Journal of Computer Applications
计算机应用,2016,36(1):113 -116
ISSN 1001-9081
CODEN JYIIDU
2016-01-10
http://www.joca.cn

文章编号:1001-9081(2016)01-0113-04
DOI:10.11772/j.issn.1001-9081.2016.01.0113

多实例云计算资源市场下超额预订决策方法

陈冬林[1],姚梦迪[1*],邓国华[1,2]
(1.武汉理工大学 电子商务与智能服务研究中心,武汉 430070; 2.江汉大学 商学院,武汉 430056)
(*通信作者电子邮箱 1126983967@qq.com)

摘 要:针对现有云供应商数据中心负载率低、云用户需求不确定及多样性的问题,为提高云供应商平均利润,建立了不确定需求下的多实例类型云服务超额预订模型。该模型结合实际云计算资源市场下超额预订对于云供应商负载均衡及云服务等级协议(SLA)的影响,给出超额预订的多重约束条件,提出了各实例类型数量最优分配策略。实验结果表明,采用该模型在预约未使用概率为0.25时,云供应商利润较高,数据中心负载率达到78%,最终确定了各实例类型的最优分配数量。

关键词:不确定需求;云计算资源市场;多实例;超额预订
中图分类号:TP393 **文献标志码**:A

Overbooking decision-making method of multiple instances under cloud computing resource market

CHEN Donglin[1], YAO Mengdi[1*], DENG Guohua[1,2]
(1. *Research Center of E-commerce and Intelligence Service, Wuhan University of Technology, Wuhan Hubei* 430070, *China*;
2. *School of Business, Jianghan University, Wuhan Hubei* 430056, *China*)

Abstract: Considering the problems of low load rate of data centers in cloud providers, uncertainty and variety of cloud user demand; in order to improve the average profit of the cloud providers, an overbooking model of multiple instances under uncertain demand was established. The proposed model combined the influences of overbooking for cloud data center load balancing and Service Level Agreement (SLA) under the actual cloud computing resource market, multi-constraint of overbooking was provided, then the optimal allocation policy of each instance type was put forward. The simulation results show that when the unused rate of reservation is 0.25, the average profit is relatively high, the load rate of data center is 78%, finally the optimal allocation of each instance type is determined.

Key words: uncertain demand; cloud computing resource market; multi-instance; overbooking

0 引言

云计算作为一种将大量网络计算资源、存储资源与软件资源统一调度以构成一个计算资源池向用户提供按需服务的互联网技术(Information Technology, IT)服务模式,逐渐被广泛运用[1]。云服务提供商为云用户提供了多种类型的云服务实例,具有代表性的是由Amazon的弹性计算云(Elastic Compute Cloud, EC2)提出的保留实例、按需实例和竞价实例这3种实例类型[2]。云计算这种全新的IT资源使用模式,使得越来越多的云用户可根据自身实际的业务需求选择不同类型的云服务实例。对于云服务提供商而言,未来云用户的需求存在高度不确定性,文献[3]指出在未采取超额预订策略时,大多数云数据中心的CPU(Central Processing Unit)平均使用率仅达到40%[3]。为此,云服务提供商将考虑采取何种实例超额预订模型来设置最佳可售实例数量,在确保数据中心负载均衡的前提下减少实例空闲损失。超额预订策略的采取可以减少预约了实例却不使用的云用户和取消预约的云用户对于云服务实例资源的浪费,以降低实例资源大量空闲的成本,提高实例资源利用率,获得更高的收益,但同时也存在超额预订的风险——当用户对于资源需求量超过数据中心服务器负载均衡时,云服务的服务质量(Quality of Service, QoS)将会受到影响,对用户而言将出现服务器宕机等问题,违反云服务等级协议(Service-Level Agreement, SLA),影响云用户体验。因此,在云用户需求不确定的前提下,如何确定实例超额预订数量,平衡好实例空闲的资源浪费和超额预订导致数据中心超载违反SLA协议这两者之间的矛盾,是云服务供应商实现云服务最大化收益的研究重点。

超额预订模型作为收益管理中重要的研究内容,在航空客运、酒店、医疗行业等领域研究较多[4-8]。朱金福等[4]提出了航空运输收益管理的以收益最大化为目标函数,应用动态规划方法建立了舱位控制和超售综合控制静态模型。赵昊天等[5]在此基础上考虑了用户需求不确定情况下的航空动态超售问题,采用可调整的鲁棒优化方法建立数学模型。周蔷等[6]则将预售期内的旅客订票看作Poisson过程,建立动态超售模型,结合枚举法进行求解。其他行业超额预订问题的研究对于研究云服务超售问题有一定的借鉴意义,但由于云计

收稿日期:2015-07-22;**修回日期**:2015-09-24。 **基金项目**:国家自然科学基金资助项目(71172043);教育部回国留学人员科研启动基金资助项目(2013-693);武汉理工大学研究生自由探索项目(2015-zy-126)。
作者简介:陈冬林(1970-),男,湖北安陆人,教授,博士,主要研究方向:云计算、商务智能、互联网经济; 姚梦迪(1991-),女,湖北安陆人,博士研究生,主要研究方向:云计算、智能推荐; 邓国华(1984-),男,湖北武汉人,博士研究生,主要研究方向:云计算、智能推荐、数据分析。

算技术、SLA 协议、多种实例类型、负载均衡等问题使得云服务超额预订模型的研究又具有一定的特殊性和难度。

目前，对于云服务超额预订决策问题的研究较少。Rachel 等[9]将虚拟机的优先级引入到 CPU 超售研究中，考虑了基础设施成本及 SLA 惩罚费用，采用 CloudSim 进行仿真对比采取超售前后的 QoS 及收益。Kim 等[8]研究面向云服务供应商的收益管理问题，提出在资源需求不确定下的基于服务接纳控制的决策模型。现有云服务收益管理模型集中研究云计算虚拟机的超额预订，并未考虑云用户需求的不确定性，不同类型的云服务实例特点和其需求分布特点的区别，未充分考虑超额预订风险对数据中心负载均衡的影响。针对此问题，本文考虑不同实例类型需求的随机分布函数，建立在数据中心负载均衡约束下的超额预订模型，得出各实例类型的数量最优分配策略，通过算例分析验证了该模型的有效性。

1 云服务多实例超额预订问题描述

现有云计算市场主要的实例类型分为：按需定价实例、保留实例和竞价实例。据相关数据分析，由于各实例类型不同，云用户对这三类实例的需求不同。现假定每类实例的数量和价格均不同。按需运行实例指用户可以根据实际应用情况调整对 EC2 实例的需求，采用按需付费(pay as you go)方式，具有灵活性好，服务有保障，但单价最高的特点。保留定制实例(Reserved Instance，RI)指用户可以提前预订 EC2 实例，一般需签订 1～3 年的购买合同，并一次性支付费用，服务质量与按需实例的服务质量相同，且费用很低。竞价实例采取低折扣、竞标租用 EC2 闲散资源的购买模式，用户设置一个对实例的要求和能接受的每小时最大费率，然后把这个请求投入“拍卖场”，市场的价格是在供需平衡的原则基础上波动的，同等条件下出价最高者获得竞价实例运行机会。根据这 3 种实例的特点，为了使得云供应商利益达到最大化，本文需合理设置各实例类型的数量，根据按需实例和竞价实例的需求分布概率，制定合适的保留实例预订策略使得云供应商获得最大的利润，且不会因为超售问题而出现 SLA 违约、降低云服务 QoS 等问题。

现根据实例类型不同，本章对于实例的超额预订作出以下假设：

1) 现有云供应商所提供的云服务实例类型为 3 种，各种类型实例类型虚拟机单价已事先确定，且不相同。

2) 云用户对于竞价实例和按需实例的需求服从参数为 λ 的泊松分布，保留定制实例采取预订模式，而对于每种实例类型的需求彼此独立。

3) 云用户取消预订保留实例的请求相互独立，并服从二项分布。

4) 假定云供应商利益用云实例销售收入扣除数据中心运行费用和超售赔偿后的利润来衡量，云服务 QoS 指标由云数据中心负载均衡为标准限制超售云实例在一定数量内为标准，最终达到这两个指标的均衡。

模型所要使用到的参数和变量定义如下：

M 为云供应商所提供的总实例数量，其中按需实例需求量 k_1 为满足参数为 λ_O 的泊松分布，竞价实例需求量 k_2 为满足参数为 λ_S 的泊松分布，且两者相互独立，则保留实例数量 $D_R = M - k_1 - k_2$；f_O 为按需实例单价(元/h)；f_R 为保留定制实例单价(元/h)；f_S 为竞价实例单价(元/h)；L_R 为可接受的保留实例预订的数量，只有当保留实例已预订的数量小于 L_R 时才可以接受预订，且满足 $L_R \geqslant D_R$ 即超售情况出现，超售实例数量为 $L_R - D_R$；N_R 为预订却未使用的保留实例数量；p_R 为每个保留实例预订却未使用的概率；b_R 为保留实例超售时，云用户获得的赔偿金；r 为单个实例的平均运行成本；d 为云数据中心当前负载率。

2 多服务实例动态超售模型

对于按需实例和竞价实例的需求由于存在不确定性，结合互联网及通信、航空[7,10]相关行业的服务中用户需求均服从泊松分布模型，本文借鉴已有的超售模型中常采用的泊松分布模型来描述单位时间(或空间)内随机事件发生的次数即单位时间内云用户需要的实例数量，则在一定时间内云用户提交 k_1 个按需实例的云任务概率为：

$$P_{k_1}(\lambda_O) = \frac{e^{-\lambda_O}\lambda_O^{k_1}}{k_1!};\ k_1 = 1,2,\cdots$$

云用户提交 k_2 个竞价型实例的云任务概率为：

$$P_{k_2}(\lambda_S) = \frac{e^{-\lambda_S}\lambda_S^{k_2}}{k_2!};\ k_2 = 1,2,\cdots$$

对于按需实例和竞价实例，云供应商的收益 s_1 是云任务个数的线性函数 $s = kf$，则云供应商从按需实例和竞价实例中获得的期望收益为：

$$E(S_1) = \sum_{s_O}^{\infty} s_O \frac{\lambda_O^{s_O/f_O}}{(s_O/f_O)!}e^{-\lambda_O} + \sum_{s_S}^{\infty} s_S \frac{\lambda_S^{s_S/f_S}}{(s_S/f_S)!}e^{-\lambda_S} = f_O\lambda_O + f_S\lambda_S$$

对于保留实例，根据 $D_R = M - k_1 - k_2$，其需求受到按需实例和竞价实例需求的影响。云供应商在保留实例中获取的收益为：

$$s_2 = \begin{cases} L_R f_R, & L_R - N_R \leqslant D_R \\ D_R f_R - (L_R - D_R - N_R)b_R, & L_R - N_R > D_R \end{cases}$$

对于每个保留实例云任务，其被云用户预订却未使用的概率服从二项分布，多个实例之间请求相互独立。设每个保留实例发生预订却未使用的概率 p_R 相同，则 N_R 个实例预订未使用的分布律为：

$$P_{N_R}(L_R) = C_{L_R}^{N_R} p_R^{N_R}(1-p_R)^{L_R-N_R};\ N_R = 0,1,\cdots,L_R$$

故云供应商保留实例获得的期望收益为：

$$E(S_2) = \sum_{N_R=0}^{L_R-D_R-1}[D_R f_R - (L_R - D_R - N_R)b_R]p_R + \sum_{N_R=L_R-D_R}^{L_R} L_R f_R / L_R$$

云供应商最终获得的平均利润为：

$$E(S) = E(S_1) + E(S_2) - Mr = \sum_{s_O}^{\infty} s_O \frac{\lambda_O^{s_O/f_O}}{(s_O/f_O)!}e^{-\lambda_O} + \sum_{s_S}^{\infty} s_S \frac{\lambda_S^{s_S/f_S}}{(s_S/f_S)!}e^{-\lambda_S} + \sum_{N_R=0}^{L_R-D_R-1}[D_R f_R - (L_R - D_R - N_R)b_R]p_R + \sum_{N_R=L_R-D_R}^{L_R} L_R f_R/L_R - Mr = f_O\lambda_O + f_S\lambda_S + \sum_{N_R=0}^{L_R-D_R-1}[D_R f_R - (L_R - D_R - N_R)b_R]p_R +$$

$$\sum_{N_R=L_R-D_R}^{L_R} L_R f_R / L_R - Mr$$

考虑到云供应商的社会声誉,以数据中心负载率作为衡量标准,根据文献[8]当数据中心的负载率超过 80% 时(即此时有j个云用户任务无法获得实例),将影响云服务 QoS,其发生的概率为 p_j。

故最终建立的超售的数学优化模型如下:

$$g = \max(E(P(L_R)))$$

$$\text{s.t. } L_R \in [0, M]$$

$$p_j \leqslant \alpha$$

对于上述模型求解最优解,即当云用户使用按需实例和竞价实例的云任务概率 λ_O 和 λ_S、数据中心负载率超过 80% 的概率 p_j 等确定的条件下,使得云供应商获得最大平均利润时,保留实例的最大预订数量 L_R。当 L_R 较低时,保留实例用户预订未使用概率较高,出现云数据中心负载率低于 60% 的情况使得云实例运行相对成本较高。当 L_R 逐渐升高时,保留实例用户预订未使用概率较低,云数据中心负载率达到均衡,云供应商平均利润逐渐升高,直到 L_R 逐渐变高使得超售的赔偿费用逐渐变多,平均利润再次降低。L_R 存在最优解,使得云供应商获得最大平均利润。

3　算例分析

以 AmazonEC2 在美国东部(弗吉尼亚北部)的数据中心为例,提供总实例数量为 5 000,分别有 3 种实例类型:按需实例、保留实例、竞价实例,云供应商规定 3 种实例平均单价(本文选定实例类型为 Linux 通用 m3. medium 实例,一年期限)分别为:0. 07 美元/h、0. 042 5 美元/h、0. 008 1 美元/h[11],其中超额预订用户只发生在保留实例类型中。取该地区 2013 年 1 月到 12 月的数据表明,平均负载率为 71%,单个实例运行成本为 0. 01 美元/h。

取按需实例和竞价实例需求的云任务个数的随机强度 $\lambda_O = \lambda_S = 4$,保留实例云用户预约未使用率 $p_R = 0 \sim 0.7$ 时,采用本文超售模型平均利润随预约未使用率变化如图 1 所示。当预约未使用率 $p_R = 0$ 时,平均利润达到理想值,即满载时平均利润为 1.956×10^3 美元/h,但此时数据中心满载,会产生数据中心 SLA 违约、云服务 QoS 降低等问题。随着 p_R 的变大,预约用户不出现,导致利润降低。超额预订可以降低预约用户未出现带来的利润损失,但同时会增加负载率,出现负载过高的风险。可知当预订未使用率为 0. 25,平均利润降低较少且负载率为 78%,说明采取超额预订可以使得平均利润和负载率达到最优状态。

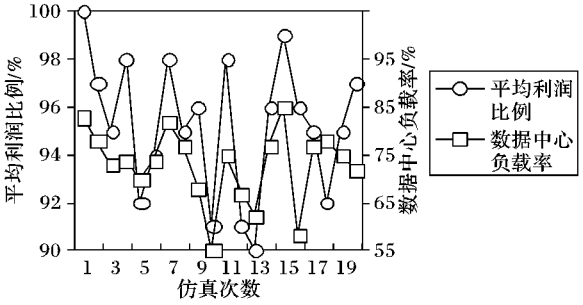

图 1　不同预订未使用率下平均利润及负载率曲线

利用本文超额预订模型对云数据中心实例进行销售控制,$M = 50\,000$,$\lambda_O = \lambda_S = 4$,$p_R = 0.2$,$b_R = 0.02$ 美元/h,$r = 0.02$ 美元,$f_O = 0.07$ 美元,$f_s = 0.008\,1$ 美元。通过计算机仿真生成满足泊松分布的实例需求,采用该超额预订模型,可得当无已购买实例却无法使用的用户且负载率刚好达到 80% 时,云数据中心平均利润为 1.956×10^3 美元/h。对该模型仿真 20 次,分析超额预订平均利润占最大平均利润的百分比如图 2 所示。其中超额预订利润 14 次超过理论最大利润的 95%,可表明该超额预订模型能够很好地降低由于预约未到所带来的损失,但当已购买实例却无法使用的情况发生时,不仅需要赔偿已购买实例却无法使用的用户导致利润降低,还会由于已购买实例却无法使用的用户导致数据中心超载运行,从而影响云服务 QoS,进而影响云服务供应商社会声誉。在该超额预订模型控制下的 20 次仿真中,数据中心超过负载均衡 3 次,即 p_j 为 15%,负载达到均衡(60% ~ 80%)为 16 次,说明超额预订模型能够将 p_j 的值控制在较小范围内,将数据中心负载率控制在均衡水平,很好地降低了由于超额预订所带来的风险,确保了云实例的服务 QoS,维护了云供应商社会声誉。

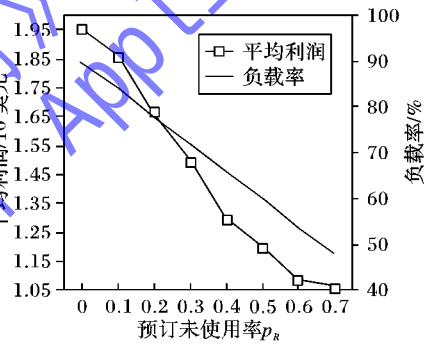

图 2　超额预订情形下平均利润百分比和负载率

根据上述实验数据和分析,可得出 3 种实例类型实例数量最优分配策略,表 1 表明按需实例和竞价实例均不需求预留,其中按需实例的最大销售实例数量为 33 750,为本数据中心销售的最大实例数量;竞价实例存在一定的销售限制,以保证保留实例存在一定的预留,在确保数据中心负载均衡的前提下,获得最大期望利润。

表 1　各实例类型最优分配数

实例类型	保留实例限制数	最大销售实例数
按需实例数	—	33 750
保留实例数	3 600	44 175
竞价实例数	—	16 500

4　结语

本文结合云计算资源市场下的实例类型特点,综合考虑不同实例类型需求的不确定性,建立动态超额预订模型,在确保云服务 QoS 的前提下提高云供应商利润。通过算例仿真验证了该模型在保留实例用户预约未使用概率达到0. 25时,数据中心负载率达到 78%;p_j 的值控制在较小范围内,很好地降低了由于超额预订所带来的风险,确保了云实例 QoS。该模型根据某一云数据中心历史数据统计得出其不同实例分布情况及超售比例,最终获得各实例类型数据的最优分配数量;但是由于本文仅选取 Linux 通用 m3. medium 一年期限的

实例进行分析,并未对存储、CPU 等不同实例进行分析,在此基础下研究不同实例类型随需求变动下的定价策略,是下一步的研究方向。

参考文献:

[1] 冯登国,张敏,张妍,等. 云计算安全研究[J]. 软件学报,2011,22(1):71 -83.(FENG D G, ZHANG M, ZHANG Y, et al. Study on cloud computing security [J]. Journal of software, 2011,22(1):71 -83.)

[2] 屈喜龙,肖鹏,白泉. 云环境中基于弹性预留机制的虚拟资源供给策略[J]. 系统工程理论与实践,2015,35(6):1573 -1581.(QU X L, XIAO P, BAI Q. Virtual resource provision policy based on elastic reservation mechanism in cloud environment [J]. System engineering theory and practice, 2015, 35(6): 1573 -1581.)

[3] HOUSEHOLDER R, ARNOLD S, GREEN R. Simulating the effects of cloud-based oversubscription on datacenter revenues and performance in single and multi-class service levels [C]// Proceedings of the 2014 IEEE 7th International Conference on Cloud Computing (CLOUD). Piscataway, NJ: IEEE, 2014: 562 -569.

[4] 朱金福,刘玮,高强. 航空客运舱位控制和超售综合静态建模研究[J]. 中国管理科学,2006,14(5):68 -72.(ZHU J F, LIU W, GAO Q. Integrated static modeling of airline seat inventory control and overbooking [J]. Chinese journal of management science, 2006, 14(5): 68 -72.)

[5] 赵昊天,贾传亮,宋砚秋,等. 不确定需求下航空超售问题的鲁棒优化研究[J]. 中国管理科学,2013,21(11):98 -101.(ZHAO H T, JIA C L, SONG Y Q, et al. Research on uncertain demands of airline overbooking model based on adjustable robust optimization [J]. Chinese journal of management science, 2013, 21(11): 98 -101.)

[6] 周蔷,刘长有. 基于随机特性的航空机票动态超售模型[J]. 系统工程理论与实践,2014,34(3):717 -722.(ZHOU Q, LIU C Y. Based on the stochastic characteristics of the airplane ticket dynamic overbooking model [J]. System engineering—theory & practice, 2014, 34(3): 717 -722.)

[7] 徐丽萍,李金林,雷俊丽,等. 基于超订的民航收益管理单航段舱位控制模型比较研究[J]. 系统工程理论实践,2014,34(1):129 -137.(XU L P, LI J L, LEI J L, et al. Comparative analysis on one-leg airline capacity control models with overbooking [J]. System engineering—theory & practice, 2014, 34(1): 129 -137.)

[8] KIM S, KIM H, LEE J, et al. Group-based memory oversubscription for virtualized clouds [J]. Journal of parallel and distributed computing, 2014, 74(4): 2241 -2256.

[9] RACHEL H, SCOTT A, ROBERT G. Simulating the effects of cloud-based oversubscription on datacenter revenues and performance in single and multi-class service levels [C]// Proceedings of the 2014 IEEE International Conference on Cloud Computing. Piscataway, NJ: IEEE, 2014: 562 -596.

[10] PÜSCHEL T, SCHRYEN G, HRISTOVA D, et al. Revenue management for cloud computing providers: decision models for service admission control under non-probabilistic uncertainty [J]. European journal of operational research, 2015, 244(2): 637 -647..

[11] Amazon EC2 instance [EB/OL]. [2014-02-05]. http://aws.amazon.com/cn/ec2/instance-types/.

[12] KHEMKA B, FRIESE R, PASRICHA S, et al. Utility maximizing dynamic resource management in an oversubscribed energy-constrained heterogeneous computing system [J]. Sustainable computing: informatics and systems, 2014, 5: 14 -30.

[13] MACÍAS M, GUITART J. SLA negotiation and enforcement policies for revenue maximization and client classification in cloud providers [J]. Future generation computer systems, 2014, 41: 19 -31.

Background

This work is partially supported by the National Natural Science Foundation of China (71172043), the Scientific Research Staring Foundation for the Returned Overseas Chinese Scholars, Ministry of Education of China (2013-693), Wuhan University of Technology Graduate Student Free Exploration Project (2015-zy-126).

CHEN Donglin, born in 1970, Ph. D., professor. His research interests include cloud computing, business intelligence, Internet economy.

YAO Mengdi, born in 1991, Ph. D. candidate. Her research interests include cloud computing, intelligent recommendation.

DENG Guohua, born in 1984, Ph. D. candidate. His research interests include cloud computing, intelligent recommendation, data analysis.

(上接第 100 页)

[8] PEDRO M, BRUNO A, RAQUEL M, et al. Self-organised middleware architecture for the Internet-of-things [C]// Proceedings of the 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber Physical and Social Computing. Washington, D. C.: IEEE Computer Society, 2013: 445 -451.

[9] 邵华钢,程劲,王辉,等. 面向物联网的系统及其中间件设计[J]. 计算机工程,2010,36(9):84 -86.(SHAO H G, CHENG J, WANG H, et al. Design of Internet of things-oriented system and its middleware [J]. Computer engineering, 2010,36(9): 84 -86.)

[10] WU J, LIU H. The study of configuration software and management information systems integration [C]// Proceedings of the 2010 International Conference on Computer Design and Applications. Washington, D. C.: IEEE Computer Society, 2010: 144 -147.

[11] 严童,谢吉华,温立超,等. 智能变电站 TCP/IP 通信网络的安全解决方案[J]. 电气自动化,2013,35(5):44 -45.(YAN T, XIE J H, WEN L C, et al. The security solutions for TCP/IP communication network of smart substation [J]. Electrical automation, 2013, 35(5): 44 -45.)

Background

This work is partially supported by the Surface Program of National Natural Science Foundation of China (41471329).

LIU Xueduo, born in 1991, M. S. candidate. His research interests include Internet of things.

JIAO Donglai, born in 1977, Ph. D., associate professor. His research interests include geographic information system, Internet of things, spatial information visualization.

JI Feng, born in 1990, M. S. candidate. His research interests include Internet of things.

YANG Hao, born in 1969, Ph. D., associate professor. His research interests include Internet of things, machine vision.