



(12)发明专利

(10)授权公告号 CN 105592134 B

(45)授权公告日 2019.04.05

(21)申请号 201510531384.4

(22)申请日 2015.08.26

(65)同一申请的已公布的文献号  
申请公布号 CN 105592134 A

(43)申请公布日 2016.05.18

(73)专利权人 新华三技术有限公司  
地址 310052 浙江省杭州市滨江区长河路  
466号

(72)发明人 董飞鸿

(74)专利代理机构 北京博思佳知识产权代理有  
限公司 11415

代理人 林祥

(51)Int.Cl.  
H04L 29/08(2006.01)

(56)对比文件

CN 104426694 A,2015.03.18,  
US 2012117563 A1,2012.05.10,  
US 2008104608 A1,2008.05.01,  
CN 103294552 A,2013.09.11,  
CN 102508718 A,2012.06.20,

审查员 胡诗婷

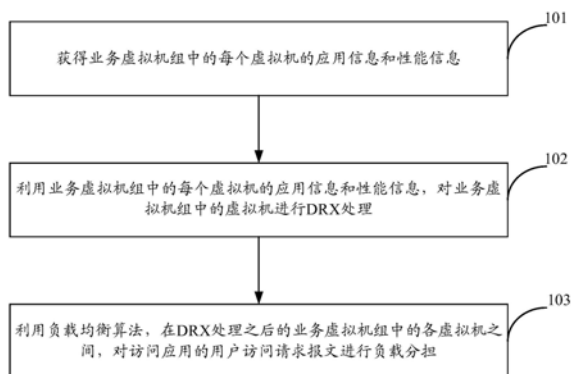
权利要求书3页 说明书10页 附图2页

(54)发明名称

一种负载分担的方法和装置

(57)摘要

本发明提供一种负载分担的方法和装置,该方法包括:获得业务虚拟机组中的每个虚拟机的应用信息和性能信息;利用所述应用信息和性能信息,对所述业务虚拟机组中的虚拟机进行DRX处理;利用负载均衡算法,在DRX处理之后的所述业务虚拟机组中的各虚拟机之间,对访问所述应用的用户访问请求报文进行负载分担。通过本发明的技术方案,通过动态调整虚拟机的数量,提高了资源利用率,并可以合理的利用虚拟机的资源。



1. 一种负载分担的方法,其特征在于,业务虚拟机组内包括多个虚拟机,且业务虚拟机组内的虚拟机用于提供同一应用,所述方法包括以下步骤:

获得所述业务虚拟机组中的每个虚拟机的应用信息和性能信息;所述应用信息包括应用运行状态和应用指标,所述应用运行状态具体为可用状态或者不可用状态,所述应用指标用于表示应用的健康状况;

利用所述业务虚拟机组中的每个虚拟机的应用信息和性能信息,对所述业务虚拟机组中的虚拟机进行动态资源扩展DRX处理;

利用负载均衡算法,在DRX处理之后的所述业务虚拟机组中的各虚拟机之间,对访问所述应用的用户访问请求报文进行负载分担。

2. 根据权利要求1所述的方法,其特征在于,

所述利用所述业务虚拟机组中的每个虚拟机的应用信息和性能信息,对所述业务虚拟机组中的虚拟机进行DRX处理的过程,具体包括:

如果所述业务虚拟机组中有虚拟机的应用运行状态为不可用状态,则对所述虚拟机进行重启处理,或者对所述虚拟机进行关闭处理;或者,

如果所述业务虚拟机组中有多个虚拟机的应用运行状态为可用状态,且所述多个虚拟机中的每个虚拟机的应用指标均小于预设第一阈值,性能信息均小于预设第二阈值,则对业务虚拟机组中的虚拟机进行关闭处理;或者,

如果所述业务虚拟机组中有多个虚拟机的应用运行状态为可用状态,且所述多个虚拟机中的每个虚拟机的应用指标均大于预设第三阈值,性能信息均大于预设第四阈值,则对业务虚拟机组中的虚拟机进行开启处理;或者,

如果所述业务虚拟机组中有多个虚拟机的应用运行状态为可用状态,且所述多个虚拟机中包括应用指标小于预设第一阈值、性能信息小于预设第二阈值的虚拟机,并包括应用指标大于预设第三阈值、性能信息大于预设第四阈值的虚拟机,则对业务虚拟机组中的虚拟机的访问权重进行调整;

预设第三阈值大于预设第一阈值,预设第四阈值大于预设第二阈值。

3. 根据权利要求2所述的方法,其特征在于,在对所述业务虚拟机组中的虚拟机进行关闭处理之后,所述方法进一步包括:

如果所述业务虚拟机组中存在处于关闭状态的虚拟机,则在符合预设虚拟机删除条件时,从所述业务虚拟机组中删除处于关闭状态的虚拟机。

4. 根据权利要求2所述的方法,其特征在于,所述对所述业务虚拟机组中的虚拟机进行开启处理的过程,具体包括:

判断所述业务虚拟机组中是否存在处于关闭状态的虚拟机;

如果存在,则开启所述业务虚拟机组中的处于关闭状态的虚拟机;

如果不存在,则创建用于提供所述应用的虚拟机,并将当前创建的虚拟机加入到所述业务虚拟机组中,并开启当前创建的虚拟机。

5. 根据权利要求2所述的方法,其特征在于,所述对所述业务虚拟机组中的虚拟机的访问权重进行调整的过程,具体包括:

增加应用指标小于预设第一阈值、性能信息小于预设第二阈值的虚拟机的访问权重,并降低应用指标大于预设第三阈值、性能信息大于预设第四阈值的虚拟机的访问权重;以

在对访问所述应用的用户访问请求报文进行负载分担时,增加发送给访问权重增加的虚拟机的用户访问请求报文的数量,并减少发送给访问权重降低的虚拟机的用户访问请求报文的数量。

6. 一种负载分担的装置,其特征在于,业务虚拟机组内包括多个虚拟机,且业务虚拟机组内的虚拟机用于提供同一应用,所述装置具体包括:

获得模块,用于获得所述业务虚拟机组中的每个虚拟机的应用信息和性能信息;所述应用信息包括应用运行状态和应用指标,所述应用运行状态具体为可用状态或者不可用状态,所述应用指标用于表示应用的健康状况;

处理模块,用于利用所述业务虚拟机组中的每个虚拟机的应用信息和性能信息,对所述业务虚拟机组中的虚拟机进行动态资源扩展DRX处理;

负载模块,用于利用负载均衡算法,在DRX处理之后的所述业务虚拟机组中的各虚拟机之间,对访问所述应用的用户访问请求报文进行负载分担。

7. 根据权利要求6所述的装置,其特征在于,

所述处理模块,具体用于当所述业务虚拟机组中有虚拟机的应用运行状态为不可用状态时,则对所述虚拟机进行重启处理,或者对所述虚拟机进行关闭处理;或者,当所述业务虚拟机组中有多个虚拟机的应用运行状态为可用状态,且所述多个虚拟机中的每个虚拟机的应用指标均小于预设第一阈值,性能信息均小于预设第二阈值时,则对业务虚拟机组中的虚拟机进行关闭处理;或者,当所述业务虚拟机组中有多个虚拟机的应用运行状态为可用状态,且所述多个虚拟机中的每个虚拟机的应用指标均大于预设第三阈值,性能信息均大于预设第四阈值时,则对业务虚拟机组中的虚拟机进行开启处理;或者,当所述业务虚拟机组中有多个虚拟机的应用运行状态为可用状态,且所述多个虚拟机中包括应用指标小于预设第一阈值、性能信息小于预设第二阈值的虚拟机,并包括应用指标大于预设第三阈值、性能信息大于预设第四阈值的虚拟机时,则对业务虚拟机组中的虚拟机的访问权重进行调整;

预设第三阈值大于预设第一阈值,预设第四阈值大于预设第二阈值。

8. 根据权利要求7所述的装置,其特征在于,

所述处理模块,还用于在对所述业务虚拟机组中的虚拟机进行关闭处理之后,如果所述业务虚拟机组中存在处于关闭状态的虚拟机,则在符合预设虚拟机删除条件时,从所述业务虚拟机组中删除处于关闭状态的虚拟机。

9. 根据权利要求7所述的装置,其特征在于,

所述处理模块,具体用于在对所述业务虚拟机组中的虚拟机进行开启处理的过程中,判断所述业务虚拟机组中是否存在处于关闭状态的虚拟机;如果存在,则开启所述业务虚拟机组中的处于关闭状态的虚拟机;如果不存在,则创建用于提供所述应用的虚拟机,并将当前创建的虚拟机加入到所述业务虚拟机组中,并开启当前创建的虚拟机。

10. 根据权利要求7所述的装置,其特征在于,

所述处理模块,具体用于在对业务虚拟机组中的虚拟机的访问权重进行调整的过程中,增加应用指标小于预设第一阈值、性能信息小于预设第二阈值的虚拟机的访问权重,并降低应用指标大于预设第三阈值、性能信息大于预设第四阈值的虚拟机的访问权重;在对访问所述应用的用户访问请求报文进行负载分担时,增加发送给访问权重增加的虚拟机

的用户访问请求报文的数量,并减少发送给访问权重降低的虚拟机的用户访问请求报文的数量。

## 一种负载分担的方法和装置

### 技术领域

[0001] 本发明涉及通信技术领域,尤其涉及一种负载分担的方法和装置。

### 背景技术

[0002] 随着数据中心业务日益增加,用户需求不断提高,数据中心的规模和功能日趋复杂,管理难度越来越高,因此,对数据中心资源进行虚拟化,成为目前数据中心整合的重要趋势。虚拟化技术通过对物理资源和提供的服务进行抽象化,让资源使用者和系统管理者不关心对象的物理特征和服务边界的细节,从而降低资源使用和管理的复杂度,提高使用效率。因而,对数据中心的虚拟化能够提高数据中心的资源利用率,并降低系统的能耗。进一步的,通过专用的虚拟化软件可以将一台物理服务器虚拟出多台虚拟机,每个虚拟机独立运行,互不影响,都有自己的操作系统、应用程序和硬件环境。

[0003] 当虚拟机对外提供应用时,如果有大量用户需要访问该应用,则提供该应用的虚拟机可能无法满足大量用户的访问需求,因此,需要增加提供该应用的虚拟机的数量。而且,如果访问该应用的用户数量较少时,如果有大量的虚拟机均提供该应用,则会浪费虚拟机的资源,因此,需要减少提供该应用的虚拟机的数量。基于此,需要知道何时增加提供该应用的虚拟机的数量,以及何时减少提供该应用的虚拟机的数量,从而动态调整虚拟机的数量。

### 发明内容

[0004] 本发明提供一种负载分担的方法,业务虚拟机组内包括多个虚拟机,且业务虚拟机组内的虚拟机用于提供同一应用,所述方法包括以下步骤:

[0005] 获得所述业务虚拟机组中的每个虚拟机的应用信息和性能信息;

[0006] 利用所述业务虚拟机组中的每个虚拟机的应用信息和性能信息,对所述业务虚拟机组中的虚拟机进行动态资源扩展DRX处理;

[0007] 利用负载均衡算法,在DRX处理之后的所述业务虚拟机组中的各虚拟机之间,对访问所述应用的用户访问请求报文进行负载分担。

[0008] 本发明提供一种负载分担的装置,业务虚拟机组内包括多个虚拟机,且业务虚拟机组内的虚拟机用于提供同一应用,所述装置具体包括:获得模块,用于获得所述业务虚拟机组中的每个虚拟机的应用信息和性能信息;

[0009] 处理模块,用于利用所述业务虚拟机组中的每个虚拟机的应用信息和性能信息,对所述业务虚拟机组中的虚拟机进行动态资源扩展DRX处理;

[0010] 负载模块,用于利用负载均衡算法,在DRX处理之后的所述业务虚拟机组中的各虚拟机之间,对访问所述应用的用户访问请求报文进行负载分担。

[0011] 基于上述技术方案,本发明实施例中,通过配置包括多个虚拟机的业务虚拟机组,并且业务虚拟机组内的虚拟机提供同一应用,从而可以利用业务虚拟机组中的每个虚拟机的应用信息和性能信息对业务虚拟机组中的虚拟机进行DRX (Dynamic Resource

eXtension,动态资源扩展)处理,当有大量用户需要访问该应用时,可以增加提供该应用的虚拟机的数量,以满足大量用户的访问需求,增强处理能力;当需要访问该应用的用户数量减少时,可以减少提供该应用的虚拟机的数量,以回收虚拟机的资源。上述方式通过动态调整虚拟机的数量,提高了资源利用率,并可以合理的利用虚拟机的资源。

### 附图说明

[0012] 图1是本发明一种实施方式中的负载分担的方法的流程图;

[0013] 图2是本发明一种实施方式中的负载分担设备的硬件结构图;

[0014] 图3是本发明一种实施方式中的负载分担的装置的结构图。

### 具体实施方式

[0015] 针对现有技术中存在的问题,本发明实施例中提出一种负载分担的方法,该方法应用于包括多个虚拟机的系统中。虚拟机是物理服务器上虚拟化的逻辑的虚拟机,不同的虚拟机可以位于相同的物理服务器上,也可以位于不同的物理服务器上。通过DRX技术,可以将多个虚拟机添加到一个业务虚拟机组中。基于此,业务虚拟机组内包括多个虚拟机,且业务虚拟机组内的虚拟机可用于提供同一应用。例如,业务虚拟机组内的多个虚拟机均用于对外提供网页应用,或业务虚拟机组内的多个虚拟机均用于对外提供邮箱应用等。

[0016] 在上述应用场景下,如图1所示,该负载分担的方法具体包括以下步骤:

[0017] 步骤101,获得业务虚拟机组中的每个虚拟机的应用信息和性能信息。

[0018] 步骤102,利用业务虚拟机组中的每个虚拟机的应用信息和性能信息,对业务虚拟机组中的虚拟机进行DRX(动态资源扩展)处理。

[0019] 步骤103,利用负载均衡算法,在DRX处理之后的业务虚拟机组中的各虚拟机之间,对访问应用的用户访问请求报文进行负载分担。

[0020] 本发明实施例中,可以启动定时任务,该定时任务用于周期性的执行上述负载分担的方法,即每隔一段时间,执行上述步骤101-步骤103,对业务虚拟机组中的虚拟机进行DRX处理,以下结合一次执行过程进行说明。

[0021] 本发明实施例中,性能信息具体包括但不限于CPU(Central Processing Unit,中央处理器)利用率和/或内存利用率。应用信息具体包括但不限于应用运行状态和应用指标,应用运行状态具体可以为可用状态或者不可用状态。

[0022] 当然,性能信息并不局限于CPU利用率和/或内存利用率,还可以包括吞吐率、带宽利用率等其它性能信息,为了方便描述,本发明实施例的后续过程中,以性能信息包括CPU利用率和/或内存利用率为例进行说明。

[0023] 在一种DRX技术中,通过将多个虚拟机添加到一个业务虚拟机组中,定时检测业务虚拟机组中的各虚拟机的CPU利用率和内存利用率。如果所有虚拟机的CPU利用率和内存利用率均小于预设阈值,则减少业务虚拟机组中的虚拟机数量,如果有虚拟机的CPU利用率和/或内存利用率大于预设阈值,则增加业务虚拟机组中的虚拟机数量。基于此,在只利用业务虚拟机组中的每个虚拟机的性能信息(如CPU利用率和/或内存利用率)对业务虚拟机组中的虚拟机进行DRX处理时,则会出现如下情况:1、虽然所有虚拟机的CPU利用率和内存利用率均小于预设阈值,但是可能有虚拟机已经无法在处理访问应用的用户访问请求报

文,其原因是该虚拟机上可能有大量的用户访问请求报文待处理,在此情况下,如果减少业务虚拟机组中的虚拟机数量,则分配到该虚拟机上的用户访问请求报文会更多,即更多的用户访问请求报文无法被及时处理。2、虽然有虚拟机的CPU利用率和/或内存利用率大于预设阈值,但是可能有虚拟机上只处理很少的用户访问请求报文,该虚拟机有大量的资源还可以使用,在此情况下,如果增加业务虚拟机组中的虚拟机数量,则分配到该虚拟机上的用户访问请求报文会更少,浪费了虚拟机的资源。

[0024] 针对上述发现,本发明实施例中,在对业务虚拟机组中的虚拟机进行DRX处理时,会综合考虑业务虚拟机组中的每个虚拟机的应用信息(如应用运行状态和应用指标)和性能信息(如CPU利用率和/或内存利用率),对业务虚拟机组中的虚拟机进行DRX处理。为了实现这一DRX处理过程,需要先统计出业务虚拟机组中的每个虚拟机的应用信息和性能信息。

[0025] 其中,在业务虚拟机组中的每个虚拟机的使用过程中,可以直接统计出该虚拟机的性能信息(如CPU利用率和/或内存利用率)。

[0026] 其中,针对业务虚拟机组中的每个虚拟机,通过检测虚拟机对外是否能够提供相应应用,确定该虚拟机的应用运行状态为可用状态或者不可用状态。当虚拟机对外不能够提供相应应用时,则确定该虚拟机的应用运行状态为不可用状态。当虚拟机对外能够提供相应应用时,则确定该虚拟机的应用运行状态为可用状态。例如,针对网页应用(Web应用),在虚拟机处理用户访问请求报文的过程中,如果虚拟机向指定的URL(Uniform Resource Locator,统一资源定位符)页面发送HTTP(Hyper Text Transfer Protocol,超文本传输协议)请求时,能够收到状态码为200的HTTP响应(状态码200的响应代表请求已成功),则认为该虚拟机对外能够提供网页应用,确定该虚拟机的应用运行状态为可用状态。如果虚拟机向指定的URL页面发送HTTP请求时,能够收到状态码为404的HTTP响应(状态码404的响应代表页面不存在),或者在最大超时时间范围内没有收到任何HTTP响应,则认为该虚拟机对外不能够提供网页应用,确定该虚拟机的应用运行状态为不可用状态。

[0027] 进一步的,在虚拟机的应用运行状态为可用状态时,还可以检测该虚拟机的应用指标。其中,该应用指标可以根据实际经验任意配置,这些应用指标用于表示该应用的健康状况。例如,针对网页应用,则可以配置应用指标为当前连接Session(会话)个数、连接响应时间等。针对Apache应用,则可以配置应用指标为每次请求字节数、每秒请求数、每秒请求字节数等。

[0028] 本发明实施例中,利用业务虚拟机组中的每个虚拟机的应用信息和性能信息,对业务虚拟机组中的虚拟机进行DRX处理的过程,具体可以包括但不限于如下方式:如果业务虚拟机组中有虚拟机的应用运行状态为不可用状态,则对该虚拟机进行重启处理,或者对该虚拟机进行关闭处理。或者,如果业务虚拟机组中有多个虚拟机(如所有虚拟机)的应用运行状态为可用状态,且这多个虚拟机中的每个虚拟机的应用指标均小于预设第一阈值,性能信息均小于预设第二阈值,则对业务虚拟机组中的虚拟机进行关闭处理。或者,如果业务虚拟机组中有多个虚拟机(如所有虚拟机)的应用运行状态为可用状态,且这多个虚拟机中的每个虚拟机的应用指标均大于预设第三阈值,性能信息均大于预设第四阈值,则对业务虚拟机组中的虚拟机进行开启处理。或者,如果业务虚拟机组中有多个虚拟机(如所有虚拟机)的应用运行状态为可用状态,且这多个虚拟机中包括应用指标小于预设第一阈值、性能信息小于预设第二阈值的虚拟机,包括应用指标大于预设第三阈值、性能信息大于预设

第四阈值的虚拟机,对业务虚拟机组中的虚拟机的访问权重进行调整。

[0029] 其中,预设第一阈值、预设第二阈值、预设第三阈值和预设第四阈值,均可以根据实际经验进行设置,且预设第三阈值大于预设第一阈值,预设第四阈值大于预设第二阈值,而且,预设第三阈值与预设第四阈值之间并没有大小关系,预设第一阈值与预设第二阈值之间也没有大小关系。

[0030] 为虚拟机的应用指标设置预设第一阈值和预设第二阈值。当虚拟机的应用指标小于预设第一阈值时,表示该虚拟机上的该应用的健康状况为优,此时该虚拟机可以处理更多的分配到该虚拟机上的用户访问请求报文。当虚拟机的应用指标大于预设第二阈值时,表示该虚拟机上的该应用的健康状况为差,此时该虚拟机已经无法继续处理分配到该虚拟机上的用户访问请求报文。当虚拟机的应用指标位于预设第一阈值与预设第二阈值之间时,表示该虚拟机上的该应用的健康状况为良,此时分配到该虚拟机上的用户访问请求报文的数量正合适,不用调整分配到该虚拟机上的用户访问请求报文的数量。

[0031] 当虚拟机的应用指标为一个应用指标时,为该应用指标设置预设第一阈值和预设第二阈值。当虚拟机的应用指标为多个应用指标时,分别为每个应用指标设置预设第一阈值和预设第二阈值,为不同应用指标设置的预设第一阈值可以相同,也可以不同,为不同应用指标设置的预设第二阈值可以相同,也可以不同。当虚拟机的应用指标为多个应用指标时,如果这多个应用指标均小于对应的预设第一阈值,则表示虚拟机的应用指标小于预设第一阈值,该虚拟机上的应用的健康状况为优。如果这多个应用指标中有任意应用指标大于对应的预设第二阈值,则表示虚拟机的应用指标大于预设第二阈值,该虚拟机上的应用的健康状况为差。对于其它情况,表示虚拟机的应用指标位于预设第一阈值与预设第二阈值之间,该虚拟机上的应用的健康状况为良。

[0032] 例如,针对Apache应用,假设应用指标包括每次请求字节数、每秒请求数、每秒请求字节数,为每次请求字节数设置预设第一阈值1和预设第二阈值1,为每秒请求数设置预设第一阈值2和预设第二阈值2,为每秒请求字节数设置预设第一阈值3和预设第二阈值3。针对虚拟机1,统计Apache应用的每次请求字节数、每秒请求数、每秒请求字节数。当每次请求字节数小于预设第一阈值1,每秒请求数小于预设第一阈值2,每秒请求字节数小于预设第一阈值3时,表示虚拟机1的应用指标小于预设第一阈值,该虚拟机1上的应用的健康状况为优。当每次请求字节数大于预设第二阈值1,和/或每秒请求数大于预设第二阈值2,和/或每秒请求字节数大于预设第二阈值3时,表示虚拟机1的应用指标大于预设第二阈值,该虚拟机1上的应用的健康状况为差。针对上述两种情况之外的其它情况,表示虚拟机1的应用指标位于预设第一阈值与预设第二阈值之间,该虚拟机1上的应用的健康状况为良。

[0033] 其中,为虚拟机的性能信息设置预设第三阈值和预设第四阈值。当虚拟机的性能信息小于预设第三阈值时,则表示该虚拟机上的性能状况为优,此时该虚拟机可以处理更多的分配到该虚拟机上的用户访问请求报文。当虚拟机的性能信息大于预设第四阈值时,则表示该虚拟机上的性能状况为差,此时该虚拟机已经无法继续处理分配到该虚拟机上的用户访问请求报文。当虚拟机的性能信息位于预设第三阈值与预设第四阈值之间时,则表示该虚拟机上的性能状况为良,此时分配到该虚拟机上的用户访问请求报文的数量正好合适,不用调整分配到该虚拟机上的用户访问请求报文的数量。

[0034] 当虚拟机的性能信息(如CPU利用率或内存利用率)为一个时,为性能信息设置预



设第三阈值和预设第四阈值。当虚拟机的性能信息(如CPU利用率和内存利用率)为多个时,分别为每个性能信息设置预设第三阈值和预设第四阈值,为不同性能信息设置的预设第三阈值可以相同,也可以不同,为不同性能信息设置的预设第四阈值可以相同,也可以不同。当虚拟机的性能信息为多个时,如果多个性能信息均小于对应的预设第三阈值,表示虚拟机的性能信息小于预设第三阈值,该虚拟机的性能状况为优。如果多个性能信息中有任意性能信息大于对应的预设第四阈值,表示虚拟机的性能信息大于预设第四阈值,该虚拟机的性能状况为差。对于其它情况,表示虚拟机的性能信息位于预设第三阈值与预设第四阈值之间,该虚拟机的性能状况为良。

[0035] 例如,为CPU利用率设置预设第三阈值1和预设第四阈值1,为内存利用率设置预设第三阈值2和预设第四阈值2。针对虚拟机1,统计CPU利用率和内存利用率。当CPU利用率小于预设第三阈值1,内存利用率小于预设第三阈值2时,表示虚拟机1的性能信息小于预设第三阈值,该虚拟机1的性能状况为优。当CPU利用率大于预设第四阈值1,和/或内存利用率大于预设第四阈值2时,表示虚拟机1的性能信息大于预设第四阈值,该虚拟机1的性能状况为差。针对上述两种情况之外的其它情况,表示虚拟机1的性能信息位于预设第三阈值与预设第四阈值之间,该虚拟机1的性能状况为良。

[0036] 下面对上述对业务虚拟机组中的虚拟机进行DRX处理的过程进行说明。

[0037] 情况一、如果业务虚拟机组中有虚拟机的应用运行状态为不可用状态,则对该虚拟机进行重启处理,或者对该虚拟机进行关闭处理。

[0038] 当虚拟机的应用运行状态为不可用状态时,基于用户的配置信息,则对该虚拟机进行重启处理,或者对该虚拟机进行关闭处理。如果用户的配置信息为仍然将用户访问请求报文负载分担到不可用状态的虚拟机或者重启不可用状态的虚拟机时,则当虚拟机的应用运行状态为不可用状态时,对该虚拟机进行重启处理。如果用户的配置信息为不将用户访问请求报文负载分担到不可用状态的虚拟机或者关闭不可用状态的虚拟机时,则当虚拟机的应用运行状态为不可用状态时,对该虚拟机进行关闭处理。

[0039] 在对虚拟机进行重启处理的过程中,为了避免连续多次重启后,仍然无法解决故障,即虚拟机的应用运行状态仍然为不可用状态的情况,则可以指定最大重启次数。在未达到最大重启次数时,如果对虚拟机进行重启处理后,虚拟机的应用运行状态仍然为不可用状态,则继续对虚拟机进行重启处理。在达到最大重启次数时,如果对虚拟机进行重启处理后,虚拟机的应用运行状态仍然为不可用状态,则停止重启虚拟机,并对虚拟机进行关闭处理。

[0040] 在对虚拟机进行重启处理的过程中,由于虚拟机的系统启动以及应用启动等均需要耗时,为了避免虚拟机尚未完全启动,该虚拟机便又被重启所导致的问题,可以根据运营维护经验来指定两次启动的最小时间间隔,即连续两次对虚拟机进行重启处理的间隔时间,需要不小于该最小时间间隔。

[0041] 在对虚拟机进行关闭处理之后,该虚拟机不再用于对外提供相应的应用,即对外停止应用,因此不会将用户访问请求报文负载分担到该虚拟机。

[0042] 在实际应用中,如果对虚拟机配置了NQA(Network Quality Analyse,网络质量分析)检测,则可以周期性检测自身与虚拟机之间的网络质量。当虚拟机的应用运行状态为不可用状态时,则检测到自身与虚拟机之间的网络质量为不可用,此时,可以将该虚拟机设置

为不可用,后续在发送用户访问请求报文时,不会将用户访问请求报文负载分担到该虚拟机。

[0043] 情况二、如果业务虚拟机组中有多个虚拟机的应用运行状态为可用状态,且这多个虚拟机中的每个虚拟机的应用指标均小于预设第一阈值,性能信息均小于预设第二阈值,则对业务虚拟机组中的虚拟机进行关闭处理。

[0044] 当业务虚拟机组中的多个虚拟机的应用指标均小于预设第一阈值,性能信息均小于预设第二阈值时,则说明业务虚拟机组中的多个虚拟机均可以处理更多的分配到该虚拟机上的用户访问请求报文。因此,为了节省虚拟机的资源,可以对业务虚拟机组中的虚拟机进行关闭处理。其中,关闭的虚拟机的数量可以根据实际需要进行选择。在一种具体应用中,可以配置业务虚拟机组中需要保留的虚拟机的数量,在关闭虚拟机时,只要未关闭的虚拟机的数量大于等于保留的虚拟机的数量即可。例如,业务虚拟机组中保留的虚拟机的数量为5,业务虚拟机组中当前存在的虚拟机的数量为8时,则可以关闭1个虚拟机,也可以关闭2个虚拟机,也可以关闭3个虚拟机。

[0045] 当业务虚拟机组中的多个虚拟机的应用指标均小于预设第一阈值,性能信息均小于预设第二阈值时,还可以记录这种情况发生的开始时间。在对业务虚拟机组中的虚拟机进行关闭处理的过程中,根据用户自定义的条件(例如持续超过2小时)以及这种情况发生的开始时间,当确定这种情况(业务虚拟机组中的多个虚拟机的应用指标均小于预设第一阈值,性能信息均小于预设第二阈值)的持续时间满足用户自定义的条件时,则可以对业务虚拟机组中的虚拟机进行关闭处理。当确定这种情况的持续时间不满足用户自定义的条件时,则不需要对业务虚拟机组中的虚拟机进行关闭处理。

[0046] 针对上述情况一和情况二,本发明实施例中,在对业务虚拟机组中的虚拟机进行关闭处理之后,如果业务虚拟机组中存在处于关闭状态的虚拟机,则在符合预设虚拟机删除条件时,还可以从业务虚拟机组中删除处于关闭状态的虚拟机。其中,该预设虚拟机删除条件可以根据实际经验任意设置,如设置预设虚拟机删除条件为处于关闭状态的虚拟机的数量超过2个,基于此,当业务虚拟机组中处于关闭状态的虚拟机的数量超过2个时,则确定符合预设虚拟机删除条件,并从业务虚拟机组中删除处于关闭状态的虚拟机。

[0047] 情况三、如果业务虚拟机组中有多个虚拟机的应用运行状态为可用状态,且这多个虚拟机中的每个虚拟机的应用指标均大于预设第三阈值,性能信息均大于预设第四阈值,则对业务虚拟机组中的虚拟机进行开启处理。

[0048] 当业务虚拟机组中的多个虚拟机的应用指标大于预设第三阈值,性能信息大于预设第四阈值时,说明业务虚拟机组中的多个虚拟机已经无法继续处理分配到该虚拟机的用户访问请求报文。为了保证用户访问请求报文被及时处理,可以对业务虚拟机组中的虚拟机进行开启处理。

[0049] 本发明实施例中,对业务虚拟机组中的虚拟机进行开启处理的过程,具体可以包括但不限于如下方式:判断业务虚拟机组中是否存在处于关闭状态的虚拟机。如果存在,则直接开启该业务虚拟机组中的处于关闭状态的虚拟机;如果不存在,则创建用于提供该业务虚拟机组对应的应用的虚拟机,并将当前创建的虚拟机加入到该业务虚拟机组中,并开启当前创建的虚拟机。

[0050] 其中,在创建用于提供应用的虚拟机时,可以根据DHCP(Dynamic Host

Configuration Protocol, 动态主机配置协议) 或者预置的IP地址范围, 分配一个空闲IP地址, 并根据预置的虚拟机模板创建新的虚拟机, 并将该空闲IP地址、子网掩码、默认网关等网络信息配置在当前创建的虚拟机上。

[0051] 其中, 开启的虚拟机的数量可以根据实际需要进行选择。在一种具体应用中, 如果业务虚拟机组中存在处于关闭状态的虚拟机, 则可以直接开启该业务虚拟机组中的所有处于关闭状态的虚拟机, 或者开启该业务虚拟机组中的部分处于关闭状态的虚拟机。如果业务虚拟机组中不存在处于关闭状态的虚拟机, 则可以创建一个新的虚拟机或者创建多个新的虚拟机。

[0052] 情况四、如果业务虚拟机组中有多个虚拟机的应用运行状态为可用状态, 且这多个虚拟机中包括应用指标小于预设第一阈值、性能信息小于预设第二阈值的虚拟机, 并包括应用指标大于预设第三阈值、性能信息大于预设第四阈值的虚拟机, 则对业务虚拟机组中的虚拟机的访问权重进行调整。

[0053] 其中, 当业务虚拟机组中的虚拟机的应用指标小于预设第一阈值, 性能信息小于预设第二阈值时, 则说明该虚拟机可以处理更多的分配到该虚拟机上的用户访问请求报文。当业务虚拟机组中的虚拟机的应用指标大于预设第三阈值, 性能信息大于预设第四阈值时, 则说明该虚拟机已经无法继续处理分配到该虚拟机的用户访问请求报文。基于此, 在本发明实施例中, 对业务虚拟机组中的虚拟机的访问权重进行调整的过程, 具体可以包括但不限于如下方式: 增加应用指标小于预设第一阈值、性能信息小于预设第二阈值的虚拟机的访问权重, 并降低应用指标大于预设第三阈值、性能信息大于预设第四阈值的虚拟机的访问权重; 以在对访问应用的用户访问请求报文进行负载分担时, 增加发送给访问权重增加的虚拟机的用户访问请求报文的数量, 并减少发送给访问权重降低的虚拟机的用户访问请求报文的数量。

[0054] 如果业务虚拟机组中的多个虚拟机的应用信息和性能信息是上述四种情况之外的其它情况, 则不会对业务虚拟机组中的虚拟机进行DRX处理。

[0055] 例如, 当业务虚拟机组中包括虚拟机1、虚拟机2、虚拟机3和虚拟机4时, 则利用负载均衡算法, 在虚拟机1、虚拟机2、虚拟机3和虚拟机4之间对访问应用的用户访问请求报文进行负载分担。针对上述情况一和情况二, 假设对业务虚拟机组中的虚拟机1进行关闭处理, 则利用负载均衡算法, 在DRX处理之后的业务虚拟机组中的虚拟机2、虚拟机3和虚拟机4之间对访问应用的用户访问请求报文进行负载分担。针对上述情况三, 假设在业务虚拟机组中添加新的虚拟机5, 并对业务虚拟机组中的虚拟机5进行开启处理, 则利用负载均衡算法, 在DRX处理之后的业务虚拟机组中的虚拟机1、虚拟机2、虚拟机3、虚拟机4和虚拟机5之间对访问应用的用户访问请求报文进行负载分担。针对上述情况四, 假设增加业务虚拟机组中的虚拟机1的访问权重, 并降低业务虚拟机组中的虚拟机2的访问权重, 则利用负载均衡算法, 在DRX处理之后的业务虚拟机组中的虚拟机1、虚拟机2、虚拟机3和虚拟机4之间对访问应用的用户访问请求报文进行负载分担。

[0056] 针对情况一、情况二和情况三, 该负载均衡算法可以使用任意的负载均衡算法, 本发明实施例中对此不做限制。针对情况四, 该负载均衡算法可以使用轮转算法或者最小链接算法, 该轮转算法或者最小链接算法可以通过调整虚拟机的访问权重, 以改变发送到该虚拟机的用户访问请求报文的数量。

[0057] 针对情况四,假设在调整访问权重之前,针对每连续的100个用户访问请求报文,有25个用户访问请求报文被发送给虚拟机1,有25个用户访问请求报文被发送给虚拟机2,有25个用户访问请求报文被发送给虚拟机3,有25个用户访问请求报文被发送给虚拟机4。而且,在调整访问权重之后,假设在增加虚拟机1的访问权重,并降低虚拟机2的访问权重之后,可以使得针对每连续的100个用户访问请求报文,有45个用户访问请求报文被发送给虚拟机1,有5个用户访问请求报文被发送给虚拟机2,有25个用户访问请求报文被发送给虚拟机3,有25个用户访问请求报文被发送给虚拟机4。经过上述处理,可以增加发送给空闲的虚拟机1的用户访问请求报文的数量,并可以减少发送给忙碌的虚拟机2的用户访问请求报文的数量。

[0058] 基于上述技术方案,本发明实施例中,通过配置包括多个虚拟机的业务虚拟机组,该业务虚拟机组内的虚拟机提供同一应用,从而利用业务虚拟机组中的每个虚拟机的应用信息和性能信息对业务虚拟机组中的虚拟机进行DRX处理,当有大量用户需要访问应用时,增加提供该应用的虚拟机的数量,以满足大量用户的访问需求,增强处理能力;当访问应用的用户数量减少时,可以减少提供应用的虚拟机的数量,以回收虚拟机的资源。上述方式通过动态调整虚拟机的数量,提高了资源利用率,并可以合理的利用虚拟机的资源。

[0059] 本发明实施例中提出的上述负载分担的方法,可以应用在负载分担设备上,也可以应用在能够管理该负载分担设备的网管设备或者其它设备上。在负载分担设备上,为应用(如应用A)创建实服务组,并在该实服务组下配置多个实服务。其中,可以将具有相同或相似应用的多个虚拟机抽象成一个实服务组,且一个实服务组可以包括多个实服务。实服务是负载分担设备上处理用户业务的实体,与虚拟机存在IP地址和端口的映射关系。针对上述过程,实服务组相当于上述的业务虚拟机组,一个实服务对应一个虚拟机。

[0060] 基于与上述方法同样的发明构思,本发明实施例中还提供了一种负载分担的装置,该负载分担的装置可以应用在负载分担设备中。该负载分担的装置可以通过软件实现,也可以通过硬件或者软硬件结合的方式实现。以软件实现为例,作为一个逻辑意义上的装置,是通过其所在的负载分担设备的处理器,将非易失性存储器中对应的计算机程序指令读取到内存中运行形成的。从硬件层面而言,如图2所示,为本发明提出的负载分担的装置所在的负载分担设备的一种硬件结构图,除了图2所示的处理器、网络接口、内存以及非易失性存储器外,负载分担设备还可以包括其他硬件,如负责处理报文的转发芯片等;从硬件结构上来讲,该负载分担设备还可能是分布式设备,可能包括多个接口卡,以便在硬件层面进行报文处理的扩展。

[0061] 针对本发明实施例提出的负载分担的装置,业务虚拟机组内包括多个虚拟机,且业务虚拟机组内的虚拟机用于提供同一应用。如图3所示,为本发明实施例提出的负载分担的装置的结构图,所述负载分担的装置具体包括:

[0062] 获得模块11,用于获得所述业务虚拟机组中的每个虚拟机的应用信息和性能信息;处理模块12,用于利用所述业务虚拟机组中的每个虚拟机的应用信息和性能信息,对所述业务虚拟机组中的虚拟机进行DRX处理;负载模块13,用于利用负载均衡算法,在DRX处理之后的所述业务虚拟机组中的各虚拟机之间,对访问所述应用的用户访问请求报文进行负载分担。

[0063] 本发明实施例中,所述应用信息包括应用运行状态和应用指标,所述应用运行状

态具体为可用状态或者不可用状态；

[0064] 所述处理模块12,具体用于当所述业务虚拟机组中有虚拟机的应用运行状态为不可用状态时,则对所述虚拟机进行重启处理,或者对所述虚拟机进行关闭处理;或者,当所述业务虚拟机组中有多个虚拟机的应用运行状态为可用状态,且所述多个虚拟机中的每个虚拟机的应用指标均小于预设第一阈值,性能信息均小于预设第二阈值时,则对业务虚拟机组中的虚拟机进行关闭处理;或者,当所述业务虚拟机组中有多个虚拟机的应用运行状态为可用状态,且所述多个虚拟机中的每个虚拟机的应用指标均大于预设第三阈值,性能信息均大于预设第四阈值时,则对业务虚拟机组中的虚拟机进行开启处理;或者,当所述业务虚拟机组中有多个虚拟机的应用运行状态为可用状态,且所述多个虚拟机中包括应用指标小于预设第一阈值、性能信息小于预设第二阈值的虚拟机,并包括应用指标大于预设第三阈值、性能信息大于预设第四阈值的虚拟机时,则对业务虚拟机组中的虚拟机的访问权重进行调整;

[0065] 预设第三阈值大于预设第一阈值,预设第四阈值大于预设第二阈值。

[0066] 所述处理模块12,还用于在对所述业务虚拟机组中的虚拟机进行关闭处理之后,如果所述业务虚拟机组中存在处于关闭状态的虚拟机,则在符合预设虚拟机删除条件时,从所述业务虚拟机组中删除处于关闭状态的虚拟机。

[0067] 所述处理模块12,具体用于在对所述业务虚拟机组中的虚拟机进行开启处理的过程中,判断所述业务虚拟机组中是否存在处于关闭状态的虚拟机;如果存在,则开启所述业务虚拟机组中的处于关闭状态的虚拟机;如果不存在,则创建用于提供所述应用的虚拟机,并将当前创建的虚拟机加入到所述业务虚拟机组中,并开启当前创建的虚拟机。

[0068] 所述处理模块12,具体用于在对业务虚拟机组中的虚拟机的访问权重进行调整的过程中,增加应用指标小于预设第一阈值、性能信息小于预设第二阈值的虚拟机的访问权重,并降低应用指标大于预设第三阈值、性能信息大于预设第四阈值的虚拟机的访问权重;以在对访问所述应用的用户访问请求报文进行负载分担时,增加发送给访问权重增加的虚拟机的用户访问请求报文的数量,并减少发送给访问权重降低的虚拟机的用户访问请求报文的数量。

[0069] 其中,本发明装置的各个模块可以集成于一体,也可以分离部署。上述模块可以合并为一个模块,也可以进一步拆分成多个子模块。

[0070] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到本发明可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件,但很多情况下前者是更佳的实施方式。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本发明各个实施例所述的方法。本领域技术人员可以理解附图只是一个优选实施例的示意图,附图中的模块或流程并不一定是实施本发明所必须的。

[0071] 本领域技术人员可以理解实施例中的装置中的模块可以按照实施例描述进行分布于实施例的装置中,也可以进行相应变化位于不同于本实施例的一个或多个装置中。上述实施例的模块可以合并为一个模块,也可进一步拆分成多个子模块。上述本发明实施例序号仅仅为了描述,不代表实施例的优劣。

[0072] 以上公开的仅为本发明的几个具体实施例,但是,本发明并非局限于此,任何本领域的技术人员能思之的变化都应落入本发明的保护范围。

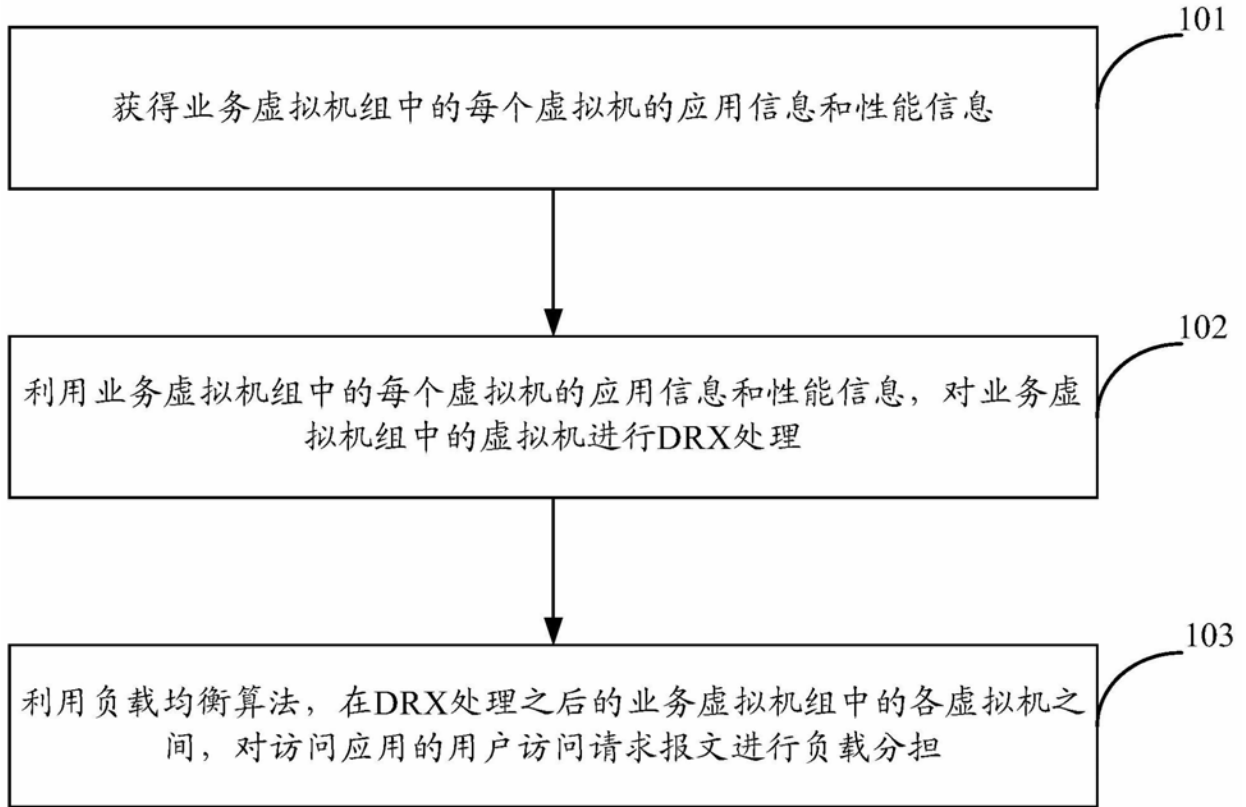


图1

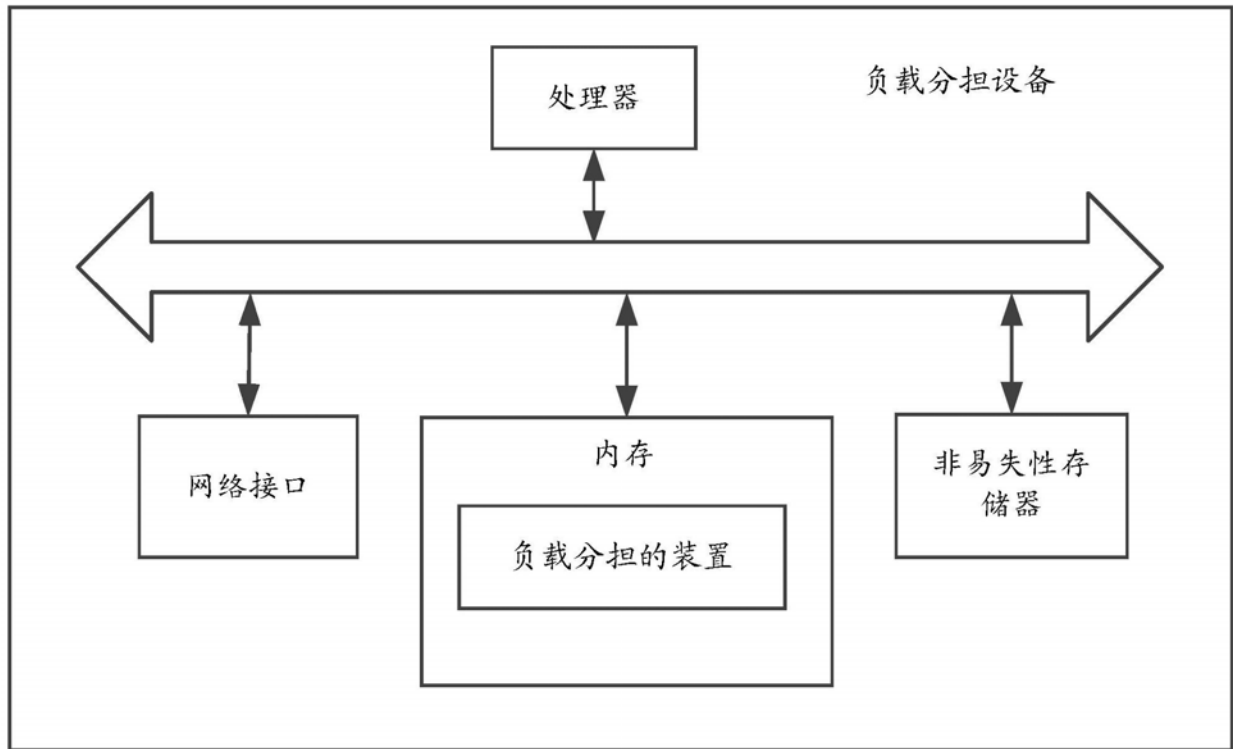


图2

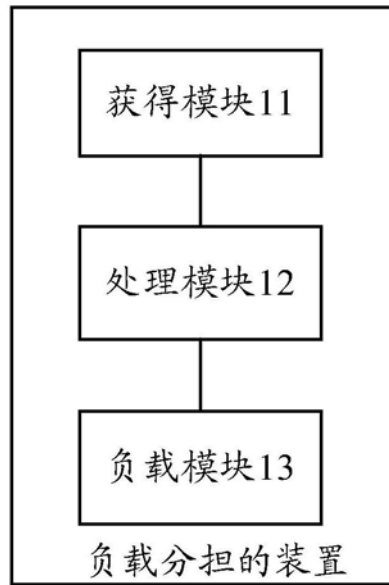


图3