## 引言

如今，随着云计算的快速发展，涌现出许多基于云资源的系统应用。这些云资源都是由云供应商提供的按需付费的资源。在复杂多变的、不确定的、弹性的云中，资源分配无疑成为一个挑战。一方面，系统用户需要良好的服务质量，这其中包含系统的响应时间、网络带宽、传输时延等；另一方面，软件开发商希望所租赁的云资源开销尽可能小。过度追求系统服务质量，则会导致云资源开销过大；相反，低的云资源开销，则会导致系统服务质量的下降。

自适应资源分配在这种背景下被提出，而实现自适应资源分配的方法有很多，传统的方法是基于规则驱动，需要专业的工程师对每种规则进行说明，这种方法花费大量的时间来编写资源分配规则，投入的管理成本过高且实现困难。

近些年，机器学习与控制论相结合的方式被人们大量研究。机器学习是通过对云中大量历史数据的训练学习，使系统从大量数据中学习专业知识，并得出QoS预测模型。该模型能够通过给定负载量和资源分配方案计算出QoS值，最后通过群智能算法搜索目标资源配置方案。但是，这需要收集大量的历史数据，不断训练，才能保证模型预测的准确性。

控制理论通过使用反馈控制器，根据测量的输出动态调整系统的行为来保证QoS。Rao等人[16]开发了具有自适应输出放大和灵活规则选择的自调整模糊控制器，并进一步设计了支持自适应多目标资源分配和服务差异化的两层QoS供应框架。Kalyvianaki等人提出了一种新的资源管理方法，该方法将卡尔曼滤波器合并到反馈控制器中，从而动态地将CPU资源分配给托管服务器应用程序的虚拟机。Farokhi等人将反馈控制回路引入到混合控制器的设计中。在设计中，他们将分配的内存作为一个控制旋钮，将应用程序性能和VM内存利用率作为反馈参数，目的是在工作负载变化的情况下满足应用程序性能约束。然而，需要大量的反馈迭代才能找到合适的资源分配计划，这可能会导致中断虚拟机的高开销。

于是，提出机器学习与控制论相结合的方法来实现自适应资源分配。该方法适用于没有太多历史数据集训练模型的情景，构造出一个迭代的QoS预测模型。在该模型中，工作负载的变化是给定的，因此框架是正交于工作负载预测的。

而我们发现，当工作负载在一段时间内固定不变的时候，如图1所示，只需在每一段负载开始阶段进行资源分配，该配置就能适应整段负载。当负载在一段时间内如图2所示复杂波动的时候，此时原来一段时间调整一次资源配置方案会导致这段时间内服务质量下降以及整体资源消费增加。于是本文基于该问题，提出一种解决方案：首先确定这段负载的资源调整点；其次构造计算整体资源配置方案的评估函数Fitness；最后通过具体的PSO算法搜索目标资源配置方案。

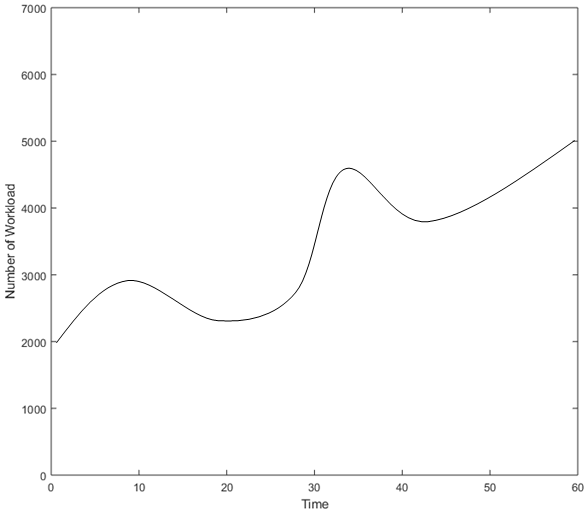
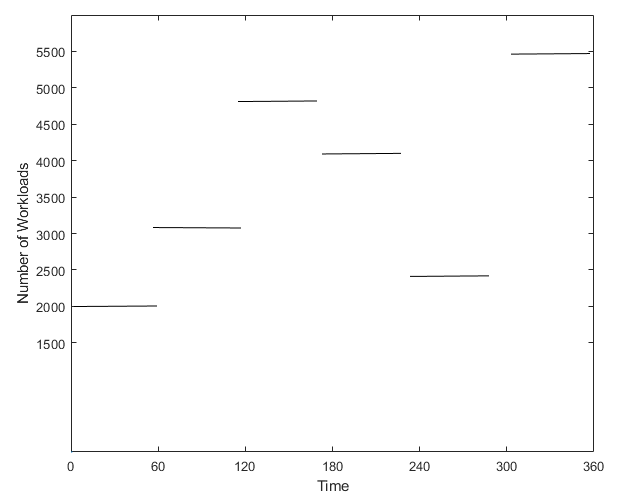


图1 负载变化图（负载平稳） 图2 负载变化图（负载波动）

## 问题定义

表1中描述了特殊标记以及每个标记的具体含义：

表1: 标记及描述

|  |  |
| --- | --- |
| 标记 | 描述 |
| workload | 包含负载量与负载类型 |
| workloadchosen | 通过选中的离散点，近似描述当前时间段的负载 |
| VM | 虚拟机资源，由三种类型的虚拟机及其数量构成 |
| VMchosen | 负载中选中的资源调整点 |
| VMobject | 目标资源配置方案 |
| QoS | 服务质量 |
| QoSmultiPoints | 多点服务总质量 |
| Cost | 虚拟机资源费用 |
| CostL | 租赁虚拟机的费用 |
| CostD | 关闭虚拟机的费用 |
| CostmultiPoints | 多点虚拟机资源总费用 |

基于云的软件质量随着云中运行时环境的改变而改变。影响云环境因素包括外在因素与内在因素，外在因素为负载量workload，而内在因素为分配的资源VM。由于负载是波动的，于是，我们选出几个点workloadchosen来近似描述负载曲线，对应这些点的资源调整点则为VMchosen。表2描述了三种虚拟机具体参数：

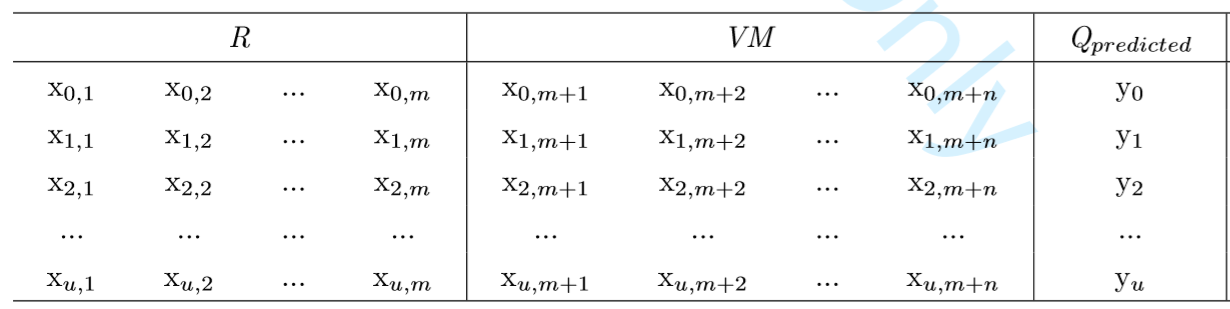
表2: 三种虚拟机类型

|  |  |  |  |
| --- | --- | --- | --- |
| **Property** | **Small** | **Medium** | **Large** |
| CPU | 1 core | 1 core | 1 core |
| Memory | 1G | 2G | 4G |
| CostL | 1.761 RMB | 1.885 RMB | 2.084 RMB |
| CostD | 0.440 RMB | 0.471 RMB | 0.521 RMB |

我们通过收集云中大量的历史数据，如表3所示，其中包括workload、VM配置以及Qpredicted。本文中QoS值主要指的是响应时间。对于这些数据，我们用Workload与VM配置方案作为模型输入，QoS值作为模型输出，通过机器学习的方法训练它们，并生成QoS预测模型。如公式(1)所示：

(1)

现在，有了QoS预测模型，在给定Workload与VM的前提下，我们就可以获取到对应的QoS预测值。而QoSmultiPoints的计算是在QoS预测模型的基础上，对所有调整点求QoS总和。

表3 运行时数据集

这里的目标资源配置方案VMobject不再是一个点的配置方案，而是一段时间内所有选中调整点组成的一个方案，如公式(2)所示，其中vm1，vm2...表示的是每个点的调整方案，n代表有n个调整点。

(2)

云资源中虚拟机的花费包括租赁与关闭消耗，如公式(3)所示：

(3)

本文中我们要考虑的CostmultiPoints表示的是所选中调整点总的租赁消费与相邻两点之间关闭虚拟机的消费。

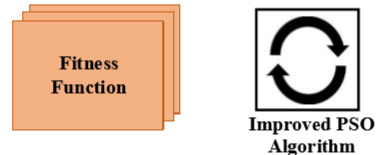
为了权衡服务质量QoS与资源消费Cost之间的关系，需要构造一个适合的评估函数，该评估函数由两个部分构成：QoSmultiPoints与CostmultiPoints

(4)

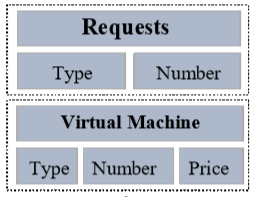
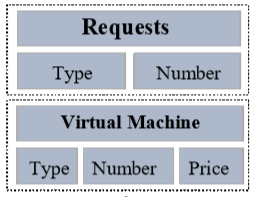
1. 方法概览

**QoS预测模型**

**负载与虚拟机资源**



**PSO Algorithm**



...

**目标配置方案**

图三 方法概览

2.1 负载与虚拟机资源配置

2.2 QoS预测模型

2.3 PSO算法流程

## 实例研究

1. 随机法
2. PSO算法
3. 两种方法的比较

## 总结