云计算环境下 SaaS 服务可伸缩性评估方法研究

张涛[1]，燕静[1]，徐照淼[1]，杨艳丽[1]，朱传曦[1]，成静[2]

（1.西北工业大学 软件与微电子学院，陕西 西安 710072；2.西北工业大学 计算机学院，陕西西安 710075）

摘 要：针对云计算环境下 SaaS 服务可伸缩性的分析评估问题，通过分析 SaaS 服务伸缩性需求和特点，给出其可伸缩性定义，并且提出一种新的可伸缩性评估模型和度量指标，具体包括：资源利用率、资源生产率、收敛性。最后，基于亚马逊 EC2 云计算平台，对评估模型和方法进行实例验证。

关 键 词：性能，服务质量，资源分配，可扩展性，稳定性，可伸缩性，云计算，云测试，服务质量等级（SLA），SaaS 服务

中图分类号：TP311 **文献标志码：**A **文章编号：**1000-2758（2014）06-0998-03

云计算提供了一个可伸缩、按需消费的 SaaS 服务平台，云计算技术基础包括 SOA 架构和软硬件资源虚拟化技术，其目的是为云服务消费者、云参与者、云提供者提供资源共享[1]。在云计算环境下，由于资源的按需动态分配和调整，使得其具备良好的伸缩性，能够动态适应和满足租户业务负载的变化需求[2-3]。因此，可伸缩性是云计算环境下 SaaS 服务的重要质量特性。

当前可伸缩性研究主要集中于传统软件领域，如并行计算、集群系统、分布式系统等[4]，主要分析计算多处理器的加速比，并不适用云计算环境。而针对云计算环境的 SaaS 服务伸缩性研究较少，缺乏对其伸缩性准确定义和有效分析方法。

1 SaaS 服务可伸缩性基本概念

云计算环境下，消费者按照服务等级合约（SLA）向 SaaS 服务提供者租赁 SaaS 服务，SaaS 服务提供者则向云计算供应商租赁软硬件资源。受业务特点、客户习惯等因素影响，SaaS 服务请求负载具有明显的动态波动特征。因此，SaaS 服务的可伸缩性，支持按需扩充和回收资源，对于保障服务质量，提高资源利用率，降低租赁费用至为重要。

云计算环境下，SaaS 服务可伸缩性主要与性能、负载、资源、费用等因素相关。对 SaaS 服务可伸缩性研究应该首先满足服务的性能要求，如满足最大请求响应时间或者最大任务完成时间等。下面给出 SaaS 服务可伸缩性相关概念的定义。

定义 1 可伸缩云计算平台，是指能够按照服务等级合约（SLA）中预先定义可伸缩性参数，为 SaaS 服务动态分配和调整资源，适应 SaaS 服务负载变化，满足服务性能要求的云计算平台。

定义 2 云平台服务在某个时间点的性能定义为多元向量 $\boldsymbol{P}=\{p_1,p_2,\cdots p_i\cdots p_n\}$。$p_i$ 表示第 i 种性能参数。云平台性能主要包括：任务吞吐量、响应时间、完成时间等。

定义 3 云平台服务在某个时间点的负载定义为多元向量 $\boldsymbol{L}=\{l_1,l_2,\cdots l_i\cdots l_n\}$。$l_i$ 表示第 i 种负载参数。云平台负载主要包括：并发服务请求数、网络流量、任务数等。

定义 4 云平台服务资源定义为多元向量 $\boldsymbol{R}=\{r_1,r_2,\cdots r_i\cdots r_n\}$。$r_i$ 表示第 i 种云平台资源，资源类型包括：CPU、内存、缓存、接入带宽、硬盘等。

定义 5 云平台费用模型定义为函数 $C(R)=C(r_1,r_2,\cdots r_i\cdots r_n)$。即根据所分配资源计算费用。

收稿日期：2014-05-10 基金项目：国家自然科学基金（61103003）与航天科技支撑计划（2014HTXGD）资助
作者简介：张涛（1976—），西北工业大学副教授、博士，主要从事移动云测试、软件安全性技术及嵌入式软件技术研究。

2　SaaS 服务可伸缩性评估模型

Amazon EC2 等一些云计算平台提供资源监控工具，允许客户收集其资源和性能数据，但缺乏有效的可伸缩性评估模型和度量指标。本文研究和定义 SaaS 服务可伸缩性评估模型和度量指标。

2.1　资源利用率

可伸缩云平台能够根据业务负载情况，动态调整资源分配，其资源利用率将不断变化和波动。资源利用率过低，则资源浪费严重; 资源利用率过高，则影响服务质量稳定性。可伸缩云平台资源利用率应该维持在 SLA 的预定范围内。

这里定义 SaaS 服务在时间 t_i 的资源利用率为 $RU_{t_i t_i} = C(R_f)/C(R_a)$，其中 R_f 为 t_i 时刻的实际使用资源; R_a 表示 t_i 时刻云平台所分配的资源。

定义资源利用率可伸缩性评价指标为: $RU_{t_i\min}$，$RU_{t_i\max}$，$RU_{t_i\text{avg}}$ 和 $£RU_{t_i}$。其中 $RU_{t_i\min}$ 表示最小资源利用率，$RU_{t_i\max}$ 为最大资源利用率，$RU_{t_i\text{avg}}$ 为平均资源利用率，$£RU_{t_i} = RU_{t_i\max} - RU_{t_i\min}$ 为资源利用率波动范围。

2.2　资源生产率

在可伸缩云平台下，定义 SaaS 服务在时间 ti 的资源生产率定义为: Prdti = Lti/C(Rti)。这里 Lti 表示 ti 时刻的系统负载，C(Rti) 表示该时刻资源租赁费用。

定义资源利用率可伸缩性评价指标为: Prdmin，Prdmax，Prdavg 和£ Prd。其中 Prdmin 表示最小资源生产率，Prdmax 为最大资源生产率，Prdavg 为平均资源生产率，£ Prd = Prdmax-Prdmin 为资源生产率波动范围。在保障性能基础上，可伸缩云平台资源生产率应该维持在 SLA 的预定范围 & Prd 内。即满足: 0《£ Prd《& Prd。

2.3　SaaS 服务收敛性

SaaS 服务请求具有明显的不确定性，使得可伸缩云平台资源动态分配滞后于其负载变化。如图 1 所示，在负载快速变化时，SaaS 服务响应时间、资源利用率等 QOS 指标均出现明显动荡。在滞后一段时间后，通过资源动态调整，SaaS 服务重新收敛恢复稳定状态。

平均收敛时间定义 MTTC 为: 针对某段时间内的特定服务请求模型，由于 SaaS 服务负载的快速变化，导致其 QOS 服务质量波动的平均时间。

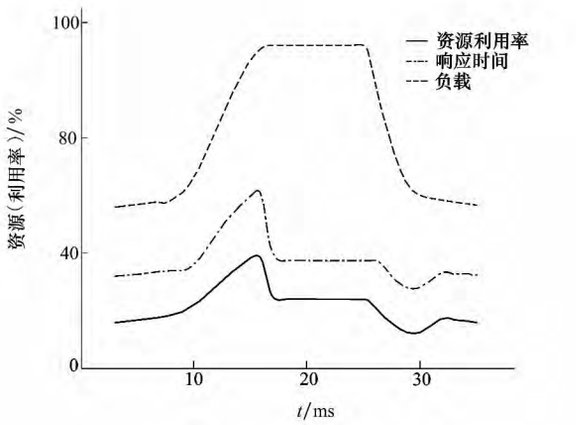

图 1　SaaS 服务收敛示意图

平均波动间隔时间定义为 MTBC: 针对某段时间内的特定服务请求模型，SaaS 服务质量连续处于稳定状态的平均时间。

SaaS 服务可收敛率定义为: SCVG = MTTC/(MTBC+MTTC)。很明显 SCVG 越小，则 SaaS 服务运行越稳定，对负载变化适应性越强。

3　SaaS 服务可伸缩性实例验证

亚马逊 EC2 提供 API 和服务工具来管理和监管资源使用情况，并且可以定义资源可伸缩性范围。本文以亚马逊 EC2 上选定的 SaaS 服务实例为测试对象，确定其业务负载模型和性能指标，测试和评估其可伸缩性。负载和资源利用率的变化关系如图 2 所示，负载和响应时间的变化关系如图 3 所示。

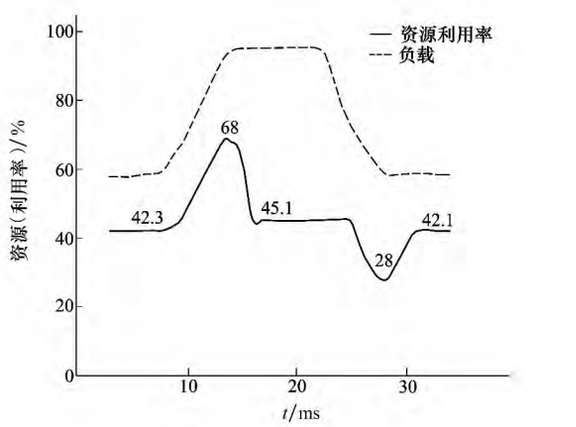

图 2　负载与资源利用率

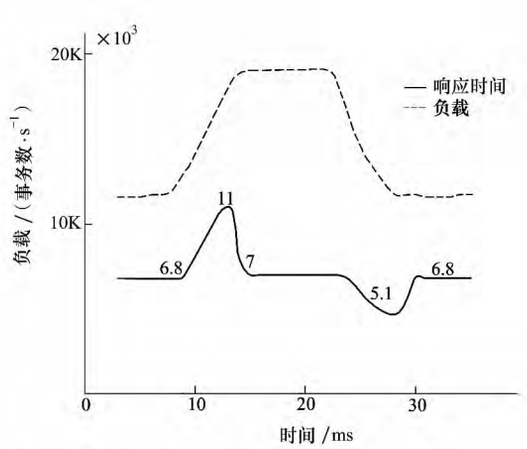

图3 负载与响应时间

4 结 论

当前对可伸缩性研究主要集中于传统软件领域,缺乏对云计算环境下SaaS服务可伸缩性的深入研究。本文系统分析云计算环境下SaaS服务可伸缩性需求,给出可伸缩性定义,并给出其评估模型与度量指标,为SaaS服务可伸缩性分析提供有效的解决方法。下一步,将研究和开发一个基于云平台的SaaS服务可伸缩性分析测试工具,实现对可伸缩性的自动分析与测试。

参考文献:

[1] Gao Jerry, Pattabhiraman P, Bai Xiaoying, Tsai W T, et al. SaaS Performance and Scalability Evaluation in Clouds [C] // IEEE 6th International Symposium on Service Oriented System Engineering, 2011

[2] Tsai W T, Yu Huang, Qihong Shao. Testing the Scalability of SaaS Applications [C] // IEEE International Conference on Service-Oriented Computing and Applications, 2011

[3] Wu Jian, Liang Qianhui, Elisa Bertino. Improving Scalability of Software Cloud for Composite Web Services [C] // IEEE International Conference on Cloud Computing, 2009

[4] Gao Jerry, Pattabhiraman, P, Xiaoying Bai, Tsai W T, et al. SaaS Performance and Scalability Evaluation in Clouds [C] // IEEE 6th International Symposium on Service Oriented System Engineering, 2011

Defining and Evaluating Scalability of SaaS in Cloud

Zhang Tao[1], Yan Jing[1], Xu Zhaomiao[1], Yang Yanli[1], Zhu Chuanxi[1], Cheng Jing[2]

(1.Department of Software Engineering, Northwestern Polytechnical University, Xi'an 710072, China
2.Department of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Cloud computing has recently emerged as a platform which provides an elasticcapability, on-demand usage for Software as a Service (SaaS). It can dynamically adjust resource assignment based on system loads to assure SaaS quality. To address the issues for scalability analysis and evaluation of SaaS in cloud computing, this paper analyzes the requirements and features of scalability and defines a notion of it. Then we propose new formal models and metrics to evaluate and analyze system scalability. Specifically, it includes resource utilization, resource productivity and convergence. In addition, the paper reports case study results based on Amazon's EC2 cloud technology using the proposed models and metrics.

Key words: bandwidth, computer software, convergence of numerical methods, cost functions, hard-disk storage, monitoring, network performance, productivity, program processors, quality of service, resource allocation, response time (control systems), scalability, stability, vectors; cloud computing, cloud testing, service level agreement (SLA), software-as-a-service (SaaS)