DOI:10.11992/tis.201703042
网络出版地址:http://kns.cnki.net/kcms/detail/23.1538.TP.20170702.1547.036.html

具有两类请求的云计算中心服务器数量的优化

张江强,赵宁,刘文奇
(昆明理工大学 理学院,云南 昆明　650500)

摘　要:为提高云计算中心的服务质量,节约系统成本,针对具有两类用户请求的云计算中心,提出云计算中心的服务器数量的优化方案。首先,建立了具有两类用户请求的排队模型,分析系统的稳态概率分布、平均队长等性能指标;然后,建立了云计算中心的能耗模型;最后,联合系统的等待成本和能耗成本,构建系统的成本函数,对系统的服务器数量进行优化,从而使系统的成本最小。数值分析结果表明最优服务器数量是用户请求到达率的非减函数,为了使系统成本最小,云计算中心需要动态调整服务器的数量。

关键词:云计算;排队系统;两类请求;性能指标;能耗;成本;服务器数量的优化
中图分类号:TP393.02　**文献标志码**:A　**文章编号**:1673-4785(2017)05-0601-07

中文引用格式:张江强,赵宁,刘文奇.具有两类请求的云计算中心服务器数量的优化[J]. 智能系统学报, 2017, 12(5): 601-607.
英文引用格式:ZHANG Jiangqiang, ZHAO Ning, LIU Wenqi. Optimization of the number of servers in a cloud computation center with two demand classes[J]. CAAI transactions on intelligent systems, 2017, 12(5): 601-607.

Optimization of the number of servers in a cloud computation center with two demand classes

ZHANG Jiangqiang, ZHAO Ning, LIU Wenqi
(Faculty of Science, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: In order to improve the service quality and to save the system cost of the cloud computing center, for a cloud computing center with two demand classes, a method to optimize the number of servers was proposed. First, a queuing model having two demand classes was established for analyzing performance measures such as distribution of the probability of stability and mean queue length; next, a power consumption model was established on the cloud computing center; finally, the wait and power-consumption cost of the system were used together to construct the cost function of the system and optimize the server quantity for realizing the lowest cost. The numerical results show that the optimal number of servers is a non-decreasing function of the arrival rate of demands. To minimize the system cost, dynamically adjusting the number of servers is necessary.

Keywords: cloud computing; queuing system; two demand classes; performance measure; power consumption; cost; optimization of the number of servers

云计算中心是基于超级计算机系统对外提供计算资源、存储资源等服务的机构或单位,以高性能计算机为基础面向各界提供高性能计算服务。当前,云计算中心主要面向大规模科学计算及工程计算应用,并在商业计算、互联网、电子政务、电子商务等领域拥有巨大发展潜力。对云计算中心的性能和能耗进行全面地分析具有十分重要的意义。

云计算中心作为服务机构,系统的性能指标(用户请求的等待时间、系统的堵塞程度等)是刻画服务质量的重要因素。将用户请求看作顾客,云计算中心的超级计算机的处理器作为服务器,而用户请求的处理过程作为服务过程,云计算中心是一个典型的排队系统。很多学者对云计算中心的性能及调度策略方面展开研究。廖倩文等[1]提出一种基于排队论的批量到达的云计算中心性能分析模型,得到系统中用户请求队长的稳态概率分布、系统的阻塞概率、立即服务概率等指标。徐小龙等[2]

收稿日期:2017-03-27.　网络出版日期:2017-07-02.
基金项目:国家自然科学基金项目(71501086,61573173).
通信作者:赵宁. E-mail:zhaoning@kmust.edu.cn.

研究了云计算系统任务调度和数据部署层面的节能机制，提出一种面向绿色云计算中心的动态数据聚集算法。许丞[3]建议将 Hadoop 云平台的任务监控和任务调度管理功能分离，从而提升云平台的工作效率。倪志伟[4]综合考虑了用户最短等待时间资源负载均衡和经济原则，提出一种离散人工蜂群算法的云任务调度优化策略。

在云计算中心，能耗开销是不容忽视的问题，著名 IT 企业如 Google、Microsoft、Amazon 等云计算中心每年能耗超过百万美元，给云计算中心长期运营带来了巨大经济负担。云计算中心的能耗问题最近得到学者的广泛关注。罗亮等[5]从处理器性能计数器和系统使用情况入手，结合多元线性回归和非线性回归的数学方法，分析不同参数和方法对服务器能耗建模的影响，并提出适合云计算中心基础架构的服务器能耗模型。现有内存能耗模型研究发现，影响内存能耗的主要因素是内存读写的吞吐量[6]。何怀文等[7]在平均响应时间受限的条件下提出云计算中心异构服务器之间的优化能耗分配方法。针对云计算中心由于服务器空闲而产生大量空闲能耗，以及由于任务调度不匹配而产生大量“奢侈”能耗的问题，文献[8-10]提出通过任务调度的方式优化管理。文献[11]研究了云计算中心的动态迁移问题。文献[12]以利润最大化为目标，分析了云计算中心的优化配置。针对多个服务器切换过程存在大量冗余信号的问题，文献[13]提出了一种改进的多业务切换机制。文献[14]运用遗传算法分析用户请求的调度策略，从而提高云计算中心能源利用率。文献[15-16]对云计算中心的能耗和性能进行了联合优化。

以上关于云计算中心的相关研究都假设系统只有一类用户请求，但实际应用中，云计算中心根据用户请求的重要程度分为不同等级[17]。例如，云计算中心将实时用户请求赋予高优先权，将非实时用户请求赋予低优先权。另外，为了吸引更多客户，云计算中心为用户提供免费体验的服务，而付费的用户相对免费用户享有高优先权。Liu 等[18]运用博弈论的方法研究具有多类用户请求的云计算中心的预约服务策略。

本文将研究具有两类用户请求的云计算中心能耗和性能的联合优化问题。假设两类用户请求的到达过程均为泊松过程，系统有多个平行的处理器，每个用户请求的处理时间服从指数分布，系统最多容纳有限的用户请求。我们将该系统构建为一个带非抢占优先权的马尔可夫过程，基于排队论对该系统的性能进行分析。将系统的能耗表示为处理器吞吐量和处理器个数的函数。最后，结合系统的性能和能耗构建系统的成本函数，对系统处理器的个数进行优化。

1 系统描述

假设云计算中心有 c 个平行的服务器，按照用户的优先级别，将用户请求分为两类，第 i 类用户请求到达过程是参数为 λ_i 的泊松过程，$i=1,2$。系统最多容纳 N 个用户请求，当系统中的用户请求个数小于 N，用户请求的到达率为 $\lambda=\lambda_1+\lambda_2$，否则到达率 $\lambda=0$。用户请求到达云计算中心后，如果系统中有空闲的服务器，则用户请求直接进入空闲服务器接受服务，反之则需要在缓冲区中排队等待接受服务。第 1 类用户请求相对第 2 类用户请求具有非抢占优先权，即系统中的第 1 类用户请求被优先服务，但是第 1 类用户请求不能打断第 2 类用户请求的服务，对于每 1 类用户请求，系统按照先到先服务的规则(FCFS)进行服务。第 i 类用户请求的服务时间服从参数为 μ_i 的指数分布，$i=1,2$。

2 系统性能分析

具有两类用户请求的云计算中心是一个非抢占优先权的 $M_1,M_2/M_1,M_2/(c/N)$ 排队模型，如图 1 所示。

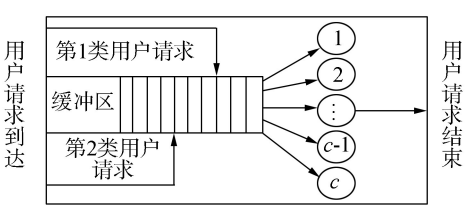


图 1 具有两类用户请求的排队系统

Fig.1 Queueing system with two demand classes

令(l_1,l_2)表示系统的状态，其中 l_i 表示系统中第 i 类用户请求的个数，该系统的状态空间为

$$E=\{(l_1,l_2),l_1,l_2=0,1,\cdots,N,l_1+l_2=N\}。$$

系统被第 i 类用户请求占用系统的概率为 $\rho_i=\dfrac{\lambda_i}{c\mu_i}$，$(i=1,2)$，令 $\rho=\rho_1+\rho_2$。当 $\rho<1$ 时，该系统存在稳态概率分布。令 $l=l_1+l_2$ 为系统的水平(即系统中请求总数)，对系统的状态按照水平划分，并且每个水平下的状态按照如下规则排序。

$l=0$ 的状态：$(0,0)$。

$l=1$ 的状态：$(1,0),(0,1)$。

$l=2$ 的状态：$(2,0),(1,1),(0,2)$。

⋮

$l=N$ 的状态：$(N,0),(N-1,1),(N-2,2),\cdots,(2,N-2),(1,N-1),(0,N)$。

任意水平 l 可能存在的状态转移为 $l\to l-1$,$l\to l$,$l\to l+1$。状态转移如图 2 所示。按照水平的顺序对所有状态排序,M_1,M_2/M_1,$M_2/(c/N)$ 排队模型的 Q 矩阵可表示为

$$\boldsymbol{Q}=\begin{bmatrix} B_{00} & \boldsymbol{B}_{01} & 0 & \cdots & 0 & 0 & 0 \\ \boldsymbol{A}_{10} & \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \boldsymbol{A}_{N-1,N-2} & \boldsymbol{A}_{N-1,N-1} & \boldsymbol{A}_{N-1,N} \\ 0 & 0 & 0 & \cdots & 0 & \boldsymbol{A}_{N-1,N} & \boldsymbol{A}_{N,N} \end{bmatrix}$$

式中:$\boldsymbol{A}_{l,l-1}$ 对应水平 l 到水平 $l-1$ 的矩阵块,$\boldsymbol{A}_{l,l}$ 对应水平 l 到水平 l 的矩阵块,$\boldsymbol{A}_{l,l+1}$ 对应水平 l 到水平 $l+1$ 的矩阵块($1\leqslant l\leqslant N-1$),$B_{00}$ 对应 $l=0$ 到 $l=0$ 的值,$\boldsymbol{B}_{01}$ 对应 $l=0$ 到 $l=1$ 的矩阵块。$\boldsymbol{Q}$ 矩阵的矩阵块随着水平 l 的增大而逐渐增大。令

$$\alpha_i=\min(i,c)$$

$$\beta_{l,i}=\min(l-i,c),i=0,1,2,\cdots,l$$

$$B_{00}=-\lambda_1-\lambda_2$$

$$\boldsymbol{B}_{01}=[\lambda_1 \quad \lambda_2]$$

$$\boldsymbol{A}_{l,l-1}=\begin{bmatrix} \beta_{l,0}\mu_1 & 0 & 0 & \cdots & 0 & 0 \\ \alpha_1\mu_2 & \beta_{l,1}\mu_1 & 0 & \cdots & 0 & 0 \\ 0 & \alpha_2\mu_2 & \beta_{l,2}\mu_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha_{l-1}\mu_2 & \beta_{l,l-1}\mu_1 \\ 0 & 0 & 0 & \cdots & 0 & \alpha_l\mu_2 \end{bmatrix}_{l+1,l}$$

$$\boldsymbol{A}_{l,l+1}=\begin{bmatrix} \lambda_1 & \lambda_2 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \lambda_1 & \lambda_2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \lambda_1 & \lambda_2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \lambda_1 & \lambda_2 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & \lambda_1 & \lambda_2 \end{bmatrix}_{l+1,l+2}$$

$$\boldsymbol{A}_{l,l}=\begin{bmatrix} -\beta_{l,0}\mu_1-\lambda & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\alpha_1\mu_2-\beta_{l,1}\mu_1-\lambda & 0 & \cdots & 0 & 0 \\ 0 & 0 & -\alpha_2\mu_2-\beta_{l,2}\mu_1-\lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & 0 & 0 \\ 0 & 0 & 0 & \cdots & -\alpha_{l-1}\mu_2-\beta_{l,l-1}\mu_1-\lambda & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\alpha_l\mu_2-\lambda \end{bmatrix}_{l+1,l+1}$$

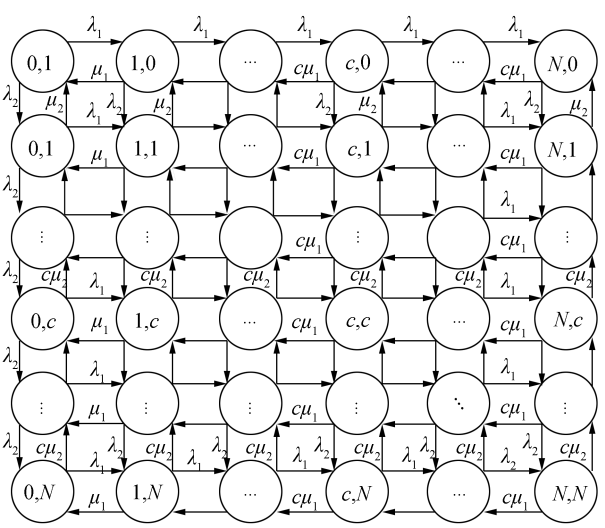


图 2　系统状态转移图

Fig.2　Transitions among the system states

令 $L_i(t)$ 表示 t 时刻系统中第 i 类用户请求的个数,$\pi_{i,j}=\lim\limits_{t\to\infty}P\{L_1(t)=i,L_2(t)=j\}$ 为系统处于状态 (i,j) 的稳态概率,令

$$\pi=(\pi_{0,0},\pi_{1,0},\pi_{0,1},\cdots,\pi_{N,0},\pi_{N-1,1},\cdots,\pi_{0,N})$$

系统的稳态概率 π 分布可以通过求解如下方程组得到:

$$\begin{cases}\pi\boldsymbol{Q}=0\\ \sum\limits_{j=0}^{N}\sum\limits_{i=0}^{N}\pi_{i,j}=1\end{cases}$$

系统中两类用户请求的平均队长分别为

$$E(L_1)=\sum_{i=1}^{N}i\sum_{j=0}^{N}\pi_{i,j}$$

$$E(L_2)=\sum_{j=1}^{N}j\sum_{i=0}^{N}\pi_{i,j}$$

根据 Little 法则,两类用户请求的平均等待时间分别为

$$E(T_1)=E(L_1)/\lambda_1$$

$$E(T_2)=E(L_2)/\lambda_2$$

系统中用户请求的平均等待时间为

$$E(T)=\frac{\lambda_1}{\lambda_1+\lambda_2}E(T_1)+\frac{\lambda_2}{\lambda_1+\lambda_2}E(T_2)$$

3　系统能耗

云计算中心服务器的能耗主要包括静态能耗和动态能耗。通常静态能耗比较稳定,假设每个服务器的静态能耗 P^* 为常数,c 个服务器的静态能耗为 $P_{静}=cP^*$。

根据文献[19],当服务器的服务速率为 μ 时,单个服务器的动态能耗为 $k\mu^{\alpha}$(单位为瓦),其中 k 为功耗比例因子,$\alpha \geqslant 3$。由于系统中服务第 i 类用户请求的平均服务器数量为 $c\rho_i(i=1,2)$,因此系统的动态能耗为

$$P_{动} = c\rho_1 k\mu_1^{\alpha} + c\rho_2 k\mu_2^{\alpha} = k\lambda_1\mu_1^{\alpha-1} + k\lambda_2\mu_2^{\alpha-1}$$

系统的总体能耗为

$$P_{总} = P_{静} + P_{动} = cP^{*} + k\lambda_1\mu_1^{\alpha-1} + k\lambda_2\mu_2^{\alpha-1}$$

4 系统成本

云计算中心作为服务系统,用户请求的等待时间反应了的系统的服务质量,较长的等待时间必然影响用户对系统的评价,从而导致系统用户的丢失。系统可以通过增加服务器的方式减少用户请求的等待时间。但是服务器的增加,必然导致系统能耗增加。下面构建系统成本,对系统服务器的数量进行优化。

系统成本包括用户的等待成本和系统能耗成本。令 h_i 表示一个第 i 类用户请求单位时间的逗留费用,则系统的等待成本为 $h_1E(L_1) + h_2E(L_2)$;令 β 为单位能耗价格,则系统的能耗成本为 $\beta P_{总}$,其中 $h_i > 0, \beta > 0, i = 1,2$。

因此,系统单位时间的成本为

$$f(c) = h_1E(L_1) + h_2E(L_2) + \beta P_{总}$$

系统最优成本可表示为如下数学规划问题:

$$\begin{cases} \min f(c) \\ \text{s.t.} \quad \dfrac{\lambda_1}{c\mu_1} + \dfrac{\lambda_2}{c\mu_2} < 1 \\ c \in \mathbf{N}^{+} \end{cases}$$

对于任意的 $c \in \mathbf{N}^{+}$,$f(c)>0$,且 $\lim\limits_{c\to+\infty} f(c) = +\infty$,因此 $f(c) \in (0, +\infty)$,在 c 的定义域内必存在最优值 c^{*} 使得 $f(c^{*}) = \min f(c)$。我们可采用边际分析法计算 c 的最优值 c^{*},即 c^{*} 满足以下两个条件:

$$\begin{cases} f(c^{*}) < f(c^{*} - 1) \\ f(c^{*}) < f(c^{*} + 1) \end{cases}$$

5 算例分析

下面针对具有两类用户请求的云计算中心的排队模型进行数值实验。假设 $h_1 = 40, h_2 = 20, \beta = 0.8, N = 100, \mu_1 = \mu_2 = 1.5$,调整参数 λ_1, λ_2 或服务器的数目 c,计算系统的性能指标,并求解系统能耗及最优的服务器的数目。

例 1 假设两类用户请求的到达率不变,分析服务器数量 c 对系统的影响。假设 $h_1 = 40, h_2 = 20, \beta = 0.8, N = 100, \mu_1 = \mu_2 = 1.5, \lambda_1 = 1, \lambda_2 = 2.5, c = 1, 2, \cdots, 7$。表 1 计算了服务器个数 c 取不同值的情况下,系统的服务强度 ρ、平均队长 $E(L_1)$ 和 $E(L_2)$、平均等待时间 $E(T)$、系统能耗 $P_{总}$ 及系统单位时间成本 $f(c)$。表 1 显示随着服务器个数 c 的增加,ρ、$E(L_1)$、$E(L_2)$、$E(T)$ 均减小,但是 $P_{总}$ 增加。随着 c 增加,系统的单位时间成本先增大后减小,在 $c = 4$ 时 $f(c)$ 取得最小值,且 $f(c^{*}) = 77.671\ 8$。实际上,对于任意的 (λ_1, λ_2),最优值 c^{*} 总是存在的,我们都可以通过表 1 的方式求解最优值 c^{*}。

表 1 云计算中心的性能、能耗及成本分析(例 1)

Table 1 The performance, power consumption and cost of cloud computing centers (example 1)

c	ρ	$E(L_1)$	$E(L_2)$	$E(T)$	$P_{总}$	$f(c)$
3	0.777 8	0.675 7	2.040 7	0.776 1	15.229 5	80.026 2
4	0.583 3	0.667 7	1.739 7	0.687 8	20.211 6	77.671 8
5	0.466 7	0.666 7	1.681 6	0.670 9	25.243 2	80.492 2
6	0.388 9	0.666 6	1.669 3	0.667 4	30.287 2	84.279 5
7	0.333 3	0.666 6	1.667 1	0.666 8	35.334 3	88.275 1

例 2 假设第二类用户请求的到达率不变,分析第一类用户请求的到达率增大对系统的影响。假设 $h_1 = 40, h_2 = 20, \beta = 0.8, N = 100, \mu_1 = \mu_2 = 1.5, \lambda_2 = 3, \lambda_1 \in [0.5, 3]$。首先,对于给定的 λ_1,通过例 1 的方法求解最优值 c^{*}。其次,令 $c = c^{*}$,分析系统的如下性能参数:ρ、$E(L_1)$、$E(L_2)$、$E(T)$、$P_{总}$ 及 $f(c^{*})$。表 2 的数值结果显示随着 λ_1 增大,c^{*} 不变或增大(如图 3所示),$E(L_1)$、$P_{总}$ 及 $f(c^{*})$ 均随着 λ_1 增大而增大。

表 2　云计算中心的性能、能耗及成本分析(例 2)

Table 2　The performance, power consumption and cost of cloud computing centers (example 2)

λ_1	c^*	ρ	$E(L_1)$	$E(L_2)$	$E(T)$	$P_{总}$	$f(c^*)$
0.5	4	0.583 3	0.333 3	2.173 5	0.716 2	20.238 3	72.992 2
0.75	4	0.625 0	0.500 1	2.173 4	0.712 9	20.432 0	79.818 6
1	4	0.666 7	0.667 5	2.173 3	0.710 2	20.596 2	86.643 5
1.25	4	0.708 3	0.835 9	2.173 3	0.708 1	20.735 2	93.492 2
1.5	4	0.750 0	1.006 5	2.173 4	0.706 6	20.853 2	100.411 6
1.75	4	0.791 7	1.180 3	2.173 2	0.706 0	20.953 3	107.437 4
2	4	0.833 3	1.358 7	2.173 1	0.706 4	21.038 4	114.642 5
2.25	4	0.875 0	1.544 1	2.173 2	0.708 1	21.110 8	122.118 0
2.5	5	0.733 3	1.681 3	2.039 6	0.676 5	26.448 1	129.201 6
2.75	5	0.766 7	1.857 8	2.039 1	0.677 7	26.514 2	136.306 8
3	5	0.800 0	2.039 7	2.039 7	0.679 9	26.570 6	143.639 7

例 3　假设第一类用户请求的到达率不变,分析第二类用户请求的到达率增大对系统的影响。假设 $h_1=40, h_2=20, \beta=0.8, N=100, \mu_1=\mu_2=1.5, \lambda_1=1, \lambda_2\in[1.25,4]$。首先,对于给定的 λ_2,通过例 1 的方法求解最优值 c^*。其次,令 $c=c^*$,分析系统的如下性能指标:ρ、$E(L_1)$、$E(L_2)$、$E(T)$、$P_{总}$ 及 $f(c^*)$。表 3 的数值结果显示,随着 λ_2 增大,c^* 不变或增大(如图 4 所示),$E(L_2)$、$P_{总}$ 及 $f(c^*)$ 随着 λ_2 增大而增大,而 ρ、$E(L_1)$、$E(T)$ 未呈现单调性。

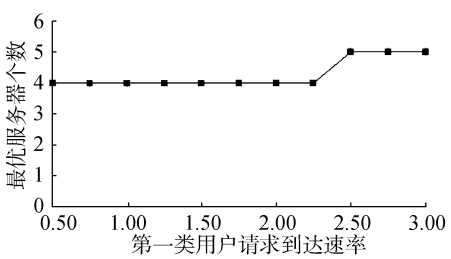


图 3　最优服务器数量(例 2)

Fig.3　The optimal number of servers (example 2)

表 3　云计算中心的性能、能耗及成本分析(例 3)

Table 3　The performance, power consumption and cost of cloud computing centers (example 3)

λ_2	c^*	ρ	$E(L_1)$	$E(L_2)$	$E(T)$	$P_{总}$	$f(c^*)$
1.25	3	0.500 0	0.675 8	0.855 5	0.680 6	13.883 9	55.247 7
1.5	3	0.555 6	0.675 7	1.045 3	0.688 4	14.239 2	59.326 4
1.75	3	0.611 1	0.675 7	1.250 5	0.700 4	14.544 1	63.671 7
2	3	0.666 7	0.675 6	1.477 4	0.717 7	14.806 5	68.417 0
2.25	4	0.541 7	0.667 5	1.544 4	0.680 6	19.968 6	73.563 0
2.5	4	0.583 3	0.667 7	1.739 7	0.687 8	20.211 6	77.671 8
2.75	4	0.625 0	0.667 5	1.947 7	0.697 4	20.418 9	81.991 1
3	4	0.666 7	0.667 5	2.173 3	0.710 2	20.596 2	86.643 5
3.25	4	0.708 3	0.667 5	2.423 2	0.727 2	20.747 8	91.761 8
3.5	5	0.600 0	0.666 6	2.422 9	0.686 6	26.049 0	95.962 8
3.75	5	0.633 3	0.666 7	2.629 8	0.694 0	26.181 1	100.211 5
4	5	0.666 7	0.666 6	2.850 9	0.703 5	26.293 5	104.717 6

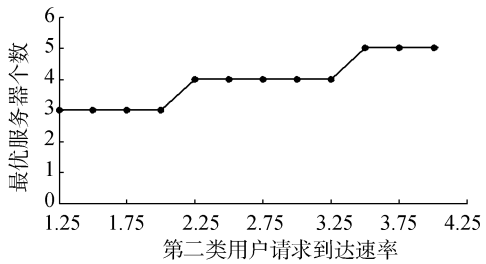


图 4 最优服务器数量(例 3)

Fig.4 The optimal number of servers (example 3)

6 结束语

本文基于排队论分别对具有两类用户请求的云计算中心建立相应的排队模型,分析系统中用户请求的稳态概率分布、平均队长等性能指标;通过引入等待成本和能耗成本,构建系统单位时间的成本函数,分析系统的最优服务器的数量。研究发现,随着用户请求到达率变化,最优服务器数量可能会发生变化,服务器数量的最优值是用户请求到达率的非减函数。对于任意的到达率,都可以得到最优的服务器数量,这为云计算中心的资源配置提供理论依据。

本文讨论的云计算中心具有两类用户请求,服务规则是非抢占优先服务,然而,在实际中存在抢占服务规则的情况及多类型用户请求的云计算中心,在后续的研究中可以讨论更多类型的用户请求问题及抢占服务规则的情况,并对多类用户请求的调度策略进行分析。

参考文献:

[1]廖倩文, 潘久辉, 王开杰. 基于排队理论的云计算中心性能分析模型[J]. 计算机工程, 2015, 41(9): 51-55.
LIAO Qianwen, PAN Jiuhui, WANG Kaijie. Performance analysis model of cloud computing center based on queueing theory[J]. Computer engineering, 2015, 41(9): 51-55.

[2]徐小龙,杨庚,李玲娟,等. 面向绿色云计算数据中心的动态数据聚集算法[J]. 系统工程与电子技术, 2012, 34(9): 1923-1929.
XU Xiaolong, YANG Geng, LI Lingjuan, el al. Dynamic data aggregation algorithm for data centers of green cloud computing[J]. Systems engineering and electronics, 2012, 34(9): 1923-1929.

[3]许丞, 刘洪, 谭良. Hadoop 云平台的一种新的任务调度和监控机制[J]. 计算机科学, 2013, 40(1): 112-117.
XU Chen, LIU Hong, TAN Liang. New mechanism of monitoring on Hadoop cloud platform[J]. Computer science, 2013, 40(1): 112-117.

[4]倪志伟, 李蓉蓉, 方清华,等. 基于离散人工蜂群算法的云任务调度优化[J]. 计算机应用, 2016, 36(1): 107-112,121.
NI Zhiwei, LI Rongrong, FANG Qinghua, et al. Optimization of cloud task scheduling based on discrete artificaial bee colony algorithm[J]. Journal of computer applications, 2016, 36(1): 107-112,121.

[5]罗亮, 吴文峻, 张飞. 面向云计算数据中心的能耗建模方法[J]. 软件学报, 2014, 25(7): 1371-1387.
LUO Liang, WU Wenjun, ZHANG Fei. Energy modeling based on cloud data center[J]. Journal of software, 2014, 25(7): 1371-1387.

[6]CAO J, LI K, STOJMENOVIC I. Optimal power allocation and load distribution for multiple heterogeneous multicore server processors across clouds and data centers[J]. IEEE transactions on computers, 2014, 63(1): 45-58.

[7]何怀文, 傅瑜, 杨亮,等. 性能受限下云中心异构服务器的能耗优化[J]. 计算机应用, 2015, 35(1): 39-42,61.
HE Huaiwen, FU Yu, YANG Liang, et al. Optimal power consumption of heterogeneous servers in cloud center under performance constraint[J]. Journal of computer applications, 2015, 35(1): 39-42,61.

[8]谭一鸣, 曾国荪, 王伟. 随机任务在云计算平台中能耗的优化管理方法[J]. 软件学报,2012, 23(2): 266-278.
TAN Yiming, ZENG Guosun, WANG Wei. Policy of energy optimal management for cloud computing platform with stochastic tasks[J]. Journal of software, 2012, 23(2): 266-278.

[9]KE M, YEH C, SU C. Cloud computing platform for real-time measurement and verification of energy performance [J]. Applied energy, 2017, 188: 497-507.

[10]SINGH S, CHANA I. EARTH: Energy-aware autonomic resource scheduling in cloud computing[J]. Journal of intelligent & fuzzy systems, 2016, 30(3):1581-1600.

[11]TAO F, LI C, LIAO T, et al. BGM-BLA: a new algorithm for dynamic migration of virtual machines in cloud computing[J]. IEEE transactions on services computing, 2016, 9(6): 910-925.

[12]MEI J, LI K, OUYANG A, et al. A profit maximization scheme with guaranteed quality of service in cloud computing[J]. IEEE transactions on computers, 2015, 64(11): 3064-3078.

[13] QI Q, LIAO J, WANG J.Integrated multi-service handoff mechanism with QoS-support strategy in mobile cloud

computing[J]. Wireless personal communications, 2016, 87(2): 593-614.

[14] HUANG Z, LU Y, OUYANG H. Scheduling strategy based on genetic algorithm for cloud computer energy optimization [C]//2015 IEEE International Conference on Communication Problem-Solving. Guilin, China, 2015:516-519.

[15] YANG B, LI Z, CHEN S, et al. Stackelberg game approach for energy-aware resource allocation in data centers[J]. IEEE transactions on parallel and distributed systems, 2016, 27(12): 3646-3658.

[16] SURESH S, SAKTHIVEL S. System modeling and evaluation on factors influencing power and performance management of cloud load balancing algorithms [J]. Journal of web engineering, 2016, 15(5/6): 484-500.

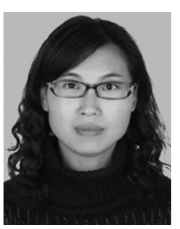
[17] 李春艳,何一舟,戴彬. Hadoop 平台的多队列作业调度优化方法研究[J]. 计算机应用研究, 2014(03): 705-707,738.
LI Chunyan, HE Yizhou, DAI Bin. Research on optimization of job scheduling based on multi-queue for Hadoop platform[J]. Application research of computers, 2014(3): 705-707, 738.

[18] LIU C, LI K, XU C, et al. Strategy configurations of multiple users competition for cloud service reservation [J]. IEEE transactions on parallel and distributed systems, 2016, 27(2): 508-520.

[19] ZHURAVLEV S, CARLOS SAEZ J, BLAGODUROV S. Survey of energy-cognizant scheduling techniques[J]. IEEE transactions on parallel and distributed systems, 2013, 24(7): 1447-1464.

作者简介:

张江强,男,1992 年生,硕士研究生,主要研究方向为排队论。



赵宁,女,1980 年生,副教授,博士,主要研究方向为排队论。发表学术论文 10 余篇。

刘文奇,男,1965 年生,教授,主要研究方向为数据挖掘和决策分析。发表学术论文 40 余篇,出版学术专著 2 部。

第四届亚洲模式识别会议
The 4th Asian Conference on Pattern Recognition (ACPR 2017)

The 4th Asian Conference on Pattern Recognition (ACPR 2017) will be held on November 26-29, 2017, Nanjing, China. The conference aims at providing one major international forum for researchers in pattern recognition and related fields to share their new ideas and achievements. Submissions from other than the Asia-Pacific regions are also highly encouraged.

Topics of interest include all aspects of pattern recognition including, but not limited to:

Computer Vision and Robot Vision

Pattern Recognition and Machine Learning

Signal Processing (signal, speech, image)

Media Processing and Interaction (video, document, medical applications, biometrics, HCI and VR)

Website: http://acpr2017.njust.edu.cn/.