中应用了具体个例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的方法及其核心思想；同时，对于本领域的一般技术人员，依据本发明的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本发明的限制。

权 利 要 求

1、一种实例节点管理的方法，其特征在于，所述方法应用于云平台的管理设备，所述方法包括:

为托管到所述云平台上的应用创建主实例节点，所述主实例节点用于运行所述应用的实例;

根据所述主实例节点的运行启动信息，确定所述应用的备实例节点所需要的最小资源配置规格; 所述应用的备实例节点用于在所述应用的主实例节点故障时接替所述应用的主实例节点运行所述应用的实例;

根据所述最小资源配置规格，创建所述应用的备实例节点，其中，所述应用的备实例节点的资源配置规格小于所述应用的主实例节点的资源配置规格。

2、根据权利要求1所述的方法，其特征在于，所述根据所述最小资源配置规格，创建所述应用的备实例节点之后，所述方法还包括:

当所述主实例节点故障时，将所述备实例节点的资源配置规格调整到与所述主实例节点的资源配置规格相同，并将所述备实例节点设置为工作实例节点，以接替所述应用的主实例节点运行所述应用的实例。

3、根据权利要求2所述的方法，其特征在于，在将所述备实例节点的资源配置规格调整到与所述应用的主实例节点的资源配置规格相同之后，所述方法还包括:

根据所述最小资源配置规格，再次为所述应用创建备实例节点。

4、根据权利要求2所述的方法，其特征在于，所述将所述备实例节点的资源配置规格调整到与所述应用的主实例节点的资源配置规格相同之后，所述方法还包括:

修复发生故障的所述主实例节点，以得到修复后的主实例节点;

将所述修复后的主实例节点的资源配置规格调整到与所述最小资源配置规格相同，并将所述修复后的主实例节点设置为所述应用的备实例节点，以便于在所述工作实例节点故障时接替所述工作实例节点运行所述应用的实例。

5、根据权利要求1-4任一所述的方法，其特征在于，所述确定所述应用的备实例节点的最小资源配置规格之后，所述方法还包括:

获取用户设备对所述应用的访问量;

所述根据所述最小资源配置规格，创建所述应用的备实例节点，包括:

当所述应用的访问量超过预置阈值时，根据所述最小资源配置规格，为所述应用创建至少两个备实例节点，其中，创建的每个备实例节点的资源配置规格与所述最小资源配置规格相同。

6、根据权利要求1-4任一所述的方法，其特征在于，所述根据所述最小资源配置规格，创建所述应用的备实例节点之后，所述方法还包括:

监测已创建的所述应用的备实例节点;

当监测到所述应用的备实例节点发生故障时，创建与发生故障的备实例节点数量相同的备实例节点。

7、根据权利要求1-4任一所述的方法，其特征在于，所述根据所述最小资源配置规格，创建所述应用的备实例节点之后，所述方法还包括:

检测运行所述应用的实例所需的资源量;

根据所述所需的资源量，调整已创建的所述应用的备实例节点的资源配置规格。

8、一种云平台的管理设备，其特征在于，所述管理设备包括:

第一创建模块，用于为托管到所述云平台上的应用创建主实例节点，所述主实例节点用于运行所述应用的实例;

确定模块，用于根据所述第一创建模块创建的所述主实例节点的运行启动信息，确定所述应用的备实例节点所需要的最小资源配置规格;所述应用的备实例节点用于在所述应用的主实例节点故障时接替所述应用的主实例节点运行所述应用的实例;

第二创建模块，用于根据所述确定模块确定的所述最小资源配置规格，创建所述应用的备实例节点，其中，所述应用的备实例节点的资源配置规格小于所述应用的主实例节点的资源配置规格。

9、根据权利要求8所述的管理设备，其特征在于，所述管理设备还包括:

第一调整模块，用于当所述主实例节点故障时，将所述第二创建模块创建的所述备实例节点的资源配置规格调整到与所述主实例节点的资源配置规格

相同，并将所述备实例节点设置为工作实例节点，以接替所述应用的主实例节点运行所述应用的实例。

10、根据权利要求9所述的管理设备，其特征在于，

所述第二创建模块，还用于根据所述最小资源配置规格，再次为所述应用创建备实例节点。

11、根据权利要求9所述的管理设备，其特征在于，所述管理设备还包括：修复模块，

所述修复模块，用于修复发生故障的所述主实例节点，以得到修复后的主实例节点；

所述第一调整模块，还用于将所述修复后的主实例节点的资源配置规格调整到所述应用的最小资源配置规格，并将所述修复后的主实例节点设置为所述应用的备实例节点，以便于在所述工作实例节点故障时接替所述工作实例节点运行所述应用的实例。

12、根据权利要求8-11任一所述的管理设备，其特征在于，所述管理设备还包括获取模块，

所述获取模块，用于获取用户设备对所述应用的访问量；

所述第二创建模块，具体用于当所述获取模块获取的所述应用的访问量超过预置阈值时，根据所述确定模块确定的所述最小资源配置规格，为所述应用创建至少两个备实例节点，其中，创建的每个备实例节点的资源配置规格与所述最小资源配置规格相同。

13、根据权利要求8-11任一所述的管理设备，其特征在于，所述管理设备还包括监测模块，

所述监测模块，用于监测所述第二创建模块已创建的所述应用的备实例节点；

所述第二创建模块，还用于当所述监测模块监测到所述应用的备实例节点发生故障时，创建与发生故障的备实例节点数量相同的备实例节点。

14、根据权利要求8-11任一所述的管理设备，其特征在于，所述管理设备还包括检测模块和第二调整模块，

所述检测模块，用于检测运行所述应用的实例所需的资源量；

所述第二调整模块，用于根据所述检测模块检测到的所述所需的资源量，调整已创建的所述应用的备实例节点的资源配置规格。

5

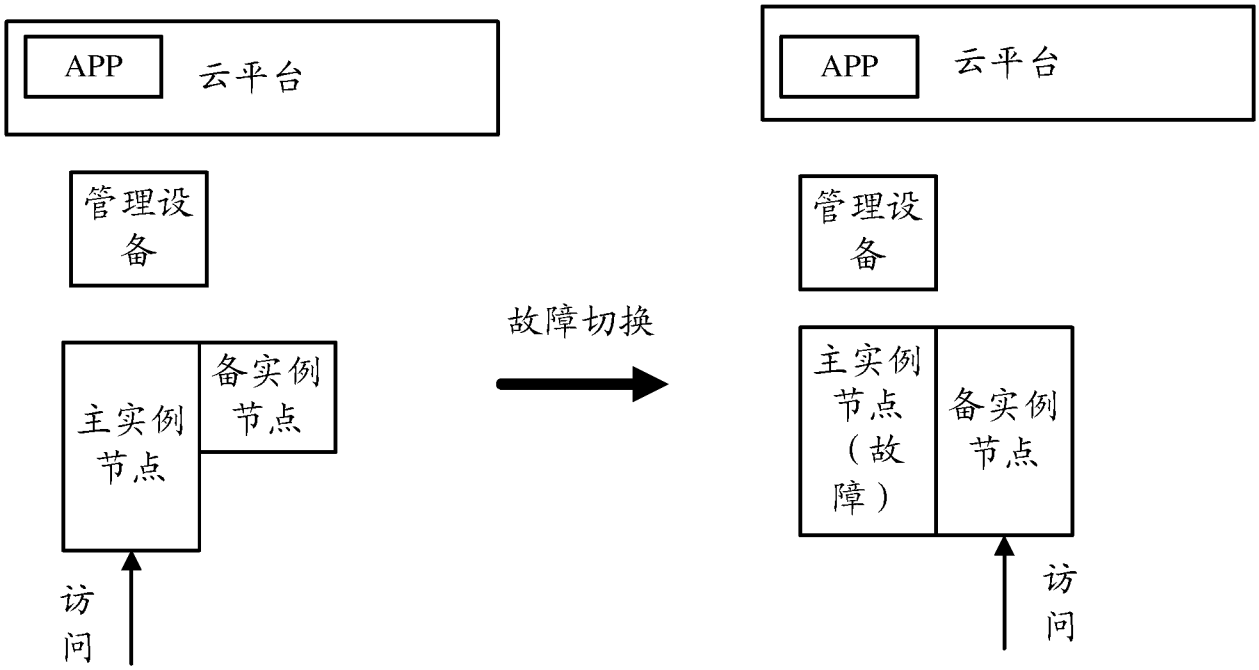
APP
云平台
管理设备
主实例节点
备实例节点
访问
故障切换
APP
云平台
管理设备
主实例节点（故障）
备实例节点
访问

图 1

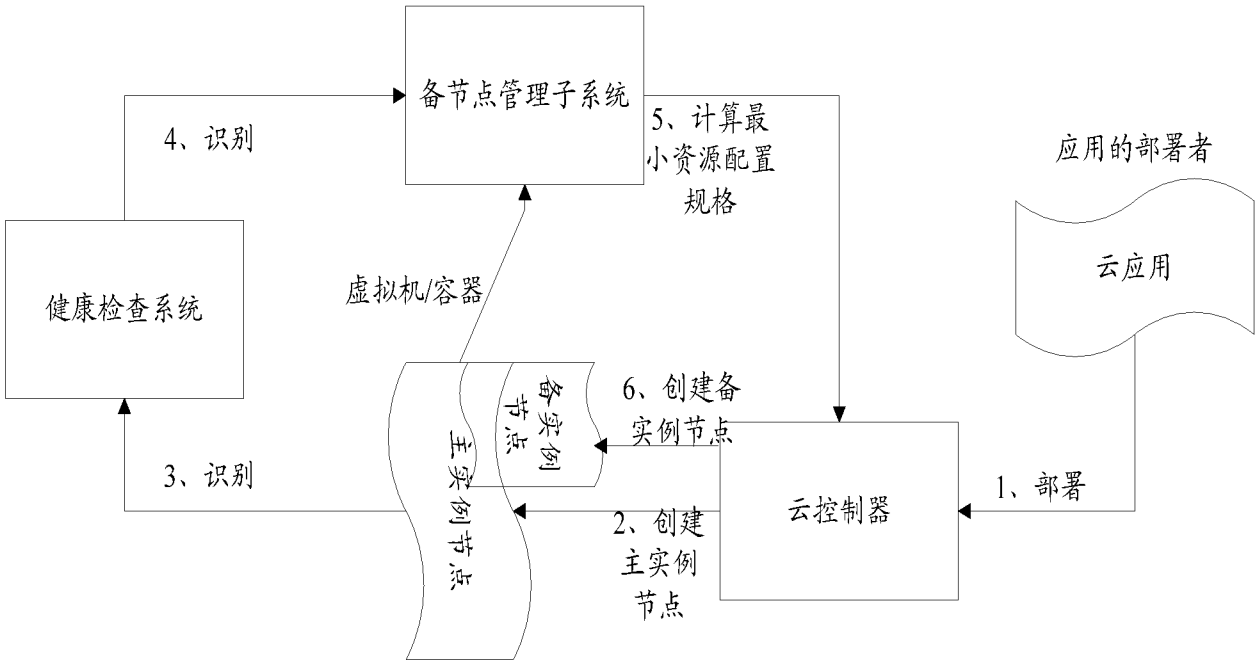
备节点管理子系统
4、识别
5、计算最小资源配置规格
应用的部署者
云应用
健康检查系统
虚拟机/容器
备实例节点
主实例节点
6、创建备实例节点
3、识别
1、部署
2、创建主实例节点
云控制器

图 2

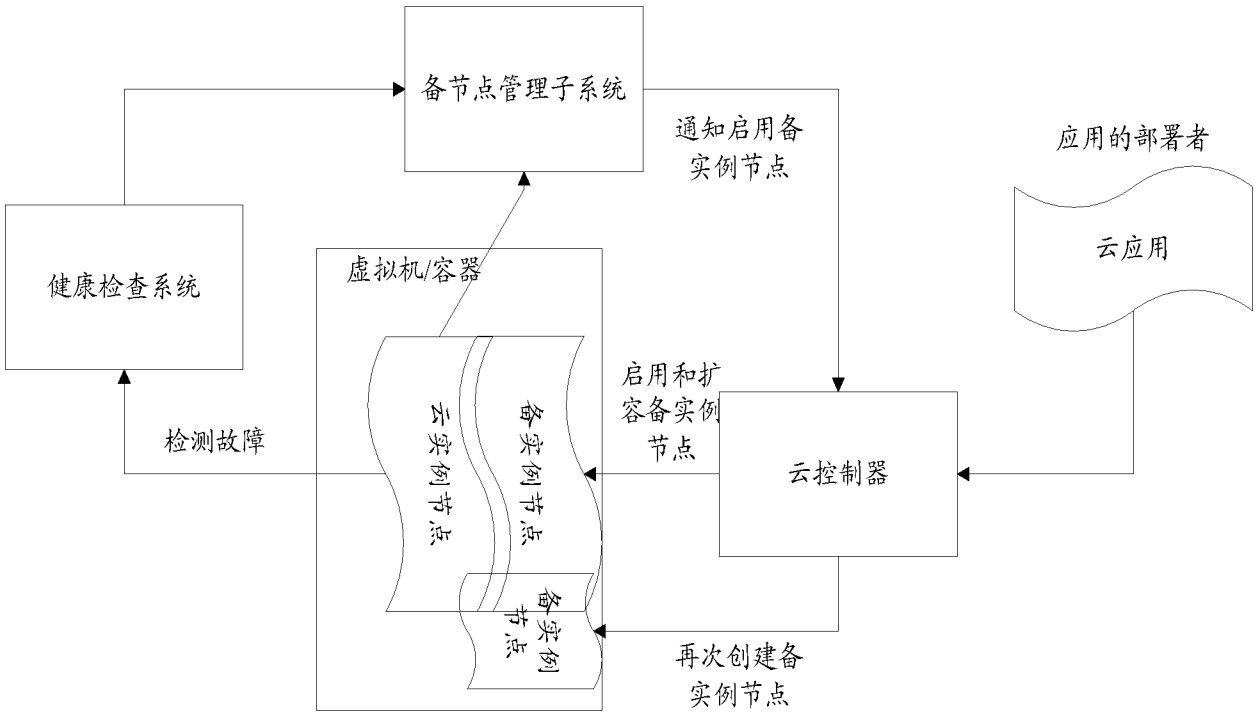
备节点管理子系统
通知启用备实例节点
应用的部署者
云应用
健康检查系统
虚拟机/容器
云实例节点
备实例节点
备实例节点
启用和扩容备实例节点
检测故障
云控制器
再次创建备实例节点

图 3

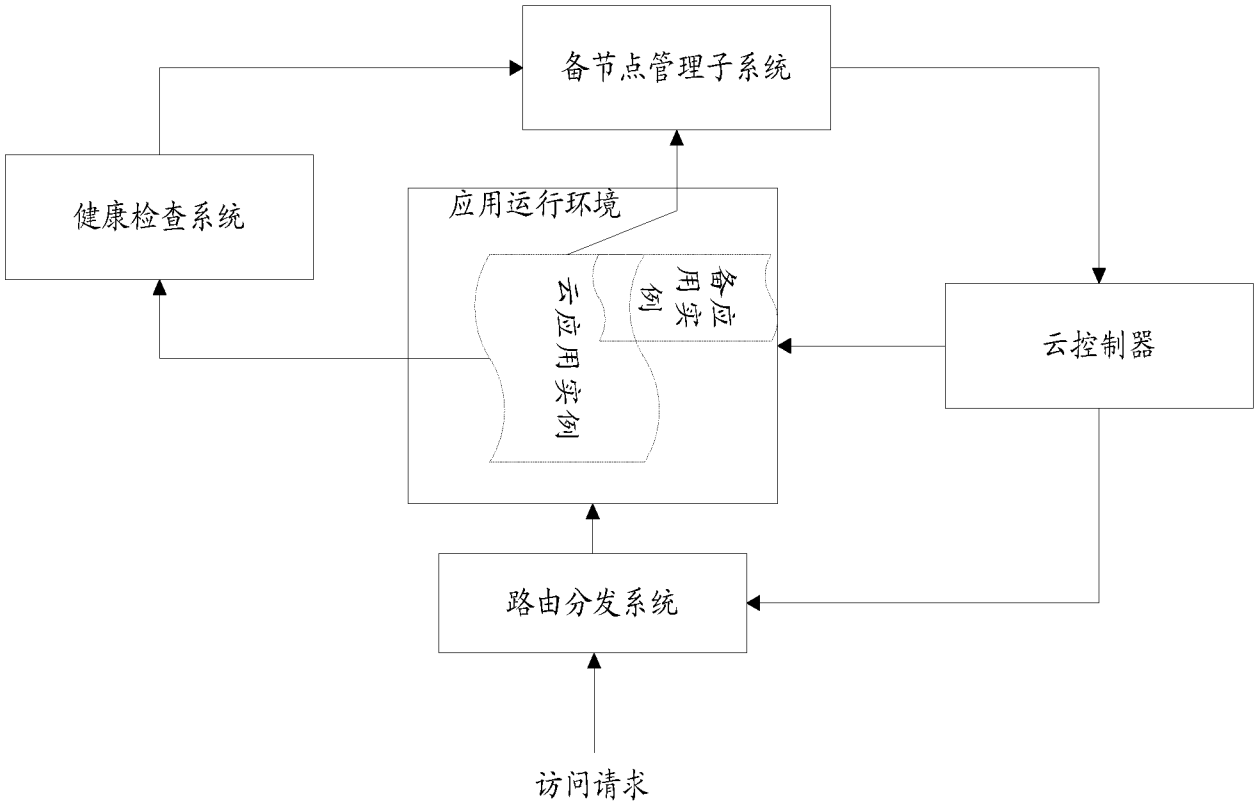
备节点管理子系统
健康检查系统
应用运行环境
云应用实例
备应用实例
云控制器
路由分发系统
访问请求

图 4

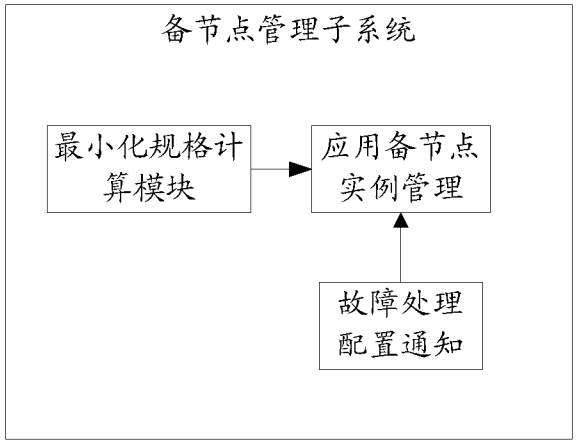
备节点管理子系统
最小化规格计算模块
应用备节点实例管理
故障处理配置通知

图 5

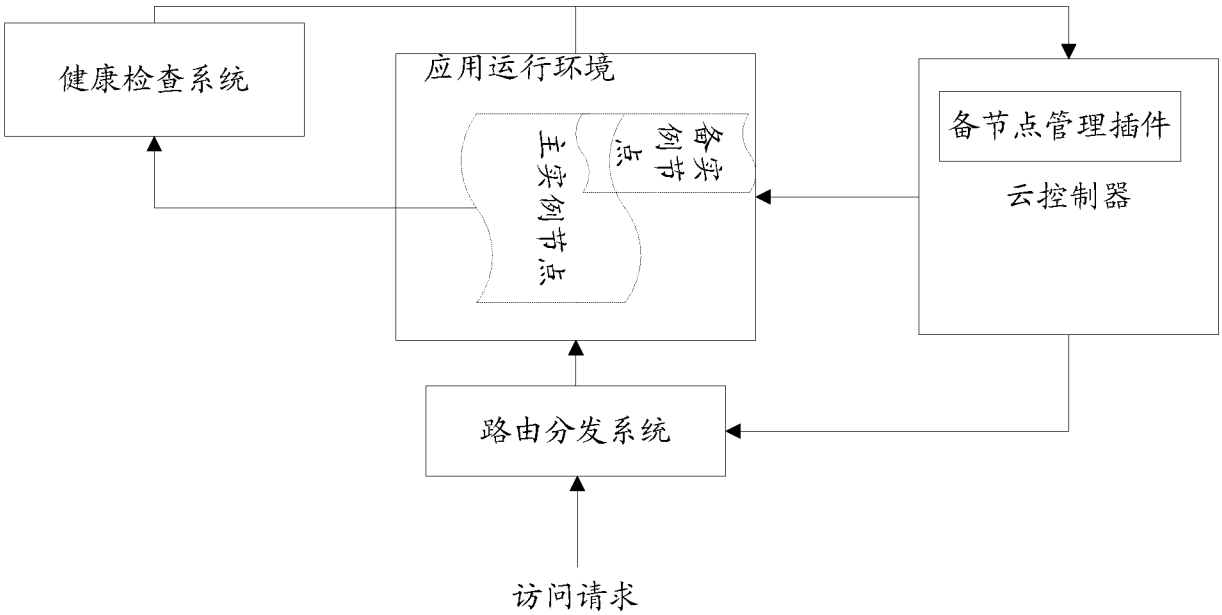
健康检查系统
应用运行环境
主实例节点
备实例节点
备节点管理插件
云控制器
路由分发系统
访问请求

图 6

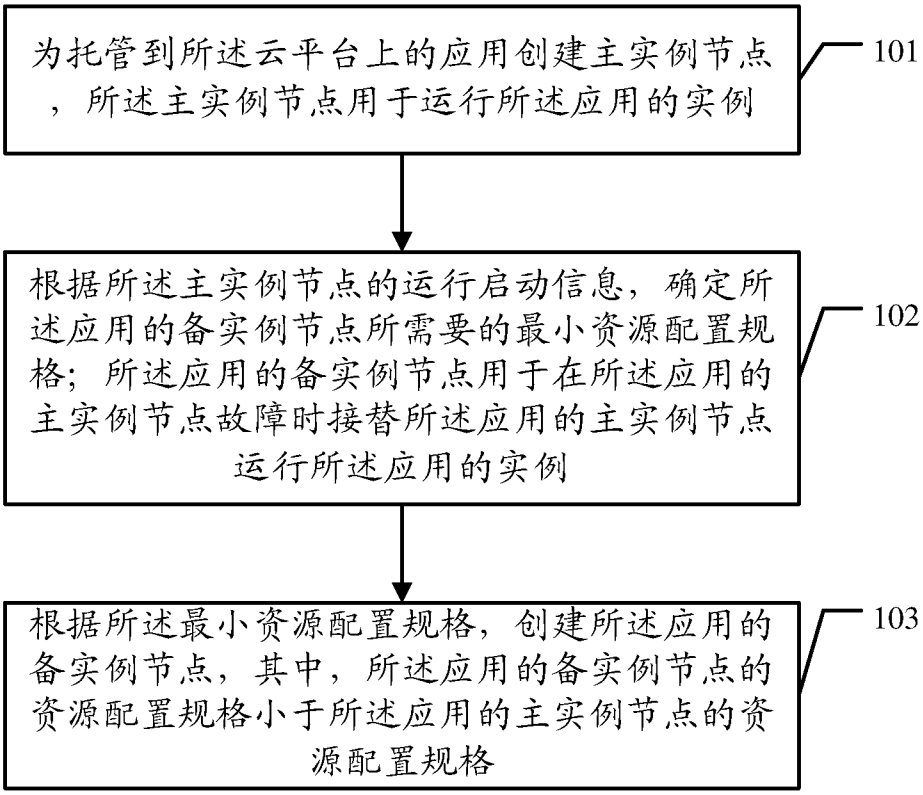
为托管到所述云平台上的应用创建主实例节点，所述主实例节点用于运行所述应用的实例
101
根据所述主实例节点的运行启动信息，确定所述应用的备实例节点所需要的最小资源配置规格；所述应用的备实例节点用于在所述应用的主实例节点故障时接替所述应用的主实例节点运行所述应用的实例
102
根据所述最小资源配置规格，创建所述应用的备实例节点，其中，所述应用的备实例节点的资源配置规格小于所述应用的主实例节点的资源配置规格
103

图 7

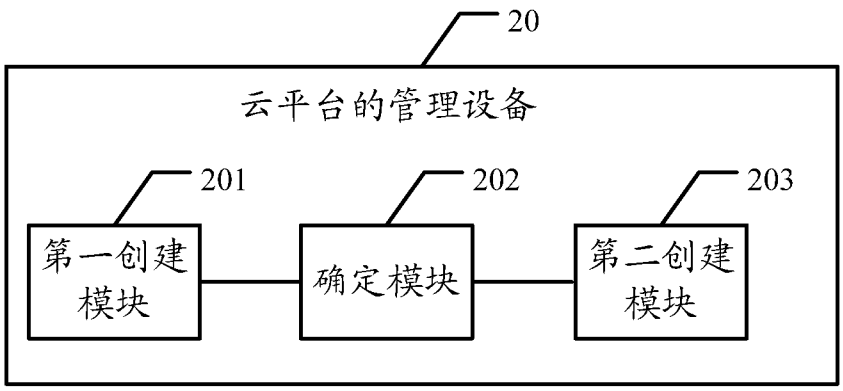
20
云平台的管理设备
201
第一创建模块
202
确定模块
203
第二创建模块

图 8

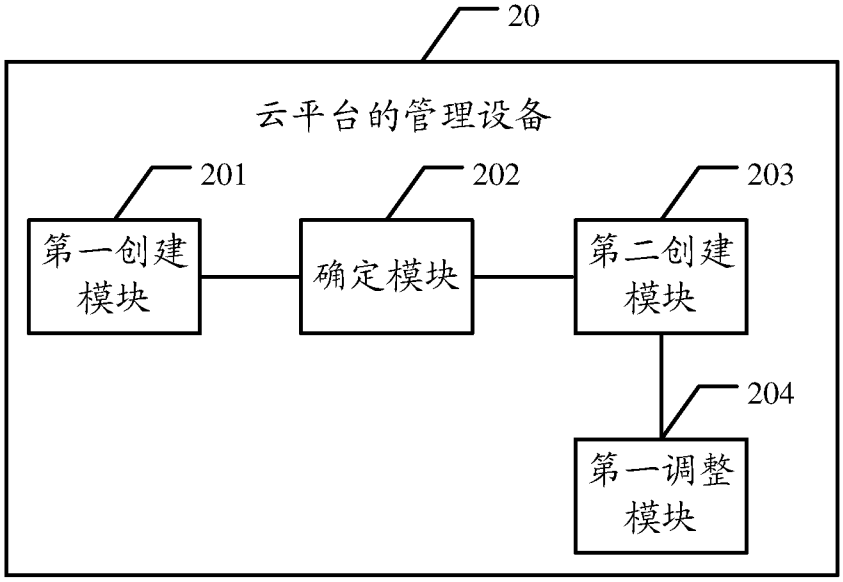
20
云平台的管理设备
201
第一创建模块
202
确定模块
203
第二创建模块
204
第一调整模块

图 9

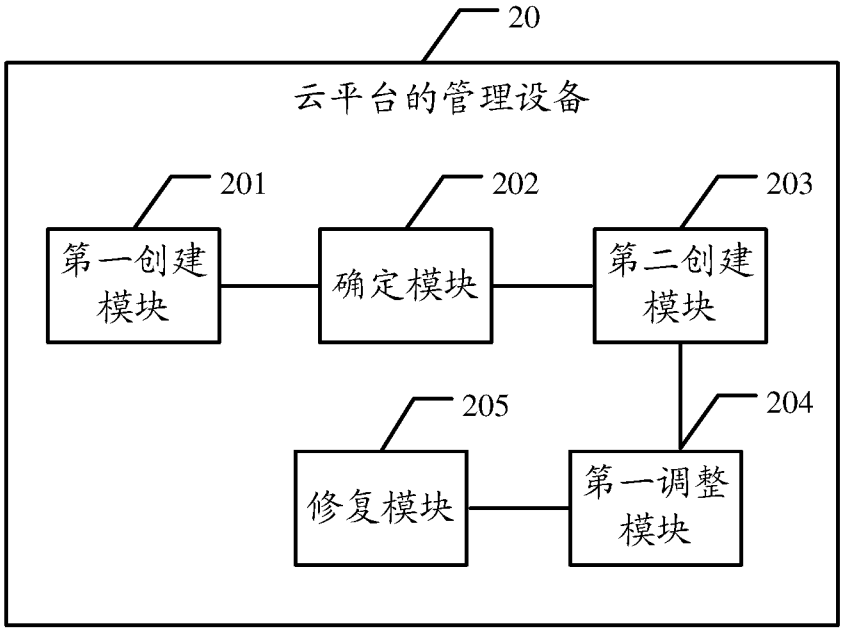
20
云平台的管理设备
201
第一创建模块
202
确定模块
203
第二创建模块
205
修复模块
204
第一调整模块

图 10

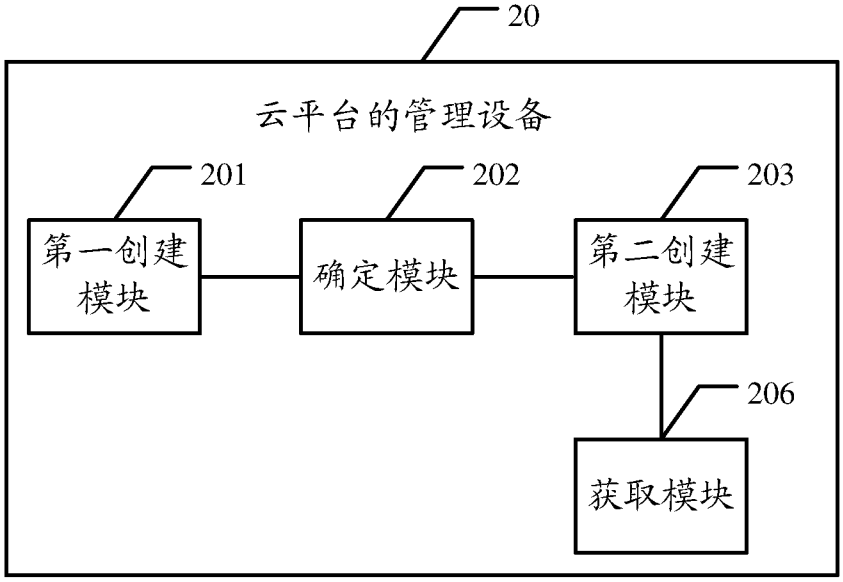
20
云平台的管理设备
201
第一创建模块
202
确定模块
203
第二创建模块
206
获取模块

图 11

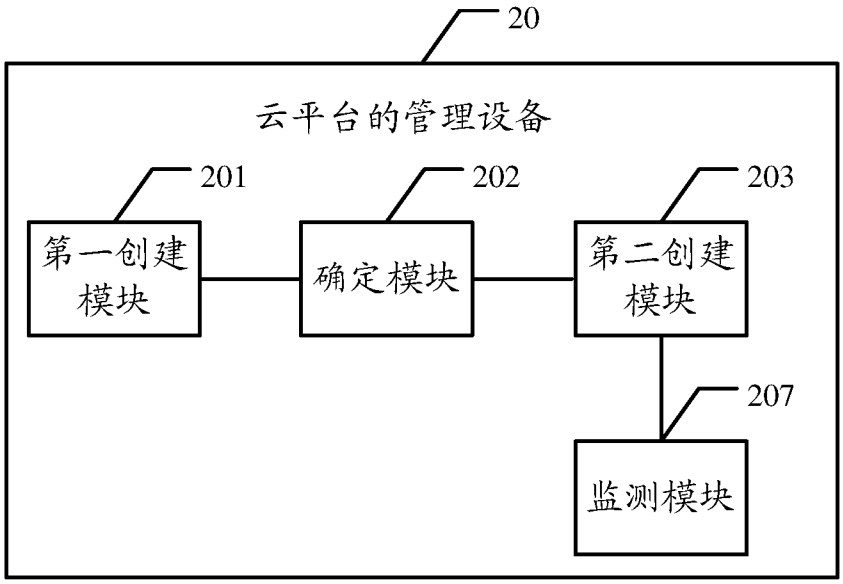
20
云平台的管理设备
201
第一创建模块
202
确定模块
203
第二创建模块
207
监测模块

图 12

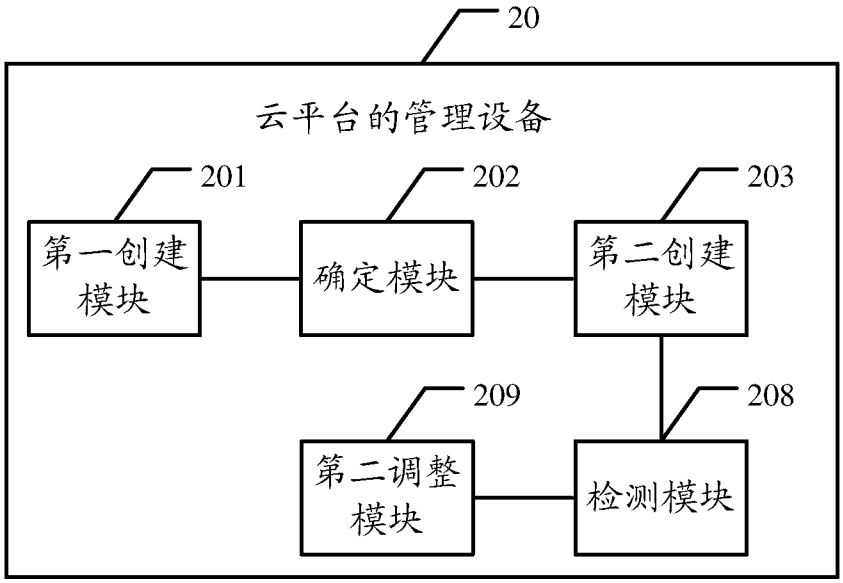
20
云平台的管理设备
201
第一创建模块
202
确定模块
203
第二创建模块
209
第二调整模块
208
检测模块

图 13

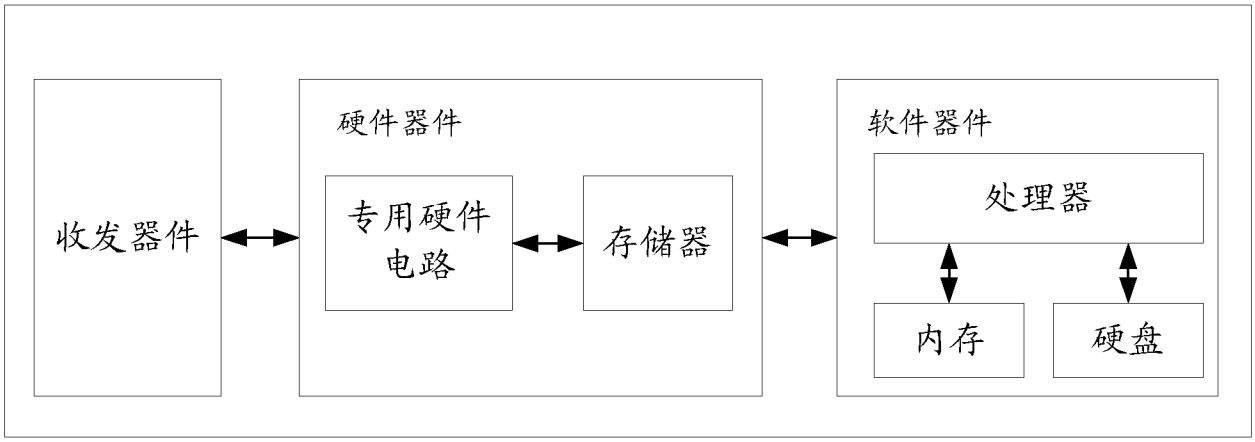
收发器件
硬件器件
专用硬件电路
存储器
软件器件
处理器
内存
硬盘

图 14

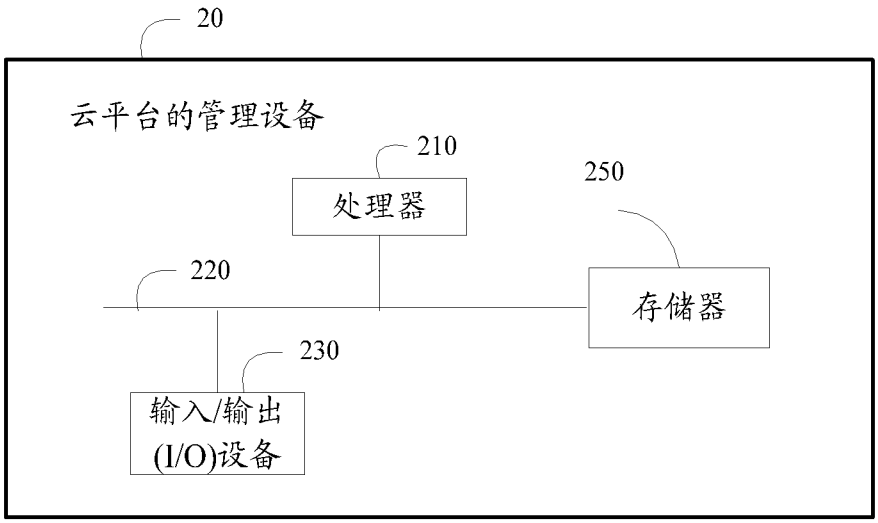
20
云平台的管理设备
210
处理器
250
存储器
220
230
输入/输出
(I/O)设备

图 15

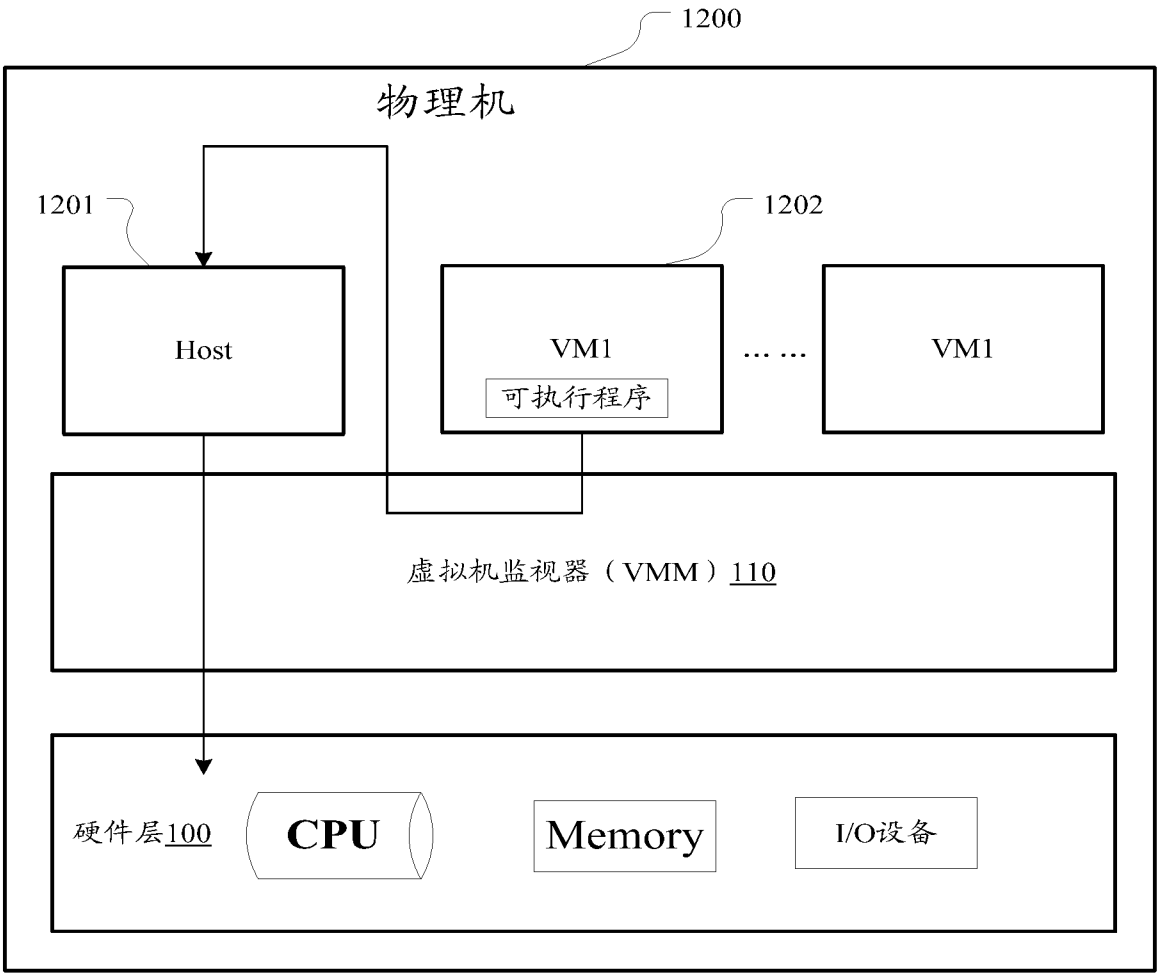
1200
物理机
1201
Host
1202
VM1
可执行程序
… …
VM1
虚拟机监视器（VMM）110
硬件层100
CPU
Memory
I/O设备

图 16

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2015/092667

A. CLASSIFICATION OF SUBJECT MATTER

H04L 29/08 (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)

H04L

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)

CNKI; CNABS; CNTXT; VEN: main/standby, main-subsidiary, node, specification, minimum, backup, master, main, instantiation, configure, deploy, resource, min

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 104836850 A (HUAWEI TECHNOLOGIES CO., LTD.), 12 August 2015 (12.08.2015), claims 1-14	1-14
X	CN 102497288 A (HUAWEI TECHNOLOGIES CO., LTD.), 13 June 2012 (13.06.2012), description, paragraphs [0033]-[0065]	1-14
A	US 2015082308 A1 (NTT DOCOMO INC.), 19 March 2015 (19.03.2015), the whole document	1-14

☐ 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
06 January 2016 (06.01.2016)	**22 January 2016 (22.01.2016)**
Name and mailing address of the ISA/CN: 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 **LI, Yan** Telephone No.: (86-10) **62088422**

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

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2015/092667

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104836850 A	12 August 2015	None	
CN 102497288 A	13 June 2012	None	
US 2015082308 A1	19 March 2015	JP 2015056182 A	23 March 2015
		EP 2849064 A1	18 March 2015
		CN 104468688 A	25 March 2015

国际检索报告

国际申请号
PCT/CN2015/092667

A. 主题的分类

H04L 29/08(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)

H04L

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））

CNKI；CNABS；CNTXT；VEN：备份，主备，主从，实例，节点，配置，规格，最小，backup，master，main，instantiation，configure，deploy，resource，min

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
PX	CN 104836850 A (华为技术有限公司) 2015年 8月 12日 (2015-08-12) 权利要求1-14	1-14
X	CN 102497288 A (华为技术有限公司) 2012年 6月 13日 (2012-06-13) 说明书[0033]-[0065]段	1-14
A	US 2015082308 A1 (NTT DOCOMO INC) 2015年 3月 19日 (2015-03-19) 全文	1-14

☐ 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
2016年 1月 6日	2016年 1月 22日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 李妍 电话号码 (86-10)62088422

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2015/092667

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	104836850	A	2015年 8月 12日	无			
CN	102497288	A	2012年 6月 13日	无			
US	2015082308	A1	2015年 3月 19日	JP	2015056182	A	2015年 3月 23日
				EP	2849064	A1	2015年 3月 18日
				CN	104468688	A	2015年 3月 25日