随着企业的业务范围越来越广泛，业务环境越来越复杂和灵活，很多企业在追求降低成本的同时，寻求更高效、灵活、按需付费的计算资源，云计算作为一种基于网络的计算模式应运而生。它通过将计算、存储和网络等资源提供给用户，使其能够按需获取、使用，大大提高了IT基础设施的灵活性、可伸缩性和成本效益。云资源的主要优势之一是其弹性和可伸缩性，用户可以根据需求随时增加或减少资源，并按照实际使用情况付费。这种灵活性使得云计算成为许多企业和个人选择的理想解决方案，能够满足不同规模和类型的工作负载[1]。但是面对海量实时动态的计算任务，资源的需求难以预测，资源的分配和调度很难有效进行，因此现有云计算生态面临能耗高，资源利用率低等缺陷。提升大规模云资源的利用效率是数字化时代的的迫切任务，更是助力国家双碳战略和绿色发展的关键组成。

业内广泛开展云资源的预测性伸缩与优化调度能力建设，时序机器学习作为其关键共性技术，负责对该复杂系统的状态实时感知、资源需求趋势预测，进而实现局部资源弹性伸缩和全局资源的规划调度，提高资源利用效率并节能减碳。

现有的云计算资源弹性调度算法包括以下类别。

从伸缩方式来看：资源缩放分为水平缩放和垂直缩放。在水平缩放过程中，以虚拟机为最小资源单位按需要添加或释放虚拟机。与此相反，垂直缩放通过改变分配给已经运行的虚拟机的资源来实现，例如增加（或减少）分配的 CPU 或内存。垂直伸缩相对于水平伸缩有更好的稳定性，但是由于最常见的操作系统均不允许在线操作（不需重启）虚拟机，因而大多数云供应商只提供水平缩放[2]。

从实现技术来看：主要可分为手动策略和自动策略两种。手动策略是由用户根据自己监测到的资源情况而进行响应的弹性行为，但由于 Web 应用负载是高度动态的，因此其自动调度过程中有两方面困难：一方面是如何应对负载上升时资源供应不足造成的资源瓶颈（尤其是 CPU 瓶颈）造成的服务失效、产生 SLA 违约、影响 QoS 甚至导致性能降级的问题；另一方面是如何应对负载降低产生的大量资源闲置，造成不必要的资源浪费的问题[3]。

传统的手动策略已经明显无法满足用户需求，针对这一问题国内外主要采用自动伸缩技术解决此类问题，其中自动伸缩技术按类别划分可以总结成如下五类：阈值规则，强化学习，排队论，控制理论和时间序列分析[4-7]。a)阈值规则，文献[8]中提出了一种简单的阈值策略，CPU 利用率超过 80%时触发器根据预先设定的伸缩规则进行调度；[9]中设置了 70%和 30%的上下阈值。但是设置合适的阈值是一个挑战，如果设置得太低，可能导致频繁的伸缩，增加系统的不稳定性；如果设置得太高，系统可能无法及时应对激增的负载。b)强化学习，强化学习中两种常见的算法包括 SARSA 和 Q-learning，文献[10]使用单纯优化方法来选择能够返回高回报的状态,[11]为每一个虚拟机都创建一个代理，每个代理都管理自身的小规模查找表,[12,13]使用一种基于 ANN 的 RL 代理来根据应用和虚拟机的性能自动调节参数。但是由于问题通常具有较大的状态空间，强化学习需要较长的学习时间并且缺乏适应性，可能无法快速适应新的工作负载模式或系统变化。 c)排队论，[15]中使用排队论进行预测平均响应时间和工作负载的值，来实现应用在服务器间的分配，是供应商的利益最大化问题。[16,17]将云计算应用建模为 G/G/n 队列，其中 n 代表服务器数量。该模型可以在给定工作负载λ或请求的平均响应时间，以及服务器配置前提下估算所需的资源。[14]中使用 G/G/1 队列构成的队列网络来建模。这种模型最大的问题是当面对峰值负载时，可能造成大量的资源浪费。而且并没有考虑到用户的服务质量与最小化费用问题。d)控制论，固定增益控制器（包括 PIDPI 和 I）是控制论中最简单的控制器类型，[18,19]使用 I 型控制器来根据平均 CPU 使用情况调整虚拟机的数量；[20]使用 PI 型控制器根据批处理作业的执行进度来管理资源需求。自适应控制器也是使用较为广泛的一种控制器。[21]提出一种 MIMO 自适应控制器，该控制器使用二阶 ARMA来对非线性的时变的资源量与性能之间的关系进行建模，该控制器能够调节 CPU 和磁盘IO的使用率。但是控制理论通常依赖于对系统的精确建模，然而系统中的许多因素可能是不确定的，例如外部环境的变化、网络延迟、硬件故障等，这些不确定性可能导致控制器的性能下降。 e）时间序列分析，文献[23,24]使用移动平均进行预测，这种模型结构简单，预测结果也较差，基于此这种方法一般都用来对时间序列进行去除燥声的操作。文献[25,26]中使用指数平滑法进行预测。自回归方法使用范围很广泛，[27-29]均使用该方法进行预测，但其一般与移动平均组合使用构成 ARMA 模型（自回归移动平均法），提高预测准确率。时间序列分析的结果依赖于选择的预测模型。通常不同策略可以结合起来形成更加结合其他方法来实现更灵活、精确的自动缩放。文献[30]提出了一种混合的方法，采用了主动机制和被动机制结合的弹性算法，在扩展时采用阈值策略，收缩时采用预测方法。通过实验表明该方法相比于现有的弹性方法减少了到达 CPU 瓶颈的数量，同时有效控制了收缩过程中发生错误的比例。但该算法处理的是稳定类型的负载，对于负载规模大幅增加的情况将发生大量的服务失效。

参考文献

[1]张继东,曹靖城,周帅.基于增强学习算法的云资源动态弹性伸缩[J].信息技术,2021,(08):

[2]Lorido-Botran T, Miguel-Alonso J, Lozano J A. A review of auto-scaling techniques for elastic applications in cloud environments[J]. Journal of grid computing, 2014, 12: 559-592.

[3]Ali-Eldin A, Kihl M, Tordsson J, et al. Efficient provisioning of bursty

scientific workloads on the cloud using adaptive elasticity control[C]// The

Workshop on Scientific Cloud Computing DATE. ACM, 2016: 31-40

[4] Caron, E, F. Desprez, and A. Muresan. "Forecasting for Grid and Cloud

Computing On-Demand Resources Based on Pattern Matching. " IEEE

Second International Conference on Cloud Computing Technology and

Science IEEE, 2010: 456-463.

[5] Galante, Guilherme, and L. C. E. D. Bona. "A Survey on Cloud Computing

Elasticity." IEEE Fifth International Conference on Utility and Cloud

Computing IEEE, 2013: 263-270.

[6] Coutinho E F, Sousa F R D C, Rego P A L, et al. Elasticity in cloud computing:

a survey[J]. Annals of Telecommunications, 2015, 70(7): 289-309.

[7] Khatua S, Ghosh A, Mukherjee N. Optimizing the utilization of virtual

resources in Cloud environment[J]. 2010: 82-87

[8] Koperek P, Funika W. Dynamic Business Metrics-driven Resource

Provisioning in Cloud Environments[C]// International Conference on Parallel

Processing and Applied Mathematics. Springer-Verlag, 2016: 171-180.

[9] Hasan M Z, Magana E, Clemm A, et al. Integrated and autonomic cloud resource scaling[J]. 2012, 104(5): 1327-1334.

[10] Bu X, Rao J, Xu C Z. Coordinated Self-configuration of Virtual Machines and Appliances using A Model-free Learning Approach[J]. IEEE Transactions on

Parallel & Distributed Systems, 2013, 24(4): 12-21.

[11] Barrett E, Howley E, Duggan J. Applying reinforcement learning towards

automating resource allocation and application scalability in the cloud[J].

Concurrency & Computation Practice & Experience, 2013, 25(12): 1656–1674.

[12] Shen Z, Subbiah S, Gu X, et al. CloudScale: elastic resource scaling for multi

tenant cloud systems[C]// ACM Symposium on Cloud Computing. ACM, 2011:

5.

[13] Tesauro G, Jong N K, Das R, et al. A Hybrid Reinforcement Learning

Approach to Autonomic Resource Allocation[C]// IEEE International

Conference on Autonomic Computing. IEEE, 2006: 65-73.

[14] Tesauro G, Jong N K, Das R, et al. A Hybrid Reinforcement Learning

Approach to Autonomic Resource Allocation[C]// IEEE International

Conference on Autonomic Computing. IEEE, 2006: 65-73.

[15] Urgaonkar B B, Shenoy P, Ch A, et al. Agile dynamic provisioning of multi

tier internet applications[C]// International Conference on Autonomic

Computing. ACM, 2010: 217-228.

[16] Menasce D A, Dowdy L W, Almeida V A F. Performance by Design: Computer Capacity Planning By Example[M]. Prentice Hall PTR, 2014.

[17] Alieldin A, Kihl M, Tordsson J, et al. Efficient provisioning of bursty

scientific workloads on the cloud using adaptive elasticity control[J].

Sciencecloud Workshop on Scientific Cloud Computing, 2012: 31-40.

[18] Kumar R, Yadav S K. Scalable Key Parameter Yield of Resources Model for

Performance Enhancement in Mobile Cloud Computing[J]. Wireless Personal

Communications, 2017: 1-32.

[19] Lim H C, Babu S, Chase J S, et al. Automated control in cloud computing:

Challenges and opportunities[J]. First Workshop on Automated Control for

Datacenters & Clouds, 2016: 13-18.

[20] Lim H C, Babu S, Chase J S. Automated control for elastic storage[C]//

International Conference on Autonomic Computing, Icac 2010, Reston, Va,

Usa, June. 2010: 1-10.

[21] Park S M, Humphrey M. Self-Tuning Virtual Machines for Predictable

eScience[C]// Ieee/acm International Symposium on CLUSTER Computing

and the Grid. IEEE, 2009: 18-21

[22] Padala P, Hou K Y, Kang G S, et al. Automated control of multiple virtualized

resources[C]// EUROSYS Conference, Nuremberg, Germany, April. 2015: 13-26.

[23] Ali-Eldin A, Tordsson J, Elmroth E. An adaptive hybrid elasticity controller

for cloud infrastructures[J]. Network Operations & Management Symposium

IEEE, 2012, 131(5): 204-212.

[24] Gong Z, Gu X, Wilkes J. PRESS: PRedictive Elastic ReSource Scaling for

cloud systems[C]// International Conference on Network and Service

Management. IEEE, 2016: 9-16.

[25] Khatua S, Ghosh A, Mukherjee N. Optimizing the utilization of virtual

resources in Cloud environment[C]// IEEE International Conference on

Virtual Environments Human-Computer Interfaces and Measurement Systems.

IEEE, 2010: 82-87.

[26] Brown R G, Meyer R F, D'Esopo D A. The Fundamental Theorem of

Exponential Smoothing[J]. Operations Research, 1961, 9(5): 673-687.

[27] Roy N, Dubey A, Gokhale A. Efficient Autoscaling in the Cloud Using

Predictive Models for Workload Forecasting[J]. 2017: 500-507.

[28] Fang W, Lu Z H, Wu J, et al. RPPS: A Novel Resource Prediction and

Provisioning Scheme in Cloud Data Center[C]// IEEE Ninth International

Conference on Services Computing. IEEE, 2015: 609-616.

[29] Islam S, Keung J, Lee K, et al. Empirical prediction models for adaptive

resource provisioning in the cloud[J]. Future Generation Computer Systems,

2012, 28(1): 155-162.

[30] Rimal B P, Choi E, Lumb I. A Taxonomy and Survey of Cloud Computing

Systems[C]// Fifth International Joint Conference on Inc, Ims and IDC. IEEE

Computer Society, 2015: 44-51.