随机供需云环境中应用提供商收益驱动的最优资源协同配置策略*

白静[1] 许建军[2] 张龙昌[3,4]
（1. 东北财经大学管理科学与工程学院，辽宁 大连 116025；
2. 东北财经大学现代供应链管理研究院，辽宁 大连 116025；
3. 宿迁学院信息工程学院，江苏 宿迁 223800；
4. 北京邮电大学深圳研究院，广东 深圳 518038）

摘 要 现有云资源配置方案，对云应用的用户访问量（即资源需求量）、资源供应量的随机性和云应用提供商收益考虑不足，本文提出了随机供需云环境中应用提供商收益驱动的最优资源协同配置策略。该策略首先建立了资源和需求的量化模型；基于云应用提供商收益最大原则，设计了随机需求确定供应、确定需求随机供应、随机供需三种最优资源协同配置策略。当用户访问量和资源供应量随机，该策略能够有效提升云应用提供商收益，无 QoS（quality of service，服务质量）约束违反并能充分利用云资源。

关键词 云应用提供商，云资源，协同配置，收益驱动，随机供需

中图分类号 C931.6

1 引言

近年来出现大量云应用向用户提供各式各样的服务，云应用提供商为了降低建设和维护成本，采用租赁基础设施的方式将应用服务/程序部署到云平台上[1]。同时云应用提供商与 IaaS（infrastructure as a service，基础设施即服务）提供商之间达成服务等级协定（service level agreement，SLA），对运行在 IaaS 上的 QoS 进行约定，如要求响应时间是 2 秒、可靠性 0.99 等，即为应用的 QoS 约束。在多租户公有云上，大量应用彼此共享同一 IaaS 基础设施，竞争资源情况频繁发生；在 IaaS 负载超负荷和过高追求资源利用率情况下不可避免会出现违反应用的 QoS 约束，甚至出现宕机情况；对用户体验造成恶劣影响，进而给应用提供商造成严重损失。应用的用户访问量是随机的（即对资源的需求量是随机的），所以资源配置量难以准确确定；IaaS 提供商过多配置资源会降低资源利用率和收益，过少配置资源又不可避免地出现违反应用 QoS 约束情况；SaaS（software as a service，软件即服务）提供商普遍采用节省计划 + 按量付费（即 subscription + on-demand，如阿里云）计费模式配置应用的资源量，部分资源量按节省计划付费长期预订，不足部分资源量按量付费，长期预订资源量配置多少能使 SaaS 提供商收益最大，目前没有有效方法确定。

* 基金项目：国家自然科学基金重点项目（72232001）、国家自然科学基金面上项目（72371059）、国家社会科学基金一般项目（19BTQ028）、辽宁省自然科学基金计划项目（2019-ZD-0496）、辽宁省教育厅科学研究一般项目（LJKZ1022）、辽宁省教育厅专项经费项目（JYTMS20230665）、宿迁学院科研启动基金资助项目（2022XRC029）。

通信作者：张龙昌，宿迁学院信息工程学院，博士，教授，E-mail：zlc_041018@163.com。

因此，在满足应用的 QoS 约束条件下，有效提高资源利用率的云资源分配成为亟须解决的难题[2]。

以往 IaaS 层面的解决方案，从部署和运行两个阶段对云应用进行资源分配，采用预测和监控方法确定用户访问量和云资源负载；先期预测方法准确度很难保证，通过监控调整资源总是滞后。如果通过多配置资源来保证应用的 QoS 约束，又会导致云资源利用率下降，增加应用提供商的成本或者是减少 IaaS 提供商收益；如果云资源分配过少导致违反 QoS 约束情况频繁发生，则会导致云应用的用户体验变差，进而降低应用提供商的收益。

在 PaaS（platform as a service，平台即服务）层，通过刻画 SaaS 的资源开销、请求率，将其分类并且搭配部署到虚拟机上，以提高资源利用率；对于请求率较大的应用采取分批转发的方法，为保证 QoS 约束，将其请求分成多批分别部署到多个虚拟机上。多应用搭配部署，或者存在部分不可利用的闲置资源，或者存在部分应用违反 QoS 约束；请求分批部署，可以减少违反 QoS 约束，但是总是存在没有被充分利用的虚拟机；也缺少对 SaaS 提供商收益的考虑；更没有考虑服务器实例的负载是随机的，进而能支撑的虚拟机数量是随机的情形。

在 SaaS 层，目前尚未见有在 SaaS 层面进行资源分配的研究成果，大部分研究基于 QoS 的服务选择和组合[3]。在应用部署初期通过选择以往 QoS 质量优的若干服务进行编排；然而服务的 QoS 是不断变化的，导致违反 QoS 约束的情况经常发生，在服务运行期进一步对那些 QoS 质量降低的服务进行重新选择，然而重选滞后导致违反 QoS 约束还是会发生，也会带来服务切换的资源开销。

在满足 QoS 约束条件下，配置适量的虚拟资源使云应用提供商收益最大是一个重要指标，用户量、应用的服务价格、虚拟资源价格[4, 5]、硬件实例的负载是影响资源配置量的重要因素，而用户量（即虚拟资源需求量）又随着时间表现出随机特性，硬件实例的负载随时间不断变化导致其能够支撑的虚拟资源数量（即虚拟资源的供给量）具有随机特性。针对这种情形，本文提出随机供需云环境中应用提供商收益驱动的最优资源协同配置策略。

理论贡献主要有：①构建随机供需云环境下最优资源配置方法，能使应用提供商期望收益最大，是对经典报童模型在随机供需云环境下的进一步扩展。②首次提出云应用提供商协同配置资源策略，该思想也可应用于库存优化中。③与以往基于 IaaS/PaaS 视角不同，基于云应用提供商期望收益最大的思想提出云资源配置策略，在随机环境下充分考虑云资源租赁价格、服务价格、资源缺货价格和资源闲置价格对云资源配置量和期望收益的影响，丰富了云资源配置理论。实践贡献主要有：①建立云应用需求和资源模型，量化虚拟资源和云资源，将需求的虚拟资源量与物理资源量进行单位统一，为最优资源量的计算提供基础。②设计协同配置资源算法，将协同应用提供商中多余资源和短缺资源进行二次匹配，减少闲置成本和缺货成本，进而增加协同提供商的收益。③考虑随机需求确定供应、确定需求随机供应，以及随机供需三种应用场景，基于用户访问量和云资源负载量随机特性设计三种云资源配置策略：需求随机并且供应确定的最优资源协同配置策略；确定需求并且随机供应的最优资源协同配置策略；需求和供应都随机的最优资源协同配置策略。

本文第 2 节总结已有相关研究；第 3 节描述问题、介绍文中涉及的符号以及假设条件；第 4 节介绍模型化资源和需求；第 5 节详细介绍云资源协同配置算法、三种最优云资源协同配置策略；第 6 节通过实验评估文中提出的策略；第 7 节总结全文并介绍下一步研究方向。

2 相关研究

早在 2008 年，Buyya 等[6]率先提出了面向市场的云计算概念。在云服务市场中，涉及 IaaS 提供商、

PaaS 提供商和应用提供商（SaaS provider）主要参与者。从此之后，国内外学者开始针对云资源分配进行了大量的研究。其中，满足 QoS 约束且能有效提高资源利用率的云资源分配问题是一个难点。本节将分别从 IaaS 提供商、PaaS 提供商和应用提供商视角简略介绍与该问题相关的研究进展情况。

2.1 基于 IaaS 提供商视角的应用 QoS 约束的云资源配置

应用部署到虚拟机上，虚拟机放置到物理节点上（即资源配置），需要考虑满足不同应用的 QoS 约束，同时能够实现资源使用的最优化。以 IaaS 为视角解决云资源分配问题主要集中在启发式算法和图匹配算法等传统经典的方法，这些方法基于预先定义的规则和策略，通过静态的分析和调整来进行资源的分配和调度。然而，随着云市场的不断成熟、细分，研究人员逐渐意识到传统方法无法很好地适应动态变化的云计算环境。因此，近几年来的研究主流逐渐转向了利用历史数据和实时监测信息以数据驱动改进的机器学习方法。

（1）启发式算法。在 2011 年，李强等[7]通过对应用 QoS 需求、负载和云资源需求分布规律建模，将资源配置定义为多约束的多目标优化问题，提出基于遗传算法的带应用 QoS 约束的多目标优化的资源配置算法。2011 年，孙大为等[8]研究基于免疫克隆的算法。随后还有一些学者提出基于蚁群算法[9]、共生生物搜索[10]等启发式算法的资源分配方案。该类方法主要目标是追求满足应用 QoS 需求的资源搜索精度和速度，尽管已经开发出了几十种该类算法来提高应用的 QoS，但是由于存在应用 QoS 需求和资源负载的高度随机性，到目前为止仍然缺乏一种能够有效适应这种动态变化的启发式多目标优化算法。

（2）图匹配算法。2014 年，匡桂娟等[11]基于图理论的云资源分配做了一些探索。2018 年，郭伟等[12]在分析应用部署时发现，云资源节点的拓扑表现为一个复杂的异构网络图，不同租户提出的应用部署需求也可以表示为带有多维性能属性的异构网络图，因此将大型应用的虚拟机放置问题映射为云资源节点拓扑图的子图查询匹配问题，基于偏序关系异构图查询匹配方法得到一组满足需求的云资源节点集合。该方法虽然能够实现应用敏捷化交付部署，但在求解精度、云资源与应用 QoS 动态变化、云资源利用率等方面仍存在很大改进空间。

（3）机器学习算法。2014 年，孙佳佳等[13]提出基于神经网络的资源分配研究。为了能更好地适应云环境下服务系统的动态变化，2017 年，闫永明等[14]应用无模型的在线学习算法解决用户并发量变化导致的系统性能保障困难的问题，该方法通过不断重复“执行—积累—学习—决策”的过程，可以不断地积累经验数据并优化决策结果。之后还有部分学者提出基于马尔可夫预测[15, 16]、直觉模糊时间序列预测[17]、聚类[18]、监督学习[19]等算法的资源分配研究。2020 年，吴悦文等[20]基于云资源共享特点，获取作业运行时监测数据和云资源配置信息，建立作业分类与优化云资源配置的启发式规则，并将该规则应用到贝叶斯优化算法，进行资源配置。2021 年，苏命峰等[21]为解决边云协同计算中的资源配置问题，在云服务中心基于二维时间序列对用户任务进行预测，分类聚合用户任务类型，推送任务资源至边缘服务器，提高用户任务命中率的平均值，减少服务器资源占用开销；边缘服务器基于随机贪心近似算法，分别对用户服务质量和系统服务效应这两个目标进行帕累托改进，寻求两个目标曲线的相切点或相交点以优化任务调度。该类方法存在的关键问题是对用户 QoS 需求和资源运行情况的预测准确度难以保证，预测结果与实际结果总是存在一定偏差，因而违反 QoS 约束无法避免；另外，算法需要经过训练、调整，进而产生了系统开销增大问题。

此外，还有基于控制论[22]、博弈论[23, 24]等的资源分配方法。2013 年，Nallur 和 Bahsoon[25]发现云服务用户并发量的产生具有很强的随机性，一次性的资源配置无法使云服务一直保持不违反 QoS 约束的运

行状态，因此需要在云服务运行时动态调整资源配置，自适应的资源调整能够更加有效地应对云环境的实时变化，如文献[26, 27]；这种方法对变化频率较高的云环境和应用 QoS 需求难以应对，并且会产生较大的额外系统开销。采用资源预留来保障 QoS 约束不失为一种有效的方法，在整个预留请求中只要有一个 QoS 指标的需求无法得到满足，整个预留请求就会被拒绝，从而导致误拒率的上升；针对该问题，在 2014 年，伍之昂等[28]改进了 QoS 偏差距离的计算方法，在资源预留协商阶段降低预留请求的误拒率。预留足够多的资源可以有效减少违反 QoS 约束的情况，但会造成资源的严重浪费。

2.2 基于 PaaS 提供商视角的应用 QoS 约束的云资源配置

云资源提供者主要任务是保证应用 QoS 需求同时有效提高资源的利用率。IaaS 层的资源配置是以虚拟机为基本对象，然而 PaaS 层主要以应用为对象，这导致其与 IaaS 层有着较大的区别。目前的资源配置机制主要集中在 IaaS 层，对 PaaS 层的应用特征考虑得不足。部署在 PaaS 平台上的应用使用资源情况差异较大，并且访问量在时间上表现出不同特性；针对这个问题，在 2016 年，魏豪等[29]通过对应用请求率变化及各项资源开销的预测，将不同类型的应用搭配部署，将请求量较大的应用划分给多个资源开销相对固定的单元处理，实现均衡、充分地使用服务器资源。该种方案中每个固定单元都会存在部分闲置资源，因此资源利用率不高。请求量较大的应用独占资源也存在资源的浪费，资源开销和应用请求率变化特征描述的精准度直接影响资源分配方案效果，而其精确度在目前的方法中难以保证。

在 PaaS 平台中，通常采用两种方式来获取资源需求量：实时获取和预测算法。第一种方式是通过监控应用的访问模式、请求频率以及资源的使用情况，估算当前应用对资源的需求程度。另一种方式是分析历史数据，构建预测模型推断应用未来可能的资源需求量。然而，由于获取实时数据调配具有一定的滞后性，因此资源需求预测成为 PaaS 平台资源分配研究的重点。在资源需求预测方法中，可以根据其时间阶段分为两类：传统的基于时间序列的预测模型和近年来兴起的基于机器学习方法的预测模型，如支持向量机和神经网络等。

（1）时间序列预测模型。在 2018 年，Jayathilaka 等[30]通过分析计算每个时间区段内回归因子的相对重要性指标变化点，提出云平台应用程序的性能异常的预测方法。在 2019 年，徐雅斌和彭宏恩[31]考虑了 PaaS 平台中应用对资源需求的多周期性特征，采用了基于时间序列的短期预测结合多元回归模型的周期性预测的综合预测模型。上述方法依赖时间序列的特征进行预测，并适用于一些稳定且周期性较强的资源需求。由于平台中应用业务类型的多样性以及服务时间的差异性，这类方法对应用资源需求预测时可能会产生较大误差，导致违反应用 QoS 约束。因此，研究学者开始探索机器学习方法来解决该问题。

（2）机器学习预测模型。传统的单值预测方法在处理并发量不确定性时存在局限性，在 2017 年，孟煜等[32]通过采用梯度下降粒子群优化的支持向量机提出了一种面向多种并发量类型（平稳型、趋势型和周期型）的云服务用户并发量区间预测模型。在 2019 年，谢晓兰等[33]针对容器云集群上应用资源供应不足和过度供应问题，提出了一种基于三次指数平滑法和时间卷积网络结合的云资源预测模型。这类方法对于不同应用负载类型和应用联合情况的考虑还存在一定的局限性。应用联合是指多个应用部署在同一台或多台服务器上共享资源。在这种情况下，应用之间的相互影响可能导致资源需求大幅度的变化。然而，现有的机器学习方法没有考虑到应用联合带来的影响，因此无法准确预测资源需求。

2.3 基于应用提供商视角的应用 QoS 约束的云资源配置

在云服务市场中，应用提供商从 IaaS 资源提供商处选择合适的资源为终端用户提供云服务，扮演着

既是资源需求者又是服务提供者的双重角色，与 IaaS 资源提供商本质是供求关系，需要从不同的角度考虑资源配置问题。目前还未见有严格意义上以应用提供商视角进行云资源配置的相关研究。现有的大多数研究采用多属性决策方法[34]、推荐方法[35]、优化方法[36-38]聚焦于 QoS 的服务选择和服务组合[39, 40]。例如，在 2014 年，Ghosh 等[34]结合可信度和 SLA 的透明度来评估交易风险帮助选择云应用服务。在 2018 年，Ding 等[35]考虑服务质量随时间变化的特性，应用相似性增强协同过滤方法捕捉用户相似性的时间特征，结合自回归综合移动平均模型提出了一种时间感知服务推荐方法。在 2019 年，Jain 和 Hazra[36]基于博弈论探讨资源需求者在选择私有云和公共云的计算能力组合时的决策问题，以及需求的不确定性对组合决策的影响。在 2020 年，Hosseini 等[37]研究在完成计算任务有时间约束的情况下，最小化总租用成本的计算资源选择问题。在 2024 年，彭高贤等[38]基于 QoS 指标兼顾制造服务的能耗经济性为任务选择组合服务，提出一种能耗感知的云制造服务选择与调度的多目标优化模型。该类问题不涉及云资源分配并且研究成果已经很多，各种不确定因素导致备选服务 QoS 不确定甚至服务失效，违反 QoS 约束的情况经常出现，并且在应用提供商收益方面考虑也存在不足。

目前基于 IaaS 提供商和 PaaS 提供商的相关研究工作，仍然不能有效解决应用 QoS 约束下的云资源利用率不高的问题，尤其是在服务用户并发量具有很强的随机性和云资源负载高度随机的条件下。该问题仍然存在是因为应用提供商不能准确提供所需的云资源量，应用提供商通常采用按量付费模式为云应用配置资源，并且提出应用的 QoS 约束，上述研究成果也都是针对此种场景进行设计。更重要的问题是以往研究较多考虑在满足 QoS 约束条件下尽可能充分利用云资源，而缺乏对应用提供商收益的考虑（更未涉及服务价格、云资源租赁价格、资源的缺货价格和资源的闲置价格对应用提供商收益的影响），从而造成其成本的增加（降低应用提供商的基础设施和平台建设成本是云平台追求的目标之一）。应用提供商可以采用节省计划付费模式配置云资源，节省计划具有更低的资源租赁成本，进而提高自身的收益。节省计划中的资源数量决定应用提供商成本，而在随机资源需求环境中，准确确定节省计划的资源量就成了一个亟须解决的问题。随机需求条件下，虽然不能做到每次资源配置都做到应用提供商收益最大，但是可以获得期望收益最大，此时的资源配置量就是节省计划付费模式的资源量。在此基础上，服务运行期间，还可能存在应用提供商资源短缺或剩余情况，可进行再调配，进一步提高应用提供商收益。本文基于应用提供商视角提出一种使应用提供商期望收益最大的云资源协同配置策略，考虑需求随机和供应随机条件下服务价格、云资源租赁价格、资源的缺货价格和资源的闲置价格对资源配置量的影响。

3 问题、假设与符号

3.1 问题描述

在云资源配置过程中（图 1）：SaaS 提供商根据用户的需求定制服务，与 PaaS 提供商达成 SLA，其中包括 QoS（如响应时间、可用性、可靠性、成功率等）限制、资源需求等约定；PaaS 提供商提供服务的运行和管理环境，将服务所需资源打包成虚拟资源，并监测服务信息（如用户访问量、QoS 等信息）；IaaS 提供商根据 SLA 和公布的资源计费模式提供基础设施服务和计费，将虚拟资源映射到相应的硬件上运行。为有效降低应用供应商在基础设施建设、运营和维护上的成本，云应用提供商租赁 IaaS 资源运行其应用，在满足用户 QoS 需求条件下向用户提供服务，并负责应用的运维。由于应用的用户访问量不仅受用户的需求、偏好、忠诚度、服务口碑、QoS 期望等主观因素影响，而且很大程度上受服务质量、自然环境、网络环境、计算环境等众多客观因素的影响，实际用户访问量具有很高的随机性。IaaS 提供

商在同时向多租户提供基础设施服务过程中，不仅受到硬件、网络、资源配置算法等因素影响，而且还很大程度受不可控因素（如用户资源需求随机、应用的资源消耗动态变化）的影响，导致其实际能够支撑的用户访问量是随机的（即在满足 QoS 约束条件下，资源量的供应随机）。

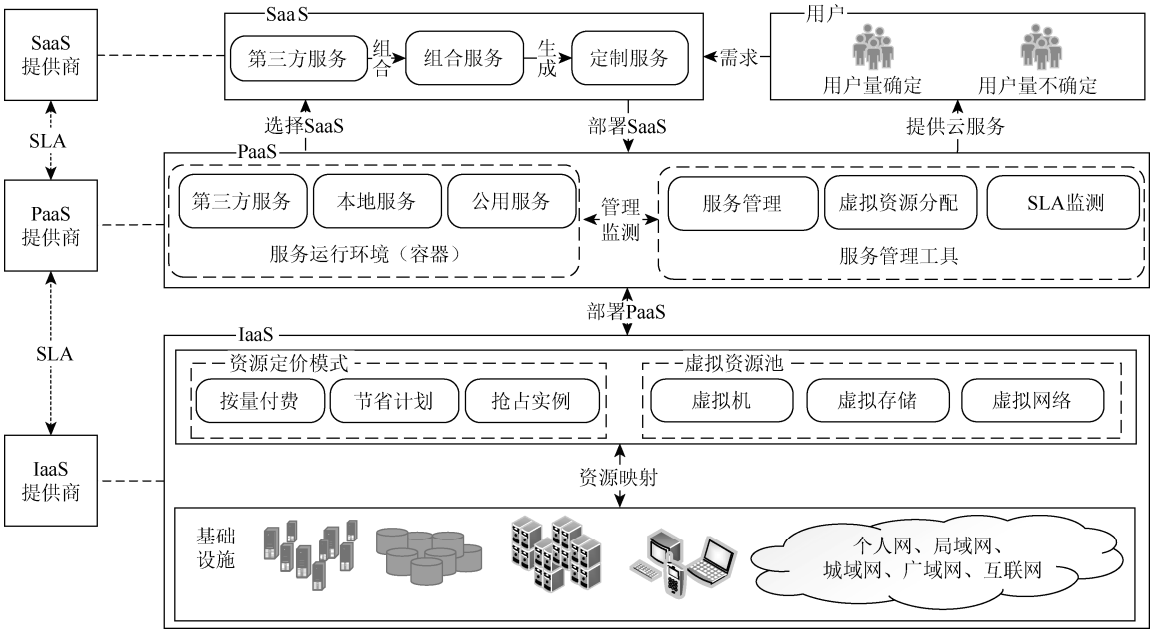


图 1 云服务架构及资源配置

根据 IaaS 资源定价模式（抢占式实例由于受到市场价格影响，IaaS 实例被释放时机不确定，不适合此状态应用场景，因此不考虑这种价格模式），应用提供商可选择按量付费、节省计划、节省计划 + 按量付费三种模式配置 IaaS 资源。①按量付费：应用提供商将 SaaS 应用部署到 IaaS 上，按照实际使用资源量付费；因为随机的用户量导致 IaaS 资源需求量是随机的，IaaS 提供商需要预留额外资源去满足应用的 QoS 约束；预留的资源多导致资源利用率降低，预留的资源少会进一步违反 QoS 约束；按量计费也是应用提供商成本最高的 IaaS 资源配置模式。②节省计划：云应用提供商预先估计应用所需 IaaS 资源量并向 IaaS 服务提供商提出资源配置量，IaaS 提供商只需按要求提供相应资源，不必考虑应用的 QoS 约束，也不用预留额外资源来保证应用 QoS，因为应用提供商在配置资源量时已经考虑了用户量随机和用户的 QoS 需求；节省计划是应用提供商成本最低的模式。③节省计划 + 按量付费：首先应用提供商根据应用的需求预订一部分资源，再根据实际情况对于不足部分按量付费；混合模式的应用提供商成本处于按量付费和节省计划之间。采用节省计划 + 按量付费混合模式可以在一定程度上提高应用提供商的收益和缓解随机的资源需求的影响，下面以某云提供商为例说明。实例 A 按量付费价格为 12 元/（台 • 小时）；“节省计划”折扣后价格为 4.75 元/（台 •小时），用户每小时承诺消费 95 元（即每小时最多抵扣 95/4.75 = 20 台实例运行）；设应用第一小时需 25 台实例 A 运行，第二小时需 15 台实例 A 运行，计费比较见表 1。

表 1 资源计费比较

计费方式	第一个小时	第二个小时	汇总费用	节省
按量付费	25×12 = 300	12×15 = 180	480	
节省计划 + 按量付费	95 +（25–95/4.75）×12 = 155	95	250	约节省 48%

节省计划付费模式有两个优势：①有利于提高应用提供商收益；②IaaS/PaaS 提供商可以准确配置资源，可以充分利用资源并且降低 QoS 违反率。

目前基于 IaaS 和 PaaS 提供商视角的资源配置主要是针对按量付费模式，关键是基于用户量提出资源配置量，同时考虑 QoS 约束和资源利用率。节省计划资源配置模式，是由应用提供商预先提出资源需求量，相对按量付费模式以较低的折扣价获取资源，服务运行期间 IaaS 和 PaaS 提供商只需提供相应数量的资源，无 QoS 约束违反，资源利用率接近 100%。应用提供商视角配置资源是基于用户量、服务价格、IaaS 资源价格、资源供应情况等因素，本文针对应用的用户访问量随机和资源供应量随机导致 IaaS 资源配置量难确定问题，以实现应用提供商期望收益最大为目标，在不违反服务 QoS 约束条件下提出最优云资源配置策略。

3.2 假设与符号

在构建供需随机的应用提供商收益驱动的最优资源协同配置策略过程中，引入下述假设：①系统中只有一个 IaaS 提供商且面向多租户提供基础设施服务，租户间存在争抢资源情况；②系统中的应用为单一产品，且服务期时间固定不变，服务期内访问应用的用户量分布可用函数描述或可统计出概率分布表；③服务期内产出的虚拟资源量分布可用函数描述，且与用户量分布相互独立；④应用提供商是理性的并且风险偏好是中性的，参与协同配置资源的应用提供商处于相同的云计算环境中；⑤各用户实例所需虚拟资源量可以量化并且数量相同或相近；⑥云资源配置量不够时，可将超出配置量部分的用户转移到其他公共计算实例上提供服务，其 QoS 下降不影响用户对应用的黏性；⑦云资源配置过量导致资源非高效利用（即闲置），冗余资源不可退回但可以提升服务的 QoS，提升用户体验同时可增加 SaaS 提供商的隐性收益，且可度量，进而闲置成本可低于租赁成本。在建立数学模型过程中所涉及的符号如表 2 所示。

表 2 重要符号说明表

	符号	说明
基本变量	S_i	参与资源协同配置的第 i 个应用提供商，$1 \leqslant i \leqslant t$，$t$ 为参与协同配置的应用提供商个数
	P_u^i	i 应用的服务价格（每用户收取的服务费）
	C_i	i 应用提供商租赁的云资源单价
	R_i	i 应用每个用户实例需要的虚拟资源量
	Q_i	i 应用实际资源需求
	k_i	i 应用单位虚拟资源收益
	L_i	i 应用虚拟资源缺货单价
	H_i	i 应用虚拟资源闲置单价
随机变量	$\tilde{D}_i$	i 应用的用户访问量，可能是随机变量或者确定变量，应用提供商统计其分布情况或数值
	u_i	i 应用虚拟资源需求量（即 $\tilde{D}_i \times R_i$），是随机变量，其概率密度函数为 $f_i(u_i)$
	y_i	i 应用的单位 IaaS 计算实例可支持的虚拟资源量，是随机变量，其概率密度函数和期望分别为 $f_i(y_i)$、Y_i

续表

	符号	说明
决策变量	Q_i^*	i 应用最优虚拟资源需求量/配置量
	X_i^*	i 应用最优云资源配置量
	$\ddot{Q}_i$	i 应用最优虚拟资源协同配置量
	$\ddot{X}_i$	i 应用最优云资源协同配置量

4 资源和需求模型

应用运行所需的硬件资源主要有 CPU、内存和外存等计算资源，在满足 QoS 约束条件下为每个用户提供服务所需资源可用一个向量描述 $R_q=(r_1,r_2,\cdots,r_n)$，总的资源需求量与用户访问量呈线性关系[29]，这里需要确定向一个用户提供服务所需的资源需求量。为了满足 QoS 约束，$r_1,r_2,\cdots,r_n$ 之间存在一定依赖关系，因而只需将一类资源的需求量转化即可确定其他资源需求量。在云计算环境，应用打包在虚拟机中，再运行在具体的 IaaS 计算实例上，因此需要将需求、虚拟资源和计算资源进行量化，三者进行映射形成衔接关系，为后续的资源分配模型建立提供基础，下面是相关概念和映射关系的定义。

定义 1 IaaS 资源。IaaS 的资源主要包括硬件资源、软件资源和网络资源。其中，应用提供商租赁的主要为硬件资源和网络资源，如 CPU/GPU（graphics processing unit，图形处理单元）、内存、外存、I/O（input/output，输入/输出）设备、交换机、带宽。IaaS 资源是这些硬件资源和网络资源的抽象，如 CPU 资源、存储资源、网络资源等。

定义 2 IaaS 计算实例。Iaas 计算实例由计算资源组成，能够独立运行软件系统的硬件平台。根据应用提供商对性能和数量的需求，配置不同等级的实例。设定某一具有基本配置的实例为标准计算实例（即设其为衡量实例计算能力的度量单位，记为 calculation instance power unit，CIPU）。如一个标准计算实例由 1G 内存、200G 外存、100M 网络带宽、频率 0.5G 的 CPU 组成，则 2 个 CIPU 的计算实例包含 2G 内存、400G 外存、200M 网络带宽、频率为 1G 的 CPU。再如 1 个 CIPU 的外存可规定为 0G 内存、200G 外存、0M 网络带宽、频率为 0G 的 CPU。

定义 3 虚拟计算实例。虚拟计算实例包含了 IaaS 计算实例中的部分计算资源，通过软件模拟的具有完整 IaaS 计算实例功能的，并且能够独立运行应用的系统。虚拟计算实例运行在某特定的一个或多个 IaaS 计算实例上，其单位为 CIPU，其 QoS 受计算实例负载制约。

定义 4 虚拟资源。虚拟资源包含了 IaaS 资源中的部分计算资源，通过软件模拟的具有完整 IaaS 资源功能的，可被视为独立的 IaaS 资源。虚拟资源通过对 IaaS 资源采用时间和空间划分方法实现一变多，而采用分布式技术实现多变一，单位与 IaaS 资源单位相一致，其 QoS 受计算资源负载制约。

定义 5 虚拟资源量。在保证用户 QoS（如响应时间、吞吐量等）需求的条件下，应用向一个用户提供服务所需的虚拟资源（也可是虚拟计算实例）数量的度量，称为虚拟资源量。给定应用向一个用户提供服务，在满足 QoS 约束条件下，在一个服务期内一个用户实例需要配置的虚拟计算实例量是 R 个 CIPU，即虚拟资源量为 R。

本文将用户访问量转换成虚拟资源需求量，再将虚拟资源需求量转换为虚拟计算实例需求量，进而

打包运行在具体计算实例上。文中虚拟资源配置量均指虚拟计算实例配置量，云资源配置量均指 IaaS 计算实例配置量。

5 最优资源协同配置策略

在租赁云资源运行应用向用户提供服务过程中，云应用提供商收益受用户访问量、服务销售收入、云资源租赁成本、缺货成本以及闲置成本等因素影响，以云应用提供商期望收益最大为目标函数，以虚拟资源需求量、云资源配置量为决策变量，建立用户访问量和虚拟资源供应量随机环境下的云资源最优配置策略。用户访问量是一段时间内云应用服务的用户数量（随机变量），通常由应用提供商进行多次统计，生成概率分布表或分布函数，其决定资源需求量（也是随机变量）；服务销售收入是云应用提供商向用户提供服务所获得的收益；云资源租赁成本是应用提供商租赁计算资源运行应用向 IaaS 提供商支出的费用，必须保证用户对应用的 QoS 要求；缺货成本是访问量到了极限，应用提供商不得不拒绝向后续用户提供服务导致的损失（通常情况将超出容量的用户请求转发到性能相对低的服务器上，从而导致 QoS 无法保证）；闲置成本是配置的云资源能够支撑的虚拟资源量超过了用户需求量，要低于租赁成本，原因有两个：①虚拟资源过量可提升 QoS 和用户体验，而某种程度上隐性地增加应用提供商收益，②IaaS 提供商以经济激励的形式回收利用租户空闲预留型资源[41]。在实际资源配置过程中，涉及的其他收入和成本可继续加入，不影响本文的云资源最优配置方法的核心思想。

在服务期内，访问应用的用户数量可能是随机变量；也存在应用的用户群体固定不变（尤其是企业用户，用户量长期保持不变），即需求确定。IaaS 基础设施服务的租户较少，有充足的计算实例，可以获得确定数量的满足 QoS 需求的虚拟资源量，即虚拟资源量供应确定；也存在 IaaS 基础设施服务的应用较多形成应用间的资源竞争，在满足 QoS 需求条件下，IaaS 提供商选择根据虚拟资源需求量和总的云资源量动态地为各应用分配资源，从而导致应用的虚拟资源量供应是随机的。下面针对存在的问题，设计三种云资源最优配置策略。

5.1 云资源协同配置算法（co-allocation algorithm for cloud resources，CA_CR）

随机的用户访问量导致资源需求量也是随机的，不可避免地出现服务期内实际的资源需求量 Q 与预定最优资源配置量 Q^* 之间存在偏差。如果 $Q>Q^*$，则配置量不能满足需求量而有收益损失；如果 $Q<Q^*$，则配置量多于需求量时退回部分资源产生收益损失。为缓解这种偏差造成的应用提供商的收益损失，在服务期内采取多应用提供商协同配置资源策略进行二次资源配置，设有 t 个应用提供商的集合 $\{S_1,S_2,\cdots,S_t\}$ 参与资源协同配置，$Q_j\in\{Q_1,Q_2,\cdots,Q_t\}$ 与 $Q_j^*\in\{Q_1^*,Q_2^*,\cdots,Q_t^*\}$ 分别为 j 提供商的资源实际需求量和最优配置量，$k_j\in\{k_1,k_2,\cdots,k_t\}$ 为 j 提供商的单位资源的收益（可用服务销售单价代替），具体的云资源协同配置算法内容如算法 1 所示。

算法 1　云资源协同配置算法（CA_CR）

输入：服务提供商集 $\{S_1,S_2,\cdots,S_t\}$，虚拟资源租赁价格集 $\{C_1,C_2,\cdots,C_t\}$，实际需求量集 $\{Q_1,Q_2,\cdots,Q_t\}$，最优配置量集 $\{Q_1^*,Q_2^*,\cdots,Q_t^*\}$，单位资源收益集 $\{k_1,k_2,\cdots,k_t\}$。

输出：协同配置量集 $\{\ddot{Q}_1,\ddot{Q}_2,\cdots,\ddot{Q}_t\}$。

步骤：

（1）建立列表-list l 包含 t 个元素，列表中元素是一个字典-dict，包含服务提供商 S、单位资源收益

k、虚拟资源租赁价格 C、资源需求量 Q、最优配置量 Q^*、协同配置量 $\ddot{Q}$ 等数据；

（2）$x=0$ //记录 $Q>Q^*$ 的服务提供商个数，初始值为 0；

（3） FOR $i=1$ TO t

//使 $Q>Q^*$ 的元素集中在低地址端（即缺货应用提供商），并且按单位资源收益倒序排列

//使 $Q>Q^*$ 的元素集中在高地址端（即有剩余应用提供商），并且按资源租赁价格升序排列

（4） $j=i$；

（5） IF（$l[i].Q>l[i].Q^*$）THEN

（6） 在 l[1]～$l[x+1]$ 中插入元素 $l[j]$，并且前 $x+1$ 个元素按 k 倒序排列；

（7） $x=x+1$；

（8） ELSE

（9） 在 $l[x+1]$～$l[t]$ 中插入元素 $l[j]$，并且后 $t-x$ 个元素按 C 升序排列；

（10） END IF

（11）END FOR

（12）FOR $i=1$ TO x //计算协作配置数量

（13） FOR $j=x+1$ TO t

（14） IF（$l[j].Q<l[j].Q^*$）THEN//将 $l[j].S$ 的剩余资源量部分或者全部配置给 $l[i].S$

（15） IF（$l[i].Q-l[i].Q^*-l[i].\ddot{Q}>l[j].Q^*-l[j].Q$）THEN

（16） $l[i].\ddot{Q}=l[j].Q^*-l[j].Q$；//全部配置

（17） CONTINUE；

（18） ELSE

（19） $l[i].\ddot{Q}=l[i].Q-l[i].Q^*$；//部分配置

（20） BREAK；

（21） END IF

（22） END IF

（23） END FOR

（24）END FOR

5.2 随机需求确定供应的最优资源协同配置策略（optimal resource co-allocation strategy with random-demand and definite-supply，ORCS_RD）

固定价格云资源最优配置策略[42]在我们前期工作中做了详细介绍，下面在此基础上介绍需求随机时的最优云资源协同配置策略，此策略包括两部分，首先是随机需求确定供应的最优资源配置策略如算法 2 所示，其次是此场景下多应用提供商最优资源协同配置策略如算法 3 所示。应用提供商 i 在尽量满足市场需求的同时保证用户 QoS 约束，而获得自身期望收益最大的收益函数为

$$\max\left\{P_i\times E\left[\min(Q_i,u_i)\right]-C_i\times Q_i/Y_i-L_i\times E[(u_i-Q_i)^+]-H_i\times E[(Q_i-u_i)^+]\right\} \quad (1)$$

向用户提供服务价格为 P_u^i，虚拟资源出售价格则为 $P_i=P_u^i/R_i$；用户量 $\tilde{D}_i$ 随机导致虚拟资源需求量随机，因此，虚拟资源需求量 u_i 是一个随机变量，概率密度函数为 $f_i(u_i)$，在处理目标函数中的随机变量 u_i 时，将目标函数转化为期望目标函数，Y_i 为 IaaS 实例产生的虚拟资源量，如式（2）所示：

$$G_i(Q_i)=\int_0^{Q_i}\left[P_i\times u_i-H_i\times(Q_i-u_i)\right]f_i(u_i)\mathrm{d}u_i+\int_{Q_i}^{+\infty}[P_i\times Q_i-L_i\times(u_i-Q_i)]f_i(u_i)\mathrm{d}u_i-C_i\times Q_i/Y_i \quad (2)$$

可得下式（见文献[42]）:

$$\int_0^{Q_i^*}f_i(u_i)\mathrm{d}u_i=\frac{P_i+L_i-C_i/Y_i}{P_i+L_i+H_i} \quad (3)$$

由式（3）可得最优虚拟资源需求量Q_i^*（这里也是最优虚拟资源配置量），由于供给确定则Y_i为确定值，所以云资源最优配置量

$$X_i=Q_i^*/Y_i \quad (4)$$

算法 2　随机需求确定供应的最优资源配置算法（optimal resource allocation algorithm with random-demand and definite-supply，ORAA_RD）

输入：云资源租赁单价C_i，每个用户实例需要消耗虚拟资源R_i CIPU，虚拟资源需求为随机变量u_i，u_i的概率分布表T_i，服务收费单价P_u^i，虚拟资源缺货单价L_i，虚拟资源闲置单价H_i，云资源产生虚拟资源量Y_i。

输出：最优云资源配置量X_i，最优虚拟资源配置量Q_i^*。

步骤：

（1）根据公式$P_i=P_u^i/R_i$得虚拟资源销售价格；

（2）根据式（3）并且查概率分布表T_i，得最优虚拟资源需求量Q_i^*；

（3）根据式（4）计算最优云资源配置量X_i。

算法 3　随机需求确定供应的最优资源协同配置策略（ORCS_RD）

输入：服务提供商集$\{S_1,S_2,\cdots,S_t\}$，实际虚拟资源需求量集$\{Q_1,Q_2,\cdots,Q_t\}$，云资源租赁单价集$\{C_1,C_2,\cdots,C_t\}$，每个用户实例需要的虚拟资源量集$\{R_1,R_2,\cdots,R_t\}$，虚拟资源需求随机变量集$\{u_1,u_2,\cdots,u_t\}$（其概率分布表集$\{T_1,T_2,\cdots,T_t\}$），每个用户服务收费集$\{P_u^1,P_u^2,\cdots,P_u^t\}$，云虚拟资源短缺单价集$\{L_1,L_2,\cdots,L_t\}$，云虚拟资源闲置单价集$\{H_1,H_2,\cdots,H_t\}$，云实例产生虚拟资源量集$\{Y_1,Y_2,\cdots,Y_t\}$。

输出：虚拟资源协同配置数量集$\{\ddot{Q}_1,\ddot{Q}_2,\cdots,\ddot{Q}_t\}$，云资源协同配置量集$\{\ddot{X}_1,\ddot{X}_2,\cdots,\ddot{X}_t\}$。

变量：最优云资源配置量集$\{X_1^*,X_2^*,\cdots,X_t^*\}$，最优虚拟资源需求量集$\{Q_1^*,Q_2^*,\cdots,Q_t^*\}$。

步骤：

（1）FOR $i=1$ TO t

（2）　Call **ORAA_RD**（C_i，R_i，u_i，T_i，P_u^i，L_i，H_i，Y_i）to get X_i^* and Q_i^*；

（3）　$k_i=P_u^i/R_i$ //单位虚拟资源的收益，这里设置为虚拟资源销售单价

（4）END FOR

（5）Call **CA_CR**（$\{S_1,S_2,\cdots,S_t\}$，$\{Q_1,Q_2,\cdots,Q_t\}$，$\{Q_1^*,Q_2^*,\cdots,Q_t^*\}$，$\{k_1,k_2,\cdots,k_t\}$）to get $\{\ddot{Q}_1,\ddot{Q}_2,\cdots,\ddot{Q}_t\}$；

（6）$\{\ddot{X}_1,\ddot{X}_2,\cdots,\ddot{X}_t\}=\{\ddot{Q}_1/Y_1,\ddot{Q}_2/Y_2,\cdots,\ddot{Q}_t/Y_t\}$；

5.3　确定需求随机供应的最优资源协同配置策略（optimal resource co-allocation strategy with definite-demand and random-supply，ORCS_DR）

下面计算供应随机并且用户访问量确定时的云资源配置量，用户访问量$\tilde{D}_i$此时为确定变量，虚拟资源需求量$Q_i^*=\tilde{D}_i\times R_i$也是确定的，单位虚拟资源出售价格$P_i=P_u^i/R_i$，应用提供商租赁尽可能多

的云资源满足市场需求，同时自身期望收益最大，此时的决策变量为云资源配置量 X_i，单位云资源能够支持的虚拟资源量为随机变量 y_i，概率密度函数为 $f_i(y_i)$，即产出虚拟资源量 $Q_i = X_i \times y_i$，有下述公式：

$$\max\left\{P_i \times E[\min(Q_i^*, Q_i)] - C_i \times X_i - L_i \times E[(Q_i^* - Q_i)^+] - H_i \times E[(Q_i - Q_i^*)^+]\right\} \tag{5}$$

由于虚拟资源供应量随机，在处理目标函数中的随机变量 X_i 时，将目标函数转化为期望目标函数，如式（6）所示：

$$G(X_i) = \int_0^{\frac{Q_i^*}{X_i}} [P_i \times Q_i - L_i \times (Q_i^* - Q_i)] f_i(y_i) \mathrm{d}y_i + \int_{\frac{Q_i^*}{X_i}}^{+\infty} [P_i \times Q_i^* - H_i \times (Q_i - Q_i^*)] f_i(y_i) \mathrm{d}y_i - C_i \times X_i \tag{6}$$

命题 1 在供应随机且需求确定条件下，单服务期的最优云资源配置策略 X_i 满足式（7）成立。

$$\int_0^{\frac{Q_i^*}{X_i}} y_i \times f_i(y_i) \mathrm{d}y_i = (C_i + 1) / (P_i + L_i + H_i) \tag{7}$$

证明： 展开 $G(X_i)$ 得

$$\begin{aligned} G(X_i) &= \int_0^{\frac{Q_i^*}{X_i}} [P_i \times X_i \times y_i - L_i \times (Q_i^* - X_i \times y_i)] f_i(y_i) \mathrm{d}y_i + \int_{\frac{Q_i^*}{X_i}}^{+\infty} [P_i \times Q_i^* - H_i \times (X_i \times y_i - Q_i^*)] f_i(y_i) \mathrm{d}y_i - C_i \times X_i \\ &= \int_0^{\frac{Q_i^*}{X_i}} P_i \times X_i \times y_i \times f_i(y_i) \mathrm{d}y_i - \int_0^{\frac{Q_i^*}{X_i}} L_i \times Q_i^* \times f_i(y_i) \mathrm{d}y_i + \int_0^{\frac{Q_i^*}{X_i}} L_i \times X_i \times y_i \times f_i(y_i) \mathrm{d}y_i + \int_{\frac{Q_i^*}{X_i}}^{+\infty} P_i \times Q_i^* \times f_i(y_i) \mathrm{d}y_i \\ &\quad - \int_{\frac{Q_i^*}{X_i}}^{+\infty} H_i \times X_i \times y_i \times f_i(y_i) \mathrm{d}y_i + \int_{\frac{Q_i^*}{X_i}}^{+\infty} H_i \times Q_i^* \times f_i(y_i) \mathrm{d}y_i - C_i \times X_i \end{aligned}$$

求收益函数 $G(X_i)$ 关于 X_i 的一阶导数得

$$\begin{aligned} \frac{\partial G(X_i)}{\partial X_i} &= \int_0^{\frac{Q_i^*}{X_i}} P_i \times y_i \times f_i(y_i) \mathrm{d}y_i + P_i \times X_i \times \frac{Q_i^*}{X_i} \times f_i\left(\frac{Q_i^*}{X_i}\right) \times \left(-\frac{Q_i^*}{X_i^2}\right) - L_i \times Q_i^* \times f_i\left(\frac{Q_i^*}{X_i}\right) \times \left(-\frac{Q_i^*}{X_i^2}\right) + \int_0^{\frac{Q_i^*}{X_i}} L_i \times y_i \\ &\quad \times f_i(y_i) \mathrm{d}y_i + L_i \times X_i \times \frac{Q_i^*}{X_i} \times f_i\left(\frac{Q_i^*}{X_i}\right) \times \left(-\frac{Q_i^*}{X_i^2}\right) - P_i \times Q_i^* \times f_i\left(\frac{Q_i^*}{X_i}\right) \times \left(-\frac{Q_i^*}{X_i^2}\right) - \int_{\frac{Q_i^*}{X_i}}^{+\infty} H_i \times y_i \times f_i(y_i) \mathrm{d}y_i + H_i \times X_i \\ &\quad \times \frac{Q_i^*}{X_i} \times f_i\left(\frac{Q_i^*}{X_i}\right) \times \left(-\frac{Q_i^*}{X_i^2}\right) - H_i \times Q_i^* \times f_i\left(\frac{Q_i^*}{X_i}\right) \times \left(-\frac{Q_i^*}{X_i^2}\right) - C_i \end{aligned}$$

即有

$$\frac{\partial G(X_i)}{\partial X_i} = \int_0^{\frac{Q_i^*}{X_i}} P_i \times y_i \times f_i(y_i) \mathrm{d}y_i + \int_0^{\frac{Q_i^*}{X_i}} L_i \times y_i \times f_i(y_i) \mathrm{d}y_i - \int_{\frac{Q_i^*}{X_i}}^{+\infty} H_i \times y_i \times f_i(y_i) \mathrm{d}y_i - C_i$$

令其一阶导函数为 0，得最优解 X_i 满足：

$$\int_0^{\frac{Q_i^*}{X_i}} y_i \times f_i(y_i) \mathrm{d}y_i = (C_i + 1) / (P_i + L_i + H_i)$$

若证明所求的最优解 X_i 满足上式，通过求函数 $G(X_i)$ 关于 X_i 的二阶导数，证明 $G(X_i)$ 是关于 X_i 的凸函数即可，其二阶导数为：$\frac{\partial^2 G(X_i)}{\partial X_i^2} = (P_i + L_i + H_i) \times \frac{Q_i^*}{X_i} \times f\left(\frac{Q_i^*}{X_i}\right) \times \left(-\frac{Q_i^*}{X_i^2}\right) < 0$。得出函数是凸函数，极大值点取一阶导数为 0 的点。

公式（7）为递增函数，只存在一个根，非常适合采用效率高的二分搜索法求解，本节采用二分搜索法求解，见算法 4。

算法 4　二分搜索求函数解（finding function solution by binary search，FFS_BS）

输入：求解的函数 $g(x)=0$，虚拟资源需求量 Q_i^*。

输出：近似解 x^*。

步骤：

（1）确定目标函数 $g(x)$ 中 x 的初始区间为 (l,r)，其中 $g(l)<0$ 且 $g(r)>0$，终止条件 ε（即 $|l-r|<\varepsilon$）；

（2）计算 $g\left((l+r)/2\right)$；

（3）WHILE（ $|l-r|\geqslant\varepsilon$ ）do

（4）　如果 $g((l+r)/2)<0$，函数近似解在 $[(l+r)/2,r]$ 区间内，$l=(l+r)/2$；

（5）　如果 $g((l+r)/2)>0$，说明极值点在 $[l,(l+r)/2]$ 区间内，$r=(l+r)/2$；

（6）END

（7）IF（ $|l-r|<\varepsilon$ ）THEN　$x^*=2Q_i^*/(l+r)$；

在上述建模、求解和算法实现基础上，下面设计确定需求随机供应的最优资源配置算法，见算法 5，以及此场景下多应用提供商最优资源协同配置算法，参见算法 6。

算法 5　确定需求随机供应的最优资源配置算法（optimal resource allocation algorithm with definite-demand and random-supply，ORAA_DR）

输入：云资源租赁单价 C_i，每个用户实例需要消耗虚拟资源 R_i CIPU，用户访问量为 $\tilde{D}_i$，每个用户服务费 P_u^i，虚拟资源短缺成本 L_i，虚拟资源闲置成本 H_i，云资源支撑虚拟资源随机变量 y_i 的概率密度函数、期望分别为 $f_i(y_i)$、Y_i。

输出：云资源最优配置量 X_i，虚拟资源需求量 Q_i^*。

步骤：

（1）根据公式 $P_i=P_u^i/R_i$ 得虚拟资源销售价格；

（2）根据公式 $Q_i^*=\tilde{D}_i\times R_i$ 得虚拟资源需求量；

（3）CALL **FFS_BS**（式（7），Q_i^*），得最优 IaaS 资源配置量 X_i。

算法 6　确定需求随机供应的最优资源协同配置策略（ORCS_DR）

输入：应用提供商集 $\{S_1,S_2,\cdots,S_t\}$，云资源租赁单价集 $\{C_1,C_2,\cdots,C_t\}$，每个用户实例需要的虚拟资源量集 $\{R_1,R_2,\cdots,R_t\}$，用户访问量集 $\{\tilde{D}_1,\tilde{D}_2,\cdots,\tilde{D}_t\}$，每个用户服务收费集 $\{P_u^1,P_u^2,\cdots,P_u^t\}$，云虚拟资源短缺成本集 $\{L_1,L_2,\cdots,L_t\}$，云虚拟资源闲置成本集 $\{H_1,H_2,\cdots,H_t\}$，云资源能支撑的虚拟资源随机变量集 $\{y_1,y_2,\cdots,y_t\}$ 的概率密度函数集 $\{f_1(y_1),f_2(y_2),\cdots,f_t(y_t)\}$ 和期望集 $\{Y_1,Y_2,\cdots,Y_t\}$。

输出：虚拟资源协同配置量集 $\{\ddot{Q}_1,\ddot{Q}_2,\cdots,\ddot{Q}_t\}$，云资源协同配置量集 $\{\ddot{X}_1,\ddot{X}_2,\cdots,\ddot{X}_t\}$。

变量：最优云资源配置量集 $\{X_1^*,X_2^*,\cdots,X_t^*\}$，虚拟资源需求量集 $\{Q_1,Q_2,\cdots,Q_t\}$，最优虚拟资源配置量集 $\{Q_1^*,Q_2^*,\cdots,Q_t^*\}$

步骤：

（1）FOR　$i=1$　TO　t

（2）　　Call **ORAA_DR**（ C_i，R_i，$\tilde{D}_i$，P_u^i，L_i，H_i，$f_i(y_i)$，Y_i ）to get　X_i^*　and　Q_i^*；

（3）　　$k_i=P_u^i/R_i$　//单位虚拟资源的收益，这里设置为虚拟资源销售单价

（4）END FOR

（5）Call **CA_CR**（ $\{S_1,S_2,\cdots,S_t\}$，$\{Q_1,Q_2,\cdots,Q_t\}$，$\{Q_1^*,Q_2^*,\cdots,Q_t^*\}$，$\{k_1,k_2,\cdots,k_t\}$ ）to get　$\{\ddot{Q}_1,\ddot{Q}_2,\cdots,\ddot{Q}_t\}$；

（6）$\{\ddot{X}_1,\ddot{X}_2,\cdots,\ddot{X}_t\}\approx\{\ddot{Q}_1/Y_1,\ddot{Q}_2/Y_2,\cdots,\ddot{Q}_t/Y_t\}$；//估计值

5.4 随机供需的最优资源协同配置策略（optimal resource co-allocation strategy with random demand and supply，ORCS_RDS）

上述两节讨论了需求随机和供应随机两种场景下的云资源配置问题，然而在很多情况下存在虚拟资源的需求和供应皆为随机的云资源配置问题，下面采用两阶段计算方法实现应用提供商期望收益最大的云资源最优配置策略：第一阶段计算需求随机下的虚拟资源需求量；第二阶段计算供应随机下的云资源配置量，具体如算法 7 所示。此外，随机供需情景的多应用提供商最优资源协同配置策略如算法 8 所示。

算法 7 随机供需的最优资源配置策略（optimal resource allocation strategy with random demand and supply，ORAS_RDS）

输入：云资源租赁单价 C_i，每个用户实例需要消耗虚拟资源 R_i CIPU，每个用户服务费 P_u^i，用户量随机导致虚拟资源需求量随机，用户访问量 $\tilde{D}_i$ 随机得出虚拟资源需求量 u_i 是随机变量，其概率密度函数为 $f_i(u_i)$，概率分布表为 T_i；虚拟资源短缺成本为 L_i，虚拟资源闲置成本为 H_i，云资源支撑虚拟资源量 y_i 是随机变量，其概率密度函数和期望分别为 $f_i(y_i)$、Y_i。

输出：云资源最优配置量 X_i。

步骤：

（1）根据公式 $P_i = P_u^i / R_i$ 得虚拟资源销售价格；

（2）根据式（3）并且查概率分布表 T_i，得最优虚拟资源需求量 Q_i^*；

（3）CALL **FFS_BS**（式（7），Q_i^*），得最优云资源配置量 X_i。

算法 8 随机供需的最优资源协同配置策略（ORCS_RDS）

输入：服务提供商集 $\{S_1, S_2, \cdots, S_t\}$，云资源租赁单价集 $\{C_1, C_2, \cdots, C_t\}$，每个用户实例需要的虚拟资源量集 $\{R_1, R_2, \cdots, R_t\}$，用户访问随机变量集 $\{u_1, u_2, \cdots, u_t\}$ 的概率密度函数集 $\{f_1(y_1), f_2(y_2), \cdots, f_t(y_t)\}$ 和概率分布表集 $\{T_1, T_2, \cdots, T_t\}$，每个用户服务收费集 $\{P_u^1, P_u^2, \cdots, P_u^t\}$，虚拟资源短缺成本集 $\{L_1, L_2, \cdots, L_t\}$，虚拟资源闲置成本集 $\{H_1, H_2, \cdots, H_t\}$，云资源支撑虚拟资源随机变量集 $\{y_1, y_2, \cdots, y_t\}$ 的概率密度函数集 $\{f_1(y_1), f_2(y_2), \cdots, f_t(y_t)\}$ 和期望集 $\{Y_1, Y_2, \cdots, Y_t\}$。

输出：虚拟资源协同配置量集 $\{\ddot{Q}_1, \ddot{Q}_2, \cdots, \ddot{Q}_t\}$，云资源协同配置量集 $\{\ddot{X}_1, \ddot{X}_2, \cdots, \ddot{X}_t\}$。

变量：最优云资源配置量集 $\{X_1^*, X_2^*, \cdots, X_t^*\}$，虚拟资源需求量集 $\{Q_1, Q_2, \cdots, Q_t\}$，最优虚拟资源配置量集 $\{Q_1^*, Q_2^*, \cdots, Q_t^*\}$。

步骤：

（1）FOR $i=1$ TO t

（2） Call **ORAS_RDS**（C_i，R_i，$f_i(u_i)$，P_u^i，T_i，L_i，H_i，$f_i(y_i)$，Y_i）to get X_i^* and Q_i^*；

（3） $k_i = P_u^i / R_i$ //单位虚拟资源的收益，这里设置为虚拟资源销售单价

（4）END FOR

（5）Call **CA_CR**（$\{S_1, S_2, \cdots, S_t\}$，$\{Q_1, Q_2, \cdots, Q_t\}$，$\{Q_1^*, Q_2^*, \cdots, Q_t^*\}$，$\{k_1, k_2, \cdots, k_t\}$）to get $\{\ddot{Q}_1, \ddot{Q}_2, \cdots, \ddot{Q}_t\}$；

（6）$\{\ddot{X}_1, \ddot{X}_2, \cdots, \ddot{X}_t\} \approx \{\ddot{Q}_1 / Y_1, \ddot{Q}_2 / Y_2, \cdots, \ddot{Q}_t / Y_t\}$；//估计值

6 实验分析

6.1 算法性能分析

硬件配置为 Intel（R）Core（TM）i7-10750H 且 2.60GHz 的 CPU、16.0GB 的 RAM（random access memory，随机存储器），操作系统为 Windows 10，算法用 Python 实现，设置协同应用提供商数为 3（通常该数量比较稳定，可视其为常数）。ORCS_RD 中最小最优资源配置量与用户访问量无关，所以实验中不必考虑用户量变化对算法执行时间的影响；ORCS_DR 和 ORCS_RDS 中需要采用二分搜索法获取最小最优配置量，与供应量有关，因此随机变量供应的均值变化对算法性能的影响见图 2。从图中可以看出三种算法的时间复杂度皆为常数。空间复杂度主要涉及需要临时空间建立概率分布表，所以空间复杂度为 $O(n)$，对于正态分布，其复杂度为 $O(320)$。

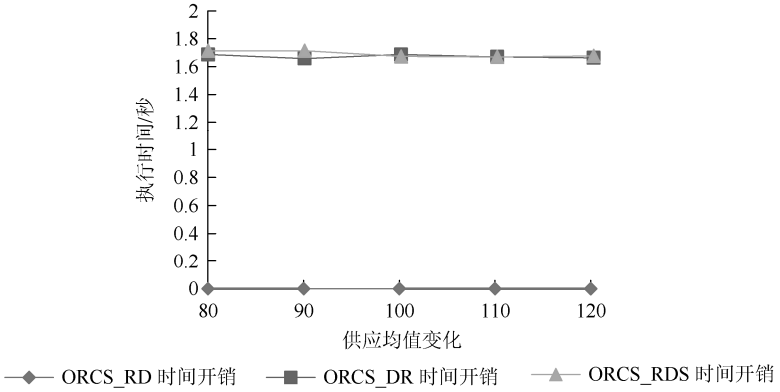


图 2 三种算法的时间复杂度分析

6.2 数值算例

6.2.1 算例基本参数设置

应用提供商 i 向用户提供服务，服务价格 $P_u^i=4$ 元，每个用户访问服务消耗的虚拟资源量 $R_i=10$ CIPU，虚拟资源价格 $P_i=0.4$ 元，云资源单位租赁成本 $C_i=10$ 元，虚拟资源单位缺货成本 $L_i=0.3$ 元，虚拟资源单位资源闲置成本 $H_i=0.1$ 元，用户访问量服从正态分布 $\tilde{D}_i \sim N(5000,500)$，则虚拟资源需求量服从正态分布 $u_i \sim N(50\ 000,5000)$，单位云资源可获虚拟资源量服从正态分布 $y_i \sim N(100,9)$。三类不同算例基本参数设置如表 3 所示。

表 3 算例基本参数

R_i	P_u^i	P_i	C_i	L_i	H_i	u_i	y_i
10	4	0.4	10	0.3	0.1	$N(50\ 000,5000)$	$N(100,9)$

6.2.2 ORCS_RD 算例计算

在该算例中只考虑需求随机情况，由式（3）得，$\int_0^X f(u)\mathrm{d}u=\dfrac{P+L-C/Y}{P+L+H}=\dfrac{0.4+0.3-0.1}{0.4+0.3+0.1}=0.75$；

查正态分布表得，$\int_0^{0.67} f(u_1)\mathrm{d}u_1 = 0.7486 < 0.75 < \int_0^{0.68} f(u_1)\mathrm{d}u_1 = 0.7517$，即有 $(Q^* / 10 - 5000) / 500 \in [0.67, 0.68]$，即 $Q^* \in [53\,350, 53\,400]$。由于本实例中供应确定，即 $Y = 100$，云资源最优配置量 $X \in [533.5, 534.0]$，取其平均值为 533.75 CIPU。

6.2.3 ORCS_DR 算例计算

在该算例中只考虑供应随机情况，需求确定即用户访问量为 5000，则虚拟资源需求量 $Q_i^* = 50\,000$，由式（7）得 $\int_0^{\frac{Q_i^*}{X_i}} y_i \times f_i(y_i)\mathrm{d}y_i = (C_i + 1)/(P_i + L_i + H_i) = 13.75$。使用 FFS_BS 算法计算等式的解，得 $X_i = 543.4783$。即需求确定供应随机下的云资源最优配置量为 543.4783 CIPU，二分搜索迭代过程见图 3。

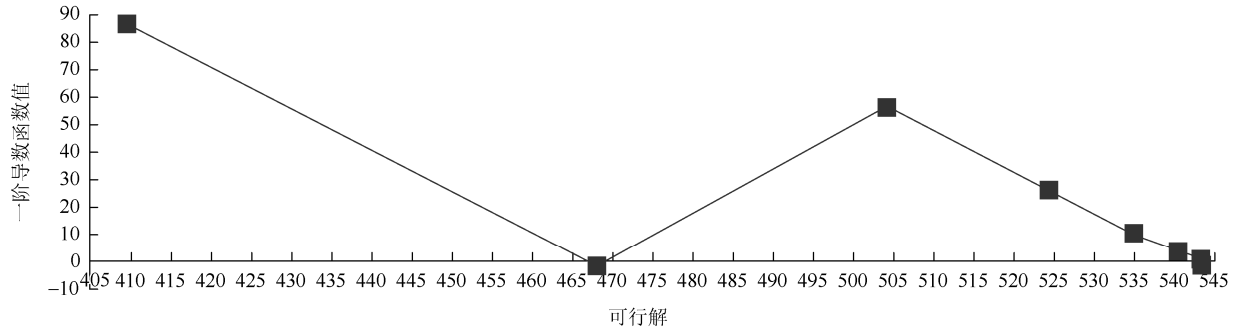


图 3 FFS_BS 算法迭代过程

6.2.4 ORCS_RDS 算例计算

在该算例中既需考虑需求随机，也需考虑供应随机。首先考虑用户访问量随机，由式（3）计算虚拟资源需求量 $Q_i^* = 53\,375$；接着考虑云计算实例能够支撑的虚拟资源数量随机，由式（7）和 FFS_BS 求解最优云资源配置量 $X_i = 580.1630$ CIPU。

6.3 算法比较

6.3.1 文中算法比较

为解决随机需求确定供应、确定需求随机供应、随机供需三种场景下的云应用的资源配置问题，文中设计了 8 个算法，其中云资源协同配置算法（CA_CR）和二分搜索求函数解（FFS_BS）被其他算法调用。因此，本文算法服务三种场景，可归为两类：基于统计信息的最优资源配置策略（ORAA_RD、ORAA_DR 和 ORAS_RDS）；资源再配置的最优资源协同配置策略（ORCS_RD、ORCS_DR 和 ORCS_RDS）。首先，分别比较三种场景下的最优资源配置策略和最优资源协同配置策略的应用提供商收益情况，即进行 ORAA_RD 与 ORCS_RD、ORAA_DR 与 ORCS_DR、ORAS_RDS 与 ORCS_RDS 的比较分析。选取 A、B、C 三个应用提供商，其服务价格分别是 3 元、4 元、5 元，供应确定时 $Y = 100$，需求确定时用户访问量为 5000，其他参数参照表 3。为了保证结果的稳健性，对实际需求和实际供应采样 10 000 次，取 10 000 次协同配置的平均收益与最优资源配置策略的期望收益进行对比，图 4、图 5、图 6 描述了协同配置资源后能更有效提升应用提供商的收益。

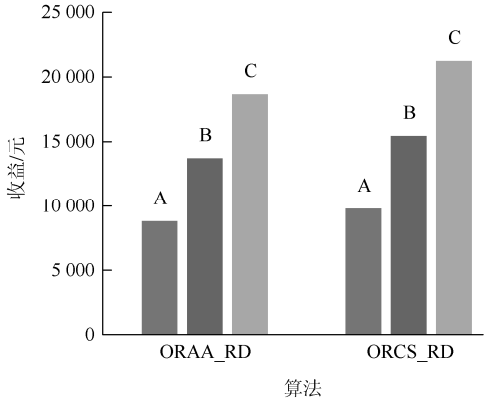


图 4　ORAA_RD 与 ORCS_RD 的收益比较

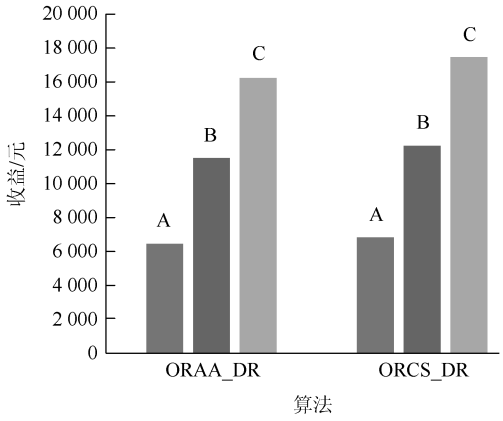


图 5　ORAA_DR 与 ORCS_DR 的收益比较

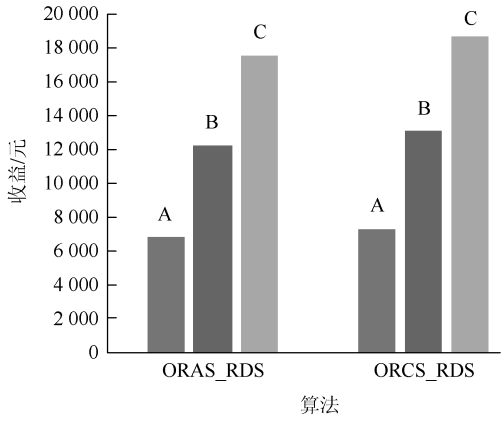


图 6　ORAS_RDS 与 ORCS_RDS 的收益比较

其次，供需随机环境下，比较三种最优资源配置策略（ORAA_RD、ORAA_DR 和 ORAS_RDS）的收益情况。同样在计算每种算法期望收益时，采用 10 000 次需求和供应数据，取其平均。图 7 描述了在供需皆随机条件下，分别采用 ORAA_RD、ORAA_DR、ORAS_RDS 三种算法的应用提供商收益情况比较（从图中可以得出，ORAS_RDS 的收益最高）。在供应确定时，ORAS_RDS 退化为 ORAA_RD；在需求确定时，ORAS_RDS 退化为 ORAA_DR；文中分别设计 ORAA_RD 和 ORAA_DR 的目的是实现 ORAS_RDS，进而实现供需随机条件下应用提供商收益最大的目标。

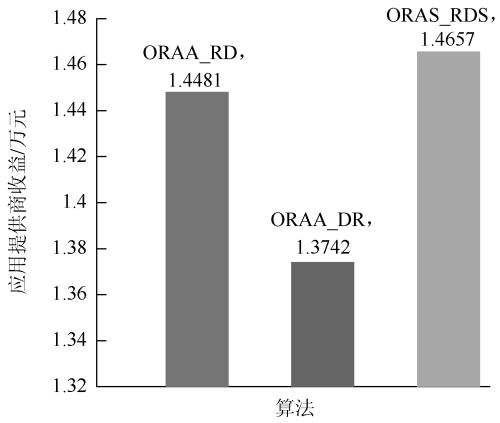
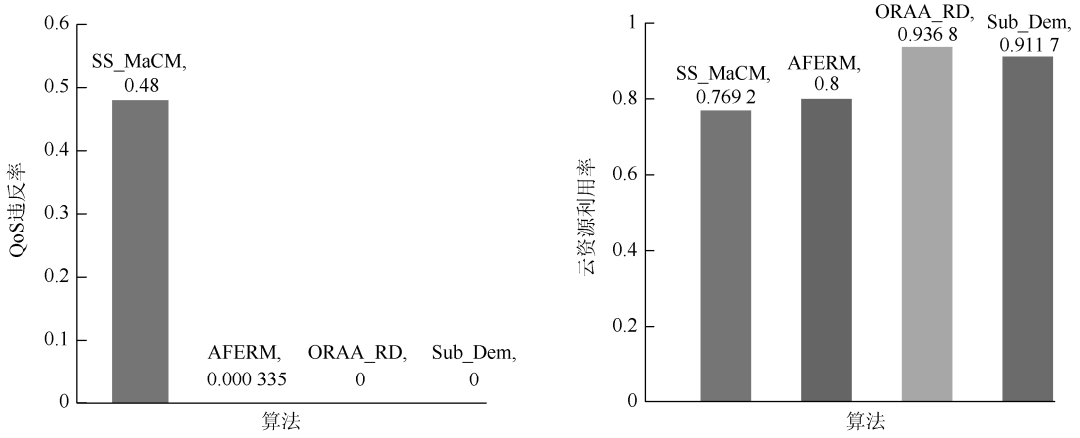


图 7　ORAA_RD、ORAA_DR、ORAS_RDS 收益比较

6.3.2 与已有算法比较

选择考虑云应用资源需求随机的有代表性的研究成果，以及能够使应用提供商有较高收益的资源租赁策略与本文算法进行比较。在文中算法比较部分已经得出协同配置策略优于最优资源配置策略和随机供需中 ORAS_RDS 的收益最高，因此只需验证 ORAA_RD 优于其他算法即可。

本实验将提出的方法 ORAA_RD 与基于云模型的 SaaS 选择算法——SS_MaCM[40]（service selection method based on Mahalanobias cloud model，基于马氏云模型的 SaaS 选择算法）、基于应用特征的 PaaS 弹性资源管理机制（application feature based elastic resource management mechanism，AFERM）进行比较[29]、节省计划 + 按量付费（即 subscription + on-Demand，记为 Sub_Dem）。SS_MaCM 不考虑服务本身及环境动态变化对 QoS 违反率的影响，只将其描述 QoS 准确率最好情况作为 QoS 违反率的参考；SS_MaCM 没有考虑服务的用户访问量的影响，设置其云资源配置为满足 QoS 约束的最大用户访问量所需资源，实际资源使用量与用户访问量的期望成正比。AFERM 方法只参考在请求率规律相对较好的应用上实验的效果，因其权衡资源开销和部署灵活性，设定虚拟资源的开销上限为 80%，不考虑没达到开销上限而剩余的虚拟资源量。Sub_Dem 由 SaaS 提供商提出资源需求，并且 IaaS 提供商采取有效措施保证其资源需求，则违反 QoS 约束率为 0；按照表 3 中的虚拟资源需求量正态分布生成 50 万个随机数中最小值为 30 312.2903，取 30 312 作为预订资源量，不足部分按量付费（按照节省计划为按量付费的 4.55 折计算，则按量付费虚拟资源单价为 0.4/0.455 = 0.88 元）；为了保证按量付费模式的资源需求，设 IaaS 提供商采用 AFERM 资源预留方法，其最高资源利用率为 80%，用虚拟资源需求量正态分布函数生成 5000 次资源需求，取其平均资源利用率。图 8 是违反 QoS 约束和资源利用率比较，ORAA_RD 具有较低的违反 QoS 约束率和较高的云资源利用率。



(a) 违反QoS约束率比较　　(b) 云资源利用率比较

图 8 违反 QoS 约束和资源利用率比较

SS_MaCM 没有考虑用户访问量对资源和应用 QoS 的影响，假定用户访问量为均值，即取算例参数（表 3）中用户访问量 $\tilde{D}_i = 5000$。AFERM 考虑了用户访问量对资源需求和服务 QoS 的影响，其基本思想是满足所有用户并发请求并且尽量减少违反 QoS 约束，取其最大用户访问量 $\tilde{D}_i = 5000 + 3 \times 500 = 6500$。Sub_Dem 的收益为服务收入–租赁虚拟资源成本，其中服务收入为 $\tilde{D}_i \times 4$，租赁虚拟资源成本为 $30\,312 \times 0.4 + \lambda \times 0.88$（其中 λ 为按量付费虚拟资源量，随机生成 5000 次 $\tilde{D}_i$，λ = 5000 次的平均虚拟资源需求量–30 312）。图 9 描述了四种算法的收益比较，图中只考虑需求随机的应用提供商收益情况。

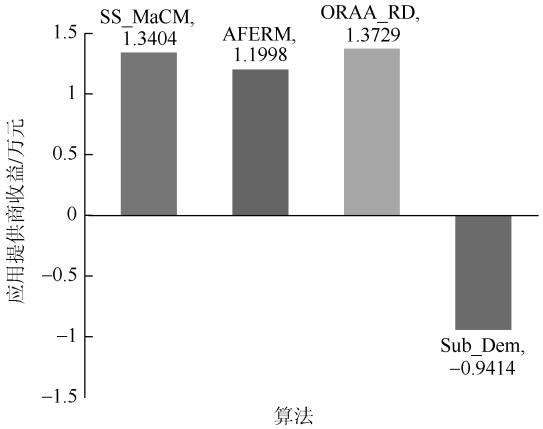
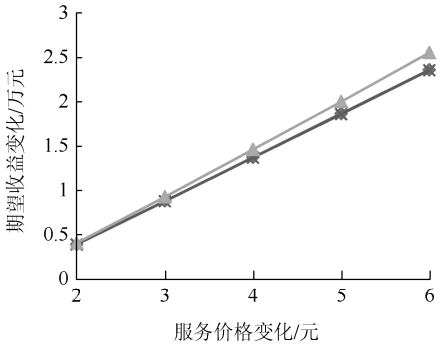


图 9　SS_MaCM、AFERM、ORAA_RD、Sub_Dem 收益比较

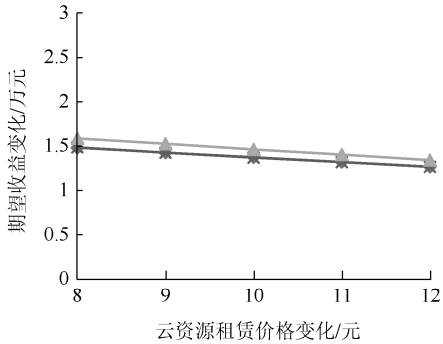
与已有算法的比较：针对用户访问量和云负载随机问题，已有的基于 IaaS 和 PaaS 层面的云资源配置方案，通常通过预测用户访问量和云资源负载进行配置，或者通过监测用户访问量和负载变化进行自适应调整，或者通过预留云资源来减少违反 QoS 约束次数。预测方法难以保证准确度，监测自适应方法不能做到及时响应，经常出现违反 QoS 约束的情况；预留云资源方法虽然可以降低违反 QoS 约束的频次，但是也降低了云系统资源利用率和提高了应用提供商的成本。本文提出的算法获得用户访问量和虚拟资源供应量随机情况下应用提供商期望收益最大的云资源最优配置数量，在 IaaS 和 PaaS 层面无违反 QoS 约束情况，按照最优配置量配置资源能够充分利用云系统资源，而不增加应用提供商成本。

6.4　参数分析

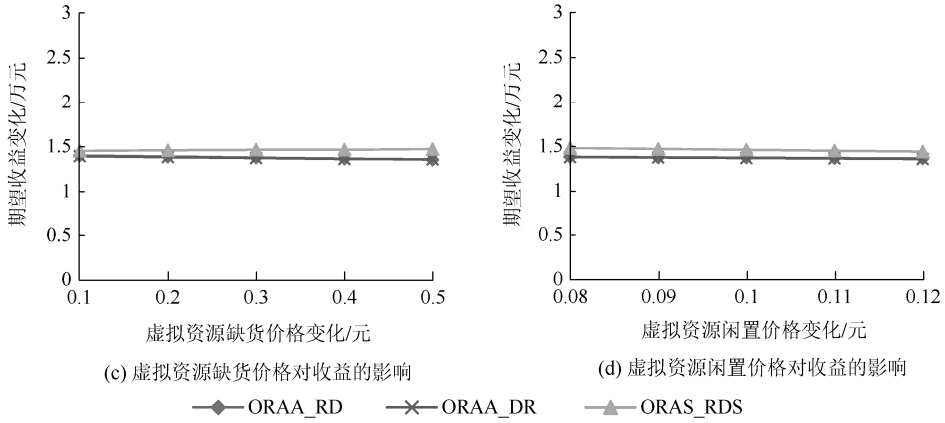
本节实验分析算法中 4 个重要参数对期望收益和最优配置量的影响。在表 3 算例参数的基础上，服务价格在 0.2～0.6 变化，云资源租赁价格在 8～12 变化，虚拟资源的缺货价格在 0.1～0.5 变化，虚拟资源的闲置价格在 0.08～0.12 变化。图 10 描述了 4 个参数的变化对期望收益的影响，从图中可以得出服务的销售价格变化对期望收益的影响最大，云资源租赁价格次之，虚拟资源缺货价格和虚拟资源闲置价格的期望收益的影响不大，总体上 ORAS_RDS 的期望收益略高于其他两种算法。图 11 描述了 4 个参数的变化对最优资源配置量的影响，从图中可以得出，最优配置量与服务价格和虚拟资源缺货价格呈正相关，而与云资源租赁价格和虚拟资源闲置价格呈负相关，总体上 ORAS_RDS 的最优资源配置量高于 ORAA_DR，ORAA_DR 高于 ORAA_RD。



(a) 服务价格对收益的影响

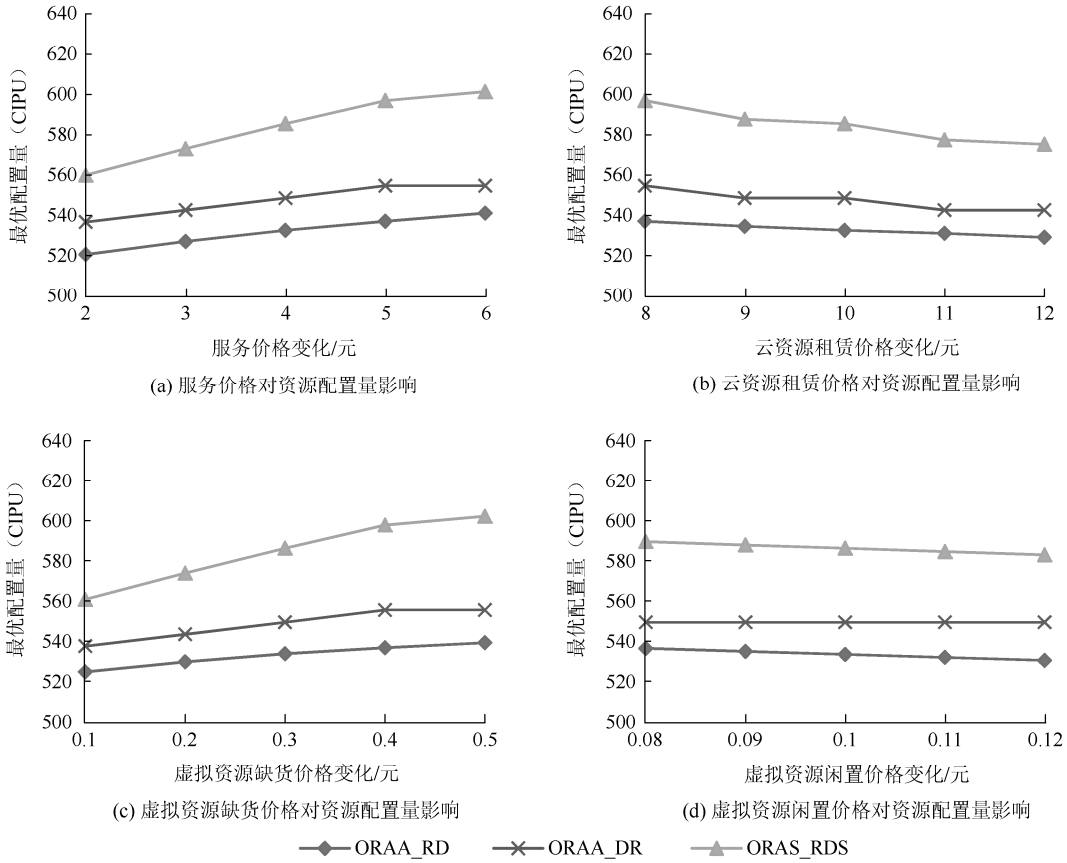


(b) 云资源租赁价格对收益的影响



(c) 虚拟资源缺货价格对收益的影响　(d) 虚拟资源闲置价格对收益的影响

图 10　参数变化对收益的影响



(a) 服务价格对资源配置量影响　(b) 云资源租赁价格对资源配置量影响

(c) 虚拟资源缺货价格对资源配置量影响　(d) 虚拟资源闲置价格对资源配置量影响

图 11　参数变化对资源配置量的影响

7　结束语

已有基于 IaaS 和 PaaS 层面的服务 QoS 约束条件下的云资源配置研究中，在用户访问量和虚拟资源供应（即云资源负载）随机条件下，违反 QoS 约束和资源利用率低问题仍然凸显，并且欠缺对应用提供商的收益考虑。针对上述难点，本文提出随机供需云环境下应用提供商收益驱动的最优资源协同配置策

略。与以往基于 IaaS/PaaS 视角不同，基于云应用提供商期望收益最大的思想提出云资源配置策略，在随机环境下充分考虑云资源租赁价格、服务价格、资源缺货价格和资源闲置价格对云资源配置量和期望收益的影响，丰富了云资源配置理论，同时也是经典报童模型在随机供需云环境下的进一步扩展。实验证明了本文的算法具有较高的运行效率，能够有效获取云资源最优配置量，服务的销售价格和资源租赁价格对收益影响较大，相比以往方法有以下优点：①能够有效确定使应用提供商期望收益最大的三种随机情况下的最优资源配置量，进而可采用节省计划付费模式配置资源，可有效提高应用提供商收益；②应用提供商确定资源配置量，在 IaaS 和 PaaS 提供商层面不存在违反 QoS 约束问题，也有利于其准确配置云资源，提升系统资源利用率；③能够有效应对用户访问量和云资源负载随机导致的资源配置不准确问题；④协同应用提供商的资源二次调配进一步提升其收益。实践中，本文提出的算法可普遍适用于初创期的中小 SaaS 提供商，面向在产品迭代或新客户市场磨合过程中不确定性用户访问量情景提供有效的决策支持，也可适用零售电商 SaaS（如淘宝、京东）在节假日面临的不确定性的资源需求等场景，有效降低云服务市场供求磨合期的运营成本，提高应用提供商的经济效益和市场竞争力。

由于本文对面向应用提供商收益的云资源配置研究目前还处于初始阶段，仍然存在很多不足需要不断完善。文中假设参与资源协同配置的应用提供商积极性高、在没有损失条件下能够接受算法的统一调配，但现实中存在一些应用提供商为了利益而选择不配合，进而出现缺货应用提供商和剩货应用提供商之间的博弈、缺货应用提供商之间竞争剩货的博弈、剩货应用提供商之间的竞价博弈，这是一个需要深入研究的问题。另外，还需进一步研究应用提供商收益驱动的价格影响需求的资源配置、多资源多期配置、云资源的统一量化，以及面向 IaaS、PaaS 和应用三级服务提供商收益的多期云资源配置策略。

参 考 文 献

[1] Zhou A，Wang S G，Zheng Z B，et al. On cloud service reliability enhancement with optimal resource usage[J]. IEEE Transactions on Cloud Computing，2016，4（4）：452-466.

[2] Zhu Q，Agrawal G. Resource provisioning with budget constraints for adaptive applications in cloud environments[J]. IEEE Transactions on Services Computing，2012，5（4）：497-511.

[3] Shi H M，Xu H C，Xu X F，et al. Service composition considering QoS fluctuations and anchoring cost[C]//2021 IEEE International Conference on Web Services. Chicago：IEEE，2021：370-380.

[4] Wu C，Toosi A N，Buyya R，et al. Hedonic pricing of cloud computing services[J]. IEEE Transactions on Cloud Computing，2021，9（1）：182-196.

[5] Alzhouri F，Agarwal A，Liu Y. Maximizing cloud revenue using dynamic pricing of multiple class virtual machines[J]. IEEE Transactions on Cloud Computing，2021，9（2）：682-695.

[6] Buyya R，Yeo C S，Venugopal S. Market-oriented cloud computing：vision，hype，and reality for delivering IT services as computing utilities[C]//2008 10th IEEE International Conference on High Performance Computing and Communications. Dalian：IEEE，2008：5-13.

[7] 李强，郝沁汾，肖利民，等. 云计算中虚拟机放置的自适应管理与多目标优化[J].计算机学报，2011，34（12）：2253-2264.

[8] 孙大为，常桂然，李凤云，等. 一种基于免疫克隆的偏好多维 QoS 云资源调度优化算法[J].电子学报，2011，39（8）：1824-1831.

[9] Addya S K，Satpathy A，Ghosh B C，et al. CoMCLOUD：virtual machine coalition for multi-tier applications over multi-cloud environments[J]. IEEE Transactions on Cloud Computing，2023，11（1）：956-970.

[10] Belgacem A，Beghdad-Bey K，Nacer H. Dynamic resource allocation method based on symbiotic organism search algorithm in cloud computing[J]. IEEE Transactions on Cloud Computing，2022，10（3）：1714-1725.

[11] 匡桂娟，曾国荪，曹洁，等. 基于图匹配理论的云任务与云资源满意“婚配”方法[J].电子学报，2014，42（8）：1582-1586.

[12] 郭伟，张凯强，崔立真，等. 支持 SaaS 应用多维异构性能需求的云资源放置方法[J].计算机学报，2018，41（6）：1225-1237.
[13] 孙佳佳，王兴伟，高程希，等. 云环境下基于神经网络和群搜索优化的资源分配机制[J].软件学报，2014，25（8）：1858-1873.
[14] 闫永明，张斌，郭军，等. 基于强化学习的 SBS 云应用自适应性能优化方法[J].计算机学报，2017，40（2）：464-480.
[15] 周平，殷波，邱雪松，等. 面向服务可靠性的云资源调度方法[J].电子学报，2019，47（5）：1036-1043.
[16] Liu C B，Tang F，Hu Y K，et al. Distributed task migration optimization in MEC by extending multi-agent deep reinforcement learning approach[J]. IEEE Transactions on Parallel and Distributed Systems，2021，32（7）：1603-1614.
[17] 任神河，郑寇全，关冬冬，等. 基于 IFTS 的云计算网络动态负载均衡方法[J]. 系统工程理论与实践，2019，39（5）：1298-1307.
[18] Denninnart C，Salehi M A. Harnessing the potential of function-reuse in multimedia cloud systems[J]. IEEE Transactions on Parallel and Distributed Systems，2022，33（3）：617-629.
[19] Chen M S，Huang S J，Fu X，et al. Statistical model checking-based evaluation and optimization for cloud workflow resource allocation[J]. IEEE Transactions on Cloud Computing，2020，8（2）：443-458.
[20] 吴悦文，吴恒，任杰，等. 面向大数据分析作业的启发式云资源供给方法[J].软件学报，2020，31（6）：1860-1874.
[21] 苏命峰，王国军，李仁发. 边云协同计算中基于预测的资源部署与任务调度优化[J].计算机研究与发展，2021，58（11）：2558-2570.
[22] 俞岭，谢奕，陈碧欢，等. 基于前馈和反馈控制运行时虚拟资源动态分配[J]. 计算机研究与发展，2015，52（4）：889-897.
[23] 苑迎，王翠荣，王聪，等. 基于非完全信息博弈的云资源分配模型[J]. 计算机研究与发展，2016，53（6）：1342-1351.
[24] Wang Y，Zhou J T，Song X Y. A utility game driven QoS optimization for cloud services[J]. IEEE Transactions on Services Computing，2022，15（5）：2591-2603.
[25] Nallur V，Bahsoon R. A decentralized self-adaptation mechanism for service-based applications in the cloud[J]. IEEE Transactions on Software Engineering，2013，39（5）：591-612.
[26] Haratian P，Safi-Esfahani F，Salimian L，et al. An adaptive and fuzzy resource management approach in cloud computing[J]. IEEE Transactions on Cloud Computing，2019，7（4）：907-920.
[27] Alsarhan A，Itradat A，Al-Dubai A Y，et al. Adaptive resource allocation and provisioning in multi-service cloud environments[J]. IEEE Transactions on Parallel and Distributed Systems，2018，29（1）：31-42.
[28] 伍之昂，罗军舟，宋爱波，等. 跨数据中心的动态资源联合预留研究[J]. 计算机学报，2014，37（11）：2317-2326.
[29] 魏豪，周扬睿，张锐，等. 基于应用特征的 PaaS 弹性资源管理机制[J]. 计算机学报，2016，39（2）：223-236.
[30] Jayathilaka H，Krintz C，Wolski R. Detecting performance anomalies in cloud platform applications[J]. IEEE Transactions on Cloud Computing，2018，8（3）：764-777.
[31] 徐雅斌，彭宏恩. 基于需求预测的 PaaS 平台资源分配方法[J]. 计算机应用，2019，39（6）：1583-1588.
[32] 孟煜，张斌，郭军，等. 云计算环境下云服务用户并发量的区间预测模型[J]. 计算机学报，2017，40（2）：378-396.
[33] 谢晓兰，张征征，王建伟，等.基于三次指数平滑法和时间卷积网络的云资源预测模型[J]. 通信学报，2019，40（8）：143-150.
[34] Ghosh N，Ghosh S K，Das S K. SelCSP：a framework to facilitate selection of cloud service providers[J]. IEEE Transactions on Cloud Computing，2014，3（1）：66-79.
[35] Ding S，Li Y Q，Wu D S，et al. Time-aware cloud service recommendation using similarity-enhanced collaborative filtering and ARIMA model[J]. Decision Support Systems，2018，107：103-115.
[36] Jain T，Hazra J. Hybrid cloud computing investment strategies[J]. Production and Operations Management，2019，28（5）：1272-1284.
[37] Hosseini L，Tang S J，Mookerjee V，et al. A switch in time saves the dime：a model to reduce rental cost in cloud computing[J]. Information Systems Research，2020，31（3）：753-775.
[38] 彭高贤，文一凭，刘建勋，等. 能耗感知的云制造服务选择与调度优化方法[J]. 计算机集成制造系统，2024，30（8）：2697.

[39] Qi L Y，Dou W C，Hu C H，et al. A context-aware service evaluation approach over big data for cloud applications[J]. IEEE Transactions on Cloud Computing，2020，8（2）：338-348.
[40] Ma H，Zhu H B，Li K Q，et al. Collaborative optimization of service composition for data-intensive applications in a hybrid cloud[J]. IEEE Transactions on Parallel and Distributed Systems，2019，30（5）：1022-1035.
[41] 周知，刘方明. 面向多租户数据中心资源回收利用的能效激励机制[J].中国科学（信息科学），2021，51（5）：735-749.
[42] 白静，张龙昌. 云应用提供商收益驱动的最佳云资源配置策略[J].计算机集成制造系统，2024，30（7）：2495.

Optimal Resource Co-allocation Strategy Driven by Application Providers Revenue in Random Supply and Demand Cloud Environment

BAI Jing[1]，XU Jianjun[2]，ZHANG Longchang[3, 4]
（1. School of Management Science and Engineering，Dongbei University of Finance and Economics，Dalian 116025，China；
2. Institute of Supply Chain Analytics，Dongbei University of Finance and Economics，Dalian 116025，China；
3. College of Information Engineering，Suqian University，Suqian 223800，China；
4. Shenzhen Research Institute，Beijing University of Posts and Telecommunications，Shenzhen 518038，China）

Abstract For the existing cloud resource configuration schemes，the randomness of user access (resource demand) and resource supply of applications，as well as the revenue of application providers，are not considered enough，the optimal resource co-allocation strategy driven by application providers revenue in random supply and demand cloud environment is proposed. A quantitative model of resources and requirements was established. Based on the principle of maximizing revenue for cloud application providers，three optimal resource co-allocation strategies were designed：random-demand and definite-supply，definite-demand and random-supply，and random-supply and random-demand. When user access and resource supply are random，this strategy can effectively improve the revenue of cloud application providers，without violating QoS constraints，and cloud resources can be fully utilized.

Key words Cloud application providers，Cloud resources，Co-allocation，Revenue-driven，Random-supply and random-demand

作者简介

许建军（1969—），男，东北财经大学管理科学与工程学院教授、博士生导师，研究方向为库存优化、动态规划、随机优化等。E-mail：xujianjun@dufe.edu.cn。

白静（1987—），女，东北财经大学管理科学与工程学院 2021 级博士研究生，研究方向为信息系统、云计算、库存优化。E-mail：bj490367659@126.com。

张龙昌（1978—），男，宿迁学院信息工程学院教授、北京邮电大学深圳研究院客座教授、硕士生导师，研究方向包括信息系统、服务计算、云计算等。E-mail：zlc_041018@163.com。