云平台虚拟资源伸缩策略及负载预测探讨

吕书林[1]，李金旭[2]

(1.河南广播电视大学，河南 郑州 450008；2.郑州幼儿师范高等专科学校，河南 郑州 450000)

摘 要：通过分析当前流行的虚拟资源伸缩技术，文章提出基于需求预测机制和实时监测相结合的资源伸缩策略，能从一定程度上解决虚拟化资源在调度过程中造成的服务时效性差和资源利用率低的问题。

关键词：云平台；负责均衡；虚拟资源伸缩

1 研究意义

云平台的优势是能提高资源的利用率，降低能源消耗，通过引入虚拟化技术等方式，来细化物理资源分配单元，从而提升系统分布的密度，提高系统使用效率，降低对物理设备的需求，从而降低IT设备投入，降低能耗节约成本。

云平台中一个很关键的技术就是资源的虚拟化，并根据不同业务在不同时间对资源需求的不同特点，来动态地调节平衡虚拟化资源的配置，从而实现在现有物理资源规模情况下资源的高效利用[1]。那么，虚拟化资源的分配调度策略便成为云平台技术研究的重点之一。虚拟化资源调度策略中研究的关键问题有两个，一是何时需要进行虚拟资源的扩充和伸缩，二是通过什么方式进行虚拟资源的伸缩。这两个问题的解决，使云平台能够在最佳时机以最优的策略进行资源的动态配置。

2 虚拟资源伸缩技术概述

虚拟化资源的管理可以通过虚拟机管理程序来实现，随着虚拟化技术的不断发展，目前针对虚拟化资源的伸缩技术研究主要分为两个层面：一是基于水平层面的水平伸缩[2]，一是基于垂直层面的垂直伸缩。

水平层面的伸缩是基于虚拟机层面来实现资源的动态配置，通过增减虚拟机的数量来完成资源的补充和释放。水平伸缩是一种非常容易实现的虚拟资源伸缩方式，在目前的各种云平台中被广泛使用。通过这种方式可以调整的资源规模相对较大，但同时因调配基于虚拟机层面，会造成不必要的资源浪费。并且部署和启动虚拟机需要较长的时间才能完成，会造成服务等待的情况，对平台服务的实效性有一定的影响。

垂直层面的伸缩是基于正在运行的虚拟机自身的资源层面来实现资源的动态调配。目前的技术已经可以做到在虚拟机正常运行的情况下完成在线虚拟机资源的动态调整。并且，基于该层面的资源伸缩过程很短，在毫秒级时间内即可完成。垂直层面资源伸缩技术的实现方式有多种，可以是同一台物理机上的虚拟机之间进行资源调度，也可以将物理机上空闲的资源分配给该物理机上的其他虚拟机。这两种方式适用的应用场景不同，在进行伸缩方式选择时需要根据具体情况分析，综合考虑。

(1) 第一种方式是基于同一台物理机上不同虚拟机之间服务类型不同、服务请求分布不均、处理能力不同等因素为出发点，通过虚拟机之间的资源调配，实现在扩充虚拟机资源要求的情况下又不占用其他空闲资源。它的特点是资源调配时间短，但同时存在调配的资源规模相对较小的问题，不能满足较大规模的资源伸缩需要。

(2) 第二种方式是利用物理机的空闲资源进行动态的调配，在虚拟机需要资源扩充时申请空闲资源完成资源的增加，在资源需求降低时释放部分空闲资源到物理机上。它的特点是充分利用物理机空闲资源，能够根据用户需求的变化做到快速响应，但同时也存在第一种方式中资源调度规模较小的问题。

3 资源伸缩策略探讨

为了实现云平台虚拟资源的可伸缩性，平台应该能够根据服务所需资源的变化动态增加或者减少虚拟资源，当需要增加虚拟资源时，最简便的方式是当监测到虚拟资源利用率超过约定上限时启动虚拟机来满足资源扩充。但是虚拟机的启动会耗费一定的时间，在这段时间内，因虚拟机尚在调配过程中不能向用户及时提供服务，将会影响服务性能，造成对用户服务请求响应的滞后。

基于上述情况，人们对云平台的资源伸缩便有了具体的要求。云平台需要能够根据具体的服务请求来动态选择适合的方式进行虚拟资源的增减。可以通过两种方式来改善动态调配过程中存在的服务时效性和资源调度成本问题。

(1) 一是为不同的应用设定资源使用底限和上限，并加入对下一阶段资源需求预测机制，根据动态的预测结果来及时补充或释放虚拟资源。由于资源的调配是基于预测结果进行，这就给资源的调度时间提供了一个比较宽松的环境，基于该策略的调度可以在水平和垂直两个层面进行。

(2) 二是通过实时监控的方式进行资源的实时扩展和实时释放。因为实时扩展过程中要尽量保证服务的实效性，那么就要求云平台能够较快地完成调度过程，我们可以选择基于垂直层面的第一种资源伸缩方式来实现。释放资源不存在影响服务时效性的问题，对调度时间的要求比较低，这使得预先释放资源的意义不大，所以可以在监测到资源使用率低于规定的下限时进行实时资源释放。

为了进一步提高资源伸缩的实际效率，可以综合考虑以上两种方法，并在资源调度过程中引入基于预测伸缩和实时伸缩相结合的方法来解决实际问题。

基于垂直层面进行资源伸缩的两种方式适合实时资源调整规模较小的情况，而虚拟机层面的水平伸缩适合对实时性要求不高的资源调整规模较大的场景。所以，在基于未来

作者简介：吕书林(1984—)，男，河南郑州；研究方向：分布式存储，云计算。

资源需要预测的伸缩过程中，根据平台负载预测结果对平台进行基于水平伸缩（虚拟机层面）和基于垂直层面第二种伸缩方式进行资源的伸缩。在实时扩展过程中，对平台进行基于垂直层面的两种方式进行扩展，在实时释放过程中，对平台进行水平层面和基于垂直层面第二种伸缩方式的资源释放。

4 云平台负载预测方法探讨

4.1 云平台负载预测研究现状

负载预测对于云平台资源伸缩来说是十分必要的，并且当前针对云平台的负载预测问题已经有了相当多的研究成果，这些研究成果对于云平台的资源伸缩技术的改进和进步都有不同的贡献。例如，有研究人员提出的一种基于自回归滑动平均模型（Autoregressive Moving Average Model, AMAM）的方法进行负载预测，方法是根据服务器负载动态变化的统计分析结果所总结出的负载变化规律，提出了基于服务器负载和时间序列的预测方法，并且在应用中使用该方法对网关服务器的负载情况进行预测。另外还有一种基于二阶自回归滑动平均模型的负载预测方法，采用了类似的策略。还有研究人员提出了一种基于负载热点检测预测法，该方法的实现分为两个步骤：第一步是依据针对负载的监测结果对负载未来发展趋势作出预判，然后依据作出的趋势判断使用相应的负载预测算法进行预测；这种基于负载热点的预测方法，能够有效预测出负载的突发性增长并及时报警，从而为资源的及时扩充提前做好准备。

这几种方法的优点是计算量相对较小，但方法中使用到的预测模型是相对固定的，模型不能够依据负载实际的变化情况而进行动态的调整，这就容易出现预测结果不够准备的结果。

除以上几种方法外，还有基于模式匹配和状态迁移的负载预测方法，如PRESS。方法首先使用信号处理技术来标识重复的负载片段，此时会出现两种情况。一是发现当前的负载片段与以往的负载出现了重合，此种情况下方法就依据重复片段进行负载预测。二是发现当前的负载与以往的负载不存在相似性，此种情况下就使用基于统计的状态驱动方法来捕捉短时间内的负载模式，并使用离散时间的马尔可夫链来预测未来的负载情况。然而，马尔可夫预测模型存在预测精度低、误差大、适用范围小等缺点。这种预测方法是通过对历史负载数据分析来预测服务负载的，负载预测结果比较准确，但是这类方法通常需要大量的计算和较长的时间，为平台带来很大计算开销。徐风苓、孟祥武等使用的基于移动任务上下文相似度的协同过滤推荐算法，也是基于字符串匹配和模式匹配的方法，首先来判断识别相似的负载片段，然后再进行负载的预测。

4.2 云平台负载的特点

云平台的负载是比较复杂的，很多情况下单靠负载分布曲线是无法准确描述的。具体来说，存在以下几方面的特点。

（1）时间关联性，上一时间段的负载对下一时间段的负载存在很大的影响。

（2）波动性，负载可能在很长一段时间内相对稳定，但也存在突然波动很大的偶发情况，不能用一般的线性规律模型进行描述。

（3）自相似性，负载在不同的时间周期内在某些特定的时间段内存在一定的相似性。

4.3 云平台负载预测方法的选取建议

通过对以上云平台负载特点的分析，我们可以得出一些建议。针对较短时间内云平台负载存在的线性相关性，我们可以基于此建立线性模型来进行预测，也就是说根据云平台当前的负载情况来预测下一段的负载是可行的；针对云平台负载变化的历史自相似性，我们也可以建立基于历史数据的负载估测模型或采用模式识别算法来预测运行过程中相似时期的负载情况；针对负载的波动性特点，我们需要在预测算法中加入相应的调整和修正策略，来尽可能地保证预测的准确性。在进行模型建立和算法设计时，同时需要考虑模型和算法的复杂度，尽量对算法本身进行优化，以减少负载预测时带来的额外开销，以降低对平台性能的影响。

5 结语

本文探讨了云平台中负载预测和资源伸缩策略，建议在实际伸缩时对平台运行状态进行监测，当平台资源利用率超出预定上限时使用基于垂直层面相关伸缩策略及时扩充资源来满足应用需求，在平台资源利用率低于约定下限时释放资源。同时，对云平台的负载预测方法提出相关的建议。从以上思路出发，使得云平台在满足更多服务请求的前提下能够尽量减少资源的使用，并降低平台成本。

[参考文献]

[1]ZHUYU Z, PEIYU L, ZHENFANG ZHU, et al. Improved visit statistical method and calculation in user's interest degree[J].Computer Engineering and Design, 2011(2): 424-426.

[2]徐风苓，孟祥武，王立才.基于移动任务上下文相似度的协同过滤推荐算法[J].电子与信息学报，2011(11)：2785-2789.

Discussion of virtual resources expansion strategy and load prediction based on cloud platform

Lyu Shulin[1], Li Jinxu[2]

(1.Henan Radio and TV University, Zhengzhou 450008, China; 2.Zhengzhou Preschool Education College, Zhengzhou 450000, China)

Abstract: Based on the analysis of the current popular virtual resources expansion technology, this paper suggests a resource expansion strategy that combines the demand forecasting mechanism and real-time monitoring, which can solve the poor timeliness problems and low resource utilization rate in the scheduling process of virtualization resources to some degree.

Key words: cloud platform; responsible for the equilibrium; virtual resource scaling