

一种通用云计算资源调度问题的快速近似算法

魏蔚 刘扬 杨卫东

(河南工业大学信息科学与工程学院 郑州 450001)

(nsyncw@126.com)

A Fast Approximation Algorithm for the General Resource Placement Problem in Cloud Computing Platform

Wei Wei, Liu Yang, and Yang Weidong

(College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001)

Abstract There are regionally distributed demands for various resources in cloud based large-scale online services. Given fixed resource budget, the service providers need to decide where to place resources to satisfy massive demands from all regions, where demands are usually represented by mean value in given time span. However, in scenarios with a large number of resources, demands are dynamic and stochastic, considering fine-grained demands and adopting stochastic model will further improve resource utilization. Compared with mean demand-based algorithm, considering demand stochasticity in algorithm will increase resource utilization ratio, but also leads to high time complexity. The time complexity of optimal algorithm is linear to total amount of resources, thus may be inefficient when dealing with a large number of resources. Based on nonlinear programming theory, we propose Fast Resource Placement (FRP), an effective resource placement method of high efficiency. In the algorithm, optimal solution is represented by continuous functions of input, and we construct approximation functions to reduce the computation complexity. The preliminary experiments show that in scenarios with general settings, compared with optimal algorithm, FRP can reduce the computation time by three orders of magnitude, and can achieve 99% effect of optimal solution. Therefore, FRP can be used to schedule large number of resources efficiently in time-tense scheduling scenarios.

Key words stochastic demand; resource scheduling; resource placement; nonlinear programming; cloud computing

摘要 在分布式云计算平台中,面向大规模用户的在线应用需处理针对海量资源的用户需求,在给定的资源预算下,服务提供商需确定最优资源放置位置,以最大程度地满足用户需求,通常需求用给定时间段内均值表示。然而真实场景中用户需求是高度动态和随机的,采用随机需求模型以考虑更多需求细

收稿日期:2014-12-08;修回日期:2015-04-08

基金项目: 国家自然科学基金项目(U1504607,61202099);河南省教育厅科学技术研究重点基金项目(2010A520008,13A413001,14A520018);河南省重点科技攻关基金项目(102102210025);新世纪优秀人才支持计划基金项目(NCET-12-0692);河南工业大学博士基金项目(2012BS011,2013BS003);河南工业大学自然科学基础研究重点培育计划基金项目(2014JCYJ04)

This work was supported by the National Natural Science Foundation of China (U1504607,61202099), the Key Science and Technology Research Project of Education Department of Henan Province (2010A520008,13A413001,14A520018), Henan Provincial Key Scientific and Technological Plan (102102210025), the Program for New Century Excellent Talents in University of Ministry of Education of China (NCET-12-0692), the Doctor Foundation of Henan University of Technology (2012BS011,2013BS003), and the Plan of Nature Science Fundamental Research in Henan University of Technology (2014JCYJ04).

节,资源利用率可得到进一步优化.但相比均值调度方法,随机需求模型会导致很高的计算复杂度.已有的最优解求解算法的时间复杂度和资源总量成正比,无法满足海量资源在线调度的效率要求.基于非线性规划理论,提出了一个快速资源分配算法,该算法可将计算复杂度降低至最优解算法的1%,并逼近最优解效果的99%,因此可用于在线应用场景中海量资源的高效调度.

关键词 随机需求;资源调度;资源放置;非线性规划;云计算

中图法分类号 TP301.6

大规模在线服务系统(如 YouTube^[1] 和 Facebook^[2])面临来自全球各个区域的高度动态的用户请求,可利用云计算平台降低费用并提升用户体验.在典型的分布式云计算平台中,需妥善分配资源到全球各个云数据中心以满足尽可能多的用户请求;且对于单个用户请求,倾向分发到较近的数据中心以提供较好的用户体验.

已有工作初步探讨了在上述场景中进行资源调度的若干问题.在流媒体分发的相关工作中,文献[3]深入分析了面向流处理服务的资源部署问题,考虑了经由中间节点响应客户请求的情况,采用集合覆盖等方法从多角度规约问题获取最优解.文献[4]分析了混合云中面向视频点播服务的资源配置问题,通过限制平均延时以保证服务质量,利用李雅普诺夫优化方法获得长时间范围内的统计平均最优.在内容分发网络研究方向上,文献[5]考虑利用空闲带宽分发容迟数据,通过延时调度充分复用带宽资源并减少平均分发时间,采用任务调度与资源整合相结合的思路.文献[6]中的 MetaCDN 架构将用户内容分发到多个不同的存储云中,最大化租用费用及服务质量相关的效用函数.文献[7]提出基于拉普洛夫优化理论的通用优化框架,在保证服务质量的前提下最小化操作费用.文献[8]研究了存储云中的内容存放问题,以及在存储云间的分发网络构建问题,并提出了多种启发式算法.文献[9]设计了一种结合树形和网状拓扑的新型混合分发架构,采用网络编码方法提高分发效率,保证服务质量同时可大幅降低骨干网流量.文献[10]提出了基于用户行为特征的资源分配策略,通过统计用户工作习惯与任务完成时间期望值的变化规律,建立用户行为特征信息表,从而动态调整云计算系统的资源分配策略.文献[11]引入用户效用的概念,建立了云环境中用户效用的描述模型,给出了用户对任务执行时间和费用满意程度的量化方法,基于线性规划理论提出了云环境中面向用户效用的任务调度优化模型.除了高度相关的云计算方向研究工作,P2P 领域中的

相关工作也研究了类似的资源调度问题^[12].

上述工作采用均值需求模型以保证调度算法效率,然而真实场景中用户需求是高度动态和随机的,通过考虑更多需求细节,采用随机需求模型,资源利用率可得到进一步优化.在均值模型中,资源需求通过给定调度时间间隔内资源需求量均值表示,参与运算的是代表均值的标量值;随机需求模型中,资源需求使用随机变量表示,参与运算的一般是需求的累积概率分布函数.文献[13-14]将基于随机需求的通用资源调度问题转化为最小费用最大流问题,并提出基于连续最短路径的最优解算法.其中,文献[13]的问题设定中,每区域有资源限制,目标是最大化满足的需求数量的期望值;文献[14]的问题设定中,没有区域资源限制,目标是资源费用和收益组合表示的整体代价最小化.但基于连续最短路径求解算法的复杂度与资源总量成正比例,无法满足大规模场景下实时快速调度的时间要求.

通过观察通用分配问题的内部结构,我们提出了快速近似算法,论文的创新点如下:1)证明该问题可转化为一个2阶段优化问题,可逐步快速求解;2)证明在每阶段的子问题中,最优结果均可表示为输入参数的函数,并给出函数的数学表达式;3)采用离散函数模拟算法中的连续函数,提出了快速近似算法.实验结果表明,该算法在达到最优解效果99%的情况下,时间复杂度可缩减至最优解算法的1%.

1 通用资源分配问题

设有来自 J 个区域的针对 I 种资源的用户请求,用 L_j^i 表示区域 j 中 i 型资源的数量,则 I 种资源到 J 个区域的1次分配称为1个资源配置方案,记为 $P=\{L_j^i\}$. 设 $L_i=\sum_{j=1}^J L_j^i$ 为 i 型资源的总量,称 $O=\{L_i\}$ 为类型配置方案,即只给出每类资源的总量.一个类型配置 O 可具体化到一组资源配置方案 P ,记为 $\{P\}^O$.

对来自区域 j 的对 i 型资源的请求,若该请求被分配给资源 i ,则该请求被标记为“已满足”.一个来自区域 j 的 i 型资源的请求被满足,则将它分配到:1)位于区域 j 的 i 型资源,称之为本地满足;或2)位于其他区域的 i 型资源,称之为远程满足.若请求没有被分配到任何资源,则将其标记为“未满足”.

对于随机需求,区域 j 中所有用户对 i 型资源的随机需求表示为随机变量 D_i^j ,所有地区的用户对 i 型资源的随机需求表示为随机变量 D_i ,累积分布函数表示为 $cdf_{i,j}$ 和 cdf_i . 给定具体的一次需求实例 d_i^j 和 d_i ,对于 i 型资源,设其本地满足需求数为 S_i^{loc} ,以及所有区域(本地和远程)满足需求的数量是 S_i^{sat} ,则有:

$$S_i^{\text{loc}} = \sum_{j=1}^J \min(L_i^j, d_i^j), \quad (1)$$

$$S_i^{\text{sat}} = \min(L_i, d_i), \quad (2)$$

则对于 i 型资源,其远程满足的需求数 $S_i^{\text{rem}} = S_i^{\text{sat}} - S_i^{\text{loc}}$. 对于一个资源配置方案 P ,我们定义其 i 型收益 R_i^P 为所有 i 型资源的总收益,表示为本地和远程满足的需求加权和:

$$R_i^P = w_{\text{sat}} S_i^{\text{sat}} + w_{\text{loc}} S_i^{\text{loc}}, \quad (3)$$

则方案 P 的 i 型收益期望值为

$$E(R_i^P) = w_{\text{sat}} E(\min(L_i, D_i)) + w_{\text{loc}} \sum_{j=1}^J E(\min(L_i^j, D_i^j)). \quad (4)$$

因对于任意连续的非负随机变量 x 与常量 C 有:

$$E(\min(C, x)) = \int_0^C (1 - cdf_x(c)) dc, \quad (5)$$

其中, $cdf_x(\cdot)$ 是 x 的累积概率分布函数,则式(4)中 i 型收益可转化为最终形式:

$$E(R_i^P) = w_{\text{sat}} \int_0^{L_i} (1 - cdf_i(n)) dn + w_{\text{loc}} \sum_{j=1}^J \int_0^{L_i^j} (1 - cdf_{i,j}(n)) dn, \quad (6)$$

则所有 i 型收益的期望值的总和为总期望收益:

$$E(R^P) = \sum_{i=1}^I E(R_i^P). \quad (7)$$

在通用的离线资源分配场景中,服务提供商通常会预订定量的全局资源(总量记为 U),有:

$$\sum_{i=1}^I \sum_{j=1}^J L_i^j = \sum_{i=1}^I L_i = U. \quad (8)$$

资源调度的目标是在给定的资源总量下找到获得最大化期望收益的资源调度方案 P^{opt} . 为获得问题的精确描述,我们引入以下记号:定义 $\{O\}^U$ 作为全局资源限制 U 下所有可能的类型配置方案集合.

首先对给定 O 下的所有 $P \in \{P\}^O$,定义 $\hat{E}_i^O$ 为最大的 i 型期望收益, $\hat{E}^O$ 为最大的总期望收益,则有:

$$\hat{E}_i^O = \max(E(R_i^P) \mid P \in \{P\}^O), \quad (9)$$

$$\hat{E}^O = \max(E(R^P) \mid P \in \{P\}^O), \quad (10)$$

则对于所有类型配置 $O \in \{O\}^U$,定义其中的最大收益 $\hat{E}^U$ 为所有 $\hat{E}^O$ 的最大值:

$$\hat{E}^U = \max(\hat{E}^O \mid O \in \{O\}^U). \quad (11)$$

因此,通用离线资源调度问题可表述为:对全局资源约束 U ,找到最优资源配置方案 P ,使得 $E(R^P) = \hat{E}^U$. 第2节我们将给出算法的具体描述.

2 快速资源分配算法

在典型的云计算系统中,通用离线资源调度问题的输入数据量会非常大,在常见场景中一般会有资源类型和区域数的乘积 $I \times J \geq 10^4$. 为寻找高效的解决算法,我们首先研究最优解的特征,首先引入定理1.

定理1. 有 N 个函数 $F_1(x_1), F_2(x_2), \dots, F_N(x_N)$, 满足 $x_n \in [0, \infty)$ 和 $F_n(x_n) \geq 0 (1 \leq n \leq N)$. 其导函数是 $f'_1(x_1), f'_2(x_2), \dots, f'_N(x_N)$, 满足 $f'_n(x_n) \in [0, Y] (Y > 0)$ 且均为连续单调递减函数.

则对于约束 $\sum_{n=1}^N x_n = U$, $\sum_{n=1}^N f_n(x_n)$ 达到最大值时,有 $f'_1(x_1) = f'_2(x_2) = \dots = f'_N(x_N)$.

证明. 反证法. 定义 $f(X) = \sum_{n=1}^N f_n(x_n)$, 假定 $f(X)$ 取最大值时并不满足上述导数相等条件, 设 $f'_a(x_a) < f'_b(x_b) (a, b \in [1, n])$, 因 $f'_a(x_a)$ 和 $f'_b(x_b)$ 均为单调

递减函数, 则必存在 x'_a 和 x'_b , 有 $x'_a > x_a$ 和 $x'_b < x_b$, 且 $x'_a + x'_b = x_a + x_b$ 并满足 $f'_a(x'_a) = f'_b(x'_b)$. 设 $f'_a(x'_a) = f'_b(x'_b) = Z$, 有:

$$f_a(x'_a) - f_a(x_a) = \int_{x_a}^{x'_a} f'_a(n) > Z(x'_a - x_a), \quad (12)$$

$$f_b(x_b) - f_b(x'_b) = \int_{x'_b}^{x_b} f'_b(n) < Z(x_b - x'_b), \quad (13)$$

可得到 $f_a(x'_a) + f_b(x'_b) > f_a(x_a) + f_b(x_b)$. 上述结果表明, 对任意不满足 $f'_1(x_1) = f'_2(x_2) = \dots = f'_N(x_N)$ 的 X , 必存在另外一个 X' 且 $f(X') > f(X)$.

证毕.

这样 $\hat{E}_i^O$ 即可转化为

$$\max(E(R_i^P)) \quad \text{s. t.} \quad \sum_{j=1}^J L_{ij}^j = L_i. \quad (14)$$

因为 $L_i (1 \leq i \leq I)$ 对给定的类型配置方案 O 是常量, 则根据定理 1, 可由以下方法找到 $\{P\}^O$ 中达到最大 i 型期望收益的资源配置方案 P' .

推论 1. 给定类型配置方案 O 和给定 $i' (1 \leq i' \leq I)$, 若对任何 $j (1 \leq j \leq J)$, 配置方案 $P' \in \{P\}^O$ 满足 $L_{i'}^j = cdf_{i',j}^{-1}(1 - G_{i'}^{-1}(L_{i'}))$, 则 P' 为最优方案.

这里对于函数 $G_i(\cdot)$, 有 $G_i(v) = \sum_{j=1}^J cdf_{i,j}^{-1}(1-v)$ ($0 \leq v \leq 1$).

在推论 1 中, $G_i^{-1}(\cdot)$ 和 $cdf_{i,j}^{-1}(\cdot)$ 是函数 $G_i(\cdot)$ 和 $cdf_{i,j}(\cdot)$ 的反函数. L_i 是类型配置方案 O 中 i 型资源的数量. 则推论 1 中最优的 i 型配置的期望收益 $\hat{E}_i^O$ 由 L_i 唯一确定. 如 $\hat{E}_i^O$ 表示为 L_i 的函数并记为 $h_i(L_i)$, 则可得如下推论:

推论 2. 给定类型配置方案 O , 有 $\hat{E}^O = \sum_{i=1}^I \hat{E}_i^O = \sum_{i=1}^I h_i(L_i)$, 则 $\hat{E}^O$ 可写为

$$\max\left(\sum_{i=1}^I h_i(L_i)\right) \quad \text{s. t.} \quad \sum_{i=1}^I L_i = U. \quad (15)$$

通过下述推论可找到限制 U 下的最优方案:

推论 3. 对给定 $U > 0$, 在类型配置方案 $O' \in \{O\}^U$ 中, 若对任意 i , 有 $L_i = h_i'^{-1}(H^{-1}(U))$, 则 O' 为最优类型配置方案. 对于函数 $H(\cdot)$ 和 $h_i'^{-1}(\cdot)$, 有 $H(s) = \sum_{i=1}^I h_i'^{-1}(s)$ 和 $h_i'(L_i) = w_{\text{sat}}(1 - cdf_i(L_i)) + w_{\text{loc}} G_i^{-1}(L_i)$.

为得到最优算法, 还需明确最优类型配置方案和最优配置方案间的内在关系, 描述如下:

推论 4. 对给定 $U > 0$, 类型配置方案 O' 满足 $\hat{E}^{O'} = \max(\hat{E}^O | O \in \{O\}^U)$. 若对于任意 i , 配置方案 $P' \in \{P\}^{O'}$ 满足 $E(R_i^{P'}) = \hat{E}_i$, 则有 $E(R^{P'}) = \max(E(R^P) | P \in \{P\}^U)$.

基于推论 1 至推论 4, 在构建需要的函数后, 原问题可转化为 2 阶段优化问题并快速求解: 阶段 1 中, 对给定的全局限制 U , 根据推论 2~3 快速定位最优类型配置方案 O' ; 阶段 2 中, 根据推论 1 在 $\{P\}^{O'}$ 中快速定位最优资源配置方案 P' . 由推论 4 可知, P' 即为满足全局资源约束 U 的最优方案.

我们通过离散函数来近似上述连续函数, 如算法 1 所示. 首先迭代计算离散函数(行⑪~⑭), 通过调用它们的反函数(行⑮~⑰)找到近优解. 该算法称为快速资源分配(fast resource placement, FRP)算法.

算法 1. 快速资源分配算法 FRP.

输入: 需求累积分布函数 $cdf_i(\cdot)$, $cdf_{i,j}(\cdot)$ 及其反函数; 资源总量 U ; 参数 $X_{i,j} = \min(cdf_{i,j}(x) = 1 | \forall x)$;

输出: 各区域资源总量 L_i^j .

```

① FOR  $i=1$  TO  $I$ 
②   IF  $x \geq \sum_{j=1}^J X_{i,j}$ 
③      $G_i^{-1}(x) = h_i'(x) = 0$ ;
④   END IF
⑤ END FOR
⑥ FOR  $k=0$  TO  $K$ 
⑦    $v = w_{\text{loc}} k / K$ ;
⑧   FOR  $i=1$  TO  $I$ 
⑨      $L_i' = \sum_{j=1}^J cdf_{i,j}^{-1}(1-v)$ ;
⑩      $G_i(v) = L_i'$ ;
⑪      $h_i'(L_i') = w_{\text{sat}}(1 - cdf_i(L_i')) + w_{\text{loc}} v$ ;
⑫   END FOR
⑬    $s = (w_{\text{sat}} + w_{\text{loc}}) k / K$ ;
⑭    $H(s) = \sum_{i=1}^I h_i'^{-1}(s)$ ;
⑮   IF  $H(s) \geq U$ 
⑯     break;
⑰   END IF
⑱ END FOR
⑲ FOR  $i=1$  TO  $I$ 
⑳    $L_i = h_i'^{-1}(H^{-1}(U))$ ;
㉑   FOR  $j=1$  TO  $J$ 
㉒      $L_{ij}^j = cdf_{i,j}^{-1}(1 - G_i^{-1}(L_i))$ ;
㉓   END FOR
㉔ END FOR

```

3 实 验

本文和文献[13]问题设定较接近, 且文献[13]的求解方法具代表性, 我们实现了文献[13]中基于连续最短路径(successive shortest path, SSP)的最

优求解算法,命名为 SSP-RP (SSP based resource placement). 基于相同的实验设定,比较均值算法、SSP-RP 和 FRP 的求解效果,并验证 FRP 求解效率和效果之间的关系. 其中,实验图表涉及的比率型数据,均表示 FRP 方法收益与文献[13]中 SSP-RP 方法收益的比值. 在实验设定中,随机用户需求符合 *Zipf* 分布(同文献[13]),即 i 型资源的平均需求占所有需求的比例如下:

$$W_i = \frac{i^{-e}}{\sum_{n=1}^I n^{-e}}, \quad (16)$$

其中, e 是实数,设总需求量期望值为 λ ,则需求 D_i^j 即为速率为 $\lambda J W_i$ 的泊松分布.

实验中,我们比较不同场景下均值方法、SSP-RP 方法和 FRP 方法的效果. 在均值方法中,区域 j' 中 i' 型资源数量 $L_{i'}^{j'}$ 和对应的需求期望成正比比例,即:

$$L_{i'}^{j'} = \frac{E(D_{i'}^{j'})U}{\sum_{i=1}^I \sum_{j=1}^J E(D_i^j)}. \quad (17)$$

首先我们比较 3 种方法在不同 *Zipf* 参数下效果的变化趋势. 参考文献[13],实验设定如下:区域数量和资源数量分别是 $J=4$ 和 $I=500$,收益权值 $w_{loc}=w_{sat}=1$,FRP 的精度 $K=100$,总请求量期望值 $\lambda=1\,000$. *Zipf* 参数取较大范围,为 $0.7\sim 1.4$. 实验中我们比较 3 种方法在资源盈余 ($U=1\,500>\lambda$) 和资源不足 ($U=500<\lambda$) 情况下的效果,如图 1 和图 2 所示,在图 1、图 2 中,横坐标为 *Zipf* 参数,纵坐标为期望收益. 我们可看到在不同的 *Zipf* 参数下,FRP 都要好于均值分配算法,且与 SSP-RP 算法的效果非常接近.

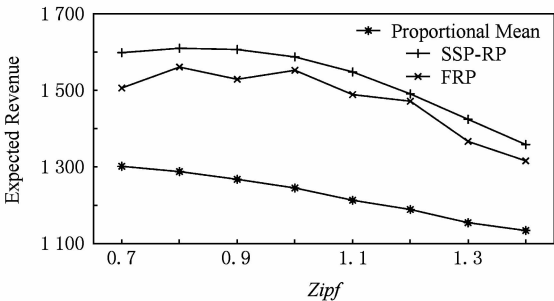


Fig. 1 Expected revenue with different *Zipf* values under deficit scenario.

图 1 资源不足时收益与 *Zipf* 的关系

为明确资料类型数量 I 对算法的影响,我们还比较了 3 种方法的效果随资源类型数变化趋势,实验设定如下:资源数量分别是 $J=4$ 和 $I=500$,收益

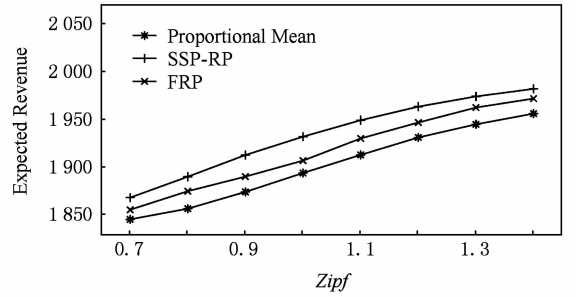


Fig. 2 Expected revenue with different *Zipf* values under surplus scenario.

图 2 资源盈余时收益与 *Zipf* 的关系

权值 $w_{loc}=w_{sat}=1$,FRP 的精度 $K=100$,总请求量期望值 $\lambda=1\,000$,*Zipf* = 1.0. 同文献[13],取 $I \in [100, 1\,000]$,以涵盖不同数量级的资源数量. 我们比较了 I 为 100~1 000 时 3 种算法的求解效果,如图 3 和图 4 所示,在不同的资源数量 I 下,FRP 的效果也始终优于均值方法,和 SSP-RP 方法接近.

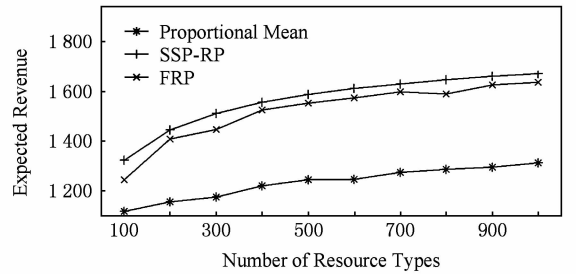


Fig. 3 Expected revenue with different number of resource types under deficit scenario.

图 3 资源不足时收益与资源数量的关系

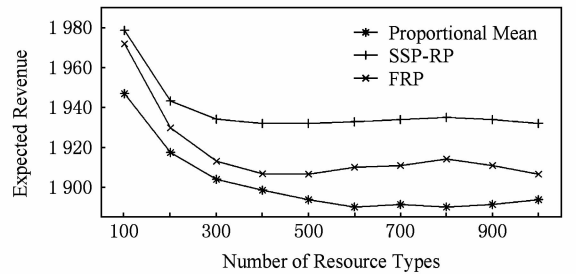


Fig. 4 Expected revenue with different number of resource types under surplus scenario.

图 4 资源盈余时收益与资源数量的关系

同时从图 1~4 的 2 个实验中,可看到随着资源数量由不足转为盈余,均值方法的效果也逐渐接近 SSP-RP 最优求解算法,使得 FRP 和 SSP-RP 的优势不明显. 分析实验数据后发现,随着资源数量的增加,由于待满足的请求没有变化,问题优化空间缩

小导致优化效果不明显. 而实际场景中资源数量一般都略有不足, 因此 FRP 相比均值算法仍有很大优势.

我们同时在通用场景中比较了 FRP 和 SSP-RP 的效果. 首先比较在不同 *Zipf* 参数和资源数量 *I* 下 FRP 和 SSP-RP 期望收益的比率, 如图 5 和图 6 所示. 我们可看到 FRP 受 *Zipf* 参数轻微影响, 同时也会受资源总量 *I* 的影响, 资源不足时资源收益略低一些. 分析实验数据后我们发现, 因为最后的分配结果会 4 舍 5 入到整数, 资源不足情况下, 舍入对结果影响较大.

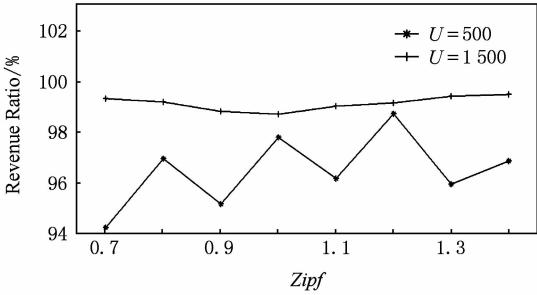


Fig. 5 Revenue ratio of FRP to SSP-RP with different *Zipf* values.

图 5 FRP 和 SSP-RP 的收益比率与 *Zipf* 的关系

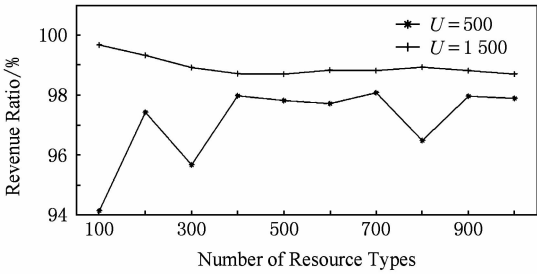


Fig. 6 Revenue ratio of FRP to S-URP with different number of resource types.

图 6 FRP 和 SSP-RP 的收益比率与资源数量的关系

FRP 算法的时间复杂度与精度 *K* 成正比, 因而我们试比较不同 *K* 下 FRP 的期望收益和 SSP-RP 算法结果的比值, 如图 7 所示. 其中横坐标为 FRP 求解精度 *K*, 取值为 10~200; 纵坐标为单个 *K* 值下不同 *U* 时多次实验得到的多个 FRP 和 SSP-RP 收益比率的均值. 我们可看到, FRP 可通过很低的时间复杂度可获得近优解, 在所有场景中都能达到 SSP-RP 的 97% 以上, 且因 SSP-RP 和 FRP 的时间复杂度分别为 $O(IJU)$ 和 $O(IJK + IK \log K)$, 当 $\lambda=10\,000$ 且 $K=10$ 时, FRP 能以降低 3 个数量级的时间复杂度接近最优解 99%, 具明显的改进效果.

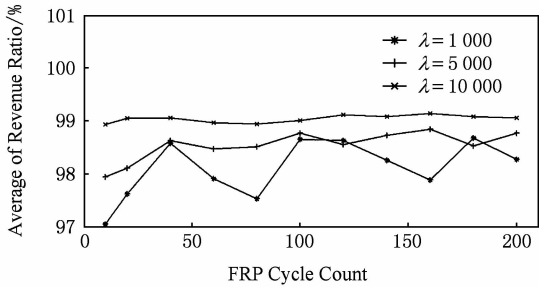


Fig. 7 Average revenue ratio of FURP to SSP-RP with different cycle counts.

图 7 FRP 和 SSP-RP 的平均收益比率与循环数的关系

4 结束语

本文研究了分布式云计算平台中基于随机需求的通用资源分配问题, 并提出了可获取近似解的快速求解方法. 相比已有文献中的最优解算法, 我们的方法可达到最优解的 99%, 同时计算复杂度只有最优解算法的 1%, 是对已有算法的有效补充.

参 考 文 献

[1] YouTube. YouTube homepage [EB/OL]. (2005-02-15) [2015-03-12]. <http://www.youtube.com>

[2] Facebook. Facebook homepage [EB/OL]. (2004-02-04) [2015-03-12]. <http://www.facebook.com>

[3] You Kun, Tang Bin, Qian Zhuzhong, et al. QoS-aware placement of stream processing service [J]. Journal of Supercomputing, 2013, 64(3): 919-941

[4] Qiu Xuanjia, Li Hongxing, Wu Chuan, et al. Dynamic scaling of VoD services into hybrid clouds with cost minimization and QoS guarantee [C] //Proc of IEEE Int Packet Video Workshop. Piscataway, NJ: IEEE, 2012: 137-142

[5] Huang Yongfeng, Dong Yongqiang, Zhang Sanfeng, et al. Leftover bandwidth-aware peer selection algorithm for inter-datacenter content distribution [J]. Journal on Communications, 2013, 34(7): 24-33 (in Chinese)
(黄永锋, 董永强, 张三峰, 等. 数据中心间空闲带宽感知的内容分发算法[J]. 通信学报, 2013, 34(7): 24-33)

[6] Broberg J, Buyya R, Tari Z. Metacdn: Harnessing storage clouds for high performance content delivery [J]. Journal of Network and Computer Applications, 2009, 32(5): 1012-1022

[7] Qiu Xuanjia, Li Hongxing, Wu Chuan, et al. Cost-minimizing dynamic migration of content distribution services into hybrid clouds [C] //Proc of INFOCOM 2012. Piscataway, NJ: IEEE, 2012: 2571-2575

- [8] Chen Fangfei, Guo Katherine, Lin John, et al. Intra-cloud lightning: Building CDNs in the cloud [C] //Proc of IEEE INFOCOM'12. Piscataway, NJ: IEEE, 2012: 433-441
- [9] He Zhicong, Gu Guangzhao, Wang Xin, et al. Novel cooperative caching strategies for video streaming distribution based on reconfiguration routers [J]. Journal on Communications, 2012, 33(6): 82-90 (in Chinese)
(何智聪, 谷光昭, 王新, 等. 基于可重构路由器上缓存的流媒体协作分发策略[J]. 通信学报, 2012, 33(6): 82-90)
- [10] Zhou Jingcai, Zhang Huyan, Zha Wenliang, et al. User-aware resource provision policy for cloud computing [J]. Journal of Computer Research and Development, 2014, 51(5): 1108-1119 (in Chinese)
(周景才, 张沪寅, 查文亮, 等. 云计算环境下基于用户行为特征的资源分配策略[J]. 计算机研究与发展, 2014, 51(5): 1108-1119)
- [11] Tang Zhuo, Zhu Min, Yang Li, et al. Random Task-oriented user utility optimization model in the cloud environment [J]. Journal of Computer Research and Development, 2014, 51(5): 1120-1128 (in Chinese)
(唐卓, 朱敏, 杨黎, 等. 云环境中面向随机任务的用户效用优化模型[J]. 计算机研究与发展, 2014, 51(5): 1120-1128)
- [12] Tan B, Massoulie L. Optimal content placement for peer-to-peer video-on-demand systems [C] //Proc of IEEE INFOCOM'12. Piscataway, NJ: IEEE, 2012: 694-702
- [13] Rochman Y, Levy H, Brosh E. Resource placement and assignment in distributed network topologies [C] //Proc of IEEE INFOCOM'13. Piscataway, NJ: IEEE, 2013: 1914-1922

- [14] Rochman Y, Levy H, Brosh E. Efficient resource placement in cloud computing and network applications [J]. ACM SIGMETRICS Performance Evaluation Review, 2014, 42(2): 49-51



Wei Wei, born in 1983. Received his BSc and PhD degrees from Zhejiang University. Associate professor. Member of China Computer Federation. His research interests include parallel computing and network security.



Liu Yang, born in 1978. Received her PhD degree from Northwestern Polytechnical University. Associate professor. Senior member of China Computer Federation. Her main research interests include grid computing and wireless sensor networks (enjoyang@126.com).



Yang Weidong, born in 1977. Received his PhD degree from Xidian University. Associate professor. Senior member of China Computer Federation. His main research interests include vehicular ad hoc network and information security (mengguyang@163.com).