(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104407925 A
(43)申请公布日 2015.03.11

(21)申请号 201410758791.4

(22)申请日 2014.12.10

(71)申请人 中国电信集团系统集成有限责任公司
地址 100035 北京市西城区西直门内大街118号冠华大厦10层

(72)发明人 胡林 赵康 王敏讷 李金岭

(74)专利代理机构 北京市京大律师事务所 11321
代理人 王凝 金凤

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

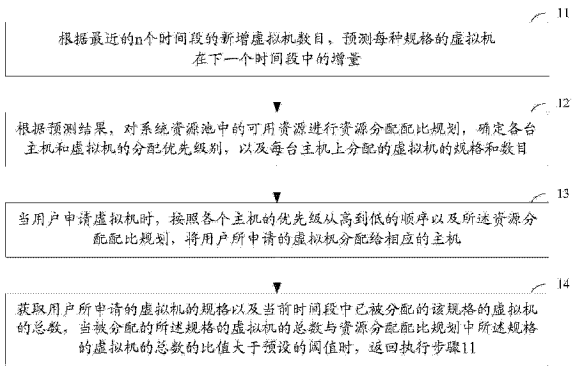
权利要求书2页 说明书8页 附图1页

(54)发明名称

一种动态的资源分配方法

(57)摘要

本发明公开了一种动态的资源分配方法，包括：根据最近的n个时间段的新增虚拟机数目，预测每种规格的虚拟机在下一个时间段中的增量；B、根据预测结果，对系统资源池中的可用资源进行资源分配配比规划，确定各台主机和虚拟机的分配优先级别，以及每台主机上分配的虚拟机的规格和数目；C、当用户申请虚拟机时，按照各个主机的优先级从高到低的顺序以及所述资源分配配比规划，将用户所申请的虚拟机分配给相应的主机；D、当被分配的所述规格的虚拟机的总数与资源分配配比规划中所述规格的虚拟机的总数的比值大于预设的阈值时，返回执行步骤A。通过上述方法，可以在IaaS环境中有效地减少资源碎片，实现物理资源的充分利用。



CN 104407925 A

1. 一种动态的资源分配方法，其特征在于，该方法包括：

A、根据最近的 n 个时间段的新增虚拟机数目，预测每种规格的虚拟机在下一个时间段中的增量；

B、根据预测结果，对系统资源池中的可用资源进行资源分配配比规划，确定各台主机和虚拟机的分配优先级别，以及每台主机上分配的虚拟机的规格和数目；

C、当用户申请虚拟机时，按照各个主机的优先级从高到低的顺序以及所述资源分配配比规划，将用户所申请的虚拟机分配给相应的主机；

D、获取用户所申请的虚拟机的规格以及当前时间段中已被分配的该规格的虚拟机的总数，当被分配的所述规格的虚拟机的总数与资源分配配比规划中所述规格的虚拟机的总数的比值大于预设的阈值时，返回执行步骤 A。

2. 根据权利要求 1 所述的方法，其特征在于，所述虚拟机的规格包括：

1 核 1G；1 核 2G；2 核 2G；2 核 4G；4 核 4G；4 核 8G；4 核 12G；8 核 8G 和 8 核 12G。

3. 根据权利要求 1 所述的方法，其特征在于：

使用二次指数平滑法预测每种规格的虚拟机在下一个时间段中的增量。

4. 根据权利要求 3 所述的方法，其特征在于，使用如下所述的公式来预测在下一个时间段中的虚拟机增量：

$$F_{k+T} = a_k + b_k T$$

其中，

$$a_k = 2S_k(1) - S_k(2)$$

$$b_k = \frac{\alpha}{1-\alpha}(S_k(1) - S_k(2))$$

$$S_k(1) = \alpha X_k + (1-\alpha) S_{k-1}(1)$$

$$S_k(2) = \alpha S_k + (1-\alpha) S_{k-1}(2)$$

其中，F_{k+T} 为第 k+T 个时间段的虚拟机增量，T 为第 k 个时间段到下一个时间段的间隔数，a_k 和 b_k 均为参数，$S_k(1)$ 和 $S_k(2)$ 分别为一次指数平滑值和二次指数平滑值，α 为平滑常数，取值范围为 [0,1]。

5. 根据权利要求 4 所述的方法，其特征在于：

所述 α 的数值为 0.9。

6. 根据权利要求 1 所述的方法，其特征在于，所述根据预测结果，对系统资源池中的可用资源进行资源分配配比规划包括：

步骤 B1，统计系统资源池中可用的资源数量以及可用资源在每天主机上的分布情况；

步骤 B2，根据统计结果将各个主机按照可用的 CPU 数量从小到大的顺序进行排序并设置优先级，CPU 数量越少的主机的优先级越高；

步骤 B3，将预测结果中的各种规格的虚拟机按照虚拟机的 CPU 数量从小到大的顺序进行排序，并根据排序结果形成虚拟机列表；

步骤 B4，根据虚拟机的 CPU 数量和内存大小确定各个虚拟机的 CPU 价值和内存价值，并为每台主机设置 CPU 容量和内存容量；

骤 B5，按照优先级从高到低的顺序依次对每台主机进行一次资源规划；在每台主机的资源规划中，按照虚拟机的 CPU 数量从小到大的顺序遍历虚拟机列表中的每台虚拟机，将

使得当前主机的总价值最大的虚拟机分配给当前的主机，并从虚拟机列表中删除已被分配的虚拟机。

7. 根据权利要求 6 所述的方法，其特征在于，使用如下所述的状态转移方程对主机进行资源规划：

fcpu[i] = max{fcpu[i-1][cpuValues][ramValues], fcpu[i][cpuValues-cpu[i]][ramValues-ram[i]]+cpu[i]}；

fram[i] = max{fram[i-1][cpuValues][ramValues], fcpu[i][cpuValues-cpu[i]][ramValues-ram[i]]+ram[i]}；

其中，fcpu[i] 表示在为主机分配第 i 台虚拟机时，主机中所分配的虚拟机的 CPU 数量的总和；fram[i] 表示在为主机分配第 i 台虚拟机时，主机中所分配的虚拟机的内存大小的总和；cpuValues 表示 cpu 的容量为 cpuValues 的主机；ramValues 表示内存容量为 ramValues 的主机；函数 fcpu[i-1][cpuValues][ramValues] 表示第 i-1 个虚机，放到 cpu 容量是 cpuValues 且内存容量是 ramValues 的主机中时，所得到的 cpu 价值；函数 fram[i-1][cpuValues][ramValues] 表示第 i-1 个虚机，放到 cpu 容量是 cpuValues 且内存容量是 ramValues 的主机中时，所得到的内存价值。

8. 根据权利要求 1 所述的方法，其特征在于：

所述阈值为 90％。

一种动态的资源分配方法

技术领域

[0001] 本发明涉及资源规划技术，特别涉及一种动态的资源分配方法。

背景技术

[0002] 目前，很多运营商和大型企业都在构建自己的基础设施即服务（IaaS，Infrastructure as a Service，），让消费者通过网络，自助获得计算机基础设施服务。因为 IaaS 能够将所有 IT 资源作为规模庞大的资源池，由云平台统一管理，所以如何提高 IT 资产的整体使用率，降低 IT 成本投入，对于运营商和企业就显得尤为重要。

[0003] 在 IaaS 中，资源分配的动态性很强，产生大量资源碎片的可能性很高。资源碎片是指散落在资源池主机上的较小空闲资源，由于其容量小，它们很难被利用而白白浪费。而资源池中碎片的多少，是影响 IT 资产使用率的一个重要因素。一般云管理系统在进行资源分配时，只是使用简单的方法去规划资源的使用，因此无法减少碎片的产生，最终导致资源池总容量有空余，但却已经无法部署指定规格的虚拟机。

[0004] 目前，现有技术中所述使用的大多数的资源分配方法，是为达到资源负载分配最优、优化资源使用性能而设计的，很少有以减少资源碎片为目的的资源分配方法。而且，即便是在资源分配方法中，涉及到减少资源碎片，也大多是仅对现有资源情况进行规划，而没有考虑资源动态变化因素对资源分配效果的影响。

发明内容

[0005] 有鉴于此，本发明提供一种动态的资源分配方法，从而可以在 IaaS 环境中有效地减少资源碎片，实现物理资源的充分利用。

[0006] 本发明的技术方案具体是这样实现的：

[0007] 一种动态的资源分配方法，该方法包括：

[0008] A、根据最近的 n 个时间段的新增虚拟机数目，预测每种规格的虚拟机在下一个时间段中的增量；

[0009] B、根据预测结果，对系统资源池中的可用资源进行资源分配配比规划，确定各台主机和虚拟机的分配优先级别，以及每台主机上分配的虚拟机的规格和数目；

[0010] C、当用户申请虚拟机时，按照各个主机的优先级从高到低的顺序以及所述资源分配配比规划，将用户所申请的虚拟机分配给相应的主机；

[0011] D、获取用户所申请的虚拟机的规格以及当前时间段中已被分配的该规格的虚拟机的总数，当被分配的所述规格的虚拟机的总数与资源分配配比规划中所述规格的虚拟机的总数的比值大于预设的阈值时，返回执行步骤 A。

[0012] 较佳的，所述虚拟机的规格包括：

[0013] 1 核 1G；1 核 2G；2 核 2G；2 核 4G；4 核 4G；4 核 8G；4 核 12G；8 核 8G 和 8 核 12G。

[0014] 较佳的，使用二次指数平滑法预测每种规格的虚拟机在下一个时间段中的增量。

[0015] 较佳的，使用如下所述的公式来预测在下一个时间段中的虚拟机增量：

[0016] $F_{k+T} = a_k + b_k T$

[0017] 其中，

[0018] $a_k = 2S_k(1) - S_k(2)$

[0019] $$b_k = \frac{\alpha}{1-\alpha}(S_k(1) - S_k(2))$$

[0020] $S_k(1) = \alpha X_k + (1-\alpha) S_{k-1}(1)$

[0021] $S_k(2) = \alpha S_k + (1-\alpha) S_{k-1}(2)$

[0022] 其中，F_{k+T} 为第 k+T 个时间段的虚拟机增量，T 为第 k 个时间段到下一个时间段的间隔数，a_k 和 b_k 均为参数，$S_k(1)$ 和 $S_k(2)$ 分别为一次指数平滑值和二次指数平滑值，α 为平滑常数，取值范围为 [0,1]。

[0023] 较佳的，所述 α 的数值为 0.9。

[0024] 较佳的，所述根据预测结果，对系统资源池中的可用资源进行资源分配配比规划包括：

[0025] 步骤 B1，统计系统资源池中可用的资源数量以及可用资源在每天主机上的分布情况；

[0026] 步骤 B2，根据统计结果将各个主机按照可用的 CPU 数量从小到大的顺序进行排序并设置优先级，CPU 数量越少的主机的优先级越高；

[0027] 步骤 B3，将预测结果中的各种规格的虚拟机按照虚拟机的 CPU 数量从小到大的顺序进行排序，并根据排序结果形成虚拟机列表；

[0028] 步骤 B4，根据虚拟机的 CPU 数量和内存大小确定各个虚拟机的 CPU 价值和内存价值，并为每台主机设置 CPU 容量和内存容量；

[0029] 骤 B5，按照优先级从高到低的顺序依次对每台主机进行一次资源规划；在每台主机的资源规划中，按照虚拟机的 CPU 数量从小到大的顺序遍历虚拟机列表中的每台虚拟机，将使得当前主机的总价值最大的虚拟机分配给当前的主机，并从虚拟机列表中删除已被分配的虚拟机。

[0030] 较佳的，使用如下所述的状态转移方程对主机进行资源规划：

[0031] fcpu[i] = max{fcpu[i-1][cpuValues][ramValues], fcpu[i][cpuValues-cpu[i]][ramValues-ra m[i]]+cpu[i]}；

[0032] fram[i] = max{fram[i-1][cpuValues][ramValues], fcpu[i][cpuValues-cpu[i]][ramValues-ra m[i]]+ram[i]}；

[0033] 其中，fcpu[i] 表示在为主机分配第 i 台虚拟机时，主机中所分配的虚拟机的 CPU 数量的总和；fram[i] 表示在为主机分配第 i 台虚拟机时，主机中所分配的虚拟机的内存大小的总和；cpuValues 表示 cpu 的容量为 cpuValues 的主机；ramValues 表示内存容量为 ramValues 的主机；函数 fcpu[i-1][cpuValues][ramValues] 表示第 i-1 个虚机，放到 cpu 容量是 cpuValues 且内存容量是 ramValues 的主机中时，所得到的 cpu 价值；函数 fram[i-1][cpuValues][ramValues] 表示第 i-1 个虚机，放到 cpu 容量是 cpuValues 且内存容量是 ramValues 的主机中时，所得到的内存价值。

[0034] 较佳的，所述阈值为 90%。

[0035] 如上可见，在本发明中的动态的资源分配方法中，由于将资源需求预测与资源规

划相结合，使用了更贴近实际情况的二次平滑指数算法预测短期内的资源需求，从而使得资源分配配比规划更具有预见性，从而减少不必要的虚拟机迁移，有效地减少了云环境中物理资源碎片，尽量满足所有虚拟机的资源请求，容纳尽可能多虚拟机，以此提高 IT 资产的使用率，降低 IT 成本投入；另外，由于本发明中所提供的上述方法中可以通过动态反馈机制，在需要的时刻，及时地重新对资源需求进行预测和进行规划，确保可以用最小的计算代价，让资源分配配比规划尽可能地贴近实际情况，以满足客户的需求；此外，由于本发明中所提供的方法更贴近 IaaS 特点，可以对主机和虚拟机进行双排序，因此可以支持任意规模的 IaaS 环境，适用性十分广泛。

附图说明

[0036] 图 1 为本发明实施例中的动态的资源分配方法的流程示意图。

[0037] 图 2 为本发明一个具体实施例中的预测结果比较示意图。

具体实施方式

[0038] 为使本发明的目的、技术方案及优点更加清楚明白，以下参照附图并举实施例，对本发明进一步详细说明。

[0039] IaaS 中每个时间段内虚拟机数目的增量可以看成一组时间序列。一般情况下，用户在短期内申请的虚拟机数量不是随机的，是有需求背景支持的，而需求背景一般是稳定的。由此可以预见，短期内某一时间段内的虚拟机增量可以看作是最近时段历史值的延续。因此，可以根据短期内的虚拟机增量的变化趋势，来预测下一时间段内虚拟机的增量。如果能够比较准确地对虚拟机的增量进行预测，则可以根据该预测值，提前对资源池的分配方案进行较好的规划，从而在无虚机迁移调整的情况下，尽量减少资源碎片的产生。

[0040] 但是在实际应用环境中，即便是在基本稳定的需求背景下，用户所申请的虚拟机数量也会有变动，因此导致某一时间段内虚拟机数目的增量可能会与预测值不同。例如，当某一时间段内某规格的虚拟机的累计增量已经达到预测值，但是用户可能还会继续申请该规格的虚拟机，从而使得原有的资源分配规则无法继续使用，此时就需要重新预测下一时间段内不同规格虚拟机的增量，重新对资源池中的资源进行规划和分配。因此，为了更加真实地反映用户申请虚拟机的需求，在本发明的技术方案中，需要设定在未来时间中，根据用户申请的实际情况，何时调整资源规划和分配规则，可以用最小的计算代价，达到减少物理资源碎片，容纳尽可能多虚拟机的目的。

[0041] 因此，在本发明的技术方案中，提出了一种动态的资源分配方法。

[0042] 图 1 为本发明实施例中的动态的资源分配方法的流程示意图。如图 1 所示，本发明实施例中的动态的资源分配方法主要包括如下所述的步骤：

[0043] 步骤 11，根据最近的 n 个时间段的新增虚拟机数目，预测每种规格的虚拟机在下一个时间段中的增量。

[0044] 在现有的实际应用环境中，用户在申请虚拟机时，可选的虚拟机的规格是确定。

[0045] 例如，较佳的，在本发明的具体实施例中，所述虚拟机的规格包括：1 核 1G；1 核 2G；2 核 2G；2 核 4G；4 核 4G；4 核 8G；4 核 12G；8 核 8G 和 8 核 12G 等 8 种规格。

[0046] 其中，1 核 1G 表示该虚拟机的 CPU 数量为 1（即 1 核），内存大小为 1G，依此类推。

[0047] 根据实际经验和用户的使用习惯，在每个时间段内，虚拟机数量的增量在短时间内会有一定规律，因此在本发明的技术方案中，通过历史数据（即最近的 n 个时间段的新增虚拟机数目）可以预测未来虚拟机的请求状况。另外，为了避免因用户操作失误，导致某一用户在大量申请某规格的虚拟机之后又删除该申请，影响预测准确度。因此在本发明的技术方案中将一个时间段内，以不同规格的虚拟机数目的增量作为预测目标。

[0048] 随着云计算技术的快速发展，用户对虚拟机的需求会越来越大。从长期来看，用户虚拟机申请数量的时间序列曲线会呈增长趋势，而从短期来看则会有略微波动，是一种总体呈增长趋势的线性时间序列。因此，经过多次实验和比较之后，在本发明的较佳实施例中，将使用更贴近实际情况的用于描述呈线性增长趋势的二次指数平滑法预测每种规格的虚拟机在下一个时间段中的增量。

[0049] 具体来说，对于每种规格的虚拟机，可以使用如下所述的公式来预测在下一个时间段中的虚拟机增量：

[0050] $F_{k+T} = a_k + b_k T$

[0051] 其中，F_{k+T} 为第 k+T 个时间段的虚拟机增量（即预测值），T 为第 k 个时间段到下一个时间段（即预测期）的间隔数，a_k 和 b_k 均为参数，且有：

[0052] $a_k = 2S_k(1) - S_k(2)$

[0053] $$b_k = \frac{\alpha}{1-\alpha}(S_k(1) - S_k(2))$$

[0054] $S_k(1) = \alpha X_k + (1-\alpha) S_{k-1}(1)$

[0055] $S_k(2) = \alpha S_k + (1-\alpha) S_{k-1}(2)$

[0056] 其中，$S_k(1)$ 和 $S_k(2)$ 分别为一次指数平滑值和二次指数平滑值，α 为平滑常数，取值范围为 [0,1]。

[0057] 在本发明的技术方案中，所述 α 的取值可以根据实际应用需要预先设置。例如，较佳的，在本发明的具体实施例中，所述 α 的数值可以为 0.9。

[0058] 通过上述的方法，可以分别预测每种规格的虚拟机在下一个时间段中的增量。

[0059] 可以通过实际的例子对于上述的方法进行验证。图 2 为本发明一个具体实施例中的预测结果比较示意图。如图 2 所示，在该具体实施例中，所采用的历史数据（即最近的 n 个时间段的新增虚拟机数目）的如下：

[0060] arr = [0 ;6 ;17 ;36 ;56 ;78 ;103 ;116 ;126 ;134 ;140 ;144 ;146 ;144 ;138 ;130 ;128 ;127 ;130 ;139 ;148 ;132 ;156 ;167 ;178 ;189 ;190]

[0061] 上述历史数据中一共有 27 个数据，即某一个规格的虚拟机最近的 27 个时间段的新增虚拟机数目，组成一组时间序列，如图 2 中所示的原始数据曲线。根据上述历史数据，使用本发明中的预测方法进行预测后，得到如图 2 中所示的预测结果曲线。

[0062] 从图 2 中可以看出，使用上述方法进行预测之后所得到的预测结果比较准确，在原始数据曲线的趋势发生变化时，预测结果曲线的反应也很迅速。

[0063] 在本发明的技术方案中，每个时间段的长度可以根据实际应用需要预先设置。例如，较佳的，在本发明的具体实施例中，每个时间段的长度为：1 个月或其它时间长度。

[0064] 在本发明的技术方案中，所述 n 的取值可以根据实际应用需要预先设置。例如，较佳的，在本发明的具体实施例中，所述 n 的数值可以为 15。

[0065] 另外，较佳的，在本发明的具体实施例中，上述步骤 11 之前还可进一步包括 ：

[0066] 从虚机表中读取每种规格的虚拟机在最近的 n 个时间段内的新增虚拟机数目 ；

[0067] 其中，所述虚机表中存储有各种规格的虚拟机在各个时间段内的新增虚拟机数目。

[0068] 较佳的，在本发明的具体实施例中，所述新增虚拟机数目为当前时间段内用户新申请的虚拟机数目与销毁的虚拟机数目之差。

[0069] 较佳的，在本发明的具体实施例中，所述虚机表存储在 CloudStack 云平台中。

[0070] 步骤 12，根据预测结果，对系统资源池中的可用资源进行资源分配配比规划，确定各台主机和虚拟机的分配优先级别，以及每台主机上分配的虚拟机的规格和数目。

[0071] 在本发明的技术方案中，本步骤的目的在于 ：利用步骤 11 中的预测结果，在预测出的每种规格的虚拟机的新增数目的情况下，如何对系统中现有的空闲资源（即可用资源）进行规划和分配，可以达到在避免迁移的同时，减少资源碎片，满足所有虚拟机的资源请求。

[0072] 所述资源分配配比规划，是指对资源池中物理机的剩余资源（CPU、内存）进行合理规划，从而保证每台主机上的资源能够得到最大化利用，尽量减少资源碎片。

[0073] 由于主机已占用的资源越多，则越容易产生碎片。因此，在本发明的技术方案中，应当先对剩余可用资源较少的主机进行资源分配，从而可以保证剩下的主机都是剩余资源较多的主机，也就是相对不容易产生碎片的主机 ；此外，考虑到占用资源越少的虚拟机，资源分配的灵活性越高，利于填充资源碎片。所以，为了保证虚拟机分配能够充分利用系统中的资源，在本发明的技术方案中，将先对规格大的虚拟机进行资源分配。综上所述，在本发明的技术方案中，系统资源的规划和分配应当遵循两个规则 ：剩余可用资源较少的主机先规划，规格较大的虚拟机先分配。

[0074] 在本发明的技术方案中，可以使用多种方式来对系统资源池中的可用资源进行资源分配配比规划。以下将以其中的一种实施方式为例，对本发明的技术方案进行介绍。

[0075] 例如，较佳的，在本发明的具体实施例中，所述根据预测结果，对系统资源池中的可用资源进行资源分配配比规划包括 ：

[0076] 步骤 121，统计系统资源池中可用的资源数量以及可用资源在每天主机上的分布情况。

[0077] 步骤 122，根据统计结果将各个主机按照可用的 CPU 数量从小到大的顺序进行排序并设置优先级，CPU 数量越少的主机的优先级越高。

[0078] 步骤 123，将预测结果中的各种规格的虚拟机按照虚拟机的 CPU 数量从小到大的顺序进行排序，并根据排序结果形成虚拟机列表。

[0079] 步骤 124，根据虚拟机的 CPU 数量和内存大小确定各个虚拟机的 CPU 价值和内存价值，并为每台主机设置 CPU 容量和内存容量。

[0080] 在本发明的技术方案中，可为每一个虚拟机设置两个价值 ：CPU 价值和内存价值。其中，所述 CPU 价值的值为该虚拟机的 CPU 数目，所述内存价值的值为该虚拟机的内存大小。例如，第 i 个虚拟机的 CPU 价值和内存价值分别为 cpu[i] 和 ram[i]。

[0081] 另外，在本发明的技术方案中，还可为每台主机设置两个容量 ：CPU 容量和内存容量。其中，所述 CPU 容量的值表示主机中所分配的虚拟机的 CPU 数目的总和，所述内存价值

的值表示主机中所分配的虚拟机的内存大小的总和。

[0082] 步骤 125，按照优先级从高到低的顺序依次对每台主机进行一次资源规划；在每台主机的资源规划中，按照虚拟机的 CPU 数量从小到大的顺序遍历虚拟机列表中的每台虚拟机，将使得当前主机的总价值最大的虚拟机分配给当前的主机，并从虚拟机列表中删除已被分配的虚拟机。

[0083] 较佳的，在本发明的具体实施例中，可以使用如下所述的状态转移方程对主机进行资源规划：

[0084] fcpu[i] = max{fcpu[i-1][cpuValues][ramValues], fcpu[i][cpuValues-cpu[i]][ramValues-ra m[i]]+cpu[i]}；

[0085] fram[i] = max{fram[i-1][cpuValues][ramValues], fcpu[i][cpuValues-cpu[i]][ramValues-ra m[i]]+ram[i]}。

[0086] 其中，fcpu[i] 表示在为主机分配第 i 台虚拟机时，主机中所分配的虚拟机的 CPU 数量的总和；fram[i] 表示在为主机分配第 i 台虚拟机时，主机中所分配的虚拟机的内存大小的总和；cpuValues 表示 cpu 的容量为 cpuValues 的主机；ramValues 表示内存容量为 ramValues 的主机；函数 fcpu[i-1][cpuValues][ramValues] 表示第 i-1 个虚机，放到 cpu 容量是 cpuValues 且内存容量是 ramValues 的主机中时，所得到的 cpu 价值；函数 fram[i-1][cpuValues][ramValues] 表示第 i-1 个虚机，放到 cpu 容量是 cpuValues 且内存容量是 ramValues 的主机中时，所得到的内存价值。

[0087] 根据上述状态转移方程可知，如果不放入预测要创建的第 i 台虚拟机，则将预测要创建前 i-1 台虚拟机放入 CPU 资源为 cpuValues，内存资源为 ramValues 的主机中 CPU 和内存能达到的最大值；如果放入第 i 台虚拟机，则转化为将前 i-1 台虚拟机放入 CPU 资源为 cpuValues-cpu[i]，内存资源为 ramValues-ram[i] 的主机中 CPU 和内存能达到的最大值。

[0088] 由此可知，在本步骤中，将依次对每台主机进行一次资源规划。在每次资源规划中，按虚拟机从小到大的顺序遍历虚拟机列表中的每台虚拟机，计算如果将该虚拟机分配给当前主机，该当前主机的 CPU 和 RAM 总价值是否将比把前一个虚拟机分配给当前主机时的总价值大。经过一轮比较之后，记录遍历过程中，该当前主机被分配哪台虚拟机时所对应的总价值最大，则选定该虚拟机分配给当前主机。当当前主机装满（即剩余资源为 0 或剩余资源已无法再分配给任何一个虚拟机列表中的虚拟机）时，根据之前的记录，找出分配到该主机的虚拟机清单，并将这些分配给当前主机的虚拟机从虚拟机列表中删除，得到的新的虚拟机列表，然后重新遍历下一台主机，直到所有主机处理完成。

[0089] 通过上述的步骤 121 ～ 125，即可确定每台主机上所分配的虚拟机的规格和数目，完成对系统资源池中的可用资源进行资源分配配比规划。

[0090] 步骤 13，当用户申请虚拟机时，按照各个主机的优先级从高到低的顺序以及所述资源分配配比规划，将用户所申请的虚拟机分配给相应的主机。

[0091] 步骤 14，获取用户所申请的虚拟机的规格以及当前时间段中已被分配的该规格的虚拟机的总数，当被分配的所述规格的虚拟机的总数与资源分配配比规划中所述规格的虚拟机的总数的比值大于预设的阈值时，返回执行步骤 11。

[0092] 在本发明的技术方案中，所述阈值可以根据实际应用需要预先设置。例如，较佳的，在本发明的具体实施例中，所述阈值可以为 90％。

[0093] 在本发明的技术方案中，在上述步骤 12 中完成资源分配配比规划之后，当用户申请虚拟机时，即可按照各个主机的优先级从高到低的顺序，并根据上述步骤 12 中制定好的资源分配配比规划将用户所申请的虚拟机放置到相应的主机上。同时，还将判断已被分配的该规格的虚拟机数目是否已接近上限，即已被分配的该规格的虚拟机的总数是否大于预设的阈值。

[0094] 例如，当已存在的该规格的虚拟机的数目，与按资源分配配比规划的资源池中可容纳该规格虚拟机数目的比值超过某个预设的阀值时，则表示如果按照之前的资源分配配比规划，继续分配该规格的虚拟机，则将无可用资源，会导致所申请的虚拟机将无法部署。所以，需要重新对资源池的资源进行分配策略调整。更进一步的，如需要调整，则还可以先判断是否物理资源已不够，若资源确实不够，则需要告知管理员；否则，重新返回执行步骤 12，对资源池进行规划和分配，制定新的分配策略，将原有的分配策略更新为调整后的虚拟机分配策略，等待下一次用户请求。

[0095] 通过上述的步骤 11 ～ 14，即在 IaaS 环境中有效地减少资源碎片，实现物理资源的充分利用。

[0096] 例如，可以将现有技术中的方法与本发明中所提供的上述方法进行比较：

[0097] 设 IaaS 环境中现有 5 台主机，各个主机中剩余可用资源分别为：

[0098] 5 核 10G，6 核 10G，7 核 10G，8 核 10G，8 核 10G。

[0099] 所需分配的虚拟机的数目为 20 台，每种规格的虚拟机的数目如下表所述：

[0100]

虚拟机规格	1 核 1G	1 核 2G	2 核 2G	2 核 4G	
虚拟机数目	4 台	3 台	6 台	7 台	总计：20 台

[0101] 如果使用现有技术中的方法，则最终的随机分配结果如下：

[0102]

主机	5 核 10G	6 核 10G	7 核 10G	8 核 10G	8 核 10G	
主机上部署的虚拟机	2 核 2G 2 核 2G 1 核 1G	2 核 4G 1 核 1G 1 核 2G 1 核 1G 1 核 1G	2 核 4G 2 核 4G 1 核 2G	2 核 4G 2 核 4G 1 核 2G	2 核 2G 2 核 2G 2 核 2G 2 核 2G	
已部署的虚拟机数目	3 台	5 台	3 台	3 台	4 台	总计： 18 台
主机上的剩余资源	0 核 5G	0 核 1G	2 核 0G	3 核 0G	0 核 2G	总计： 5 核 8G

[0103] 如上表所示可知，一共部署了 18 台虚拟机，剩余两台 2 核 4G 的虚拟机无法部署。此时主机所剩余的资源总和完全可以满足这两台虚拟机，但是是每台主机都不满足需求，无法部署剩余的两台虚拟机。此时系统资源的碎片率为：CPU：14.7％，RAM：16％。

[0104] 而如果使用本发明中所提供的方法，则可以得到如下所述的分配结果：

[0105]

主机	5 核 10G	6 核 10G	7 核 10G	8 核 10G	8 核 10G	
主机上部署的虚拟机	2 核 4G 2 核 4G 1 核 2G	2 核 4G 2 核 4G 2 核 2G	2 核 4G 2 核 2G 2 核 2G 1 核 2G	2 核 4G 2 核 2G 2 核 2G 2 核 2G	2 核 4G 1 核 2G 1 核 1G 1 核 1G 1 核 1G 1 核 1G	
已部署的虚拟机数目	3 台	3 台	4 台	4 台	6 台	总计: 20 台
主机上的剩余资源	无剩余	无剩余	无剩余	无剩余	1 核 0G	总计: 1 核 0G

[0106] 如上表所示,20 台虚拟机全部被部署,主机剩余的资源总和为 1 核 0G,系统资源的碎片率为 :CPU:2.94%,RAM:0%。

[0107] 对比上述两种结果结果,按照本发明中所提供的方法分配虚拟机所产生的资源碎片比现有技术中的方法减少了 15%左右,远低于现有技术中的方法中的资源碎片率。

[0108] 综上可知,在本发明中的动态的资源分配方法中,由于将资源需求预测与资源规划相结合,使用了更贴近实际情况的二次平滑指数算法预测短期内的资源需求,从而使得资源分配配比规划更具有预见性,从而减少不必要的虚拟机迁移,有效地减少了云环境中物理资源碎片,尽量满足所有虚拟机的资源请求,容纳尽可能多虚拟机,以此提高 IT 资产的使用率,降低 IT 成本投入 ;另外,由于本发明中所提供的上述方法中可以通过动态反馈机制,在需要的时刻,及时地重新对资源需求进行预测和进行规划,确保可以用最小的计算代价,让资源分配配比规划尽可能地贴近实际情况,以满足客户的需求 ;此外,由于本发明中所提供的方法更贴近 IaaS 特点,可以对主机和虚拟机进行双排序,因此可以支持任意规模的 IaaS 环境,适用性十分广泛。

[0109] 以上所述仅为本发明的较佳实施例而已,并不用以限制本发明,凡在本发明的精神和原则之内,所做的任何修改、等同替换、改进等,均应包含在本发明保护的范围之内。

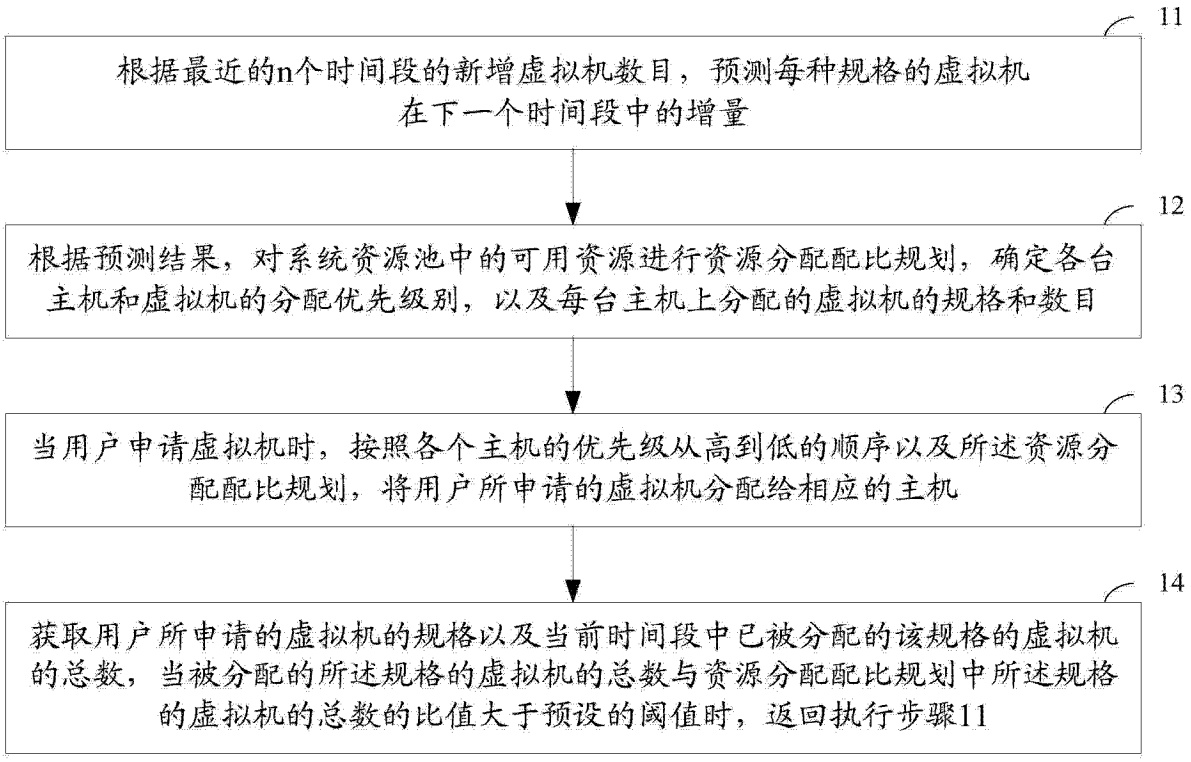

11
根据最近的n个时间段的新增虚拟机数目，预测每种规格的虚拟机在下一个时间段中的增量
12
根据预测结果，对系统资源池中的可用资源进行资源分配配比规划，确定各台主机和虚拟机的分配优先级别，以及每台主机上分配的虚拟机的规格和数目
13
当用户申请虚拟机时，按照各个主机的优先级从高到低的顺序以及所述资源分配配比规划，将用户所申请的虚拟机分配给相应的主机
14
获取用户所申请的虚拟机的规格以及当前时间段中已被分配的该规格的虚拟机的总数，当被分配的所述规格的虚拟机的总数与资源分配配比规划中所述规格的虚拟机的总数的比值大于预设的阈值时，返回执行步骤11

图 1

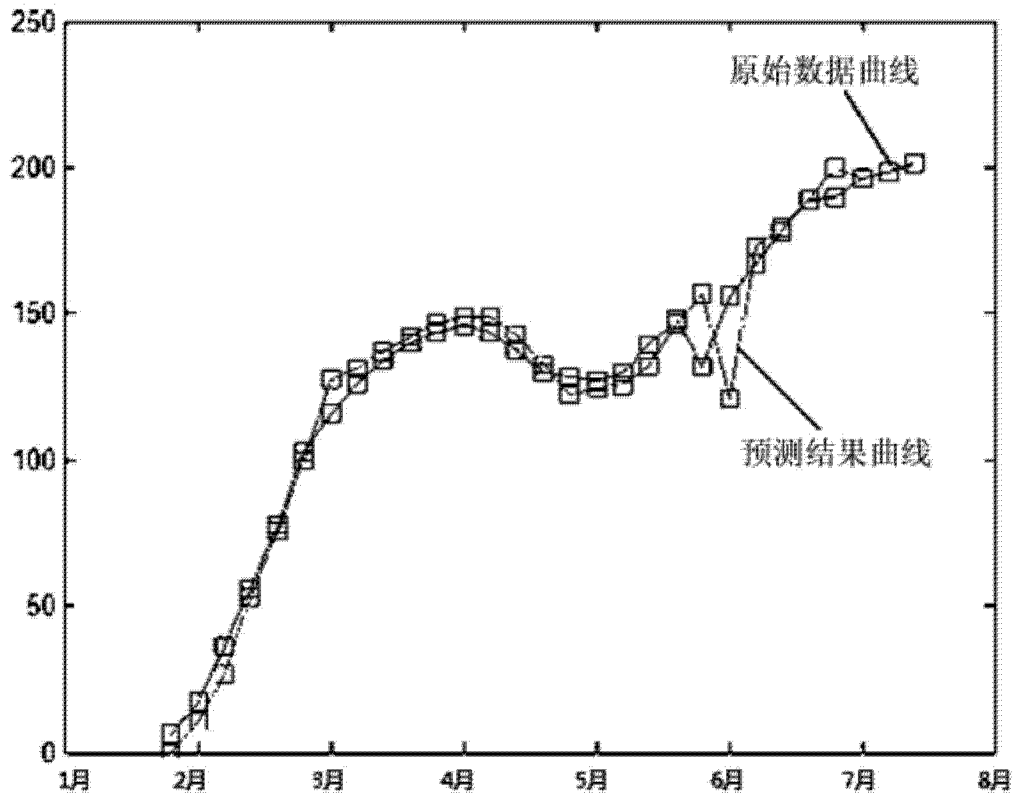

250
200
150
100
50
0
1月
2月
3月
4月
5月
6月
7月
8月
原始数据曲线
预测结果曲线

图 2