高技术通讯 2017 年 第 27 卷 第 9 – 10 期:790 ~ 799

doi:10.3772/j.issn.1002-0470.2017.09 – 10.003

融合区分服务和速率调整的请求调度策略①

金顺福②[*,**,***] 薛元铮[*,**,***] 顾 蕊[*,**,***]

([*]燕山大学信息科学与工程学院 秦皇岛 066004)

([**]河北省计算机虚拟技术与系统集成重点实验室 秦皇岛 066004)

([***]河北省软件工程重点实验室 秦皇岛 066004)

摘 要 针对异构用户服务质量(QoS)需求不同及云数据中心能源浪费等问题,提出了一种融合区分服务和速率调整的请求调度策略。该策略将 IaaS 云数据中心分为提供速率调整机制的快速云和提供差错恢复机制的可靠云,然后研究新型混合云计算模式。构建了混合排队网络系统模型,从延迟敏感请求的平均响应时间、快速云中虚拟机的节能水平及可靠云中虚拟机的利用率等方面评估了请求调度策略的系统性能。进行了系统实验,验证了策略的有效性及模型的合理性。通过构造系统成本函数,改进 Jaya 智能寻优算法,给出了请求调度策略的优化方案。

关键词 云计算,请求调度,区分服务,速率调整,排队网络,智能优化

0 引 言

云计算技术中的服务模式可以分为软件即服务(software as a service, SaaS)、平台即服务(platform as a service, PaaS)和基础设施即服务(infrastructure as a service, IaaS)[1,2]。当用户通过互联网向 SaaS 提供商请求服务时,SaaS 可以利用本地资源提供服务,也可以利用 IaaS 云资源提供服务,形成一种混合云计算模式。这种模式能够动态处理用户请求,缓解本地服务器负载压力,提高用户响应性能。

Li[3]等为了减少 SaaS 提供商的运营成本,分别考虑长期调度、短期调度和动态供应,基于混合云计算模式设计了一种动态在线监控算法(online dynamic provision algorithm, ODPA)。黄旭[4]等以混合云计算模式为基础,提出了一种基于排队理论的云计算服务分流博弈模型,用以解决 SaaS 服务系统的博弈问题。上述研究专注于基于同构用户的请求调度策略。考虑请求的异构性,Nan[5]等研究了云计算环境中多媒体服务的资源分配问题,建立了一种排队网络模型描述云计算环境下的区分服务。Ding[6]等在云计算环境中提出了一种基于拍卖的资源调度(Abacus)框架,指定用户请求优先权及请求属性,实现了用户请求的区分服务。相关文献主要面向单一云对请求及服务进行了抽象分类。本文将基于混合云模式的分流技术及面向延迟敏感和差错敏感请求的区分服务相结合,考虑系统的节能水平,提出了一种融合区分服务和速率调整的新的请求调度策略。本研究建立了混合排队网络系统模型,进行了延迟敏感请求平均响应时间、快速云中虚拟机节能水平及可靠云中虚拟机利用率等系统性能指标的理论分析。按照请求调度策略的工作原理进行了系统仿真,给出了相关性能指标的统计结果。综合多性能指标构造系统成本函数,改进了 Jaya 智能寻优算法,给出了快速云中虚拟机速率差值与 SaaS 提供商的本地服务器系统容量的联合优化方案。

① 国家自然科学基金(61472342)和河北省自然基金(F2017203141)资助项目。

② 女,1966 年生,博士,教授;研究方向:计算机通信网络的系统建模与性能分析,排队论应用等;联系人,E-mail:jsf@ysu.edu.cn

(收稿日期:2017-06-08)

1 请求调度策略及系统模型

1.1 请求调度策略

对于 SaaS 提供商,大规模用户请求的涌入会造成本地服务器压力过大,从而造成响应性能恶化[7]。联合使用本地服务器与 IaaS 云数据中心,形成混合云计算模式。SaaS 提供商拥有本地小型数据中心,同时还可以从 IaaS 云数据中心获取基础设施资源。考虑到云计算环境中用户 QoS 需求的多样性,引入区分服务机制。将用户请求划分为延迟敏感请求和差错敏感请求,对应地,将 IaaS 数据中心划分为具有速率调整机制的快速云和具有差错恢复机制的可靠云。本文提出的新型的融合区分服务和速率调整的请求调度策略的工作原理如图 1 所示。

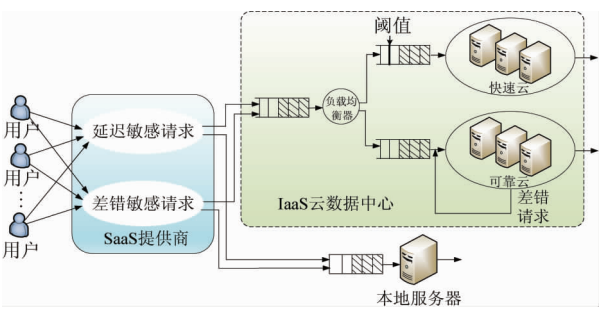

图 1 请求调度策略工作原理

SaaS 提供商拥有 1 台本地服务器。所有用户请求,包括延迟敏感请求和差错敏感请求,首选本地服务器接受服务。由于 SaaS 提供商的本地服务器处理能力有限,通过设置有限系统容量对到达系统的请求进行接入控制。无法进入本地服务器的请求将分流到 IaaS 云数据中心中。

IaaS 云数据中心中的负载均衡器实际上是 1 个虚拟机。根据 IP(Internet protocol)数据包头的服务类型(type of service, ToS)字段,负载均衡器将延迟敏感请求调度至快速云中,差错敏感请求调度至可靠云中。

快速云中的虚拟机具有速率调整机制。根据当前系统的工作负载动态设置处理器的电压和频率,调整虚拟机运行速度,实现能耗和性能的平衡。当快速云系统中的延迟敏感请求数低于或等于某一阈值时,虚拟机处于低速运行状态,以减少能源浪费。当快速云系统中的延迟敏感请求数超过该阈值时,虚拟机处于高速运行状态,从而提高用户的响应性能。

可靠云中的虚拟机具有差错恢复机制。用户请求可能由于信道噪声等影响造成处理失败或传输错误。针对差错敏感请求引入了差错恢复机制,以保证请求成功处理。例如在文件传输过程中,为了检测差错敏感请求处理正确与否,需要在链路层的数据帧中插入一个校验和字段,并对每个请求进行备份。当传输正确时,备份便在传输结束后丢弃。当传输错误时,重新传输备份数据,从而达到差错恢复的目的。

1.2 系统模型

系统遵循先到先服务原则。无法进入 SaaS 提供商本地服务器的用户请求由负载均衡器送入 IaaS 云数据中心。延迟敏感请求分流至快速云,差错敏感请求分流至可靠云,从而构成了一种新型混合云模式。

假设整个系统用户请求的到达是参数为 $\lambda(\lambda > 0)$ 的泊松过程,本地服务器及负载均衡器对一个请求的处理时间分别服从参数为 $\mu_0(\mu_0 > 0)$ 和 $\mu_1(\mu_1 > 0)$ 的指数分布;快速云中虚拟机在低速和高速运行状态对一个请求的处理时间分别服从参数为 $\mu_2^L(\mu_2^L > 0)$ 和 $\mu_2^H(\mu_2^H > 0)$ 的指数分布 $(\mu_2^H \geqslant \mu_2^L)$,可靠云中虚拟机对一个请求的处理时间服从参数为 $\mu_3(\mu_3 > 0)$ 的指数分布。在所提出的策略下,到达 SaaS 提供商本地服务器的请求及到达 IaaS 云数据中心的请求均服从泊松过程。假设 SaaS 提供商本地服务器的系统容量(简称本地系统容量)为 K,差错敏感请求占用户请求总数的比例为 $p(0 \leqslant p \leqslant 1, \bar{p} = 1 - p)$,快速云及可靠云中虚拟机数分别为 c_1 和 c_2,快速云中虚拟机的速率切换阈值为 $\alpha(\alpha > c_1 + 1)$,差错敏感请求一次处理失败的概率为 $\theta(0 \leqslant \theta \leqslant 1, \bar{\theta} = 1 - \theta)$。

对于 SaaS 提供商的本地服务器,建立了有限缓存 M/M/1/K 排队。对于 IaaS 云数据中心的负载均衡器,建立无限缓存 M/M/1/K 排队。根据 Burke 定理[8],M/M/1 排队的输出过程仍是与输入过程具有

相同参数的泊松过程。负载均衡器的输出分流至快速云和可靠云,分别构成其输入流。对于快速云,建立服务速率可变的无限缓存 M/M/ c_1 排队。对于可靠云,建立带反馈的无限缓存 M/M/ c_2 排队。基于上述 4 个排队系统,构造混合排队网络系统模型。

2 模型解析

令 $N(t)$ 表示 t 时刻 SaaS 提供商本地服务器系统中的请求数量。$N(t)$ 的稳态分布 P_i 定义为

$$P_i = \lim_{t\to\infty} P\{N(t) = i\},\ i = 0,1,\cdots,K \tag{1}$$

令 $X(t)$, $Y(t)$ 和 $Z(t)$ 分别表示 t 时刻 IaaS 云数据中心中负载均衡器、快速云和可靠云中的请求数量。令 $\pi_{n_1}^{(1)}$、$\pi_{n_2}^{(2)}$ 和 $\pi_{n_3}^{(3)}$ 分别表示 $X(t)$、$Y(t)$ 和 $Z(t)$ 的稳态分布,则有

$$\begin{aligned}
\pi_{n_1}^{(1)} &= \lim_{t\to\infty} P\{X(t) = n_1\},\ n_1 = 0,1,\cdots \\
\pi_{n_2}^{(2)} &= \lim_{t\to\infty} P\{Y(t) = n_2\},\ n_2 = 0,1,\cdots \\
\pi_{n_3}^{(3)} &= \lim_{t\to\infty} P\{Z(t) = n_3\},\ n_3 = 0,1,\cdots
\end{aligned} \tag{2}$$

2.1 稳态条件

本文建立的混合排队网络系统模型由 4 个排队系统构成。结合排队论相关知识,逐个分析各系统模型的流通强度。

基于 SaaS 提供商本地服务器建立的有限缓存 M/M/1/K[9] 排队的流通强度 ρ_0 表示为

$$\rho_0 = \frac{\lambda}{\mu_0} \tag{3}$$

到达 IaaS 云数据中心的用户请求是被 SaaS 提供商本地服务器阻塞的用户请求。则 IaaS 云数据中心中负载均衡器的到达率为 $P_K\lambda$。基于 IaaS 云数据中心中的负载均衡器建立的 M/M/1[9] 排队的流通强度 ρ_1 表示为

$$\rho_1 = \frac{P_K\lambda}{\mu_1} \tag{4}$$

其中, P_K 为本地服务器的阻塞率,即系统中有 K 个请求的概率, $P_K = \dfrac{\rho_0^K(1-\rho_0)}{1-\rho_0^{K+1}}$。

已知延迟敏感请求占总用户请求的比例为 $\bar{p}$,则快速云的到达率为 $P_K\lambda\bar{p}$。基于 IaaS 云数据中心中快速云建立的服务速率可变的 M/M/c_1 排队的流通强度 ρ_2 表示为

$$\rho_2 = \frac{P_K\lambda\bar{p}}{c_1\mu_2^L} \tag{5}$$

已知差错敏感请求占总用户请求的比例为 p,则可靠云的到达率为 $P_K\lambda p$。可靠云中虚拟机对一个请求的实际处理时间服从参数为 $\bar{\theta}\mu_3$ 的指数分布。基于 IaaS 云数据中心中可靠云建立的带反馈的 M/M/ c_2 排队的流通强度 ρ_3 表示为

$$\rho_3 = \frac{P_K\lambda p}{c_2\bar{\theta}\mu_3} \tag{6}$$

为确保整个混合排队网络达到稳态,需要 IaaS 云数据中心中每个具有无限缓存的排队系统均达到稳态。系统的稳态条件为

$$\max\{\rho_1, \rho_2, \rho_3\} < 1 \tag{7}$$

2.2 稳态分布

在具有速率调整机制的快速云中,阈值 α 控制虚拟机的速率调整。在系统稳态的条件下,建立系统平衡方程如下:

$$\begin{cases}
P_K\lambda\bar{p}\,\pi_0^{(2)} = \mu_2^L\,\pi_1^{(2)} \\
(P_K\lambda\bar{p} + n_2\mu_2^L)\,\pi_{n_2}^{(2)} = P_K\lambda\bar{p}\,\pi_{n_2-1}^{(2)} + (n_2+1)\mu_2^L\,\pi_{n_2+1}^{(2)}, \\
\qquad\qquad n_2 = 1,2,\cdots,c_1 \\
(P_K\lambda\bar{p} + c_1\mu_2^L)\,\pi_{n_2}^{(2)} = P_K\lambda\bar{p}\,\pi_{n_2-1}^{(2)} + c_1\mu_2^L\,\pi_{n_2+1}^{(2)}, \\
\qquad\qquad n_2 = c_1+1, c_1+2,\cdots,\alpha-1 \\
(P_K\lambda\bar{p} + c_1\mu_2^L)\,\pi_{\alpha}^{(2)} = P_K\lambda\bar{p}\,\pi_{\alpha-1}^{(2)} + c_1\mu_2^H\,\pi_{\alpha+1}^{(2)} \\
(P_K\lambda\bar{p} + c_1\mu_2^H)\,\pi_{n_2}^{(2)} = P_K\lambda\bar{p}\,\pi_{n_2-1}^{(2)} + c_1\mu_2^H\,\pi_{n_2+1}^{(2)}, \\
\qquad\qquad n_2 = \alpha+1, \alpha+2,\cdots
\end{cases} \tag{8}$$

结合归一化条件 $\sum_{n_2=0}^{\infty}\pi_{n_2}^{(2)} = 1$, 可得到服务速率可变的 M/M/ c_1 排队系统的稳态分布:

$$\pi_{n_2}^{(2)} = \begin{cases}
\dfrac{\pi_0^{(2)}}{n_2!}\left(\dfrac{P_K\lambda\bar{p}}{\mu_2^L}\right)^{n_2}, & n_2 = 1,2,\cdots,c_1 \\
\dfrac{\pi_0^{(2)}}{c_1!}\left(\dfrac{P_K\lambda\bar{p}}{\mu_2^L}\right)^{c_1}\left(\dfrac{P_K\lambda\bar{p}}{c_1\mu_2^L}\right)^{n_2-c_1}, & n_2 = c_1+1, c_1+2,\cdots,\alpha \\
\dfrac{\pi_0^{(2)}}{c_1!}\left(\dfrac{P_K\lambda\bar{p}}{\mu_2^L}\right)^{c_1}\left(\dfrac{P_K\lambda\bar{p}}{c_1\mu_2^L}\right)^{\alpha-c_1}\left(\dfrac{P_K\lambda\bar{p}}{c_1\mu_2^H}\right)^{n_2-\alpha}, & n_2 = \alpha+1, \alpha+2,\cdots
\end{cases} \tag{9}$$

其中,

$$\pi_0^{(2)} = \left(1 + \left(\sum_{n_2=1}^{c_1}\frac{1}{n_2!}\left(\frac{P_K\lambda\bar{p}}{\mu_2^L}\right)^{n_2} + \sum_{n_2=c_1+1}^{\alpha}\frac{1}{c_1!}\left(\frac{P_K\lambda\bar{p}}{\mu_2^L}\right)^{c_1}\left(\frac{P_K\lambda\bar{p}}{c_1\mu_2^L}\right)^{n_2-c_1} + \sum_{n_2=\alpha+1}^{\infty}\frac{1}{c_1!}\left(\frac{P_K\lambda\bar{p}}{\mu_2^L}\right)^{c_1}\left(\frac{P_K\lambda\bar{p}}{c_1\mu_2^L}\right)^{\alpha-c_1}\left(\frac{P_K\lambda\bar{p}}{c_1\mu_2^H}\right)^{n_2-\alpha}\right)\right)^{-1} \tag{10}$$

具有差错恢复机制的可靠云的到达率为 $P_K\lambda p$,实际服务率为 $\bar{\theta}\mu_3$。基于可靠云建立的带反馈的 M/M/c_2排队系统稳态分布表示为

$$\pi_{n_3}^{(3)} = \begin{cases} \dfrac{(c_2\rho_3)^{n_3}\pi_0^{(3)}}{n_3!}, & n_3 = 1,2,\cdots,c_2 \\ \dfrac{c_2^{c_2}\rho_3^{n_3}\pi_0^{(3)}}{c_2!}, & n_3 = c_2+1, c_2+2,\cdots \end{cases} \tag{11}$$

其中, $\pi_0^{(3)} = \left(\sum_{n_3=0}^{c_2}\frac{(c_2\rho_3)^{n_3}}{n_3!} + \frac{\rho_3(c_2\rho_3)^{c_2}}{c_2!(1-\rho_3)}\right)^{-1}$。

3 系统性能指标

关注延迟敏感请求的平均响应时间、快速云中虚拟机的节能水平和可靠云中虚拟机的利用率,评估融合区分服务和速率调整的请求调度策略的性能指标。

定义一个用户请求的响应时间为其到达 SaaS 提供商至处理结束离开系统的总时间。

令 $E[T_0]$ 表示服务于 SaaS 提供商本地服务器的延迟敏感请求的平均逗留时间。由 Little 公式[10], $E[T_0]$ 的表达式为

$$E[T_0] = \frac{\rho_0 - (K+1)\rho_0^{K+1} + K\rho_0^{K+2}}{\lambda(1-P_K)(1-\rho_0)(1-\rho_0^{K+1})} \tag{12}$$

令 $E[T_1]$ 和 $E[T_2]$ 分别表示调度至 IaaS 云数据中心的延迟敏感请求在负载均衡器和快速云的平均逗留时间。$E[T_1]$ 和 $E[T_2]$ 的表达式分别为

$$E[T_1] = \frac{1}{\mu_1(1-\rho_1)},\ E[T_2] = \frac{1}{P_K\lambda\bar{p}}\sum_{n_2=0}^{\infty}n_2\pi_{n_2}^{(2)} \tag{13}$$

其中, $\pi_{n_2}^{(2)}$ 表示服务速率可变的 M/M/ c_1 排队系统的稳态分布,由式(9)给出。

延迟敏感请求既可以在本地服务器也可以在快速云中处理。延迟敏感请求的平均响应时间 $E[T_{ds}]$ 的表达式为

$$E[T_{ds}] = (1-P_K)E[T_0] + P_K(E[T_1] + E[T_2]) \tag{14}$$

定义快速云中虚拟机节能水平 S 为虚拟机处于低速运行状态而降低的能量消耗。虚拟机的节能水平与虚拟机处于低速运行状态的概率及快速云中虚拟机速率差值有关。令快速云中虚拟机的速率差值 φ 为高速 μ_2^H 与低速 μ_2^L 的差,即 $\varphi = \mu_2^H - \mu_2^L$。快速云中虚拟机节能水平 S 的表达式为

$$S = f_s\varphi\sum_{n_2=1}^{\alpha}\pi_{n_2}^{(2)} \tag{15}$$

其中, f_s 为节能水平参数,表示虚拟机每降低一个速率单位所节省的能耗。

定义可靠云中虚拟机的利用率 U 为虚拟机被占用的概率。当系统中的请求数小于或等于虚拟机数时,虚拟机的利用率为 n_3/c_2; 当系统中的请求数大于虚拟机数时,虚拟机的利用率为 1。可靠云中虚拟机的利用率 U 表达式为

$$U = \sum_{n_3=1}^{c_2}\frac{n_3\pi_{n_3}^{(3)}}{c_2} + \sum_{n_3=c_2+1}^{\infty}\pi_{n_3}^{(3)} \tag{16}$$

其中, $\pi_{n_3}^{(3)}$ 表示带反馈的 M/M/ c_2 排队系统的稳态分布,由式(11)给出。

4 数值与仿真实验

在融合区分服务和速率调整的请求调度策略下,进行数值实验及仿真实验,量化系统性能的变化趋势。Matlab 具有高效的数值计算功能,Java 语言是一种面向对象的编程语言,可以通过类的定义表示请求的类别、等待及执行等多种属性。理论分析实验在 Matlab 环境下运行,仿真统计实验在 MyEclipse2014 环境下使用 Java 语言运行。进行实验的计算机硬件环境为: Intel(R) Core(TM), i7-4790 CPU @ 3.60GHz,8.00GB RAM。

在系统保持稳态的前提下,设定系统参数如下,系统到达率为 λ =18.0个/s,SaaS 提供商本地服务器的服务率为 μ_0 =5.0个/s,IaaS 云数据中心中负载均衡器的服务率为 μ_1 =14.0个/s,快速云中虚拟

机低速状态的服务率为 μ_2^H = 2.8 个/s，可靠云中虚拟机的服务率为 μ_3 = 0.8 个/s。差错敏感请求占总请求的比例 p = 0.4，快速云中虚拟机数量为 c_1 = 10 台，可靠云中虚拟机数量为 c_2 = 10 台。在满足式(7)表示的稳态条件下，参数设置不会影响实验结果的趋势。

针对不同的本地系统容量 K 和速率切换阈值 α，延迟敏感请求的平均响应时间 $E[T_{ds}]$ 随速率差值 φ 的变化趋势如图 2 所示。

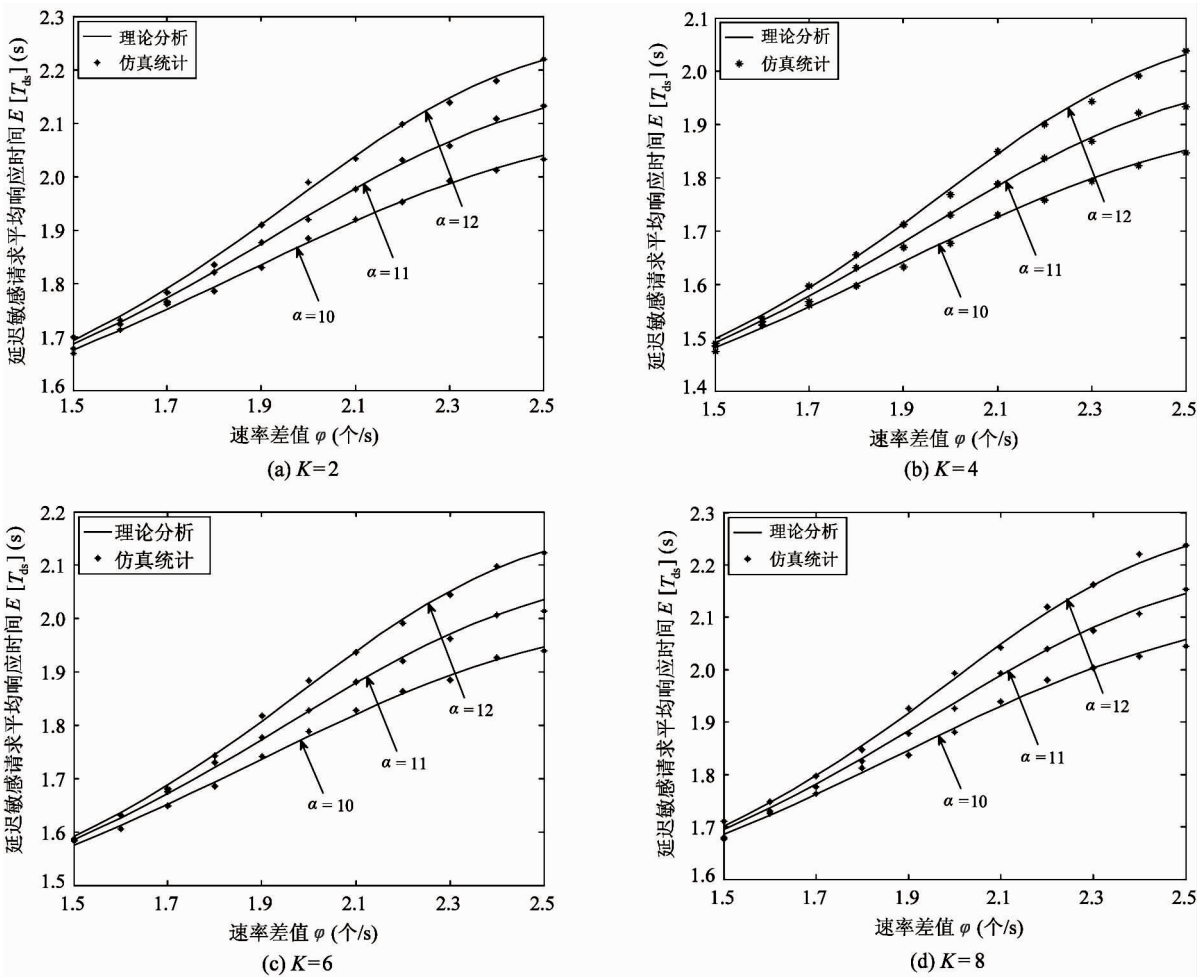

图 2 延迟敏感请求平均响应时间的变化趋势

给定本地系统容量 K，分析速率差值 φ 及速率切换阈值 α 对延迟敏感请求平均响应时间 $E[T_{ds}]$ 的影响。由图 2(a) ~ (d)可以看出，当速率差值一定时，平均响应时间随着速率切换阈值的增大呈现上升趋势。速率切换阈值越大，快速云中虚拟机处于高速运行状态时间越短，更多请求在虚拟机低速状态下处理，因而平均响应时间增大。当速率切换阈值一定时，平均响应时间随速率差值的增大呈现上升趋势。速率差值越大，虚拟机在低速运行状态下的服务率越低，请求的平均响应时间随之增大。

给定速率差值 φ 和速率切换阈值 α，分析本地系统容量 K 对延迟敏感请求平均响应时间 $E[T_{ds}]$ 的影响。由图 2(a)和(b)可知，当本地系统容量较小时，平均响应时间随着本地系统容量的增大而呈现下降趋势。当本地系统容量较小时，随着系统容量的增大，本地服务器的阻塞率降低，调度至本地服务器上的请求较多。本地服务器有较好的处理速度，而且此时的服务资源是充足的，所以平均响应时间会随着本地系统容量的增大而减小。由图 2(c)和(d)可知，当本地系统容量较大时，平均响应时间

随本地系统容量的增大而呈现上升趋势。当本地系统容量较大时,本地服务器的阻塞率逐渐变成一个较小的值,使得大量任务涌入本地服务器,造成服务资源匮乏,从而增加了请求的等待时间,所以平均响应时间会随着本地系统容量的增大而提高。对于每一个速率差值及速率切换阈值,都存在一个最优的本地系统容量,使延迟敏感请求的平均响应时间达到最小值。

设定节能水平参数 f_s = 1mJ/个/s,针对不同的本地系统容量 K 和速率切换阈值 α,快速云中虚拟机节能水平 S 随速率差值 φ 的变化趋势如图 3 所示。

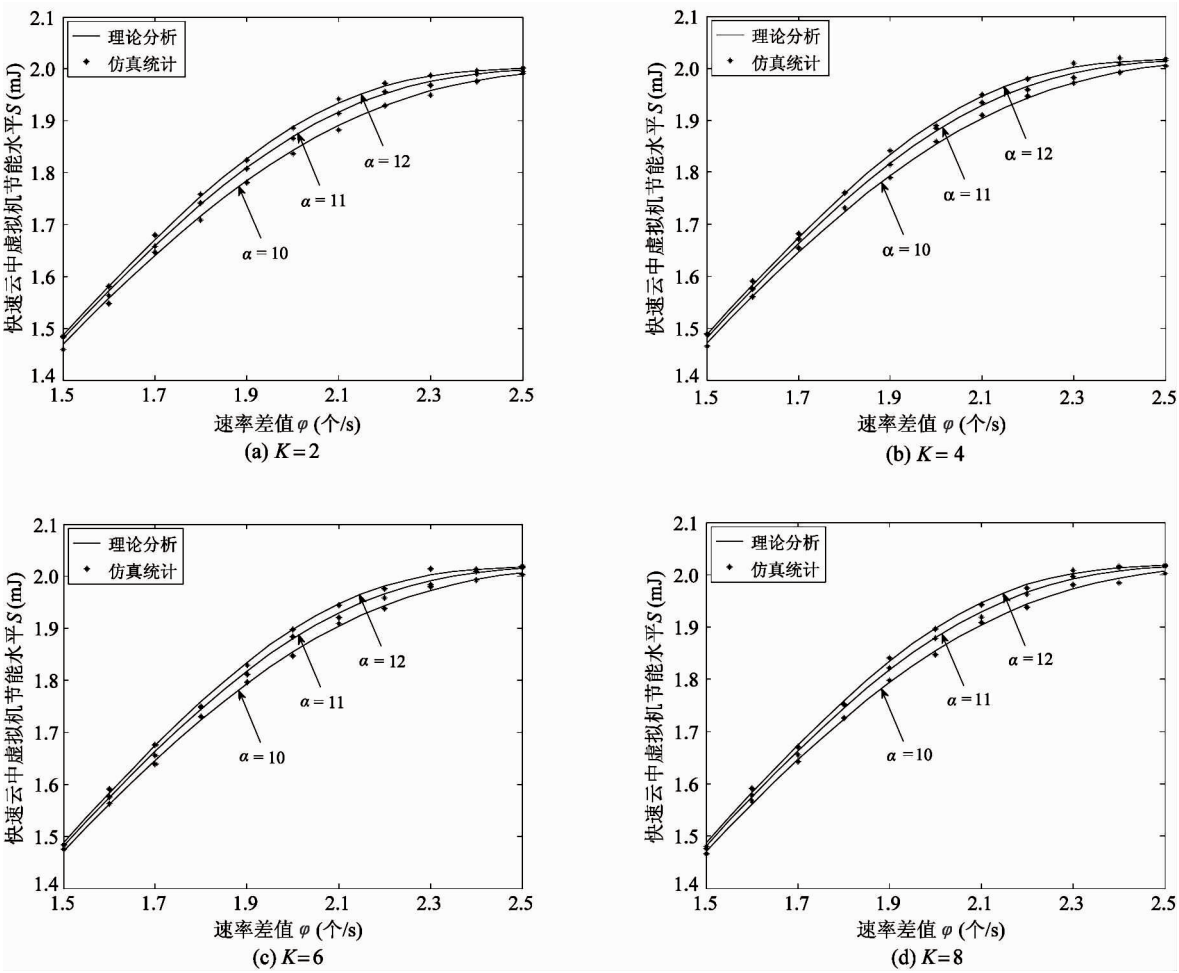

图 3　快速云中虚拟机节能水平的变化趋势

给定本地系统容量 K,分析速率差值 φ 及速率切换阈值 α 对快速云中虚拟机节能水平 S 的影响。由图 3(a) ~ (d)可以看出,当速率差值一定时,节能水平随着速率切换阈值的增大呈现上升趋势。速率切换阈值越大,快速云中虚拟机处于低速运行状态时间越长,快速云中虚拟机节能水平越高。当速率切换阈值一定时,快速云中虚拟机节能水平随速率差值增大呈现出上升趋势。速率差值越大,虚拟机在低速运行状态下的服务率越低,节能水平随之增大。

给定速率差值 φ 和速率切换阈值 α,分析本地系统容量 K 对快速云中虚拟机的节能水平 S 的影响。由图 3(a)和(b)可知,当本地系统容量较小时,节能水平随本地系统容量的增大呈现小幅上升趋势。随着本地系统容量的增大,本地服务器的阻塞率降低,调度至快速云的请求减少,虚拟机处于低速运行状态时间增加,快速云中虚拟机节能水平随之提高。由图 3(c)和(d)可知,随着本地系统容量

不断增大,本地服务器的阻塞率逐渐变成一个比较小的定值,调度至快速云的请求数量基本保持不变,因而,快速云中虚拟机节能水平基本保持不变。

针对在不同的差错敏感请求的错误率 θ, 本地系统容量 K 对可靠云中虚拟机利用率 U 的影响如图 4 所示。

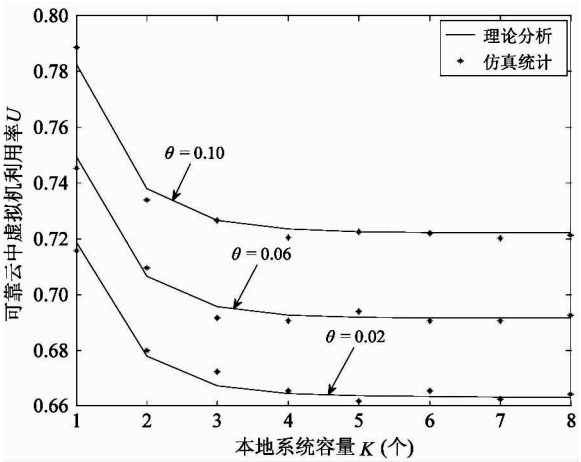

图 4 可靠云中虚拟机利用率的变化趋势

由图 4 可以看出,当本地系统容量一定时,可靠云中虚拟机利用率随着错误率的增大呈现上升趋势。错误率的增大会使反馈到可靠云中重新接受服务的请求增多,从而提高了虚拟机的利用率。当错误率一定时,虚拟机利用率随着本地系统容量的增大呈现出下降趋势。本地系统容量越大,在本地服务器的请求越多,而调度到可靠云中的请求越少,虚拟机利用率随之减少。

以上的实验结果表明,快速云中虚拟机速率差值及本地系统容量对系统性能有着不容忽视的影响。快速云中虚拟机速率差值的增大会引起延迟敏感请求的平均响应时间的增加,同时快速云中虚拟机节能水平也有所增加。另一方面,本地系统容量对平均响应时间的影响是复杂的。当本地系统容量较小时,其增大会引起平均响应时间的减小,而当本地系统容量较大时,其增大会引起平均响应时间的增大。对应于最小的平均响应时间的本地系统容量的增大会引起快速云中虚拟机节能水平的增加及可靠云中虚拟机利用率的降低。在混合云计算模式下,需要对快速云中虚拟机速率差值及本地系统容量进行联合优化。

5 系统优化

为了权衡系统中平均响应时间、节能水平及利用率三者之间的折中关系,构造系统成本函数 F 为:

$$F = R_1 E[T_{ds}] - R_2 S - R_3 U \tag{17}$$

函数 F 的参数说明如下:

(1) 令 R_1 表示延迟敏感平均响应时间对系统成本的影响因子。

(2) 令 R_2 表示快速云中虚拟机节能水平对系统成本的影响因子。

(3) 令 R_3 表示可靠云中虚拟机利用率对系统成本的影响因子。

为了更好地探究系统成本函数的变化规律,进行数值实验。沿用第 4 节给出的实验参数,并设定错误率 $\theta = 0.02$, 系统成本影响因子 $R_1 = 1.0$, $R_2 = 0.5$, $R_3 = 0.5$。针对不同的本地系统容量 K 和速率切换阈值 α, 系统成本函数 F 随着速率差值 φ 的变化趋势如图 5 所示。

给定本地系统容量 K, 分析速率差值 φ 对系统成本 F 的影响。由图 5(a) ~ (d)可以看出,系统成本随着速率差值的增大呈现出先下降后上升的趋势。当速率差值较小时,随着速率差值的增大,快速云中虚拟机节能水平的上升趋势强于延迟敏感请求平均响应时间的上升趋势,快速云中虚拟机节能水平成为影响系统成本的主导因素,因此系统成本呈现下降趋势。当速率差值较大时,随着速率差值的增大,延迟敏感请求平均响应时间的上升趋势强于快速云中虚拟机节能水平的上升趋势,延迟敏感请求平均响应时间成为影响系统成本的主导因素,因此系统成本呈现上升趋势。

给定速率差值 φ, 分析本地系统容量 K 对系统成本 F 的影响。由图 5(a)和(b)可知,当本地系统容量较小时,系统成本随着本地系统容量的增大呈现下降趋势。当本地系统容量较小时,随着本地系统容量的增大,延迟敏感请求平均响应时间的下降趋势及快速云中虚拟机节能水平的上升趋势强于可

靠云中虚拟机利用率的下降趋势。延迟敏感请求平均响应时间及快速云中虚拟机节能水平成为影响系统成本的主导因素,因此系统成本呈现下降趋势。由图5(c)和(d)可知,当本地系统容量较大时,系统成本随着本地系统容量的增大呈现上升趋势。当本地系统容量较大时,随着本地系统容量的增大,延迟敏感请求平均响应时间增大,而快速云中虚拟机节能水平及可靠云中虚拟机利用率基本保持不变。延迟敏感请求平均响应时间成为影响系统成本的主导因素,因此系统成本呈现上升的趋势。

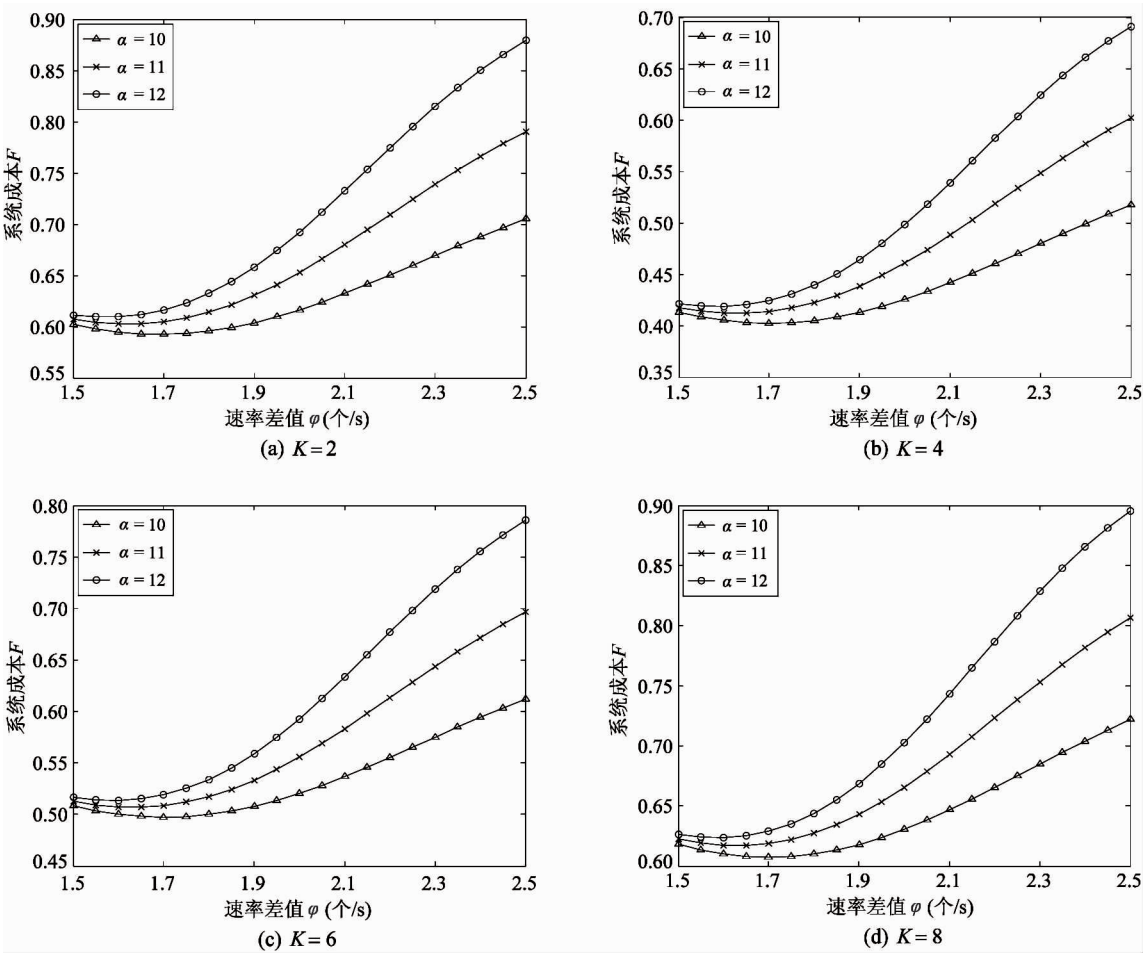

图5　系统成本的变化趋势

由于延迟敏感请求平均响应时间 $E[T_{ds}]$、快速云中虚拟机节能水平 S 及可靠云中虚拟机利用率 U 均是非线性的,系统成本函数 F 的严格单调性无法确定,基于数值分析的优化结果很难给出。

智能优化算法为解决复杂问题提供了新的思路与手段。为了加快算法搜索速度,利用混沌方程进行种群初始化,改进Jaya算法[11]求解速率差值及本地系统容量的最优组合。算法的主要步骤如下所示。

步骤1　初始化最大迭代次数 $iter_{max}$,混沌方程参数 b(以 $b=4$ 为例),种群规模 N,速率差值上限 φ_{up} 及下限 φ_{low},缓存容量上限 K_{up} 及下限 K_{low}。初始化本地系统容量为 $K=K_{low}$,迭代次数为 $iter=0$。

步骤2　利用正弦混沌方程初始化种群中各成员 $(\varphi, K)_i$, $i \in \{1,2,\cdots,N\}$。

$\varphi_1 = rand\,\%$, $rand$ 是(0,1)内的随机数

$$\varphi_{i+1} = \frac{b}{4}\sin(\pi\varphi_i)$$

$K_i = K$

步骤3　修正初始速率差值 $\varphi_i \in [\varphi_{low}, \varphi_{up}]$, $i \in \{1,2,\cdots,N\}$。

$$\varphi_i = \frac{\varphi_i - \min\limits_{j\in\{1,\cdots,N\}}\{\varphi_j\}}{\max\limits_{j\in\{1,\cdots,N\}}\{\varphi_j\} - \min\limits_{j\in\{1,\cdots,N\}}\{\varphi_j\}} \times (\varphi_{\text{up}} - \varphi_{\text{low}})$$

步骤 4 计算最优组合 $(\varphi, K)_{\text{best}}$ 与最差组合 $(\varphi, K)_{\text{worst}}$。

$$F((\varphi, K)_i) = R_1 E[T_{ds}] - R_2 S - R_3 U,$$
$$i \in \{1,2,\cdots,N\}$$
$$(\varphi, K)_{\text{best}} = \underset{i\in\{1,\cdots,N\}}{\text{argmin}}\{F((\varphi, K)_i)\}$$
$$(\varphi, K)_{\text{worst}} = \underset{i\in\{1,\cdots,N\}}{\text{argmax}}\{F((\varphi, K)_i)\}$$

步骤 5 更新种群中各成员的速率差值 φ_i, $i \in \{1,2,\cdots,N\}$。

$$\varphi_i^u = rand \times (\varphi_{\text{best}} - \varphi_i) - rand \times (\varphi_{\text{worst}} - \varphi_i)$$
$$+ \varphi_i \times \left(1 + \exp\left(-\left(\frac{F((\varphi, K)_i)}{\max\limits_{j\in\{1,\cdots,N\}}\{F((\varphi, K)_j)\}}\right)^{\text{iter}}\right)\right)^{-1}$$

if $F((\varphi, K)_i) > F((\varphi_i^u, K_i))$

then $(\varphi, K)_i = (\varphi_i^u, K_i)$

endif

步骤 6 找出局部最小成本。

$iter = iter + 1$。

if $iter \leqslant iter_{\max}$

then 跳转到**步骤 4**

else $F((\varphi, K)_q^{\text{T}}) = \min\limits_{i\in\{1,\cdots,N\}}\{F((\varphi, K)_i)\}$

$q = q + 1$

endif

步骤 7 找出全局最小成本。

$K = K + 1$。

if $K \leqslant K_{up}$

then 跳转到**步骤 2**

else $(\varphi, K)^* = \underset{s\in\{1,\cdots,q\}}{\text{argmin}}\{F((\varphi, K)_s^{\text{T}})\}$

endif

步骤 8 输出最优参数组合 $(\varphi, K)^*$。

利用改进的 Jaya 算法,数值求解不同速率切换阈值下系统最小成本,给出快速云中虚拟机速率差值与本地系统容量优化组合。优化结果如表 1 所示。

表 1 虚拟机速率差值与本地系统容量的联合优化结果

速率切换阈值 α	虚拟机速率差值与本地系统容量的最优组合 $(\varphi, K)^*$	最小系统成本 F^*
10	(1.6990, 3)	0.3978
11	(1.6277, 3)	0.4076
12	(1.5818, 3)	0.4142

6 结 论

权衡用户的响应性能、系统的节能水平与云资源利用率进行请求调度是云计算技术中的关键问题。本文面向混合云计算模式提出了一种新型的融合区分服务及速率调整的请求调度策略。通过构建混合排队网络系统模型,给出了延迟敏感请求平均响应时间、快速云中虚拟机节能水平及可靠云中虚拟机利用率等系统性能指标。结合理论分析实验及仿真统计实验,刻画速率切换阈值、速率差值和本地系统容量对系统性能的影响,揭示了不同性能指标之间的折中关系。改进智能寻优算法,给出速率差值及本地缓存容量的联合优化方案,实现了系统成本的最小化。

参考文献

[1] Lin F, Zhou X, Huang D, et al. Service scheduling in cloud computing based on queuing game model[J]. *KSII Transactions on Internet and Information Systems*, 2014, 8(5): 1554-1566

[2] Huang Q. Development of a SaaS application probe to the physical properties of the Earth's interior: an attempt at moving HPC to the cloud[J]. *Computers and Geosciences*, 2014, 70: 147-153

[3] Li S, Zhou Y, Jiao L, et al. Towards operational cost minimization in hybrid clouds for dynamic resource provisioning with delay-aware optimization[J]. *IEEE Transactions on Services Computing*, 2015, 8(3): 398-409

[4] 黄旭, 吴小红, 马小龙. 基于排队论的云服务分流博弈及均衡分析[J]. 兰州理工大学学报, 2015, 41(3): 96-101

[5] Nan X, He Y, Guan L. Towards optimal resource allocation for differentiated multimedia services in cloud computing environment[C]. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2014. 684-688

[6] Ding J, Zhang Z, Richard T B M, et al. Auction-based cloud service differentiation with service level objectives [J]. *Computer Networks*, 2016, 94(C): 231-249

[7] Madni S H H, Latiff M S A, Coulibaly Y, et al. Resource scheduling for infrastructure as a service (IaaS) in cloud computing: Challenges and opportunities[J]. *Journal of Network and Computer Applications*, 2016, 68 (C): 173-200

[8] Ibe O C. Fundamentals of Stochastic Networks[M]. New York: Wiley, 2011

[9] 孙荣恒,李建平. 排队论基础[M].北京:科学出版社, 2015

[10] 金顺福,肖玉鹏,赵媛. 认知无线网络频谱分配策略的社会最优问题研究[J]. 高技术通讯, 2014, 24(4): 379-385

[11] Rao R V. Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems[J]. *International Journal of Industrial Engineering Computations*, 2016, 7(1): 19-34

A request scheduling strategy integrating differentiation service and speed adjustment

Jin Shunfu[*,**,***], Xue Yuanzheng[*,**,***], Gu Rui[*,**,***]

([*]School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004)

([**]Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao 066004)

([***]Key Laboratory for Software Engineering of Hebei Province, Qinhuangdao 066004)

Abstract

Considering the difference of heterogeneous user demands for quality of service (QoS) and the energy consumption of cloud data centers, a request scheduling strategy integrating differentiation service and speed adjustment is proposed. The strategy divides an IaaS (infrastructure as a service) cloud data center into the rapid cloud with speed adjustment and the reliable cloud with fault restoration, then a novel hybrid cloud computing paradigm is investigated. Accordingly, a system model based hybrid queueing network is established. From the view of the average response time of delay sensitive requests, the energy saving level of virtual machines in rapid clouds and the utilization of virtual machines in reliable clouds, the system performance of the request scheduling strategy is estimated. The effectiveness of the proposed strategy and the rationality of the established system model are verified by experiments. By constructing a system cost function and improving the Jaya intelligent searching algorithm, the proposed request scheduling strategy is optimized.

Key words: cloud computing, request scheduling, differentiation service, speed adjustment, queueing network, intelligent optimization