参赛作品说明书

云平台弹性建模及量化计算算法

学　　校：长沙学院

学　　院：计算机工程与应用数学学院

专 业 班 级：18 级物联网工程 02 班

姓　　名：张鹏，明晓星

指 导 教 师：周舟

完 成 日 期：2020 年 8 月

目录

一、 研究背景

市场上大数据环境下的文本挖掘存在个性化需求，这种个性化需求导致文本挖掘算法并行化存在很多不确定因素，例如并行化数据规模的不确性，并行化数据对象的不确定，并行化策略的不确定性等。同时，随着大数据应用的推广普及，越来越多用户加入到大数据应用的行列。对于其中大部分的非专业用户来说，与其克服各种技术难题自行构建大数据处理系统，不如直接选择通过大数据处理平台提供的共享资源完成大数据环境下的文本挖掘服务。在此背景下，弹性云计算平台悄然走热。

弹性云计算平台服务的主要特征是其能按需提供、弹性规模、底层设施透明地让用户利用计算资源。弹性云计算平台的云计算资源将以自适应伸缩的方式来提供，且随着任务负载和用户请求的大小来弹性地调整资源的配置。但云计算平台尚未能在资源易用性上为用户解决资源配置难题，文本挖掘并行化的不确定性所带来的计算任务变更，可能需要对集群资源配置进行频繁更改，这将给用户带来极大的挑战：一方面是用户该如何在众多的弹性云平台中选择合适的弹性云平台，另一方面是用户该如何根据任务负载和请求的大小来配置并行计算所需的资源数量。这对云计算平台上数量众多且缺乏专业知识的普通用户而言，这是难以逾越的技术门槛。

弹性云计算平台的性能基础是弹性，可以被认为是云计算的关键优势。弹性是系统以自主方式提供或减少供应计算资源从而适应工作负载变化的程度，使得在每个时间点，可用资源尽可能匹配当前需求。通过动态优化获取的资源总量，弹性用于各种目的。从服务提供商的角度来看，弹性确保能更好地利用计算资源和更多地节约能源，并允许同时服务多个用户。从用户的角度来看，弹性已被用于避免资源供应不足和系统性能降级，并实现成本降低。此外，弹性可以用于其他目的，例如增加本地资源的能力。

弹性的大小也决定了用户所需的并行计算资源，同时用户可以通过弹性的测量来评估弹性云计算平台。最近很多学术界和商业领域的研究人员探讨了弹性机制，并且已经投入了巨大的努力来使云平台能够以弹性方式工作。然而，关于弹性的研究中没有常见的和精确的公式来计算弹性值。现有的研究文献中的弹性定义都是模糊的概念，并没有抓住弹性资源配置的本质。这些弹性公式不适合于量化和测量弹性。此外，没有提出用于量化弹性行为的系统方法。只有定量弹性值才可以在不同的云平台。

二、 国内外研究现状与发展趋势

目前，国内外学者在这方面做了一定的研究，这些研究可分为三类：分别为云平台弹性定量分析[1]-[7]，云平台弹性定性分析[8]-[10]，弹性云平台的发展[11]-[12]，如表 1 所示：

表 1 云平台弹性研究

文献编号	类型	主要思想
[1]	云平台弹性定量分析	通过研究物理信息系统中的资源供给，提出了一种“滞弹性”定义。
[2]		对云计算中的“弹性”进行量化建模和分析计算。
[3]		利用“重配置效果”，“重配置的频率”和“重配置的时间”度量云平台的弹
[4]		弹性被定义为所消耗的时间与平均资源量乘积的倒数。
[5]		所定义的弹性中考虑了对资源过度供给与供应不足的惩罚。
[6]		考虑了“成本弹性”和“质量弹性”。
[7]		根据四个特征如“范围”，“策略”，“目标”和“方法”对“弹性”进行分
[8]	云平台弹性定性分析	“弹性”被认为是快速响应用户请求，接收和释放资源的能力。
[9]		“弹性”被看成是系统能够快速适应负载的度。
[10]		介绍了一种基于 Actor 模型的弹性可伸缩的流处理框架。
[11]	弹性云平台的发展	描述了一个由服务质量驱动的云平台支持弹性应用。
[12]		提出了一种基于应用特征的 PaaS 弹性资源管理机制 AFERM。

第一类为云平台弹性定量分析[1]-[7]。这一类能给予“弹性”形式化的定义，如在文献[1]中，龚红仿等作者提出了滞弹性系统中的滞弹性计算模型和计算方法，给出了 FSDMC 模型中的滞弹性量化的定义;在文献[2]中，李克勤教授对云计算中的“弹性”进行量化建模和分析计算，并使用五个指标度量云平台中的弹性；在文献[3]中，Kuperberg 等作者在考虑云平台三方面特征的情况下，使用单阈值去度量云平台的弹性。这三个方面分别为重配置效果（增加或者删除的资源量，表示资源的自适应粒度），重配置的频率（一段时间内重配置的密度）和重配置的时间（需要多长时间完成资源的重配置）；在文献[4]中，弹性被定义为系统从不正常状态到正常状态所用平均时间与系统平均资源量的倒数;在文献[5]中，Islam 等作者定义了弹性，该弹性考虑了对资源过度供给与资源供给不足的惩罚；在文献[6]中，弹性考虑到了成本弹性（Cost elasticity）和质量弹性(Quality elasticity)，成本弹性指快速响应应用请求所用的开销，质量弹性指资源利用率的服务质量；在文献[7]中，弹性可根据下述四个特征进行分类，这四个特征分别为范围（基础设施，应用，平台），策略（手动的，主动的，被

动的），目标（性能，容量，成本，能耗）和方法（复制，迁移或者重配置）。

第二类云平台弹性定性分析[8]-[10]。这一类在分析云平台弹性特征时，这些特征不能量化。如在文献[8]中，“弹性”被定义为快速响应用户请求，接收和释放资源的能力。同时，文章中也提到，在提供计算资源满足变化负载时，“弹性”具有动态变化的特征；在文献[9]中，“弹性”被定义成系统能够快速适应负载的度，系统以自动的方式提供资源供给与部署，以使得在每个时间点上，可用的资源都尽可能的匹配目前的需求；在文献[10]中，詹杭龙等作者介绍了一种基于 Actor 模型的弹性可伸缩的流处理框架。

第三类为弹性云平台的发展[11]-[12]。在这一方面，国内外学者也进行了一定的探讨，如在文献[11]中，Calheiros 等作者描述了一个由服务质量驱动的云平台支持弹性应用。针对云计算环境中弹性资源管理所面临的问题，即在保障服务质量的前提下，尽可能地节省服务器资源；在文献[12]中，魏豪等作者在考虑不同应用具体特征的情况下，提出了一种基于应用特征的 PaaS 弹性资源管理机制 AFERM (Application Feature based Elastic Resource Management Mechanism)，实验证明了作者所提出算法的正确性和有效性。

综上所述，目前有一些关于云弹性的定义和研究，但这些定义具有不可追踪性，且缺乏统一的共识量化和精确计算云平台的弹性值。本课题主要研究点正是以研究云平台的“弹性”，给出了云平台“弹性”的形式化定义，然后将云平台建模为排队系统，使用时间连续的马尔科夫链去精确计算云平台的弹性值，从而帮助资源提供者预测和优化云平台性能，进而降低能耗成本。

三、 研究目的

（1）揭示影响云平台“弹性”的因素；

（2）给出云平台“弹性”的形式化定义，该定义只需测量几个关键指标，即可计算出平台的“弹性”值，为云资源提供者预测和优化平台性能提供理论指导。

四、 研究内容

（1）研究影响云平台“弹性”的关键因素，并探讨这些因素之间的关联；

（2）给出云平台“弹性”的形式化定义，拟将云平台建模为排队系统，使用时间连续的马尔科夫链去精确计算云平台的弹性值,从而有利于云资源提供者预测和优化云平台性能。

五、 创新点与特色

针对云平台“弹性”难以定量评估的问题，本课题首先拟给出云平台“弹性”的形式化定义，然后将云平台建模为排队系统，使用时间连续的马尔科夫链去精确计算云平台的弹性值。该项研究内容为本课题在理论方面的创新。

六、 技术路线

6.1 云平台弹性值的定义

云平台弹性值的高低直接反映所使用节能策略或者算法的好坏，一个好的弹性云平台，意味着其资源能得到充分利用，能够快速响应用户的请求，提供高质量的服务。设变量 i 表示服务器中虚拟机的数量，变量 j 为系统中任务的请求数量，则定义四种状态：

1）资源严重供应不足状态。如果 $j > 5i$,云平台资源处于“严重”供应不足状态，T_y 定义为资源严重供应不足状态下的累积时间。

2）资源轻微供应不足状态。如果 $3i < j \leq 5i$，云平台资源处于“轻微”供应不足状态，T_u 定义为资源轻微供应不足状态下的累积时间。

3）资源平衡状态。如果 $i < j \leq 3i$，云平台资源处于“平衡”供应状态，T_j 定义为资源供应平衡状态下的累积时间。

4）资源过剩状态。如果 $0 < j \leq i$，云平台资源处于“过剩”供应状态，T_o 定义为资源供应过剩状态下的累积时间。

定义 1：设 T_m 为测量时间，云平台的弹性 E 为平台处理供需平衡状态的时间百分比，则云平台的弹性 E 满足如下关系：

$$E = \frac{T_j}{T_{\mathrm{m}}} = 1 - \frac{T_y}{T_{\mathrm{m}}} - \frac{T_u}{T_{\mathrm{m}}} - \frac{T_o}{T_{\mathrm{m}}} \quad (1)$$

上式中，其中$\frac{T_y}{T_{\mathrm{m}}}$指平台处于资源严重供应不足状态的百分比，$\frac{T_u}{T_{\mathrm{m}}}$指平台处于资源轻微供应不足状态的百分比，$\frac{T_o}{T_{\mathrm{m}}}$指平台处于资源过剩状态的百分比，变量 T_m,T_y,T_u,T_o 都可以通过虚拟机实例的数量以及资源进行缩放的时间测量得到。

定义 2：设 P_y,P_u,P_j,P_o 分别代表资源严重供应不足状态下的累积概率，资源轻微供应不足状态下的累积概率，资源平衡状态下的累积概率和资源过剩状态下的累积概率。如果 T_m 足够长，则有 $P_y=\frac{T_y}{T_{\mathrm{m}}}$，$P_u=\frac{T_u}{T_{\mathrm{m}}}$，$P_j=\frac{T_j}{T_{\mathrm{m}}}$和 $P_o=\frac{T_o}{T_{\mathrm{m}}}$。因此，可以定义如下：

$$E=P_j=1-P_y-P_u-P_o \quad (2)$$

公式（2）中参数 P_y,P_u 和 P_o 的值需要进一步量化，才能得到平台的“弹性”值。如何量化 P_y,P_u 和 P_o 的值,将在下面进行建模计算。

6.2 与弹性相关的属性

在本节中，本文将云弹性与其他一些相关概念（如可恢复性，可扩展性和效率）进行比较。

可恢复性：Laprie[130]定义可恢复性作为服务交付的持久性，使得当面临变化时可以被合理地信任。因此，云可恢复性意味着 1）云系统能承受外部工作负载变化的程度，并且在该程度下不需要计算资源重新供应，以及 2）及时重新配置云系统的能力。本文认为后者的含义定义了云的弹性而前者的含义仅存在于云可恢复性中。在本文的弹性研究中，则关注后者。

可扩展性：弹性通常与可扩展性混淆。可扩展性反映了云资源重新配置时的性能加速比。换句话说，可扩展性表征了新的计算集群（大或小）处理给定工作负载的性能方面有多好。另一方面，弹性解释了计算集群可以准备好处理工作负载的重新配置时间有多快。云可扩展性受到很多因素的影响，如计算节点类型和数量，工作负载类型和数量。例如，Hadoop MapReduce 应用程序通常比其他单线程应用程序的扩展性更好。它可以根据线程，进程，节点，甚至数据中心的扩展数量来定义。另一方面，云弹性不仅受云服务提供商提供的能力的限制而且与云弹性相关的其他因素也有关，包括：备用机器的类型和数量，需要重新配置的计算资源。与云可扩展性不同，云弹性不涉及工作负载/应用程序类型和数量。

效率：效率表征了云资源如何在其扩大或缩小时可以有效地利用。此概念来自加速比，该术语定义了在重新配置计算资源后的相对性能。弹性与云的效率密切相关。效率定义为可实现的最大性能（加速或利用率）的百分比。高弹性导致更高的效率。然而，这种暗示并不总是真实的，因为效率可以受与系统弹性机制

无关的其他因素（例如，相同操作的不同实现）的影响。同时，可扩展性也受云效率的影响。因此，效率可以增强弹性，但不完全。这是因为弹性取决于资源类型，但效率不受资源类型的限制。例如，对于多租户架构，用户可能超过其资源配额，他们可能竞争资源或干扰对方的工作执行。

6.3 云平台弹性值的建模计算

上一部分给出了云平台“弹性”的形式化定义，这一部分通过将云平台建模为排队系统，使用时间连续的马尔科夫链去精确计算云平台的弹性值;假设用户请求以泊松分布到达，其速率为 λ，每个虚拟机的服务时间，启动时间和关闭时间服从指数分布，其速率分别为 μ，α 和 β，设变量i表示当前正在服务的虚拟机的数量，且变量j表示正在接收服务或正在等待的请求数。此时，扩展的M/M/m排队系统构建了一个二维连续时间马尔科夫链（CTMC），如图1所示。

图1实际上是一个对所有m=1, 2, 3，...的混合M/M/m排队系统。CTMC模型记录了虚拟机的数量和接收到的用户请求的数量，其最终用来计算弹性值E。图2中，模型中的每个状态，被标记为(i, j)，其中$i(i \in \{1, \ldots, m\})$表示当前正在处理请求的虚拟机的数量，$j(j \in \{0, 1, \ldots, m\})$表示正在接受服务的请求数量。为了数值计算的方便，将可部署的最大虚拟机数设置为m，其足够大以保证更高的精度。类似地，j的最大值也是m。令 μ 是每个虚拟机的服务速率。因此，每个状态的总服务率是运行的虚拟机数量和 μ 的乘积。弹性云计算模型中的状态由于用户请求到达，服务完成，虚拟机启动或虚拟机关闭发生转换。

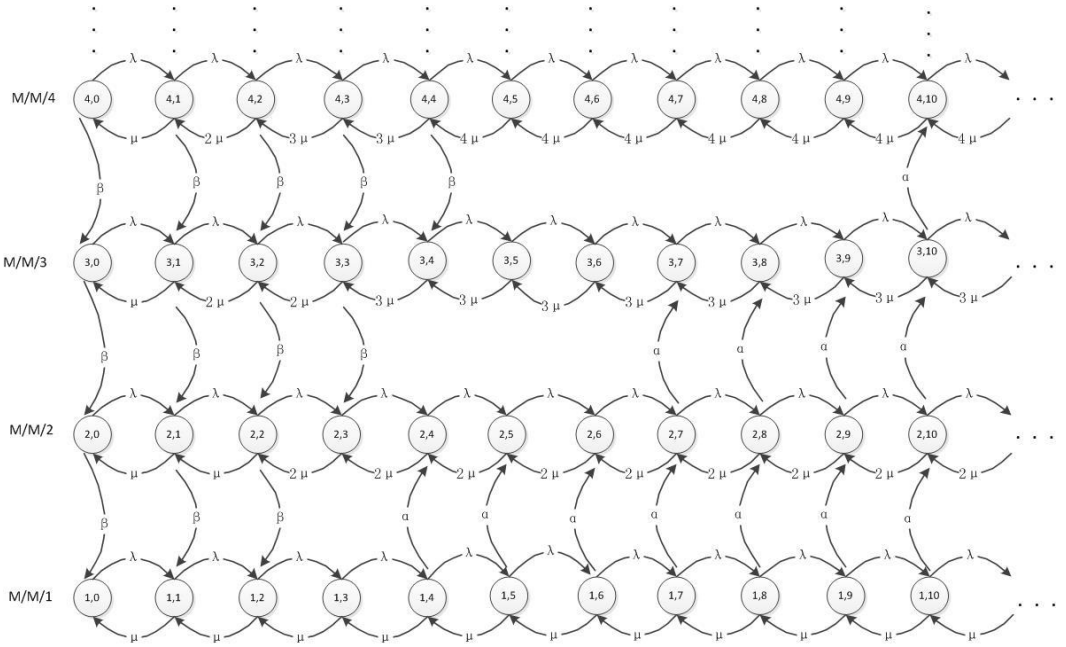

图 1 扩展 M/M/m 排队系统状态转换率图

CTMC模型的输入和输出参数总结如下：

（1）输入：请求到达率 λ，服务率 μ，虚拟机启动率 α，和虚拟机关闭率 β。

（2）输出：

•云平台的累积“严重”供应不足状态概率为P_y：

$$P_y = \sum_{i=1}^{m} \sum_{5i+1}^{m+1} P_{i,j} \ (3)$$

其中 $P_{i,j}$ 是稳态状态的概率。

•云平台的累积“轻微”供应不足状态概率为P_u：

$$P_u = \sum_{i=1}^{m} \sum_{3i+1}^{5i} P_{i,j} \ (4)$$

其中 $P_{i,j}$ 是稳态状态的概率。

•云平台的累积供应“过剩”状态概率为 P_o：

$$P_o = \sum_{i=1}^{m} \sum_{0}^{i} P_{i,j} \quad （5）$$

其中 $P_{i,j}$ 是稳态状态的概率。

云平台通过公式（2）（3）（4）（5），就可以获得该平台的弹性值。至此，该课题组云平台弹性值的度量已完成。

七、 结果展示

图 1 云平台下资源供给方式的统计

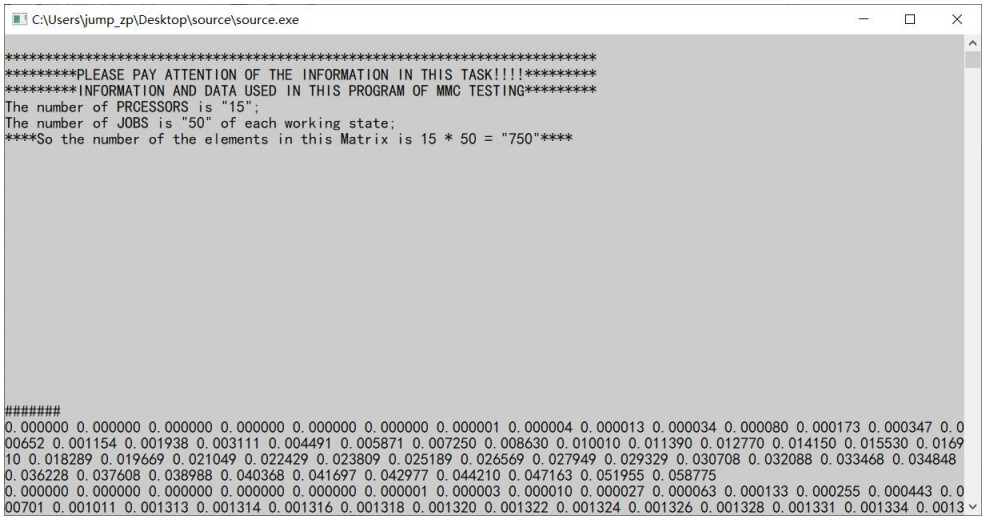

图 2 云平台下资源统计

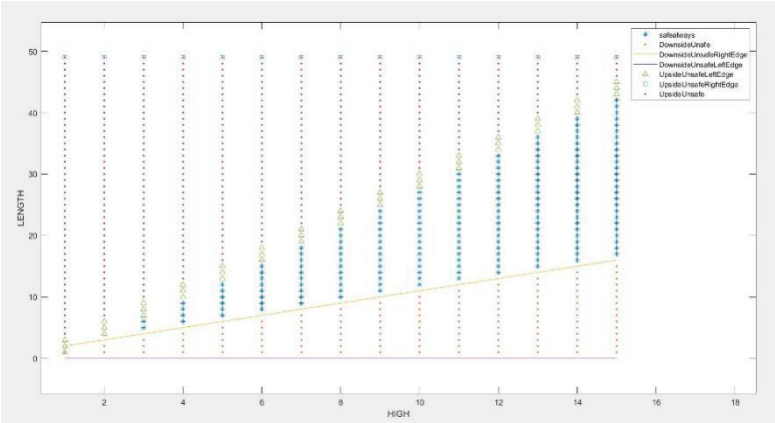

图 3 云平台下资源状态划分

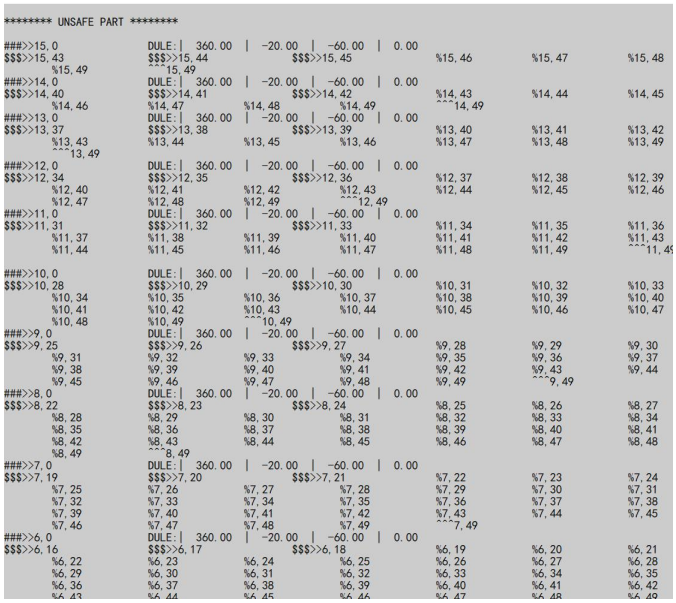

图 4 云平台下资源供给计算

参考文献

[1] Hongfang G, Renfa L, Jiyao A, et al. Quantitative Modeling and Analytical Calculation of Anelasticity for a Cyber-Physical System[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018:1-16. DOI: 10.1109/TSMC.2018.2861918.

[2] Li K. Quantitative modeling and analytical calculation of elasticity in cloud computing [J]. IEEE Transactions on Cloud Computing, 2018. DOI: 10.1109/TCC.2017.2665549.

[3] Kuperberg M, Herbst N R, Kistowski J G, et al. Defining and Quantifying Elasticity of Resources in Cloud Computing and Scalable Platforms[M]. Karlsruhe Institute of Technology, Karlsruhe, Karlsruhe Reports in Informatics, Tech. Rep. 16, 2011.

[4] Herbst N R, Kounev S, Reussner R. Elasticity in cloud computing: What it is, and what it is not[C]// International Conference on Autonomic Computing. 2013.

[5] Islam S, Lee K, Fekete A, et al. How a consumer can measure elasticity for cloud platforms[C]// Proceedings of the 3rd ACM/SPEC International Conference on Performance Engineering. ACM, 2012:85-96.

[6] Dustdar S, Guo Y, Satzger B, et al. Principles of Elastic Processes[J]. IEEE Internet Computing, 2011, 15(5): 66-71.

[7] Galante G, Bona L C E D. A Survey on Cloud Computing Elasticity[C]// IEEE Fifth International Conference on Utility and Cloud Computing. IEEE, 2013:263-270.

[8] 詹杭龙,刘澜涛,康亮环,曹东刚,谢冰.一种基于 Actor 模型的弹性可伸缩的流处理框架[J].计算机研究与发展,2017,54(05):1086-1096.

[9] Badger M L, Grance T, Patt-Corner R, et al. Cloud Computing Synopsis and Recommendations[M]. IEEE Standard 800‐146, National Institute of Standards and Technology, 2012.

[10] R. Buyya, J. Broberg, and A. Goscinski, Eds., Cloud Computing Principles and Paradigms [M]. Hoboken, NJ, USA: Wiley, 2011, 28(6): 1-10.

[11] Calheiros R N, Vecchiola C, Karunamoorthy D, et al. The Aneka platform and QoS-driven resource provisioning for elastic applications on hybrid Clouds[J]. Future Generation Computer Systems, 2012, 28(6):861-870.

[12] 魏豪,周抒睿,张锐,杨挺,王千祥.基于应用特征的 PaaS 弹性资源管理机制[J].计算机学报,2016,39(02):223-236.