批量到达下的IaaS云计算中心服务性能评价

何怀文[1,2]，傅 瑜[1]

(1. 电子科技大学中山学院计算机学院 广东 中山 528402; 2. 中山大学信息科学与技术学院 广州 510275)

【摘要】针对请求批量到达下基础设施即服务(IaaS)云计算中心性能分析问题，提出基于排队系统的云计算中心分析模型，并获得平稳状态时重要的服务性能参数：阻塞概率、立即服务概率、响应时间百分比、平均队长等。通过数值仿真实验分析了缓冲区和批量大小变化对系统性能的影响。数值仿真结果表明：同等排队强度下，缓冲区的增加对批量到达系统性能的改善优于单个到达系统；每批到达请求数的突发度越大，系统性能越差。

关 键 词 批量到达； 云计算中心； 性能评价； 响应时间

中图分类号 TP393.02 **文献标志码** A doi:10.3969/j.issn.1001-0548.2015.03.022

Service Performance Evaluation of IaaS Cloud Computing Center Under Batch Arrivals

HE Huai-wen [1,2] and FU Yu[1]

(1.School of Computer, Zhongshan Institute, University of Electronic Science and Technology of China Zhongshan Guangdong 528402;
2. School of Information Science and Technology, Sun Yat-sen University Guangzhou 510275)

Abstract An analytical model based on queue system is proposed to deal with performance analysis of cloud center under batch arrivals. Some important performance indicators are acquired at steady status; these indicators include blocking probability, instance service probability, percentile response time, average queue length, and so on. The system performance influenced by changing buffer size and batch arrivals size is analyzed through numerical simulation. The numerical simulation results indicate that when buffer size is increased, the system performance with batch arrivals is better than single arrival under the same queuing intensity, and the system performance decreases as the burstiness of the number of every batch arrivals increase.

Key words batch arrivals; cloud computing center; performance evaluation; response time

云计算通过Internet为用户提供各种弹性计算资源，主要包括基础设施即服务(infrastructure as a service, IaaS)、平台即服务(platform as a service, PaaS)和软件即服务(software as a service, SasS)等不同层次的服务[1]。其中IaaS提供在云计算中心虚拟机(virtual machine, VM)实例化部署的服务，如Amazon EC2[2]、IBM Cloud[3]、GoGrid[4]等。云计算中心需要评估系统的性能和用户的需求，以期以最小的资源成本来保证用户的服务质量(quality of service, QoS)[5]。但是由于云计算服务环境的动态性、用户需求的多样性，给精确评价云计算中心的性能带来了较大的困难[6]。

由于云计算中心的扩展性和可变性，基于仿真和测量的传统性能评价方法并不适用于大规模的云计算中心[7]。文献[8]提出基于随机回报网(stochastic reward nets, SRNs)的IaaS云计算中心分析模型，分析了云计算中心的利用率、可用性、等待时间和响应度等性能指标；文献[9]在考虑节点和链路故障恢复情况下分析了云计算的服务响应时间；文献[10]则提出了基于$M/G/m/m+r$排队系统的云计算中心性能分析模型；文献[7]针对IaaS端到端性能分析，提出一种交互式的随机模型分析方法，分析了服务的可用性和响应延时等关键的QoS参数。但是大部分研究都假设用户请求为单个到达的泊松流，而在IaaS云计算中，请求往往是批量到达(如用户需要部署多个VM实例)。文献[11-12]基于$M^{(x)}/G/m/m+r$模型进一步分析了批量到达和完全拒绝策略下云计算服务响应时间、阻塞概率和理解服务概率和批量大小之间的关系，但是$M^{(x)}/G/m$模型只能获取性能参数的近似解，无法获得精确数学表示式。

收稿日期：2013－09－17；修回日期：2014－03－20
基金项目：国家自然科学基金(61300095)；广东省自然科学基金(S2012010010508)；中山市科技计划项目(2014A2FC396, 2013A3FC0285)
作者简介：何怀文(1980－)，男，博士生，主要从事云计算、资源分配调度及绿色计算方面的研究.

本文从IaaS用户请求的特性和流量特征出发，提出基于$M^{(x)}/M/n/n+r$模型的云计算中心分析模型，首次分析了IaaS云计算中心常见性能参数的精确表达式，同时分析了云计算中心中重要QoS参数——响应时间百分比和服务台数量以及服务速率之间的关系。本文从阻塞概率、立即服务概率、等待队长、响应时间等多个角度讨论了批量到达IaaS云计算中心的性能，通过大量数值仿真实验，同时比较了不同缓冲区大小以及批量到达数在不同分布下系统性能的变化，并分析批量到达大小对系统性能的影响。研究数据可以为云计算中心进行合理的资源配置提供理论依据和参考数据。

1 批量到达下IaaS云计算中心模型

1.1 模型描述

批量到达下IaaS云计算中心模型为$M^{(x)}/M/n/n+r$排队模型，符合以下假设条件：1) 请求批量到达，每批到达请求数X可以服从任意概率分布，$P(X=i)=a_i$，$E(X)=\overline{a}$；2) 每批到达的时间间隔符合参数为λ的负指数分布；3) 服务台数量为n，请求所需服务时间服从参数为μ的负指数分布，系统采用先到先服务排队规则(first come first server, FCFS)；4) 系统缓冲区大小为r。同批次到达的请求是可分割的，当前批次请求无法全部进入缓冲区时，则允许部分请求进入缓冲区，直至缓冲区占满，剩余请求被阻塞。

1.2 模型分析

$M^{(x)}/M/n/n+r$排队模型属于马尔科夫过程，假设队长为k的概率为π_k，系统容量$N=n+r$，流量强度$\rho=\lambda\overline{a}/n\mu$。已知当$\rho<1$时，系统存在平稳状态[13]，平稳状时概率转移如图1所示。

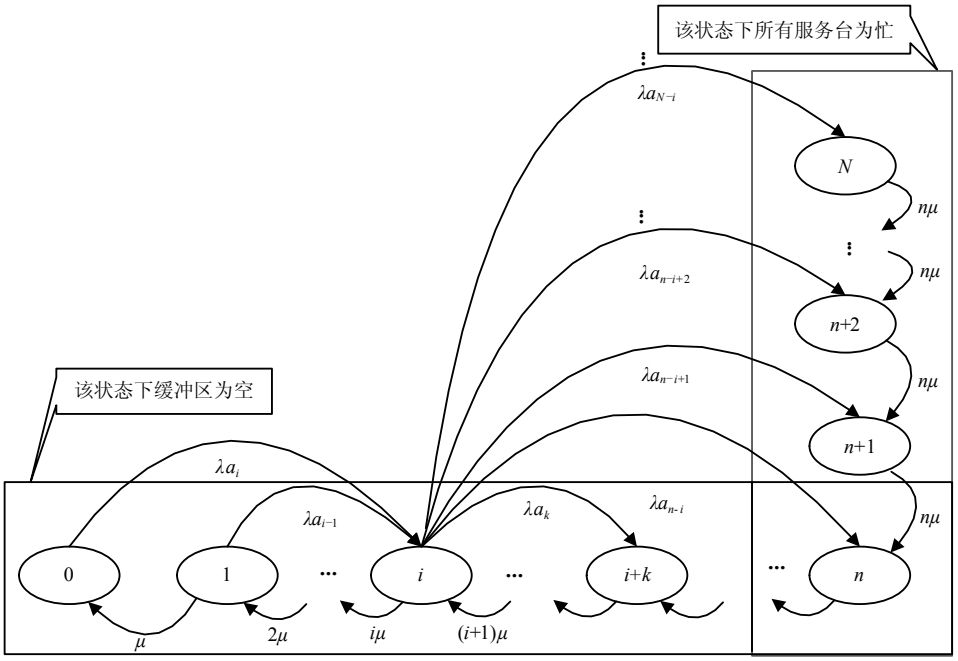


图1 系统稳定状态概率转移图

根据Chapman-Kolmogorov方程，可以得到：

$$
\begin{aligned}
&-\lambda\pi_0+\mu\pi_1=0\\
&-(\lambda+i\mu)\pi_i+(i+1)\mu\pi_{i+1}+\lambda\sum_{j=0}^{i-1}a_{i-j}\pi_j=0 \qquad 0<i<n\\
&-(\lambda+n\mu)\pi_i+n\mu\pi_{i+1}+\lambda\sum_{j=0}^{i-1}a_{i-j}\pi_j=0 \qquad n\leqslant i<N\\
&-\lambda\left(\pi_0\sum_{j=N}^{\infty}a_j+\pi_1\sum_{j=N-1}^{\infty}a_j+\cdots+\pi_i\sum_{j=N-i}^{\infty}a_j+\cdots+\pi_{N-1}\sum_{j=1}^{\infty}a_j\sum_{j=1}^{\infty}a_j\right)+n\mu\pi_N=0
\end{aligned}
\tag{1}
$$

由式(1)可知，π_{i+1}可以由π_i递推得到，即可获得$\pi_1,\pi_2,\cdots,\pi_N$和π_0之间的关系，由$\sum_{i=0}^{N}\pi_i=1$可获得

所有队长的分布概率。例如在$M^{(4)}/M/3/3+2$模型中，每批到达请求数为4，流量强度ρ=0.9，由式(1)可得队长分布概率为：π_0=0.256 8，π_1=0.173 4，π_2=0.145 2，π_3=0.129 5，π_4=0.158 6，π_5=0.136 5。

2　IaaS云计算中心性能分析

2.1　阻塞概率

阻塞概率指请求无法进入缓冲区的概率，是衡量系统可用性的重要指标。由于系统采用部分接收策略，当该批次的请求无法全部进入系统时，可能存在两种不同结果：进入缓冲区排队或者被拒绝。考虑某个单个请求R，假设请求R在该批次中的位置符合均匀分布，则R属于大小为j批次的概率为[14]：

$$\Pr\{\text{任意请求}R\text{属于批量大小为}j\text{的概率}\}=\frac{j\times a_j}{\overline{a}}$$

假设系统当前队长为i，批量到达请求数为j。当$j>N-i$时，则该批次中位于$N-i$之后的请求将无法进入缓冲区，即该批次中将有$\dfrac{j-N+i}{j}$个请求被阻塞。因此，阻塞概率为：

$$\begin{aligned}P_{\text{block}}=&\pi_0\sum_{j=N+1}^{\infty}\frac{ja_j}{\overline{a}}\times\frac{j-N}{j}+\cdots+\\&\pi_i\sum_{j=N-i+1}^{\infty}\frac{ja_j}{\overline{a}}\times\frac{j-N+i}{j}+\cdots+\\&\pi_N\sum_{j=1}^{\infty}\frac{ja_j}{\overline{a}}\times\frac{j}{j}=\\&\frac{1}{\overline{a}}\sum_{i=0}^{N}\pi_i\sum_{j=N-i+1}^{\infty}(j-N+i)a_j\end{aligned}\tag{2}$$

2.2　立即服务概率

如果请求到达时存在空闲的服务器，则请求可能无须等待，立即获得服务。假设系统队长为$i(i<n)$，批量到达数为j，当$j\leqslant n-i$时，所有请求均获得立即服务；当$j>n-i$时，位置在$[1,j-n+i]$的请求可获得立即服务。因此，立即服务概率为：

$$\begin{aligned}P_s=&\sum_{i=0}^{n-1}\pi_i\left(\sum_{j=1}^{n-i}\frac{ja_j}{\overline{a}}+\sum_{j=n-i+1}^{\infty}\frac{ja_j}{\overline{a}}\times\frac{j-n+i}{j}\right)=\\&\frac{1}{\overline{a}}\sum_{i=0}^{n-1}\pi_i\left(\sum_{j=1}^{n-i}ja_j+\sum_{j=n-i+1}^{\infty}a_j\times(j-n+i)\right)\end{aligned}\tag{3}$$

2.3　响应时间分布函数和响应时间百分比

平均响应时间是传统系统性能重要参数之一。但在云计算服务中，用户更多的是关注响应时间百分比。响应时间百分比是指响应时间在指定时间段中的概率分布百分比，是云计算中用户关注的一个重要QoS指标[5]。响应时间比的定义如下：

$$F_T(t)=\int_0^{\Delta T}f_T(t)dt\geqslant\gamma\%\tag{4}$$

式中，$f(t)$是响应时间t的概率密度函数；ΔT表示用户使用服务的时间段；$\gamma\%$为响应时间百分比。式(4)表示在响应时间小于ΔT时间段的概率不能小于$\gamma\%$。

响应时间由等待时间和服务时间两部分组成。假设请求到达时系统队长为i，请求数量为j，$W_q(t)$为等待时间t的概率分布函数，则等待时间为0的概率为：

$$W_q(t=0)=\frac{P_s}{P\{\text{请求能进入系统的概率}\}}=\frac{P_s}{1-P_{\text{block}}}\tag{5}$$

假设请求R在同批次中的位置为k，$k\in[1,j]$。如果R进入系统时无法获得立即服务，则需要排队等待前面的请求完成离开后才能获得服务。记R需要等待离开的请求数为l，因为每个请求所需的服务时间为参数为μ的负指数分布，服务台数量为n，l个请求的离去流则符合$n\mu$的l阶的Erlang分布。离开请求数l可以分为下面两种情况。

1) 当$i<n$，即请求到达时存在空闲的服务台。当$j\in(n-i,N-i]$时，而$k\in(n-i,j]$时，R会排队等待。当$j\in(N-i,\infty]$时，而$k\in(n-i,N-i]$时亦会等待。有l=k–n+i，等待时间的分布函数$W_q^1(t)$为：

$$\begin{aligned}&W_q^1(t)=P\{W_q^1\leqslant t\}=\\&\frac{1}{1-P_{\text{block}}}\left(\sum_{i=0}^{n-1}\pi_i\left(\sum_{j=n-i}^{N-i}\sum_{k=n-i}^{j}(1-\sum_{l=0}^{k-n+i}\frac{1}{n!}e^{-n\mu t}(n\mu t)^l)+\right.\right.\\&\left.\left.\sum_{j=N-i+1}^{\infty}\sum_{k=n-i}^{N-i}(1-\sum_{l=0}^{k-n+i}\frac{1}{n!}e^{-n\mu t}(n\mu t)^l)\right)\right)\end{aligned}\tag{6}$$

2) 当$i\geqslant n$，即请求到达时不存在空闲的服务台。当$j\in[1,N-i]$时，而$k\in[1,j]$时，请求会排队等待。当$j\in(N-i,\infty]$时，而$k\in[1,N-i]$时亦会等待。有l=k+i–n，等待时间的分布函数$W_q^2(t)$为：

$$\begin{aligned}&W_q^2(t)=P\{W_q^2\leqslant t\}=\\&\frac{1}{1-P_{\text{block}}}\left(\sum_{i=n}^{N-1}\pi_i\left(\sum_{j=1}^{N-i}\sum_{k=1}^{j}(1-\sum_{l=0}^{k-n+i}\frac{1}{n!}e^{-n\mu t}(n\mu t)^l)+\right.\right.\\&\left.\left.\sum_{j=N-i}^{\infty}\sum_{k=1}^{N-i}(1-\sum_{l=0}^{k-n+i}\frac{1}{n!}e^{-n\mu t}(n\mu t)^l)\right)\right)\end{aligned}\tag{7}$$

因此，等待时间$W_q(t)$为：

$$W_q(t)=W(t=0)+W_q^1(t)+W_q^2(t)=$$
$$\frac{1}{1-P_{\text{block}}}\left(P_s+\sum_{i=0}^{n-1}\pi_i\left(\sum_{j=n-i}^{N-i}\sum_{k=n-i}^{j}(1-\sum_{l=0}^{k-n+i}\frac{1}{n!}e^{-n\mu t}(n\mu t)^l)+\right.\right.$$
$$\left.\sum_{j=N-i+1}^{\infty}\sum_{k=n-i}^{N-i}\left(1-\sum_{l=0}^{k-n+i}\frac{1}{n!}e^{-n\mu t}(n\mu t)^l\right)\right)+$$
$$\sum_{i=n}^{N-1}\pi_i\left(\sum_{j=1}^{N-i}\sum_{k=1}^{j}(1-\sum_{l=0}^{k-n+i}\frac{1}{n!}e^{-n\mu t}(n\mu t)^l)+\right.$$
$$\left.\left.\sum_{j=N-i}^{\infty}\sum_{k=1}^{N-i}\left(1-\sum_{l=0}^{k-n+i}\frac{1}{n!}e^{-n\mu t}(n\mu t)^l\right)\right)\right) \tag{8}$$

响应时间 $W=W_q+\chi$，其中 χ 为请求所需的服务时间。而 W_q 和 χ 相互独立。根据卷积公式可以得到响应时间的分布函数为：

$$W(t)=P\{W\leqslant t\}=\int_0^t P\{W_q\leqslant$$
$$t-x\}\mathrm{d}P\{\chi\leqslant x\}=\int_0^t W_q(t)\mu e^{-\mu t}\mathrm{d}x \tag{9}$$

时间段 ΔT 的响应时间百分比可以通过 $W(t\leqslant\Delta T)\geqslant\gamma\%$ 计算得到，通过式(9)可以获得响应时间百分比和服务台数量以及服务速率之间的关系。

2.4 其他指标

系统其他的重要性能指标如下：1) 平均队长为 $\bar{N}=\sum_{i=0}^{N}i\pi_i$；2) 平均等待队长为 $\bar{N}_q=\sum_{i=n+1}^{N}(i-n)\pi_i$。根据Little公式，可以得到平均等待时间为 $\bar{W}_q=\bar{N}_q/\lambda_e$，平均响应时间为 $\bar{W}=\bar{N}/\lambda_e$。其中 λ_e 为有效到达速率为 $\lambda_e=\lambda\times\bar{a}\times(1-P_{\text{block}})$。

3 数值仿真与结果分析

本文使用离散事件仿真软件Arena[15]对批量到达的云计算中心模型在进行模拟仿真，分别对缓冲区大小和到达批量大小的变化对系统性能的影响进行了实验分析。

1) 缓冲区大小对性能的影响

通过改变缓冲区的大小可以在一定程度上改善系统性能。本文实验假设云计算中心服务台n=200，流量强度ρ=0.85，服务速率μ=0.2，缓存大小r从0每次增加10递增到100。分别考察以下不同情形系统性能的变化：① 单个到达；② 批量大小X服从几何分布 $a_i=\theta^{i-1}(1-\theta)$，$\bar{a}=10$；③ 批量大小$X$服从泊松分布 $a_i=\frac{\lambda^i}{i!}e^{-\lambda}$，$\bar{a}=10$，结果如图2所示。

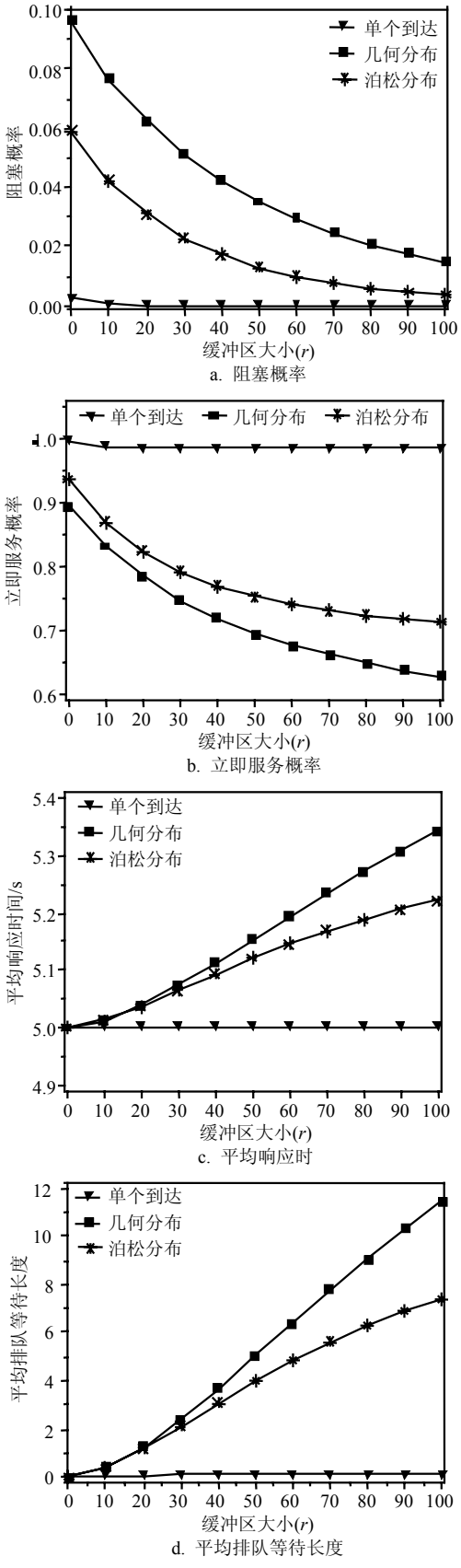


图2 缓冲区大小变化对云计算中心性能的影响

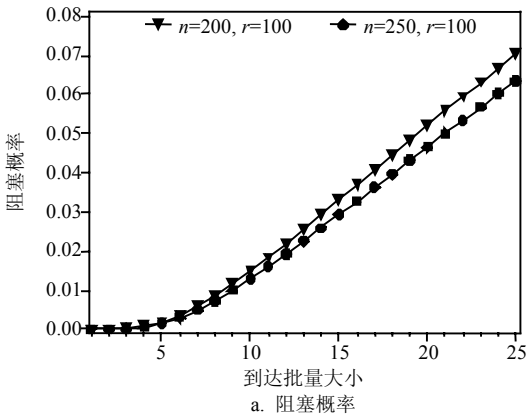
从图2可见，随着缓冲区的增加，单个到达系统的阻塞概率、立即服务概率、平均响应时间和平均

排队长度变化非常平缓，而批量到达系统的各项性能参数均有明显变化，说明在相同的排队强度下，缓冲区的增加对批量到达系统性能的改善明显优于单个到达系统。随着缓冲区的增大，批量到达系统的阻塞概率下降较快，但是由于进入系统的请求数增加，立即服务概率、平均响应时间和平均排队长度均随之增高，意味着系统吞吐量提高。而当批量大小符合泊松分布时，系统的各项性能指标要优于批量大小符合几何分布的情况。由上述分析可知，在批量到达系统中，通过增加缓冲区将能带来更好的性能提升。

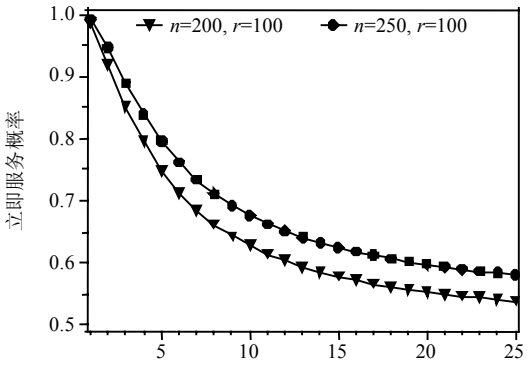
2) 批量到达数对性能的影响

本文在两种不同规模的云计算中心环境考察批量到达数对系统的影响。云计算中心1的服务台n=200；云计算中心2的服务台n=250。缓冲区的大小r=100，流量强度ρ=0.85，服务速率μ=0.2，批量到达请求数假设符合最常见的几何分布，$\bar{a}$ 从0递增到25，实验结果如图3所示。

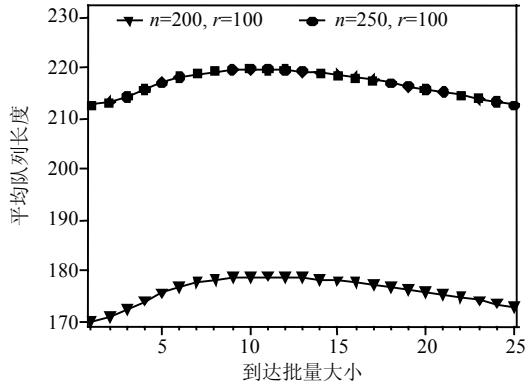
在图3a中，阻塞概率随到达批量的大小增加而增加，在$\bar{a}>5$后几乎呈线性递增；在图3b中，立即服务概率则随$\bar{a}$的增加而下降，在$\bar{a}=10$的前后均呈现线性关系；在图3c中，平均队列长度开始随着$\bar{a}$的增加而增加，在$\bar{a}=12$之后便逐渐下降；在图3d～图3f中，平均等待队长、平均等待时间和平均响应时间均随着$\bar{a}$的增加明显增大，说明当批量到达数量的的增多，系统性能的会急剧下降。同时由图3可见，云计算中心2的阻塞概率、平均等待时间、平均响应时间均小于云计算中心1，而立即服务概率、平均队列长度、平均等待长度要高于云计算中心1，说明云计算中心2系统性能明显优于云计算中心1，即服务器数量的增加有利于性能的改善，与现实情况相符。因此，在面对突发量大的请求时，系统需要通过增加服务器数量或提高服务速率方能保证用户的QoS。



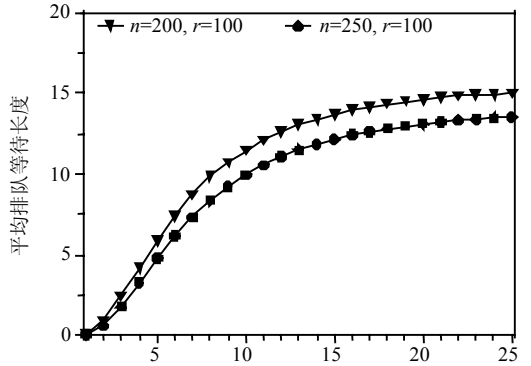
a. 阻塞概率



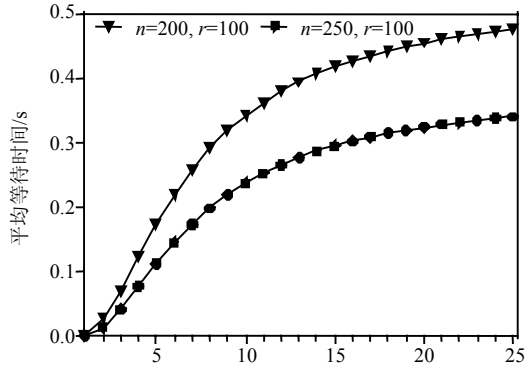
b. 立即服务概率



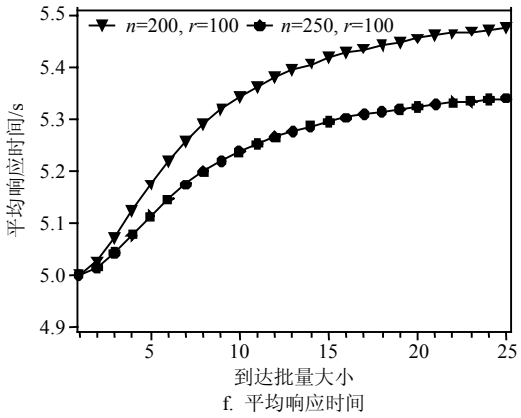
c. 平均队列长度



d. 平均排队等待长度



e. 平均等待时间



f. 平均响应时间

图3 到达批量大小变化对云计算中心性能的影响

4 结 束 语

论文利用排队模型对批量到达下云计算中心IaaS服务性能进行了分析，获取了云计算中心重要的QoS参数如阻塞概率、立即服务概率和响应时间百分比的表达式，并通过数值仿真实验对结果进行了验证，获取了各项性能参数和批量到达数之间的关系。结果表明缓冲区容量的增加，对批量到达系统下性能的改进要优于单个到达系统；同时，随着批量到达数的增加，系统各项性能会急剧下降，需要通过增加系统资源配置才能保证云计算QoS服务质量。实验数据和结论将对云计算中心运营商为了保证QoS，优化资源配置和避免配置过载提供有用的参考依据。

参 考 文 献

[1] ARMBRUST M, FOX A, GRIFFITH R, et al. A view of cloud computing[J]. Communications of the ACM, 2010, 53(4): 50-58.

[2] Amazon Web Services. Amazon EC2[EB/OL]. [2013-12-03]. http://aws.amazon.com/ec2.

[3] KOCHUT A, DENG Y, HEAD M R, et al. Evolution of the IBM cloud: Enabling an enterprise cloud services ecosystem[J]. IBM Journal of Research and Development, 2011, 55(6): 1-7.

[4] LI Xin-fu, GONDI C. Cloud Computing hosting[C]// 2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT). [S.l.]: IEEE: 2010: 194-198.

[5] XIONG K, PERROS H. Service performance and analysis in cloud computing[C]//2009 World Conference on Services-I. [S.l.]: IEEE: 2009: 693-700.

[6] XIONG K, PERROS H. SLA-based resource allocation in cluster computing systems[C]//IEEE International Symposium on Parallel and Distributed Processing. [S.l.]: IEEE: 2008: 1-12.

[7] GHOSH R, TRIVEDI K S, NAIK V K, et al. End-to-end performability analysis for infrastructure-as-a-service cloud: an interactingstochastic models approach[C]//2010 IEEE 16th Pacific Rim International Symposium on Dependable Computing (PRDC). [S.l.]: IEEE, 2010: 125-132.

[8] BRUNEO D. A stochastic model to investigate data center performance and QoS in IaaS cloud computing systems[J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(3): 560-569.

[9] YANG B,TAN F, DAI Y-S. Performance evaluation of cloud service considering fault recovery[J]. The Journal of Supercomputing, 2013, 65(1): 426-444.

[10] KHAZAEI H, MISIC J, MISIC V B. Performance analysis of cloud computing centers using *M*/*G*/*m*/*m*+*r* queuing systems[J]. IEEE Transactions on Parallel and Distributed Systems, 2012, 23(5): 936-943.

[11] KHAZAEI H, MISIC J, MISIC V B. Performance analysis of cloud centers under burst arrivals and total rejection policy[C]//Global Telecommunications Conference. [S.l.]: IEEE, 2011: 1-6.

[12] KHAZAEI H, MISIC J, MISIC V B. Performance of cloud centers with high degree of virtualization under batch task arrivals[J]. IEEE Transactions on Parallel and Distributed Systems, 2013, 24(12): 2429-2438.

[13]唐应辉, 唐小我. 排队论——基础与分析技术[M]. 北京: 科学出版社, 2006.
TANG Ying-hui, TANG Xiao-wo. Queuing theory: Basic and analysis[M]. Beijing: Science Press, 2006.

[14] BURKE P. Technical note: Delays in single-server queues with batch input[J]. Operations Research, 1975, 23(4): 830-833.

[15] ALTIOK T, MELAMED B. Simulation modeling and analysis with Arena[M]. [S.l.]: Academic Press, 2010.

编 辑 蒋 晓