(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105052111 B
(45)授权公告日 2019.08.16

(21)申请号 201480004300.5

(22)申请日 2014.01.08

(65)同一申请的已公布的文献号
申请公布号 CN 105052111 A

(43)申请公布日 2015.11.11

(30)优先权数据
13/737,745 2013.01.09 US

(85)PCT国际申请进入国家阶段日
2015.07.08

(86)PCT国际申请的申请数据
PCT/US2014/010571 2014.01.08

(87)PCT国际申请的公布数据
WO2014/110062 EN 2014.07.17

(73)专利权人 微软技术许可有限责任公司
地址 美国华盛顿州

(72)发明人 S·P·里瓦斯卡 M·U·阿扎德 S·塞耶德 C·P·阿尔米达 A·玛尼

(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 杨洁

(51)Int.Cl.
H04L 29/08(2006.01)
G06F 9/48(2006.01)

(56)对比文件
US 2009/0276771 A1,2009.11.05,
WO 2011/150195 A2,2011.12.01,
CN 102521009 A,2012.06.27,

审查员 林桂荣

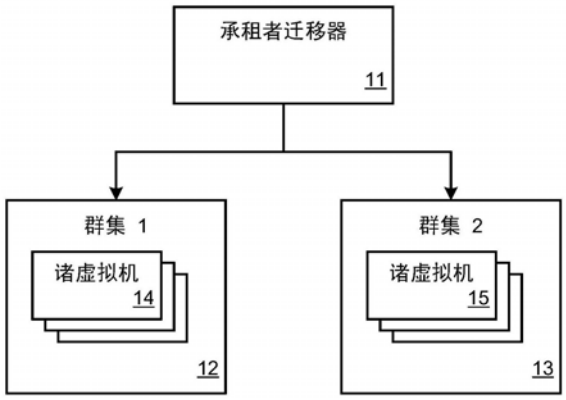
权利要求书1页 说明书5页 附图5页

(54)发明名称

跨群集边界的服务迁移

(57)摘要

各实施例提供了跨不同群集的服务迁移以平衡利用率并且满足顾客需求。不同的服务迁移选项可以有停机时间或没有停机时间地被执行。服务的人工产物被移至新的目的地群集。服务在新目的地群集上被创建并且被发起,以使服务几乎准备好启动。在一实施例中,服务在旧群集上被停止并且在新群集上被启动。在停止服务之后,DNS被更新以指向新群集上的服务。在另一实施例中,服务在旧群集上被停止并且以同一IP地址在新群集上被启动,以避免DNS重新编程及相关联的延迟。在进一步实施例中,通过将服务逐部分地自一个群集移至另一个群集而无停机时间地执行迁移。



CN 105052111 B

1.一种计算机实现的方法,包括:

复制在计算环境中的第一群集上运行的服务的人工产物,所述第一群集包括第一组虚拟机;

使用所述人工产物在所述计算环境中的第二群集上创建所述服务,所述第二群集包括第二组虚拟机;

在所述第一群集上停止所述服务的所选部分;

在所述第二群集上启动所述服务的相对应的所选部分;

将所述停止和启动步骤执行两次或更多次,直到所选服务的所有部分已在所述第一群集上被停止且所述服务的所有相对应部分已在所述第二群集上被启动;以及

删除所述第一群集上的服务;

其中,所述方法还包括更新网络以连接所述第一群集上的服务和所述第二群集上的服务的已启动部分。

2.如权利要求1所述的方法,还包括:

向所述服务指派IP地址;以及

使用所述IP地址在所述第一群集和所述第二群集两者上同时支持所述服务。

3.如权利要求1所述的方法,还包括:

向所述第二群集上的服务指派IP地址;以及

更新网络位置服务以便将所述服务与所述第二群集上的IP地址相关联。

4.如权利要求3所述的方法,其特征在于,所述网络位置服务是域名系统DNS。

5.如权利要求1所述的方法,还包括:

在停止所述第一群集上的服务之后,从所述第一群集移除与所述服务相关联的IP地址;以及

向所述第二群集上的服务指派IP地址。

6.如权利要求1所述的方法,其特征在于,所复制的人工产物包括代码、证书和型号中的一个或多个。

7.如权利要求1所述的方法,其特征在于,所述第一群集和所述第二群集位于数据中心内。

8.如权利要求1所述的方法,其特征在于,承租者迁移器执行复制、创建、停止、启动和删除步骤。

跨群集边界的服务迁移

[0001] 背景

[0002] 大规模数据中心一般包括运行标准软件包的集合的硬件机器的有组织群集，诸如web服务器、数据库服务器等等。出于容错和管理原因，数据中心中的机器一般被分成多个群集，所述多个群集独立地由协调各软件应用的各资源的一框架监控和管理。在一实施例中，框架可以是例如供应、支持、监控和命令构成数据中心的各虚拟机(VM)和物理服务器的Windows Azure™结构控制器。

[0003] 在现有的数据中心中，每个承租者在其整个生命周期被部署至单个群集，允许承租者的部署被单个框架管理。然而，该配置可以限制承租者的成长，因为扩展被限制于该单个群集内的机器。承租者和群集间的紧密耦合要求数据中心操作者将群集的容量维持在一级别，该级别将满足部署在该群集上的承租者的潜在的将来要求。通常，这导致群集在预期可能的将来需求时以低电流利用率进行操作。即使多余容量被维持时，这仅仅改进了承租者的将来需求将被支持的可能性。不保证承租者规模请求将被限制于已保留的容量并且，因此，有时承租者可能不能获得所需的容量。

[0004] 将服务限制于一个群集也为该服务创建单个故障点。如果控制该群集的框架发生故障，则整个群集将发生故障，并且该群集上所支持的所有服务都将不可用。

[0005] 概述

[0006] 提供本概述是为了以精简的形式介绍将在以下详细描述中进一步描述的一些概念。本概述并不旨在标识所要求保护主题的关键特征或必要特征，也不旨在用于限制所要求保护主题的范围。

[0007] 本发明的各实施例允许承租者的服务有停机时间或无停机时间地在多个群集间移动。服务与群集上的特定IP地址相关联。用户使用一域名接入服务，该域名通过域名系统(DNS)或其他网络位置服务转换成IP地址。在服务在各群集间移动时，服务的IP地址可能改变或可能不改变。

[0008] 服务可以通过以下步骤有停机时间地被迁移：在新群集中发起服务的新实例；等待该新实例准备就绪；然后停止原始实例；并将服务的DNS名称指向与服务在新群集上的新部署相对应的IP地址。

[0009] 或者，服务可以有停机时间地被迁移至新群集，并且保留原始IP地址。这会避免在DNS高速缓存被重新填充的同时对DNS及相关延迟重新编程的需求。

[0010] 迁移服务的进一步替代方案是通过以下步骤来无停机时间地执行迁移：逐部分地移动服务使得服务在迁移过程期间总是在群集中的一者或两者中运行。

[0011] 附图简述

[0012] 为了进一步阐明本发明的各实施例的以上和其他优点和特征，将参考附图来呈现本发明的各实施例的更具体的描述。可以理解，这些附图只描绘本发明的典型实施例，因此将不被认为是对其范围的限制。本发明将通过使用附图用附加特征和细节来描述和解释，附图中：

[0013] 图1是图示用于跨不同群集移动服务的承租者迁移器的框图。

[0014] 图2图示了具有服务停机时间且需要DNS重新编程的服务迁移。

[0015] 图3图示了具有服务停机时间但保留服务的IP地址的服务迁移。

[0016] 图4图示了消除服务停机时间并保留服务的IP地址的服务迁移。

[0017] 图5图示了用于承租者迁移的适当的计算和联网环境的示例。

[0018] 详细描述

[0019] 图1是图示用于跨不同群集12、13移动服务的承租者迁移器11的框图。承租者迁移器11连接至数据中心中的所有群集。一旦数据中心操作者决定在各群集间移动服务，例如，为了平衡利用率或为了满足承租者需求，承租者迁移器11就标识该服务的正确目的地群集。目的地群集的选择可以基于各因素，诸如潜在目的地群集的利用、服务所作的当前需求等等。一旦标识了目的地群集，承租者迁移器11就通过在原始群集和新群集上创建/删除VM 14、15上的实例来移动服务。

[0020] 承租者迁移器11控制如操作者所选择的那样是有停机时间还是没有停机时间地执行迁移。如果新IP地址被指派给服务则承租者迁移器11可以请求对DNS记录的更新，或者如果服务保持相同地址时则承租者迁移器11可以将IP地址移至新群集。服务存在性在迁移期间是互斥的。例如，当服务被迁移时，承租者迁移器11确保从顾客角度来看服务的两个实例绝不会都在运行。

[0021] 图2图示了根据一实施例的具有服务停机时间且需要DNS重新编程的服务迁移。承租者迁移器21已标识在群集22上运行的要被移至群集23的服务。旧服务被指派群集22上的一个旧IP地址。在步骤201中，承租者迁移器21标识并复制来自群集22的服务人工产物，诸如代码、比特、证书、型号等。通过使用这些人工产物，在步骤202中在群集23上创建新服务，但该服务未被启动。

[0022] 承租者迁移器21在步骤203中指示新群集23来发起新服务群集23在步骤204中选择适当的节点并且设立VM来运行该服务。群集23上的新IP地址被指派给该新服务。群集23在此时不启动该服务。承租者迁移器21在步骤206中等待该服务在新群集上被发起，这例如在步骤205中指示。

[0023] 一旦新服务已被发起，承租者迁移器21就在步骤207中停止旧服务，并接着在步骤208中启动新服务。在步骤209中从群集22删除旧服务，这为该群集上运行的其他服务打开空间以扩展或被添加。

[0024] 然后，承租者迁移器在步骤210中更新中央DNS记录以使该服务的域名指向群集23上的适当的新IP地址。DNS记录更新可以用步骤207和208同时执行，而同时旧服务被停止且新服务被启动。

[0025] 在步骤207中停止旧服务和在步骤208中启动新服务之间有一时间段服务将对于用户不可用。此外，如果用户使用域名来接入服务，则在DNS记录从服务的域名的旧IP地址被更新至新IP地址的同时，可能有附加延迟。由于DNS支持跨互联网分布的许多本地高速缓存，因此需要时间来更新全部这些高速缓存。一旦中央DNS记录被更新，则本地DNS高速缓存被清除并且用新IP地址来更新。在这些更新发生之前，用户将被定向至旧群集22，该旧群集22不再运行服务并，因此，使用该服务的尝试将失败

[0026] 图3图示了根据一实施例的具有服务停机时间但保留服务的IP地址的服务迁移。承租者迁移器31已标识在群集32上运行的要被移至群集33的服务。旧服务被指派群集32上

的一个IP地址。在步骤301中，承租者迁移器31标识并复制来自群集32的服务人工产物，诸如代码、比特、证书、型号等。通过使用这些人工产物，在步骤302中在群集33上创建新服务，但该服务未被启动。

[0027] 承租者迁移器31在步骤303中指示新群集33来发起新服务。群集33在步骤304中选择适当的节点并且设立VM来运行该服务。群集33在此时不启动该服务。承租者迁移器31在步骤306中等待该服务在新群集上被发起，这例如在步骤305中指示。

[0028] 一旦新服务已被发起，则承租者迁移器31在步骤307中停止旧服务。在步骤308中，服务的IP地址从群集32移除。

[0029] 服务的IP地址在步骤309中被添加至群集33，并且群集33上的新服务在步骤310上被启动。

[0030] 最后，在步骤311中从群集32删除旧服务，这为该群集上运行的其他服务打开空间以扩展或被添加。

[0031] 由于服务的IP地址尚未改变，因此承租者迁移器不需要如同图2所示的过程中所需的那样更新DNS记录。因此，在步骤307中停止旧服务和在步骤310中启动新服务之间有一时间段服务将对于用户不可用。然而，一旦新服务被启动，用户可能仍使用域名接入该服务，而不等待任何DNS记录更新延迟。本地DNS高速缓存将是准确的，因为服务的域名将仍旧与服务的相同IP地址相关联。

[0032] 图4图示了根据一实施例的消除服务停机时间且保留服务的IP地址的服务迁移。承租者迁移器41已标识在群集42上运行的要被移至群集43的服务。旧服务被指派群集42上的一个旧IP地址。在步骤401中，承租者迁移器41标识并复制来自群集42的服务人工产物，诸如代码、比特、证书、型号等。通过使用这些人工产物，在步骤402中在群集43上创建新服务，但该服务未被启动。

[0033] 承租者迁移器41在步骤403中指示新群集43来发起新服务。群集43在步骤404中选择适当的节点并且设立VM来运行该服务。同一IP地址在群集42和群集43两者上用于该服务。承租者迁移器41在步骤406中等待该服务在新群集上被发起，这例如在步骤405中指示。

[0034] 一旦新服务已被发起，则承租者迁移器41在步骤407中停止旧服务的一部分。然后，承租者迁移器41在步骤408中启动新服务的相应部分。网络也在步骤408中按需被更新以连接旧服务和新服务的已启动部分以及负载平衡器及其他路由组件，以允许它们跨群集42、43指向已启动的服务。不像图2和3所示的过程，服务的仅仅一部分(例如，所选数量的VM或实例)在步骤407中被停止且然后在步骤408中被启动。承租者迁移器在步骤409中等待在新群集上被启动的该部分准备就绪供使用。

[0035] 一旦新部分在步骤409中准备就绪，则承租者迁移器对于服务的下一部分重复(步骤410)步骤407－409。这些步骤在循环410中继续，直到该服务的全部已经零碎地从旧群集42移至新群集43。在一实施例中，在每次经过循环410期间逐一地移动值得服务的一个更新域。承租者会准备好在对服务的升级期间丢失升级域，因此那些分段可用于分割该服务用于群集间的迁移。

[0036] 在服务的所有部分已经在循环410中被移动之后，在步骤411中从群集42删除旧服务。

[0037] 由于服务的IP地址尚未改变，因此承租者迁移器不需要如同图2所示的过程中进

行的那样更新DNS记录。不存在服务在两个群集上均被停止的时间段。因此，服务将没有停机时间而总是对于用户可用。

[0038] 图5解说了其上可以实现图1-4的示例的适当的计算和联网环境的示例。例如，承租者迁移器11和/或VM 14、15可以主存在一个或多个计算系统500上。计算系统环境500只是合适的计算环境的一个示例，而非意在暗示对本发明的使用或功能性范围有任何限制。例如，多个这样的计算系统500可以被分组以支持数据中心中的群集11、12。本发明可用众多其他通用或专用计算系统环境或配置来操作。适用于本发明的公知计算系统、环境、和/或配置的示例包括但不限于：个人计算机、服务器计算机、手持式或膝上型设备、平板设备、多处理器系统、基于微处理器的系统、机顶盒、可编程消费电子产品、网络PC、微型计算机、大型计算机、包括任何以上系统或设备的分布式计算环境等等。

[0039] 本发明可在诸如程序模块等由计算机执行的计算机可执行指令的通用上下文中描述。一般而言，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等。本发明也可被实践在分布式计算环境中，分布式计算环境中任务是由通过通信网络链接的远程处理设备执行的。在分布式计算环境中，程序模块可以位于包括存储器存储设备在内的本地和/或远程计算机存储介质中。

[0040] 参考图5，用于实现本发明的各个方面的示例性系统可以包括计算机500形式的通用计算设备。组件可包括但不限于诸如处理单元501之类的各种硬件组件、诸如系统存储器之类的数据存储502、以及将包括数据存储502在内的各种系统组件耦合到处理单元501的系统总线503。系统总线503可以是若干类型的总线结构中的任一种，包括存储器总线或存储器控制器、外围总线和使用各种总线体系结构中的任一种的局部总线。作为示例而非限制，这样的体系结构包括工业标准体系结构(ISA)总线、微通道体系结构(MCA)总线、增强型ISA(EISA)总线、视频电子技术标准协会(VESA)局部总线和外围部件互连(PCI)总线(也称为夹层(Mezzanine)总线)。

[0041] 计算机500通常包括各种计算机可读介质504。计算机可读介质504可以是能由计算机500访问的任何可用介质，并同时包含易失性和非易失性介质以及可移动、不可移动介质，但不包括传播信号。作为示例而非限制，计算机可读介质504可包括计算机存储介质和通信介质。计算机存储介质包括以存储诸如计算机可读的指令、数据结构、程序模块或其他数据之类的信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括，但不仅限于，RAM、ROM、EEPROM、闪存或其他存储器技术、CD-ROM、数字多功能盘(DVD)或其他光盘存储、磁带盒、磁带、磁盘存储或其他磁存储设备，或可以用来存储所需信息并可以被计算机500访问的任何其他介质。通信介质通常以诸如载波或其他传输机制之类的已调制数据信号来体现计算机可读指令、数据结构、程序模块或其他数据，并且包括任何信息传送介质。术语“已调制数据信号”是指使得以在信号中编码信息的方式来设置或改变其一个或多个特性的信号。作为示例而非限制，通信介质包括诸如有线网络或直接线连接之类的有线介质，以及诸如声学、RF、红外及其他无线介质之类的无线介质。上面各项中的任何项的组合也包括在计算机可读介质的范围内。计算机可读介质可被实现为计算机程序产品，诸如存储在计算机存储介质上的软件。

[0042] 数据存储或系统存储器502包括诸如只读存储器(ROM)和/或随机存取存储器(RAM)之类的易失性和/或非易失性存储器形式的计算机存储介质。基本输入/输出系统

(BIOS)包含有助于诸如启动时在计算机500中元件之间传递信息的基本例程，它通常被存储在ROM中。RAM通常包含处理单元501可立即访问和/或当前正在操作的数据和/或程序模块。作为示例而非限制性，数据存储502保存操作系统、应用程序、其他程序模块、和程序数据。

[0043] 数据存储502还可以包括其它可移动/不可移动、易失性/非易失性计算机存储介质。仅作为示例，数据存储502可以是对不可移动、非易失性磁介质进行读写的硬盘驱动器，对可移动、非易失性磁盘进行读写的磁盘驱动器，以及对诸如CD ROM或其它光学介质等可移动、非易失性光盘进行读写的光盘驱动器。可在示例性操作环境中使用的其它可移动/不可移动、易失性/非易失性计算机存储介质包括但不限于，磁带盒、闪存卡、数字多功能盘、数字录像带、固态RAM、固态ROM等。上文所描述的并且在图5中所显示的驱动器以及它们的关联的计算机存储介质，为计算机500提供对计算机可读取的指令、数据结构、程序模块及其他数据的存储。

[0044] 用户可通过用户接口505或诸如平板、电子数字化仪、话筒、键盘和/或定点设备(通常指的是鼠标、跟踪球或触摸垫)等其它输入设备输入命令和信息。其他输入设备可以包括操纵杆、游戏垫、圆盘式卫星天线、扫描仪等等。另外，语音输入、使用手或手指的手势输入、或其它自然用户接口(NUI)也可与适当的输入设备(诸如话筒、相机、平板、触摸垫、手套、或其它传感器)一起使用。这些及其他输入设备常常通过耦合到系统总线501的用户输入接口505连接到处理单元503，但是，也可以通过其他接口和总线结构，如并行端口、游戏端口、通用串行总线(USB)端口来进行连接。监视器506或其他类型的显示设备也通过诸如视频接口之类的接口连接至系统总线503。监视器506也可以与触摸屏面板等集成。注意到监视器和/或触摸屏面板可以在物理上耦合至其中包括计算设备500的外壳，诸如在平板型个人计算机中。此外，诸如计算设备500等计算机还可以包括其他外围输出设备，诸如扬声器和打印机，它们可以通过输出外围接口等连接。

[0045] 计算机500可使用至一个或多个远程设备(诸如远程计算机)的逻辑连接507在网络化或云计算环境中操作。远程计算机可以是个人计算机、服务器、路由器、网络PC、对等设备或其它常见的网络节点，并且一般包括上面关于计算机500所述的许多或全部元件。图5中所描述的逻辑连接包括一个或多个局域网(LAN)和一个或多个广域网(WAN)，但是，也可以包括其他网络。此类联网环境在办公室、企业范围的计算机网络、内联网和因特网中是常见的。

[0046] 当在联网或云计算环境中使用时，计算机500可通过网络接口或适配器507连接至公共或私有网络。在一些实施例中，使用调制解调器或用于在网络上建立通信的其它装置。调制解调器可以是内置或外置的，它经由网络接口503或其它适当的机制连接至系统总线507。诸如包括接口和天线的无线联网组件可通过诸如接入点或对等计算机之类的合适的设备耦合到网络。在联网环境中，相关于计算机500所示的程序模块或其部分可被存储在远程存储器存储设备中。可以理解，所示的网络连接是示例性的，也可以使用在计算机之间建立通信链路的其他手段。

[0047] 尽管用结构特征和/或方法动作专用的语言描述了本主题，但可以理解，所附权利要求书中定义的主题不必限于上述具体特征或动作。相反，上述具体特征和动作是作为实现权利要求的示例形式公开的。

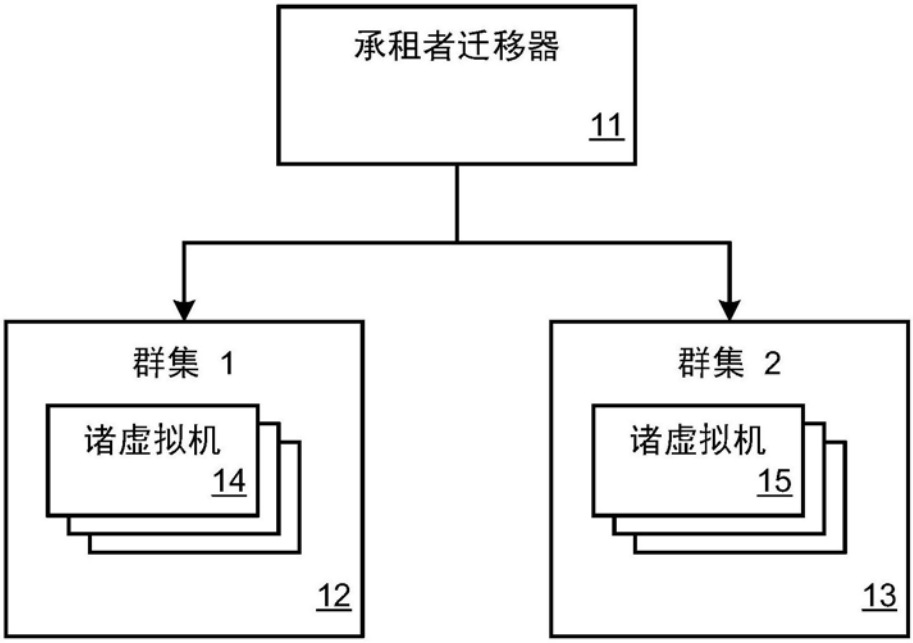


图1

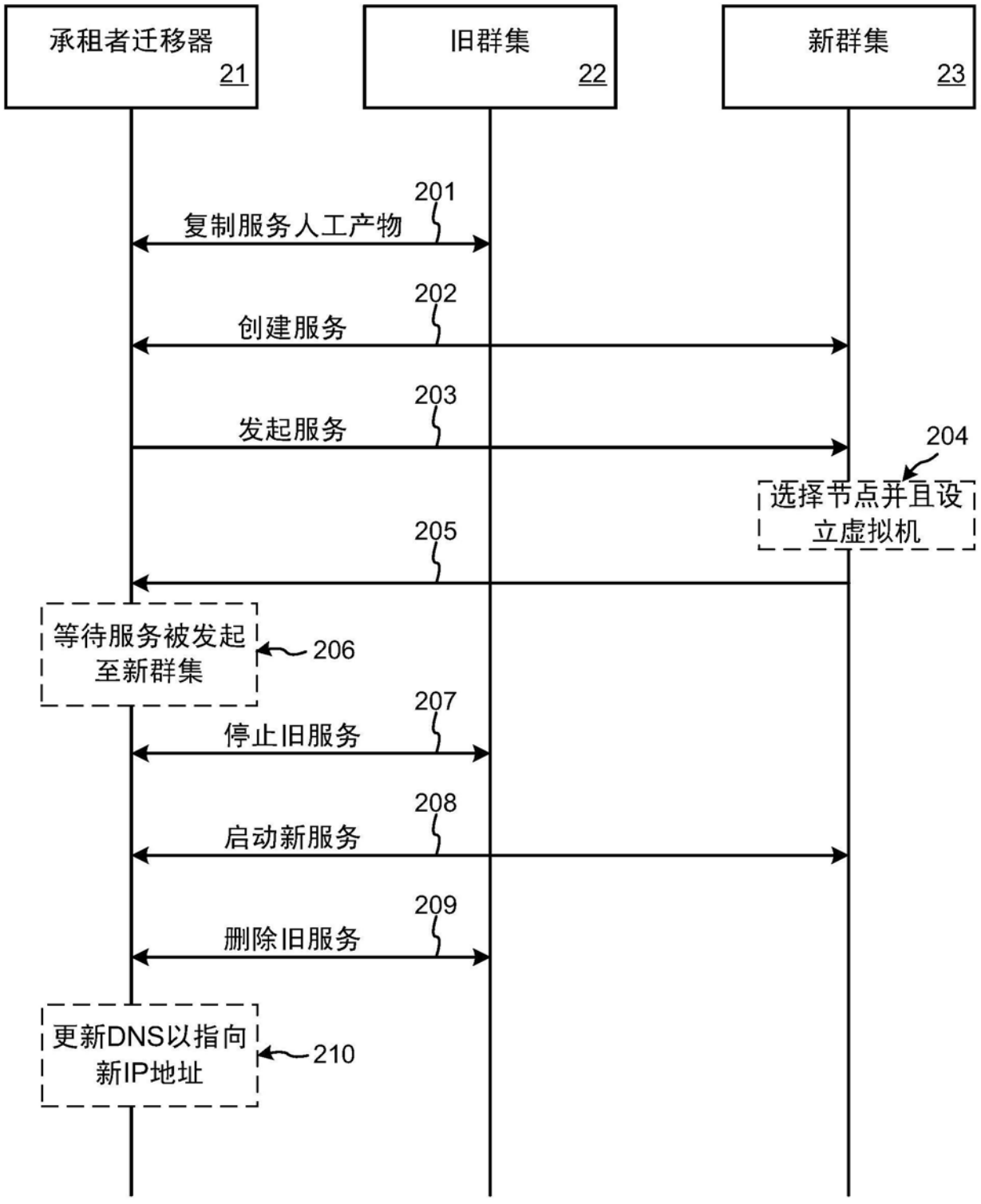

承租者迁移器
21
旧群集
22
新群集
23
201
复制服务人工产物
202
创建服务
203
发起服务
204
选择节点并且设立虚拟机
205
等待服务被发起至新群集
206
207
停止旧服务
208
启动新服务
209
删除旧服务
更新DNS以指向新IP地址
210

图2

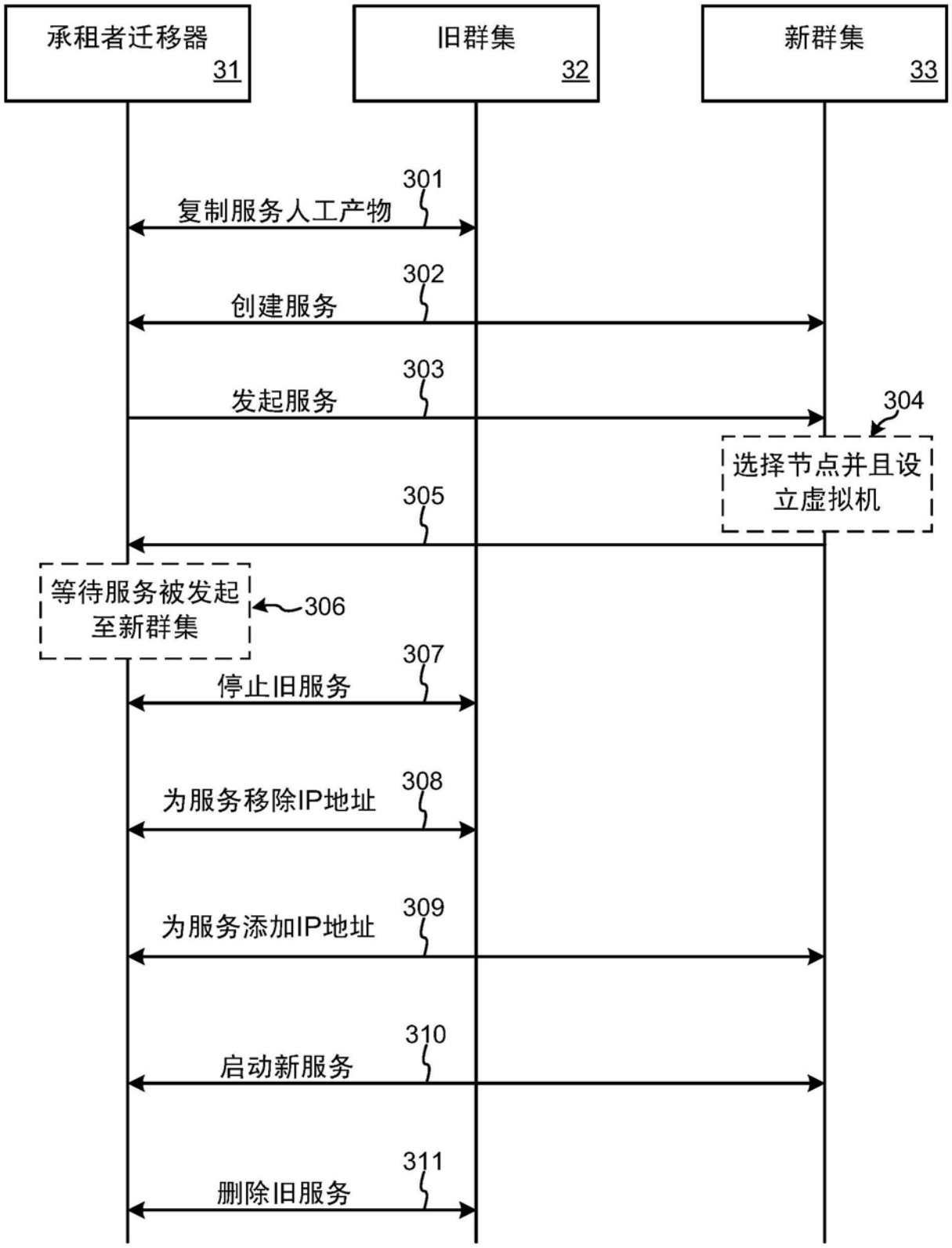

承租者迁移器
31
旧群集
32
新群集
33
301
复制服务人工产物
302
创建服务
303
发起服务
304
选择节点并且设立虚拟机
305
等待服务被发起至新群集
306
307
停止旧服务
308
为服务移除IP地址
309
为服务添加IP地址
310
启动新服务
311
删除旧服务

图3

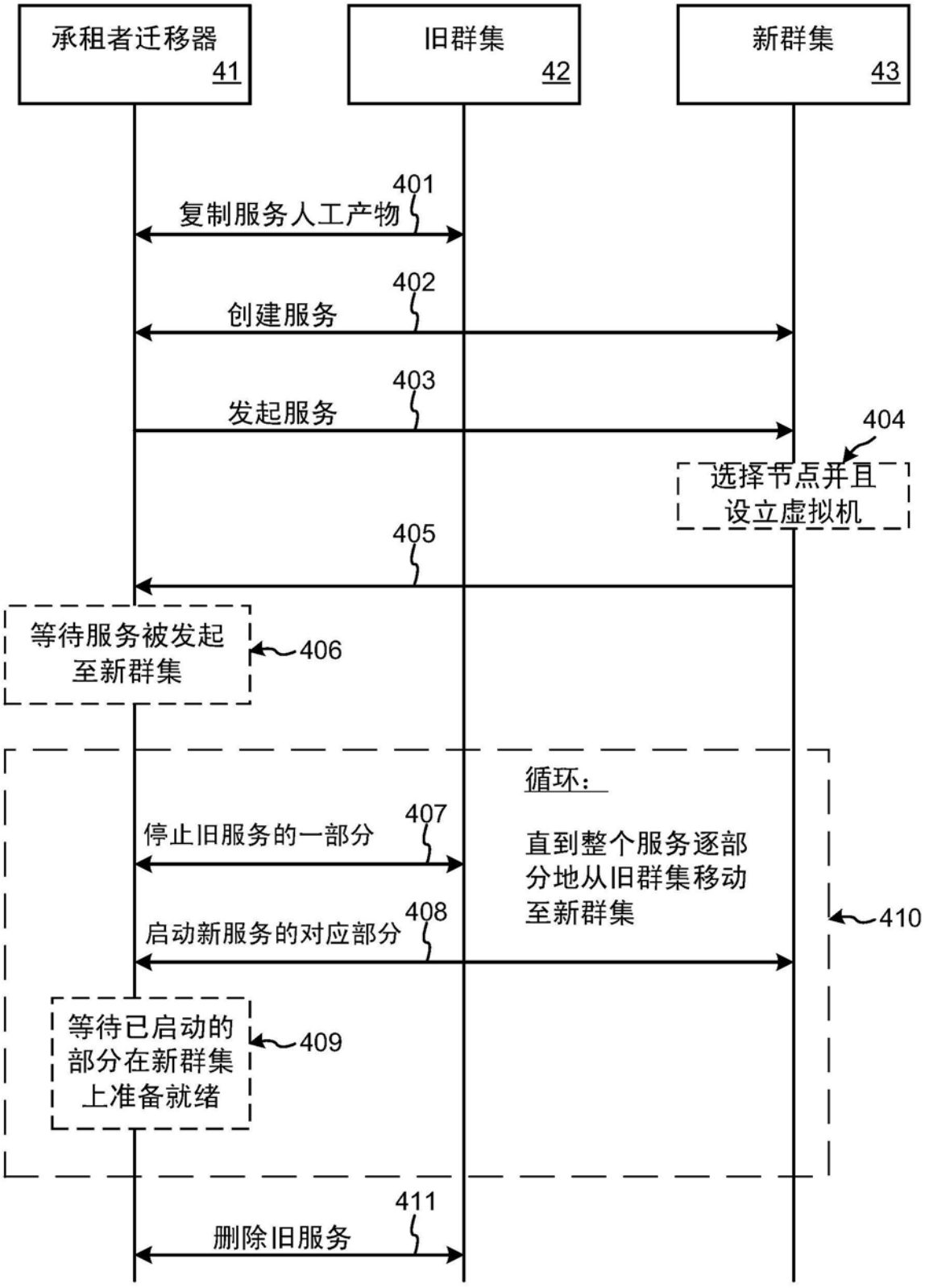

承租者迁移器
41
旧群集
42
新群集
43
401
复制服务人工产物
402
创建服务
403
发起服务
404
选择节点并且
设立虚拟机
405
等待服务被发起
至新群集
406
循环：
直到整个服务逐部
分地从旧群集移动
至新群集
407
停止旧服务的一部分
408
启动新服务的对应部分
410
等待已启动的
部分在新群集
上准备就绪
409
411
删除旧服务

图4

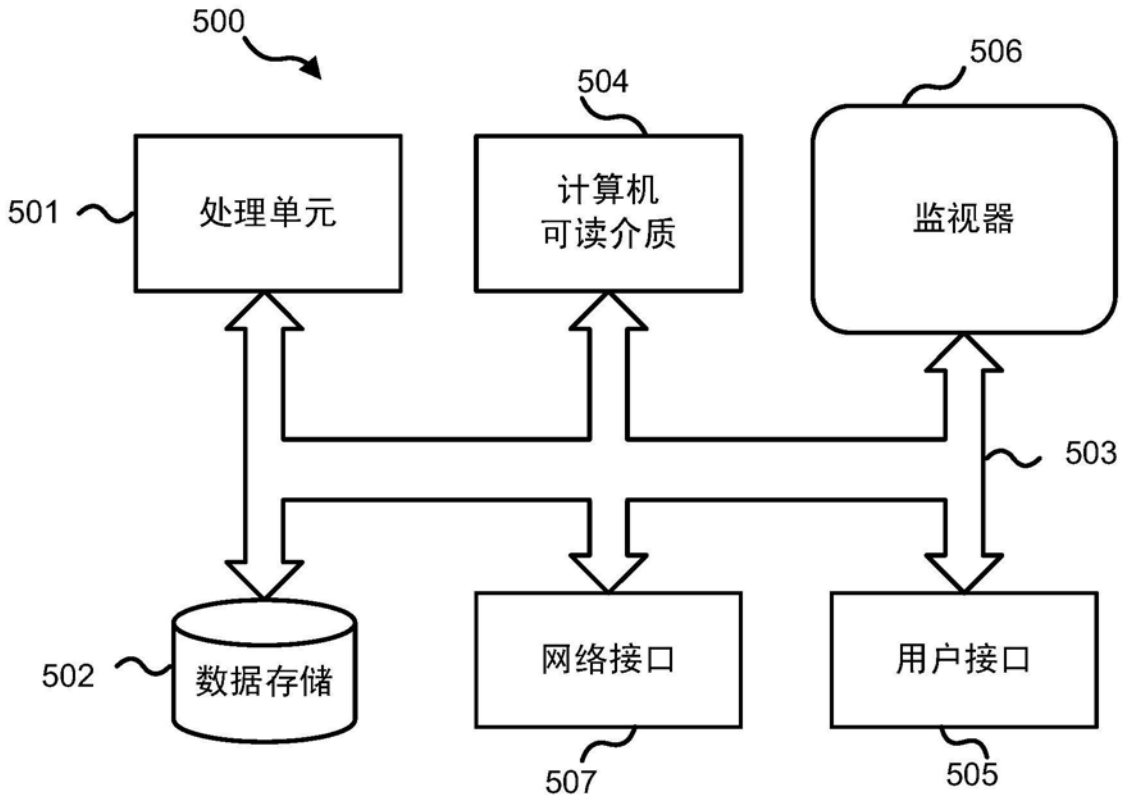

500
501
处理单元
504
计算机
可读介质
506
监视器
503
502
数据存储
网络接口
507
用户接口
505

图5