



保证响应时间百分比下的云中心资源优化配置

何怀文^{1,2}

(1.电子科技大学中山学院计算机学院 中山 528402;

2.中山大学信息科学与技术学院 广州 510275)

摘要:为了保证响应时间百分比 SLA 的同时,最小化商用云中心资源配置成本,基于排队理论提出了云中心模型,从经济角度分析云中心资源配置成本,并推导出云中心响应时间百分比的数学表达式。给出了计算最小配置成本算法(MPCA),通过大量实验分析了负载流量、响应时间百分比和基准响应时间对配置成本的影响。实验结果表明,通过 MPCA 对云中心服务器数量和运行速率进行优化,可以使配置成本随着负载流量的增加而呈线性增长。

关键词:响应时间百分比;云中心;资源配置;SLA

doi: 10.11959/j.issn.1000-0801.2015197

Optimal Resource Provision of Cloud Center with Guarantee Percentile Response Time

He Huaiwen^{1,2}

(1.School of Computer, University of Electronic Science and Technology of China Zhongshan

Institute, Zhongshan 528402, China;

2. School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: In order to minimize resource provisioning cost of commercial cloud center, while guarantee percentile response time SLA, a cloud center model based on queuing theory was proposed. Resource provisioning cost was analyzed from an economic point of view, and the mathematical expressions of cloud center percentile response time were deduced. Meanwhile the minimum provisioning cost algorithm (MPCA) was developed to calculate the minimum cost of cloud center, the affect on cloud center cost by load flow, percentile response time and based response time was analyzed via a large number of experiments. The results show that with MPCA to optimized server number and speed of cloud center, it can achieve cost linear increase with increasing traffic load.

Key words: percentile response time, cloud center, resource provision, SLA

收稿日期:2015-03-11;修回日期:2015-07-31

基金项目:国家自然科学基金青年科学基金资助项目(No.61300095),广东省自然科学基金博士启动项目基金资助项目(No.S2012040011123),中山市科技计划基金资助项目(No.2013A3FC0285, No.2014A2FC396)

Foundation Items: The National Natural Science Foundation of China(No.61300095), The Ph D Start-up Fund of Natural Science Foundation of Guangdong Province (No.S2012040011123), The Science and Technology Project of Zhongshan (No.2013A3FC0285, No.2014A2FC396)

论文引用格式:何怀文. 保证响应时间百分比下的云中心资源优化配置. 电信科学, 2015197

He H W. Optimal resource provision of cloud center with guarantee percentile response time. Telecommunications Science, 2015197



1 引言

云计算作为一种新的计算范式,通过分布式计算和虚拟化技术,以资源集中管理的方式通过 Internet 为用户提供弹性计算服务^[1]。CSP(cloud service provider,云服务提供商)的成功运营,需要提供 QoS(quality of service,服务质量)保证。QoS 是影响用户对云服务满意程度的关键因素,通常以 SLA(service layer agreement,服务水平协议)形式描述^[2]。目前,在面向性能的 SLA 参数中,大部分研究均采用平均响应时间作为主要评价指标^[3]。然而,仅满足平均响应时间带来的服务体验对大部分用户而言是不够的。用户更关心的是响应时间的概率统计边界,而不仅是平均响应时间^[2,4],例如用户期望响应时间小于期望时间的概率为 95%。Amazon 公司提出了使用第 99.9 百分点概率分布来表示和度量 SLA,其商业产品的成功表明,选择百分比方法定义 SLA 带来了更好的用户体验,从而吸引了广泛的用户^[5]。

云计算环境的动态性、用户请求的多样性和负载的时变性给云中心的配置优化资源带来了很大困难^[6]。对于最小化资源配置成本和满足 SLA 的权衡问题,目前已经开展了广泛的研究^[7-9],通常采用动态电压频率调整(dynamic voltage frequency scaling, DVFS)改变服务器运行速率或者使用动态能耗管理(dynamic power management, DPM)调整活动服务器数量^[10],以使云中心计算资源被最大化利用。由于响应时间百分比的计算困难,多数文献采用平均响应时间作为云服务 SLA 的性能指标。参考文献[11]研究了异构多核服务器的云中心负载分发和能耗分配优化问题,以优化平均响应时间为目标。参考文献[12]基于多服务台排队模型,考虑了平均响应时间超时的 SLA 违规惩罚,提出了云中心利润最大化的经济模型。参考文献[13]则首次提出了基于响应时间百分比的云中心性能分析,以概率统计方式精确刻画了用户的性能需求。参考文献[14]基于排队网络理论,提出了在集群计算环境下,保证响应时间百分比时主机数量最小化的源配置方法。参考文献[3]使用 M/G/1/PS 模型,提出了保证响应时间百分比的云中心能耗优化的 PowerTail 策略。但是,以上研究没有给出响应时间百分比的数学表达公式,且多数文献只采用平均响应时间作为 SLA 性能参数或只从云中心能耗优化角度出发,没有从经济上考虑云中心运营成本的优化。

基于上述原因,本文从经济角度出发,综合考虑云中

心的能耗成本和服务器租用成本,提出在保证用户响应时间百分比 SLA 下,调整服务器数量和运行速率,以优化云中心资源配置问题。本文基于 $M/M/n$ 排队模型,提出在响应时间百分比 SLA 受限下,通过优化云中心服务器数量和运行速率,以达到资源配置成本最小化;推导了云中心响应时间百分比的数学表达式;分析了负载流量、响应时间百分比、基准时间变化对配置成本的影响。

2 系统模型

2.1 响应时间百分比

定义 1 (响应时间) 响应时间定义为从请求进入系统到完成服务后离开系统的时间间隔,记为 W_r 。包括排队等待时间 W_q 和服务时间 τ 。

定义 2 (响应时间百分比) 响应时间百分比指响应时间小于期望时间的概率百分比,表示为 $P\{W_r < \hat{T}\} \leq \gamma\%$,其中 $\hat{T}$ 为用户期望时间, $\gamma\%$ 为百分比。

云中心响应时间百分比也可表示为 $W_r \{t \leq \hat{T}\} = \int_0^{\hat{T}} f_{w_r}(t) dt \geq \gamma\%$,其中 $W_r(t)$ 为响应时间的概率分布函数, $f_{w_r}(t)$ 为响应时间的概率密度函数。例如在 $M/M/1$ 排队模型中,响应时间概率密度函数为 $f_{w_r}(t) = \mu(1-\rho)e^{-\mu(1-\rho)t}$ ^[15],其中 μ 为服务速率, λ 为请求到达速率, $\rho = \lambda/\mu$,则有:

$$\int_0^{\hat{T}} \mu(1-\rho)e^{-\mu(1-\rho)t} dt = 1 - e^{-\mu(1-\rho)\hat{T}} \geq \gamma\% \quad (1)$$

由式(1)可知,响应时间百分比 $\gamma\%$ 随着服务速率 μ 的增加而增加。为了保证约定的响应时间百分比,服务速率 μ 必须足够大,满足:

$$\mu \geq \frac{-\ln(1-\gamma\%) + \lambda}{\hat{T}} \quad (2)$$

2.2 云中心模型

假设云中心为一个 $M/M/n$ 多服务台的排队模型系统,拥有 n 台运行速率为 s 的服务器,记为 $S=\{S_1, S_2, \dots, S_n\}$, $|S|=n$ 。与模型相关的符号符号见表 1。用户请求到达为泊松流,速率为 λ 。当请求到达时,如果没有空闲服务器,则进入缓冲区排队等待,缓冲区大小为无穷大。服务系统按照先到先服务(FCFS)规则进行调度。假设完成每个请求所需计算量 r 为独立同分布的随机变量,符合参数为 $\bar{r}$ 的负指数分布。因此,请求执行时间 $x=r/s$ 符合参数为 $\bar{x}=\bar{r}/s$ 的负指数分布。服务速率 $\mu=1/\bar{x}=s/\bar{r}$,服务器利用率 $\rho=\lambda/n\mu=\lambda\bar{r}/ns$ 。

表 1 模型相关的符号描述

符号	意义
λ	负载到达速率
r	完成请求所需的计算量
x	完成请求所需的时间
n	云中心服务器数量
ρ	服务器利用率, 等于 $\lambda/(n\mu)$
s	服务器运行速率
μ	服务速率
$\hat{T}$	基准响应时间
ψ	单位能耗价格
k	能耗比例系数
$\gamma\%$	响应时间百分比
$\bar{r}$	请求所需的平均计算量
$\bar{x}$	完成请求所需的平均时间
π_k	队长为 k 的概率
P^*	服务器基本能耗
C	云中心配置总成本
C_r	能耗成本
C_e	服务器租用成本
ω	单位时间服务器租用价格
η	能耗比例因子

当 $\rho < 1$, $M/M/n$ 排队系统可到达平稳状态, 其中 $\pi_k = \lim_{t \rightarrow \infty} P\{N(t)=k\}$, 其中 $N(t)$ 为时刻 t 的系统队长^[16]。根据参考文献[16], 有:

$$\pi_k = \begin{cases} \frac{(n\rho)^k}{k!} \pi_0, & 0 \leq k < n \\ \frac{n^n \rho^k}{n!} \pi_0, & k \geq n \end{cases} \quad (3)$$

因为 $\sum_{k=1}^{\infty} \pi_k = 1$, 可得 $\pi_0 = \left[\sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!} + \frac{(n\rho)^n}{n!} \frac{1}{1-\rho} \right]^{-1}$, $\rho < 1$ 。

请求到达时, 进入缓冲区排队等待的概率记为:

$$P_w = \sum_{k=n}^{\infty} \pi_k = \sum_{k=n}^{\infty} \frac{n^n \rho^k}{n!} \pi_0 = \frac{\pi_n}{1-\rho} \quad (4)$$

定理 1 在平稳状态下, $M/M/n$ 云中心模型请求响应时间的分布函数表示为:

$$W_r(t) = \begin{cases} 1 - e^{-\mu t} [1 + \mu t P_w], & \rho = 1 - \frac{1}{n} \\ 1 - e^{-\mu t} - \frac{P_w}{n - n\rho - 1} [e^{-\mu t} - e^{-\mu t(n-\rho)}], & \rho \neq 1 - \frac{1}{n} \end{cases} \quad (5)$$

$$W_r(t) = \begin{cases} 1 - e^{-\mu t} [1 + \mu t P_w], & \rho = 1 - \frac{1}{n} \\ 1 - e^{-\mu t} - \frac{P_w}{n - n\rho - 1} [e^{-\mu t} - e^{-\mu t(n-\rho)}], & \rho \neq 1 - \frac{1}{n} \end{cases} \quad (6)$$

证明 响应时间为等待时间 W_q 和服务时间 τ 之和, 即 $W_r = W_q + \tau$, 而 W_q 和 τ 相互独立, τ 服从参数为 μ 的负指数分布。

由参考文献[17]得到:

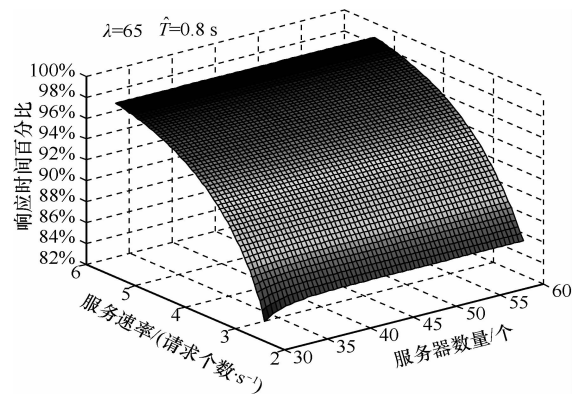
$$W_q(x) = \begin{cases} 1 - P_w, & x=0 \\ 1 - P_w \cdot e^{-\mu x(1-\rho)}, & x>0 \end{cases} \quad (7)$$

由卷积公式, 可得:

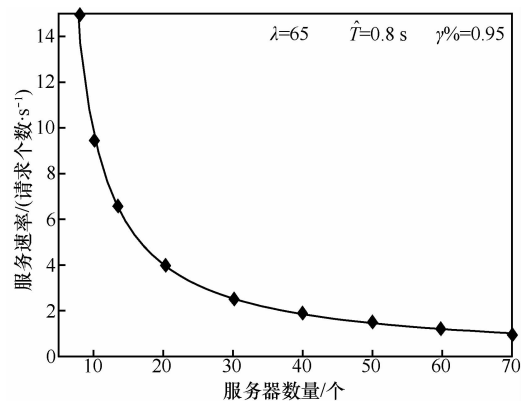
$$W_r(t) = \int_0^t P\{W_q \leq t-x\} dP\{\tau \leq x\} = \int_0^t (1 - P_w e^{-\mu(n-\rho)(t-x)}) \mu e^{-\mu x} dx \quad (8)$$

对式(8)求解积分, 即可得到 $W_r(t)$ 的详细表达式, 详细推导过程省略, 定理 1 得证。

当 $\lambda=65$, $\hat{T}=0.8$ s 时, 响应时间百分比与服务器数量、服务速率的关系如图 1(a) 所示。由图 1(a) 可见, 响应时间百分比 $\gamma\%$ 随着服务器数量 n 、服务速率 μ 的降低单调递减。当 $\lambda=65$, $\hat{T}=0.8$ s, $\gamma\%=0.95$ 时, 服务器数量 n 和服务速率 μ 的关系如图 1(b) 所示。由图 1(b) 可知, 在响应时间百分比 $\gamma\%$ 受限的情况下, 随着服务器数量 n 的增加, 服务速率 μ 单调下降。但是当 μ 降低到边界值附近时, 即使增加再多的服务器, 对响应时间百分比 $\gamma\%$ 的影响也极低。



(a) 响应时间百分比与服务器数量、服务速率的关系



(b) 满足固定响应时间百分比时服务器数量、服务速率的关系

图 1 响应时间百分比和服务器数量、服务速率的关系



3 最小化配置成本问题

3.1 问题描述

CSP 的运营成本主要由以下两部分组成:从云基础设施提供商租用服务器的租用成本、服务器运行时产生的能耗成本。其中,服务器租用成本与服务器数量和租用时间相关,假设单位时间内一台服务器的租用成本为 ω ,则云中心服务器租用总成本为 $C_r=n\omega$ 。

服务器单位时间能耗可表示为 $P=Cfv^2$ (单位为 W),其中 C 为电路交换能力, f 为处理器时钟频率, v 为处理供应电压^[79]。由于处理器运行速率 $s \propto f$, 而 $f \propto V^\phi$, 其中 $0 < \phi < 1$, 可知 $P \propto s^d$, 其中 α 是 2~3 的常数, 表示固定的比例指数。因此, 单位时间服务器能耗可表示为 $P=Cfv^2=ks^\eta$, 通常取 $\eta=3$ ^[8]。在 $M/M/n$ 排队系统中, 服务器利用率为 $\rho=\lambda/(\eta\mu)$ 。假设服务器空闲时仍然以较低的速度运转, 此时单位时间的基本能耗为 P^* 。服务器忙时能耗为 $P_{\text{busy}}=\rho ks^\eta=\frac{\lambda r}{ns}ks^\eta=\frac{\lambda r}{n}ks^{\eta-1}$ 。则服务器单位时间总能耗为 $P=n(P_{\text{busy}}+P^*)=\lambda r \bar{k}s^{\eta-1}+nP^*$ 。假设每瓦特的能耗价格为 ψ , 能耗成本可以表示为 $C_e=\psi(\lambda r \bar{k}s^{\eta-1}+nP^*)$ 。

因此, 云服务提供商的总成本可表示为:

$$C=C_r+C_e=n\omega+\psi(\lambda r \bar{k}s^{\eta-1}+nP^*) \quad (9)$$

保证响应时间百分比下最小化云中心配置成本的问题可描述为:

$$\text{Minimize } C(n,s) \quad (10)$$

约束条件为:

$$\begin{cases} W(t|\hat{T}) \geq \gamma\%, \gamma\% < 1 \\ \rho = \frac{\lambda r}{ns} < 1 \end{cases} \quad (11)$$

3.2 问题分析

由式(5)和式(6)可知, 当 ρ 取值不同时, $W(t)$ 具有不同的表达式。当服务器数量 $n > 10$, 使用式(3)计算 $W(t)$ 时, 利用率 $\rho=1-1/n > 0.9$ 。随着 n 的增多, $\rho=1-1/n$ 将非常接近 1。根据参考文献[17]可知, 当 $\rho > 0.9$ 时, 用户的等待时间会变得非常长; 当 ρ 接近 1 时, 系统在高负荷下出现系统状态空间崩溃, 因此本文忽略了 ρ 接近 1 的情形, 只考虑 $\rho \neq 1-1/n$ 的情况, 即只考虑使用式(6)计算响应时间百分比。

推论 1 响应时间百分比可以近似表示为:

$$W(t) \approx 1 - e^{-\mu t} - \frac{1}{(\sqrt{2\pi n} (1-\rho)(e^\rho/e\rho)^n + 1)(n-n\rho-1)} [e^{-\mu t} - e^{-\mu(n-n\rho)t}] \quad (12)$$

证明 根据泰勒公式, 得到 $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots +$

$\frac{x^n}{n!}$, 可知 $\sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!} = e^{n\rho}$, 同时由 Stirling 公式可知 $n! =$

$\sqrt{2\pi n} \left(\frac{n}{e}\right)^n$, 可得:

$$\begin{aligned} \pi_0 &= \left(e^{n\rho} + \frac{(n\rho)^n e^n}{\sqrt{2\pi n} (n)^n 1-\rho} \right)^{-1} \\ &= \left(e^{n\rho} + \frac{(e\rho)^n}{(1-\rho)\sqrt{2\pi n}} \right)^{-1} \end{aligned} \quad (13)$$

$$\pi_n = \frac{n^n \rho^n}{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n} \pi_0 = \frac{1-\rho}{\sqrt{2\pi n} (1-\rho) \left(\frac{e^\rho}{e\rho}\right)^n + 1} \quad (14)$$

$$P_w = \frac{\pi_n}{1-\rho} = \frac{1}{\sqrt{2\pi n} (1-\rho) \left(\frac{e^\rho}{e\rho}\right)^n + 1} \quad (15)$$

因此, 可得出式(2), 推论 1 得证。

推论 2 当 $\eta=3$ 时, 式(10) $C(n,s)=n\omega+\psi(\lambda r \bar{k}s^{\eta-1}+nP^*)$ 存在唯一最小值。

证明 当 $\eta=3$ 时, $C(n,s)=n\omega+\psi(\lambda r \bar{k}s^2+nP^*)$, 为 n, s 的单调递增函数。为了满足响应时间百分比, 由式(3)和式(4)可知, $\gamma\% = W(t) < 1 - e^{-\mu \hat{T}} = 1 - e^{-\hat{T}r}$, 因此得到 s 的最小边界为 $s_{\text{Low}} = \frac{\bar{r} \ln(1-\gamma\%)}{\hat{T}}$, 而服务器数量 $n=1$ 时, 可以求出 s 的最大

值 s_{Max} 。根据图 1(b)可知, 在 $(s_{\text{Low}}, s_{\text{Max}}]$, n 随着 s 的减少而单调增加。令 $n=\mathcal{G}(s)$, 得到:

$$\begin{aligned} C(n,s) &= n(\omega + P^*) + \psi \lambda r \bar{k} s^2 \\ &= \mathcal{G}(s)(\omega + P^*) + \psi \lambda r \bar{k} s^2 \end{aligned} \quad (16)$$

$\mathcal{G}(s)$ 为单调递减函数, 同时也为凹函数, 即 $\frac{\partial^2 \mathcal{G}}{\partial s^2} > 0$ 。因

$\frac{\partial^2 C}{\partial s^2} = \frac{\partial^2 \mathcal{G}}{\partial s^2} (\omega + P^*) + 2\psi \lambda r \bar{k} > 0$, 即函数 $C(n,s)$ 在 $(s_{\text{Low}}, s_{\text{Max}}]$ 为凹函数, 存在最小值。因此, 存在一个恰当的 (n,s) , 可以满足响应时间百分比 $\gamma\%$, 同时使得 $C(n,s)$ 获得最小值。

3.3 问题求解

对于式(11), 令 $R = (\frac{e^\rho}{e\rho})$, $P_1 = e^{-\hat{T}r} - e^{-\lambda(n-n\rho)\hat{T}r}$, $P_2 = \sqrt{2\pi n}$

$$(1-\rho)R+1, P_3 = n-n\rho-1, Q = \frac{P_1}{P_2 P_3}.$$

则配置成本最小化问题可以转化为:

$$\text{Minimize } \mathcal{C}(n,s) \quad (17)$$

满足:

$$\begin{cases} \mathcal{T}(n,s)=\gamma\% \\ \rho=\frac{\lambda\bar{r}}{ns}<1 \end{cases} \quad (18)$$

其中, $\mathcal{T}(n,s)=1-e^{-s\hat{T}/\bar{r}}-Q=\gamma\%$ 。

该问题为一个二元非线性约束优化问题,假设拉格朗日乘子函数为:

$$H(n,s,\xi)=\mathcal{C}(n,s)+\xi(\mathcal{T}(n,s)-\gamma\%) \quad (19)$$

分别对 n, s, ξ 求偏导数,得到以下方程组:

$$\begin{cases} \frac{\partial \mathcal{C}}{\partial n} - \xi \frac{\partial \mathcal{T}}{\partial n} = 0 \\ \frac{\partial \mathcal{C}}{\partial s} - \xi \frac{\partial \mathcal{T}}{\partial s} = 0 \\ \mathcal{T}(n,s) - \gamma\% = 0 \end{cases} \quad (20)$$

由式(20)的前两个等式可得到:

$$\frac{\partial \mathcal{C}}{\partial n} \times \frac{\partial \mathcal{T}}{\partial s} = \frac{\partial \mathcal{C}}{\partial s} \times \frac{\partial \mathcal{T}}{\partial n} \quad (21)$$

由式(20)的第三个等式和式(21)组成新的方程组为:

$$\begin{cases} \frac{\partial \mathcal{C}}{\partial n} \times \frac{\partial \mathcal{T}}{\partial s} = \frac{\partial \mathcal{C}}{\partial s} \times \frac{\partial \mathcal{T}}{\partial n} \\ \mathcal{T}(n,s) = \gamma\% \end{cases} \quad (22)$$

式(22)的解即问题的最优解。

在式(21)中,对于函数 $\mathcal{C}(n,s)$, $\frac{\partial \mathcal{C}}{\partial n} = \omega + \psi P^*$, $\frac{\partial \mathcal{C}}{\partial s} = \psi(\eta -$

$1)\lambda ks^{\eta-2}/\bar{r}$ 。对于函数 $\mathcal{T}(n,s)$,有:

$$\frac{\partial \mathcal{T}}{\partial n} = -\frac{\partial Q}{\partial n} = \frac{1}{P_2 P_3} \left(\frac{P_1}{P_3} \frac{\partial P_3}{\partial n} + \frac{P_1}{P_2} \frac{\partial P_2}{\partial n} - \frac{\partial P_1}{\partial n} \right) \quad (23)$$

$$\begin{aligned} \frac{\partial \mathcal{T}}{\partial s} &= \hat{T} e^{-s\hat{T}/\bar{r}} - \frac{\partial Q}{\partial s} \\ &= \hat{T} e^{-s\hat{T}/\bar{r}} - \frac{1}{P_2 P_3} \left(\frac{P_1}{P_3} \frac{\partial P_3}{\partial s} + \frac{P_1}{P_2} \frac{\partial P_2}{\partial s} - \frac{\partial P_1}{\partial s} \right) \end{aligned} \quad (24)$$

在式(23)中,有:

$$\frac{\partial P_1}{\partial n} = (s\hat{T}/\bar{r}) e^{-s(n-n\rho)\hat{T}/\bar{r}} \quad (25)$$

对于 $R=(\frac{e^{\rho}}{e\rho})^n$,两边取自然对数,得到 $\ln R=n(\rho-1\ln\rho-1)$,

再对 n 求偏导数,得到:

$$\frac{\partial R}{\partial n} \frac{1}{R} = \rho - \ln \rho - 1 + n \left(\frac{\partial \rho}{\partial n} - \frac{1}{\rho} \frac{\partial \rho}{\partial n} \right) \quad (26)$$

而 $\frac{\partial \rho}{\partial n} = -\frac{\rho}{n}$,化简得到 $\frac{\partial R}{\partial n} = -R \ln \rho$ 。

因此,得到:

$$\frac{\partial P_2}{\partial n} = \sqrt{2\pi} \left(\frac{1}{2\sqrt{n}} (1+\rho) R - R \sqrt{n} (1-\rho) \ln \rho \right) \quad (27)$$

同时,有:

$$\frac{\partial P_3}{\partial n} = 1 \quad (28)$$

对于式(24),使用相同的方法,可以得到:

$$\frac{\partial P_1}{\partial s} = -s e^{-s\hat{T}} + n \hat{T} e^{-s(n-n\rho)\hat{T}} \quad (29)$$

$$\frac{\partial P_2}{\partial s} = \frac{1}{s} \sqrt{2\pi n R (n(1-\rho)^2 + \rho)} \quad (30)$$

$$\frac{\partial P_3}{\partial s} = \frac{\lambda}{s^2} \quad (31)$$

根据推论 2,当 $\eta=3$ 时, $\mathcal{C}(n,s)$ 存在唯一的最小值,即方程组(22)存在唯一的实数解。由于 $\mathcal{T}(n,s)$ 是关于服务器数量 n 和运行速率 s 的单调递增函数,对于给定 n ,可以使用二分法求解出方程 $\mathcal{T}(n,s)=\gamma\%$ 关于 s 的唯一解。因此,对方程组(22)也可以使用嵌套二分法思想求解。求解方程组(22)的 MPCA (minimum provisioning cost algorithm,最小配置成本算法)的详细描述如下。

输入: $\lambda, k, \bar{r}, \gamma\%, \hat{T}, P^*, \psi, \omega$ 。

输出: opt_n, opt_s, min_cost 。

$$s_{Low} \leftarrow \frac{\bar{r} \ln(1-\gamma\%)}{\hat{T}};$$

使用二分法求解当 $n=1$ 方程 $\mathcal{T}(n,s)=\gamma\%$ 关于 s 的解,为

s_{Max} ;

$left \leftarrow s_{Low}, right \leftarrow s_{Max}$;

do

$s' \leftarrow opt_s$;

$opt_s \leftarrow (left + right)/2$;

使用二分法求出 $s=opt_s$ 时,方程 $\mathcal{T}(n,s)=\gamma\%$ 关于 n 的解;

if $(\frac{\partial \mathcal{C}}{\partial n} \times \frac{\partial \mathcal{T}}{\partial s} - \frac{\partial \mathcal{C}}{\partial s} \times \frac{\partial \mathcal{T}}{\partial n} > 0)$

$left \leftarrow opt_s$;

else

$right \leftarrow opt_s$;

end if

while $(|s' - opt_s| \leq \delta)$;



使用二分法求出 $s=opt_s$ 时, $T(n,s)=\gamma\%$ 关于 n 的解 opt_n ;

$mini_cost=n\omega+\psi(\lambda ks^{n-1}+nP^b)$;

return $opt_n, opt_s, mini_cost$

4 实验结果

为了验证服务器数量 n 、运行速率 s 、响应时间百分比 $\gamma\%$ 对云中心配置成本的影响, 本文采用了离散事件仿真软件 Arena^[20] 对云中心模型进行模拟仿真实验。首先, 由于采用多服务台排队模型 $M/M/n$ 来模拟云计算中心, 负载到达速率 λ 服从负指数分布。在给定 λ 和 $\gamma\%$ 之下, 先由式(19)求解出服务台数 n 和服务速率 s 之间的关系。然后在 Areal 软件中根据不同的 (n, s, λ) 数值设置排队模型相关参数进行仿真, 同时观察请求的响应时间以及响应时间的分布, 验证其是否符合响应时间百分比。最后, 将 (n, s, λ) 代入式(9)即可计算出不同配置下的云中心的配置成本。根据参考文献[12,19], 云中心各项参数设置见表 2。

表 2 云中心参数设置

参数	数值
能耗比例系数 k	1
能耗比例因子 η	3
服务器租用价格 $\omega/(\text{€} \cdot \text{s}^{-1})$	0.5
能耗价格 $\psi/(\text{\$} \cdot \text{kW} \cdot \text{h}^{-1})$	0.06
服务器基本能耗 P^b/W	2
基准响应时间 $\hat{T}/\text{s}$	0.8
响应时间百分比 $\gamma\%$	0.98
请求所需的平均计算量 $\bar{r}$ /十亿级指令	1

4.1 云中心配置成本最小化

当云中心负载 λ 和响应时间百分比 $\gamma\%$ 固定时, 云中心配置成本 C 和服务器数量 n 关系如图 2 所示。由图 2 可以看出, 配置成本曲线随着服务器数量的增加先是急剧下降, 到达最低点后缓慢上升。其意味着当服务器数量 n 较少时, 需要较高的运行速率才能保证用户响应时间百分比的需求, 进而导致能耗成本的增加, 造成配置成本较大。当服务器到达一定数量时, 只需要较低的运行速率即可满足

用户需求。因此, 即使增加再多的服务器, 对保证用户响应时间百分比的影响也不大, 但是由此产生的服务器租用成本和空闲能耗成本会影响总配置成本。同时, 由图 2 可知, 当负载 λ 固定, 存在服务器数量 n 和运行速率 s , 使得配置成本 C 最小化, 同时满足用户 SLA。当 $\lambda=45, 55, 65, 75$ 时, 最优配置成本 C 分别为 74.049 4、90.017 52、105.954 93、121.868 51, 具体各项参数见表 3。

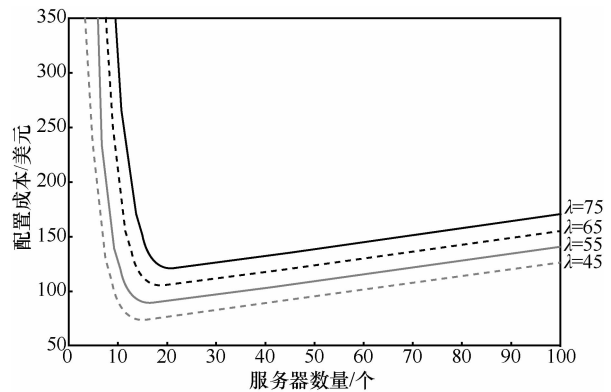


图 2 云中心配置成本和服务器数量关系

由表 3 可知, 云中心的资源配置成本始终随着负载的增加而增多, 与现实相符。同时, 资源配置优化使利用率 ρ 不能过高。较高的利用率容易造成等待时间过长, 违反 SLA 协议, 但是利用率过低则会产生资源的浪费。

4.2 配置成本的影响因素

由第 2.3 节可知, 云中心的配置成本与服务器数量和运行速率有关。其中, 服务器数量影响租用成本和能耗成本, 而运行速率则主要对能耗成本产生影响。同时, 在满足相同的用户 SLA 前提下, 可通过本文的 MPCA 计算出最优的服务器数量和最优的运行速率, 使得云中心成本最优化。当云中心需要提供更好的 SLA 服务(如更快的响应时间或者更高的响应时间百分比)时, 需要增加更多的计算资源, 其最优配置成本也将提高; 同时, 当用户请求速率增加时, 为了处理更多的用户请求, 同时保证 SLA, 云中心也需要增加配置资源。

在本文仿真实验中, 只考虑满足 SLA 下云中心配置成本受到的影响因素, 主要包括云中心请求速率 λ 、响应时

表 3 $\lambda=45, 55, 65, 75$ 下的成本优化

λ	s	n	C	C_e	C_r	ρ
45	4.914 27	14.264 89	74.049 40	66.916 95	7.132 44	0.641 92
55	4.912 12	16.761 26	90.017 52	81.636 87	8.380 63	0.668 02
65	4.910 46	19.218 55	105.954 93	96.345 65	9.609 27	0.688 76
75	4.909 13	21.645 46	121.868 51	111.045 78	10.822 72	0.705 81

间百分比 $\gamma\%$ 、基准响应时间 $\hat{T}$ 等因素。它们之间的关系如图 3 所示。

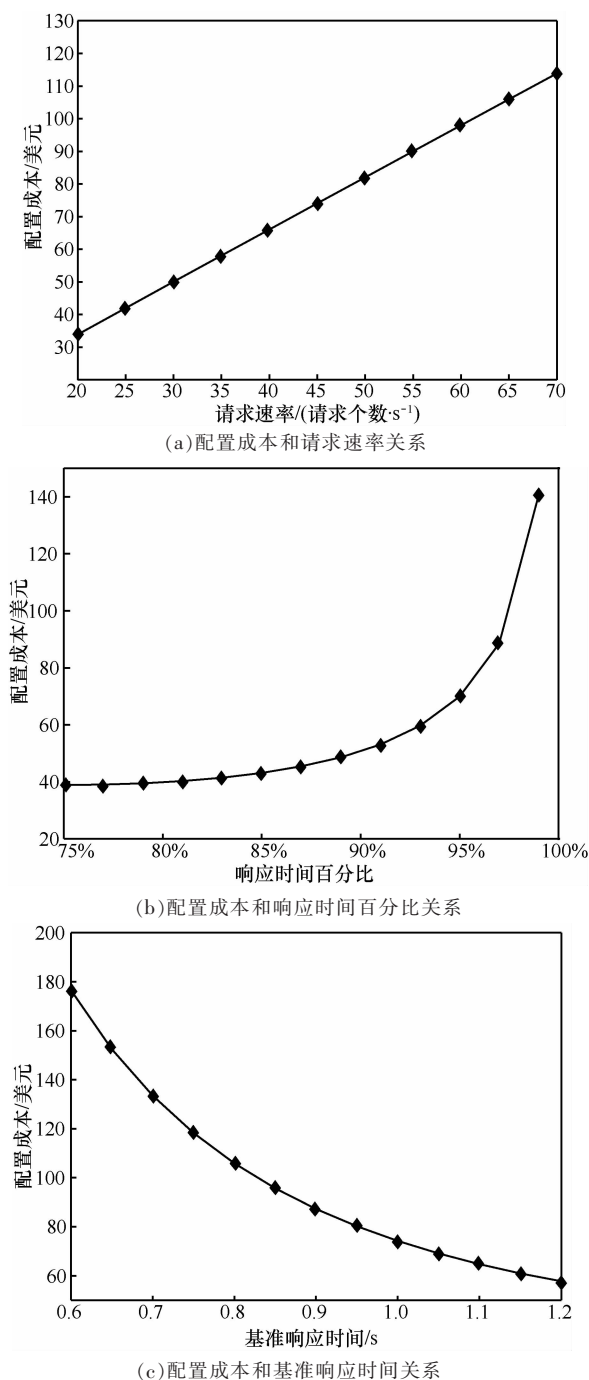


图 3 配置成本和请求速率、响应时间百分比、基准响应时间的关系

用户请求速率 λ 对最优配置成本的影响如图 3(a) 所示。由图 3(a) 可见, C 随着 λ 的增加而递增。在图 3(a) 中, 经过优化配置的云中心, 成本随着请求速率的增加几乎呈线性增长, 达到资源配置和请求速率 λ 按比例增加的良好

效果。响应时间百分比对配置成本的影响如图 3(b) 所示, 其中请求速率和基准响应时间分别设定为 $\lambda=65, \hat{T}=0.8$ s。由图 3(b) 可见, C 随着 $\gamma\%$ 增加而呈非线性递增。图 3(b) 可以认为是对不同级别 SLA 的模拟 (如不同类型的云用户对响应时间百分比的需求是不同的)。由图 3(b) 可知, 当响应时间百分比小于 90% 时, 配置成本曲线变化平缓, 而当响应时间百分比大于 90% 时, 配置成本急剧增加, 意味着满足更高级别的 SLA 时, 云中心需要配置更多的计算资源。图 3(c) 则反映了基准响应时间 $\hat{T}$ 的改变对配置成本的影响, 此时请求速率和响应时间百分比分别设定为 $\lambda=65, \gamma\%=0.98$ 。由图 3(c) 可见, 随着 $\hat{T}$ 的增加, 配置成本 C 递减。图 3(c) 意味着, 当需要更快基准响应时间的云服务 (如软实时计算) 时, 云中心需要更多的资源; 而对于时延容忍的云服务 (如 Web 服务、MapReduce 计算), 可以使用较低的资源配置。

5 结束语

本文从云用户角度出发, 分析响应时间百分比受限下云中心配置成本, 同时考虑了云中心的能耗和服务器租用成本, 并推导出了云中心响应时间百分比的数学表达式, 给出求解云中心最优配置成本的 MPCA。还进一步分析了请求速率、响应时间百分比、基准响应时间对配置成本的影响。实验结果表明, 通过采用 MPCA, 实现了在保证响应时间百分比的前提下, 云中心配置成本随着系统负载的增长而线性增加的目的。下一步工作包括: 研究满足响应时间百分 SLA 的动态资源配置; 结合负载均衡, 研究在异构云计算环境下的资源分配; 研究与虚拟化技术结合的资源配置模型。

参考文献

- 1 Armbrust M, Fox A, Griffith R, *et al.* A view of cloud computing. *Communications of the ACM*, 2010, 53(4): 50~58
- 2 Bruneo D. A stochastic model to investigate data center performance and QoS in IaaS cloud computing systems. *IEEE Transactions on Parallel and Distributed Systems*, 2014, 25(3): 560~569
- 3 Wang S, Munawar W, Liu J, *et al.* Power-saving design for server farms with response time percentile guarantees. *Proceedings of the 18th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, Beijing, China, 2012
- 4 Xiong K Q, Perros H. QRP01-6: resource optimization subject to a percentile response time SLA for enterprise computing.



- Proceedings of Global Telecommunications Conference, San Francisco, CA, USA, 2006
- 5 Decandia G, Hastorun D, Jampani M, *et al.* Dynamo: amazon's highly available key-value store. SIGOPS Oper Syst Rev, 2007, 41(6): 205~220
- 6 Khazaei H. Performance modeling of cloud computing centers (doctor dissertation). The University of Manitoba, 2012
- 7 Zhuravlev S, Saez J C, Blagodurov S, *et al.* Survey of energy-cognizant scheduling techniques. IEEE Transactions on Parallel and Distributed Systems, 2013, 24(7): 1447~1464
- 8 Deng W, Liu F, Jin H, *et al.* Reliability-aware server consolidation for balancing energy-lifetime tradeoff in virtualized cloud datacenters. International Journal of Communication Systems, 2014, 27(4): 623~642
- 9 Berl A, Gelenbe E, Di Girolamo M, *et al.* Energy-efficient cloud computing. The Computer Journal, 2010, 53(7): 1045~1051
- 10 Bilal K, Malik S U R, Khalid O, *et al.* A taxonomy and survey on green data center networks. Future Generation Computer Systems, 2014, 36(7): 189~208
- 11 Cao J, Li K, Stojmenovic I. Optimal power allocation and load distribution for multiple heterogeneous multicore server processors across clouds and data centers. IEEE Transactions on Computers, 2014, 63(1): 45~58
- 12 Cao J W, Kai H, Li K Q, *et al.* Optimal multiserver configuration for profit maximization in cloud computing. IEEE Transactions on Parallel and Distributed Systems, 2013, 24(6): 1087~1096
- 13 Xiong K Q, Perros H. Service performance and analysis in cloud computing. Proceedings of 2009 World Conference on Services-I, Los Angeles, CA, USA, 2009: 693~700
- 14 Xiong K Q, Suh S. Resource Provisioning in SLA-Based Cluster Computing Job Scheduling Strategies for Parallel Processing. Berlin: Springer Berlin Heidelberg, 2010: 1~15
- 15 陆传赓. 排队论. 北京: 北京邮电学院出版社, 1994
- Lu C L. Queuing Theory. Beijing: Beijing University of Posts and Telecommunications Press, 1994
- 16 唐应辉, 唐小我. 排队论: 基础与分析技术. 北京: 科学出版社, 2006
- Tang Y H, Tang X W. Queuing Theory: Basic and Analysis. Beijing: Science Press, 2006
- 17 Gross D, Shortle J F, Thompson J M, *et al.* Fundamentals of Queueing Theory. Hoboken: John Wiley & Sons, 2013
- 18 Zhai B, Blaauw D, Sylvester D, *et al.* Theoretical and practical limits of dynamic voltage scaling. Proceedings of the 41st Design Automation Conference, San Diego, USA, 2004
- 19 Mazzucco M, Dyachuk D. Balancing electricity bill and performance in server farms with setup costs. Future Generation Computer Systems-The International Journal of Grid Computing and Escience, 2012, 28(2): 415~426
- 20 Altiock T, Melamed B. Simulation modeling and analysis with Arena. New York: Academic Press, 2010

[作者简介]



何怀文,男,电子科技大学中山学院讲师,主要研究方向为随机优化、云计算资源分配、绿色计算。