doi:10.11887/j.cn.201605008 http://journal.nudt.edu.cn

云计算系统虚拟机内存资源预留方法*

阚运奇[1,2],刘宏伟[1],左德承[1],张 展[1]
(1. 哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001;
2. 东北电力大学 信息工程学院,吉林 吉林 132400)

摘 要:为降低消费者租借云计算系统资源的开销,提出了成本约束的内存预留随机整数线性规划模型及方法。结合预留计划和按需计划的内存资源价格,设计包含成本及资源总量约束条件的随机开销函数,并以函数期望值最小化为目标,基于内存消耗量概率分布求出优化的内存预留量。试验表明,消费者利用该方法租借资源的开销比利用预留计划、按需计划及同类方法租借资源的开销更小。

关键词:云计算;虚拟机;内存资源预留;成本优化
中图分类号:TN95 **文献标志码**:A **文章编号**:1001-2486(2016)05-045-07

Memory resources reservation method for virtual machine in cloud computing system

KAN Yunqi[1,2], *LIU Hongwei*[1], *ZUO Decheng*[1], *ZHANG Zhan*[1]
(1. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China;
2. School of Information Engineering, Northeast Dianli University, Jilin 132400, China)

Abstract: In order to reduce the consumers' costs of renting resources in cloud computing system, a random integer linear programming model and a method for memory reservation were presented on the basis of cost constraints. Combined with the memory resource price of the reservation plan and the on-demand plan, the random cost function which consists of costs and total amount of resources constraints was designed. Aimed at minimizing the expected value of the cost function, the optimal amount of memory reserved was obtained on the basis of the probability distribution of memory consumption. The experiments show that the cost of renting recourses by utilizing this method is less than the cost of renting recourses by adopting reservation plan, on-demand plan and other similar methods.

Key words: cloud computing; virtual machine; memory resources reservation; cost optimization

云计算是一种大规模的分布式计算模式[1],按使用量付费,这种模式提供可用的、便捷的、按需的网络访问,云提供商将CPU、内存、硬盘等复合资源整合到虚拟机(Virtual Machine, VM),并通过资源配置机制租借给消费者[2]。但多数云服务提供商提供的虚拟机型号有限且规格固定,即每个类型的虚拟机包含的资源量是固定的,而不同的消费者对各类资源需求各异,规格固定、型号有限的虚拟机无法满足各类消费者的需求,这将导致消费者无法充分地利用虚拟机各项资源,造成资源浪费及成本增加,尤其是价格较高的CPU、内存等资源。

另外,包括Amazon EC2[3-4]、Go Grid[5]和阿里云平台等在内的云计算系统向消费者提供的虚拟机租借方案有多种,其中两个主要的租借方案是:短期按需计划和长期预留计划。一般来说,按需计划是指云计算提供商按消费者需求提供资源的计划,并按资源使用时间收费。采用按需计划的消费者可以灵活地申请和退订资源,但资源的单价较高。预留计划是指消费者长期租借定量的资源并预付费用,如1年期,计费标准如表1所示。消费者采用预留计划租借虚拟机资源单价较低,但租借方式不够灵活。租借虚拟机来执行持续任务的消费者无法预知任务负载,多采用复合型租借方案,即:预留满足普通日常负载需求的资源,在负载量突增时按需租借资源。消费者采用该方案可降低由于预留过量资源造成的开销,同时也可避免全部采用按需计划租借资源造成的高额开销。如何求得一个优化的资源预留量来降低成本并满足负载运行的需求成为亟须解决的问题。

* **收稿日期**:2015-11-25
基金项目:国家863计划资助项目(2013AA01A215)
作者简介:阚运奇(1981—),男,吉林吉林人,副教授,博士研究生,E-mail: kanyunqi@ftcl.hit.edu.cn

表1　阿里云虚拟机计费标准

Tab.1　VM billing standard in Aliyun cloud computing system

CPU/GHz	内存/GB	存储/GB	按需计费/(元/h)	1年期预留计费/(元/a)	预留单价/(元/h)
1	1	0	0.280	520.0	0.060
1	1	512	0.741	2363.0	0.273
1	2	0	0.440	960.0	0.111
1	2	512	0.901	2803.0	0.324
2	2	0	0.560	1868.0	0.216
2	2	512	1.021	3693.2	0.427
2	4	0	1.341	2480.0	0.287
2	4	512	0.690	4323.2	0.500

国内外有较多针对云计算系统资源预订及配置的相关研究[6-15]：

文献[10]提出了云资源优化配置(Optimal Cloud Resource Provisioning, OCRP)算法，通过制定一个随机规划模型解决由于消费者需求不确定造成的预订资源困难问题，降低了消费者的成本；文献[11]提出了具有容错功能的考虑负载均衡的成本开销最小化的云计算系统资源分配方案。以上研究均以虚拟机为整体作为研究对象来考虑资源预定计划及资源分配方案，并没有以虚拟机各项资源为对象细粒度地研究资源分配方案。

文献[14]设计了内存均衡管理器实现了虚拟机内存均衡(Virtual Machine Memory Balancing, VMMB)机制，动态监测各虚拟机内存需求，定期重新平衡虚拟机的内存配额；文献[15]提出了两种动态虚拟机调度技术，减少了系统响应时间且平衡了CPU、内存等资源利用率。以上方法侧重于从云计算服务商角度进行资源配置管理，但没有从消费者角度考虑开销问题。

1　问题描述

1.1　云计算系统及相关定义

云计算系统由高性能硬件资源构成基础设施，由云操作系统管理资源组建虚拟机为消费者提供安全可靠[16]的服务：计算、数据存储、文件处理[17]等。为方便描述，下面定义一些符号及集合：

设定云计算系统的消费者集合为 $U=\{U_1,U_2,\cdots,U_i\}$，U_i 代表第 i 个消费者；设定任务集合为 $D=\{D_1,D_2,\cdots,D_j\}$，D_j 代表第 j 个任务，用 D_{ij} 表示 i 用户的第 j 个任务；用 $\boldsymbol{R}(D_{ij})$ 表示执行 D_{ij} 需要的资源量，$\boldsymbol{R}(D_{ij})=(C_{ij}^r,\ M_{ij}^r,\ SP_{ij}^r,\ I_{ij}^r)^{\mathrm{T}}$，$C_{ij}^r$ 代表任务 D_{ij} 对CPU的需求量，M_{ij}^r 代表 D_{ij} 对内存的需求量，SP_{ij}^r 代表 D_{ij} 对磁盘资源的需求量，I_{ij}^r 代表 D_{ij} 对I/O资源的需求量；用 $V=\{V_1,V_2,\cdots,V_n\}$ 表示虚拟机集合，V_n 代表 n 类型虚拟机，设置虚拟机身份标识为 V_{mn}，其中 m 代表虚拟机的编号，该编号是唯一的；用 $\boldsymbol{S}(V_n)$ 代表该类虚拟机资源总量，$\boldsymbol{S}(V_n)=(C_n,M_n,S_n,N_n)^{\mathrm{T}}$，$C_n$ 代表该类型虚拟机中CPU资源的总量，M_n 代表内存总量，S_n 代表存储空间总量，N_n 代表网络带宽容量。

1.2　云计算系统付费模式

云提供商向消费者出租虚拟机，双方遵守SLA协议(Service-Level Agreement)[18]。典型的如阿里云计算平台，其向消费者提供两种虚拟机租借计划——预留计划和按需计划，按这两种计划出租各种规格虚拟机的单价如表1所示。

如果消费者租借虚拟机用来执行长期不间断性任务，那么他们一般采用混合方式租借资源，下面以这类消费者为对象研究虚拟机资源配置及预留方案，这类消费者的开销可表示为：$Cost_{\mathrm{total}}=Cost_{\mathrm{on}}+Cost_{\mathrm{r}}$，$Cost_{\mathrm{r}}$ 代表以预留计划租借资源的开销，$Cost_{\mathrm{on}}$ 代表以按需计划租借资源的开销。消费者采用预留计划比采用按需计划租借资源的价格要低，但如果用户预留过多的资源会造成用户资金浪费；而如果预留资源过少，会导致任务负载频繁超过预订资源的承载量，消费者会采用按需计划以较高的价格频繁租借资源，最终其总开销 $Cost_{\mathrm{total}}$ 增加。

1.3　各项资源利用不均衡问题

为降低用户开销可以以虚拟机为对象制订虚拟机预留方案，在保障日常任务正常运行的前提下降低开销，但由于虚拟机规格固定所以不可避免地造成单项资源(如：内存)浪费。

在VMware云计算系统中验证了这种现象的存在，首先参考阿里云计算系统基础型虚拟机(1核CPU、1 GB内存)设置并启动了一个小型虚拟机，在虚拟机中安装Windows Server 2003操作系统并搭建网站，然后利用Loadrunner系统[19]模拟产生了峰值为500的网站访问负载，通过VMware虚拟机性能监测器监控获得各项资源消耗量。

试验中虚拟机CPU资源配额为1 GHz，为提高实验的直观性增加了1 GHz的预备扩展容量。用 V_{mn} 代表该虚拟机，截取某 t_i 时刻资源消耗量数据可以看到，当CPU资源利用率 $L(U_{t_i}^c)$ 到达

97.66% 的临界状态时,其他资源利用率并不高,内存资源利用率 $L(U_{t_i}^M)$ 仅为 24.20%,磁盘资源的利用率 $L(U_{t_i}^S)$ 仅为 26.56%,如表 2 所示。

表 2　某时刻虚拟机资源利用率

Tab. 2　Utilization rates of VM resources at one moment

资源类型	$L(U_{t_i}^c)$	$L(U_{t_i}^M)$	$L(U_{t_i}^S)$
利用率	97.66%	24.20%	26.56%

为充分展现各项资源消耗对比情况,设置虚拟机资源量可以弹性扩展,但在商业云计算系统中,当任意一项资源消耗量达到配额量时需要启动新的虚拟机或延迟服务,这样会导致消费者开销增大或服务质量下降。通过实验获得了 CPU、内存资源消耗率的对比,如图 1 所示。

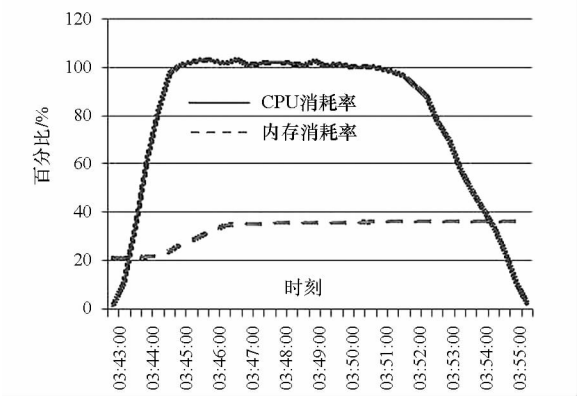

图 1　高负载下主要资源消耗对比图

Fig. 1　Comparison of main resources consumption rates under heavy load

当任务负载对 CPU 资源的需求持续增加时,为保障任务正常执行,消费者需要申请启动新的虚拟机 $V_{(m+1)n}$,虽然满足了任务负载对 CPU 资源量的需求,但也造成内存及其他资源更大的浪费。通过实验表明消费者很难充分地利用资源规格固定的虚拟机中的各项资源,最终导致开销增加及资源浪费。

2　基于成本约束的内存预留随机整数线性规划模型

2.1　VM 资源均衡预订问题描述

不同的商业云计算系统对于资源单位的定义不同,如:亚马逊 EC2 以 ECU 为单位配置 CPU 资源,1 个 ECU 相当于 1 GHz 的计算单元;阿里云平台以核为单位配置 CPU 资源,1 核相当于 1 GHz的计算单元。一般云计算系统为虚拟机配置 CPU 资源量均为整数单位 1 GHz,并且 CPU 资源单价最高,所以资源预留方案中可先确定 CPU 预留量然后围绕 CPU 资源量来确定内存的用量,但几乎所有的云提供商给出的可供选择的内存配额规格都较少,如:500 M,1 GB 等。

消费者采用复合内存预留方案后,可以记录其租借的虚拟机在一段较长时期的 CPU、内存消耗量,然后选取合理的 CPU 资源量,围绕 CPU 资源量进一步为消费者预留适量的内存,降低用户开销。

2.2　内存预留量随机整数规划期望值模型

某消费者在一段时间内的内存消耗量是随机整数[20],通过数据分析可以确定内存消耗量的概率分布,结合两种租借计划内存资源的单价,采用随机整数规划求得优化的内存预留量。

为研究问题方便直观,首先以基础型 VM 为研究对象,该类型 VM 具有 1 核 CPU、1 GB 内存。

设 t_i 时刻至 t_{i+1} 时刻间内存开销函数为 $c_{t_i}^M(x,\varepsilon_{t_i})$,其中 x 代表内存预订量,ε_{t_i} 代表 t_i 时刻内存实际消耗量。$T=\{t_0,t_1,\cdots,t_n\}$ 代表一个较大的时间段。

$$c_{t_i}^M(x,\varepsilon)=\begin{cases}e_n^r\cdot x_n^r, & x_n^r>\varepsilon_{t_i}\\ e_n^r\cdot x_n^r+(\varepsilon_{t_i}-x_n^r)\cdot e_n^0, & x_n^r\leqslant\varepsilon_{t_i}\end{cases}\tag{1}$$

式中,ε 表示不同场景下内存消耗量,x_n^r 代表按预留计划租借 n 类型虚拟机内存资源的数量,e_n^r 代表按预留计划租借 n 类型的虚拟机内存的费用,e_n^0 表示采用按需计划租借内存的费用。在时间段 T 内存总开销为:

$$C_T^M(x,\varepsilon)=\sum_{t_0}^{t_n}c_{t_i}^M(x,\varepsilon)\tag{2}$$

ε 是随机变量,定义开销随机函数 $K_\Omega(x_n^r,\varepsilon)$ 来代表在 T 时间内不同场景 Ω 下超出预订量内存资源的开销,由式(1)得到:

$$K_\Omega(x_n^r,\varepsilon)=\begin{cases}\sum_{t_0}^{t_n}(\varepsilon_{t_i}-x_n^r)\cdot e_n^0, & x_n^r\leqslant\varepsilon_{t_i}\\ 0, & x_n^r>\varepsilon_{t_i}\end{cases}\tag{3}$$

由式(2)、式(3)推导得到总开销公式为:

$$C_T^M(x,\varepsilon)=\sum_{t_0}^{t_n}e_n^r\cdot x_n^r+K_\Omega(x_n^r,\varepsilon)\tag{4}$$

在一定约束条件下使总开销最小化,由式(4)推导出:

$$\begin{cases}\text{Min} & C_T^M(x,\varepsilon)\\ \text{s.t.} & x\in\mathbf{N},\ x\leqslant T_C\end{cases}\tag{5}$$

式中，T_C 代表物理节点的内存总数量。由式(4)、式(5)推导得到：

$$\begin{cases}\text{Min} & \sum_{t_0}^{t_n} e_n^r \cdot x_n^r + K_\Omega(x_n^r, \varepsilon) \\ \text{s.t.} & x \in \mathbf{N},\ x \leqslant T_C\end{cases} \tag{6}$$

由于内存的需求量 ε 是随机变量，因而开销函数 $C_T^M(x,\varepsilon)$ 也是随机函数，不能准确地计算出内存预留量 x。一个自然的方法就是考虑期望开销最小化来求得 x。因为在实际的问题中预订的内存数 x 是整数，得到预留问题的随机整数规划期望值模型为：

$$\begin{cases}\text{Min} & E\left[\sum_{t_0}^{t_n} e_n^r \cdot x_n^r + K_\Omega(x_n^r, \varepsilon)\right] \\ \text{s.t.} & x \in \mathbf{N},\ x \leqslant T_C\end{cases} \tag{7}$$

3 基于内存消耗分布概率的预留内存定量求解方法

内存消耗量 ε_{t_i} 是随机变量，值取决于任务负载量，对于执行稳定服务的 VM，ε_{t_i} 值在一定区间波动。为求 x_n^r 的值，首先要获取内存消耗 ε_{t_i} 历史数据集。开启 1 台基础型 VM(1 GHz CPU、1 GB 内存、512 GB 硬盘)，在 VM 中搭建 WEB 服务，并采用 Loadrunner[19] 系统向 VM 服务器产生一定量的负载，并且负载的数量是变化的，通过 VMware 虚拟机监控器获取在 T 时间段内每 20 s 间隔的内存消耗量 ε_{t_i}。服务特征决定了 ε_{t_i} 的分布区间 $[\varepsilon_{\min}, \varepsilon_{\max}]$ 及分布概率 $P(\varepsilon)$，ε_{t_i} 分布概率的精度由 i 的大小决定。

VMware 虚拟机监控器以 20 s 为时间间隔记录监测数据，给出一组资源消耗量实验数据，如表 3 所示。由于篇幅关系，该表仅列出部分测试数据，表中灰色区域中的数据是 CPU 消耗量超配额(1 GHz)的数据。

定义任务 $D = \{d_{t_0}, d_{t_1}, \cdots, d_{t_i}, \cdots\}$，在 d_{t_0} 未到达时各项资源利用率较低，定义该时间段为 t_{ideal}。任务到达时刻定义为 t_0，通过表 3 发现 t_0 时刻后 CPU 资源消耗量 $U_{t_i}^C$ 上升速度较快，内存资源消耗量 $U_{t_i}^M$ 上升速度较慢，并且在负载到达顶峰 $U_{t_i}^C > R_n^C$ 时，内存用量 $U_{t_i}^M < \frac{1}{2} R_n^M$。为直观反映资源用量对比情况，实验前设置虚拟机的 CPU 资源配额可弹性增加。由于研究的是基础型虚拟机内存预订量，所以在求解式(7)中 x_n^r 最优解时需要将 CPU 资源消耗量 $U_{t_i}^C > 1024$ 的数据(表 3 中灰色数据)剥离。在采集的数据足够多的条件下剥离部分数据并不影响内存预留量优化解的精度。

表 3 虚拟机资源消耗量实验数据

Tab. 3 Experiment data of VM resources consumption

资源量 时刻	$U_{t_i}^C$/ MHz	$U_{t_i}^M$/ KB	$U_{t_i}^S$/ MB	$U_{t_i}^N$/ (MB/s)
3:43:00	22	224 804	10	0
3:43:20	17	224 720	2	0
3:43:40	114	225 320	34	7843
3:44:00	368	226 476	54	33 973
3:44:20	632	228 752	73	68 794
3:44:40	831	232 568	90	99 832
3:45:00	1000	253 804	136	117 494
3:45:20	1037	283 816	127	117 625
3:45:40	1048	299 836	127	117 559
3:46:00	1059	324 684	148	117 522
3:46:20	1054	347 872	131	117 735
3:46:40	1045	369 344	122	117 605
3:47:00	1060	370 952	134	117 664
3:47:20	1037	373 004	128	117 593
3:47:40	1043	374 332	125	117 675
3:48:00	1046	376 408	138	117 389
3:48:20	1049	376 832	132	117 677
3:48:40	1043	377 752	126	117 608
3:49:00	1039	377 832	129	117 689
3:49:20	1052	379 264	134	117 734
3:49:40	1030	380 224	121	117 610
3:50:00	1042	380 036	140	117 697
3:50:20	1033	380 928	127	117 638
3:50:40	1026	381 144	126	117 716
3:51:00	1032	382 220	135	117 563
3:51:20	1018	383 560	126	117 521
3:51:40	1008	383 992	125	117 517
3:52:00	989	384 224	139	117 572

通过多次实验获取内存用量概率分布 $P(\varepsilon)$ 用来对式(7)进行求解。日常环境中人们更倾向于以 MB 作为内存单位，预留内存资源量也为整数，所以在统计 ε 分布概率时将监测数据取整处理。通过多组随机负载实验，获取了随机变量 ε 的分布概率如图 2 所示，该分布概率可作为计算最优预订量 x_n^r 的基础数值。

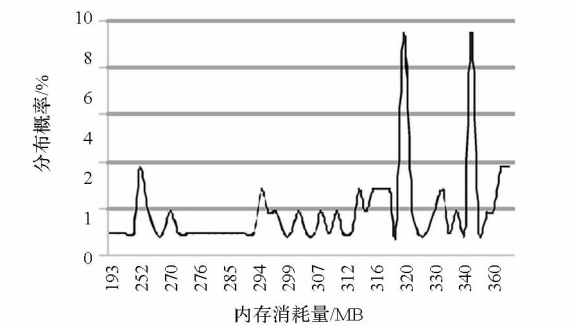

图 2　内存消耗量概率分布图

Fig. 2　Memory consumption probability distribution map

4　实验

4.1　实验环境及方案

运行本次实验的云计算实验平台配置如表 4 所示。采用阿里云计费体系作为实验的计费参考标准(如表 1 所示)。

表 4　云计算实验系统硬件配置

Tab. 4　Hardware configuration of cloud computing test system

实验系统组件	设备	数量
云计算服务器	浪潮刀片机	8 台
云存储服务器	HP EVA4400	2 台
云计算管理系统	VMware	1 套
压力发生软件	Loadrunner	1 套
压力负载发生器	HP 网络负载发生器	15 台

通过阿里云不同规格虚拟机的对比计算得到内存的预留计费单价及按需计费单价分别为:$Cost^m_{re}$ = 0.051 元/Gb/h, $Cost^m_{on}$ = 0.16 元/Gb/h。该价格作为求解内存优化预留量的常量,在 LINGO 计算软件中结合图 2 中内存消耗量分布概率计算得到优化预订量 x^r_n = 338,每个消费者租借虚拟机运行的任务负载不同,内存优化预留量也不同。

实验由 1 台安装 Loadrunner 系统的服务器控制 10 台 HP 瘦客户机向虚拟机端发送 WEB 网络负载,模拟不同数量的网络请求 *Requests*,产生不同规模的负载,每次测试时间为 20 min。

4.2　与商业云计算系统资源租借方案开销对比分析

根据监测得到的内存消耗数据,可计算求得采用多种租借方案的最终开销。首先与阿里云计算系统提供的两种资源租借方案进行了对比。

值得注意的是,针对虚拟机内存调度问题,Min 等[14]提出了虚拟机内存均衡机制,其方法侧重于从服务器负载均衡、提高资源利用率等角度进行内存调度,不考虑消费者预订资源的开销,所以在该机制下消费者以预留计划租借内存的开销 $Cost^m_{re}$、以按需计划租借内存的开销 $Cost^m_{on}$ 并没有受到影响。采用预留计划租借内存的开销 $Cost^m_{re}$、采用按需计划租借内存的开销 $Cost^m_{on}$ 与采用内存复合预留方案租借内存的开销 $Cost^m_{mix}$ 对比如表 5 所示。

表 5　不同负载下用各方案租借内存 20 min 开销

Tab. 5　Cost of renting memory resources by different plans under different load in 20 min

Requests(峰值)	$Cost^m_{mix}$/分	$Cost^m_{on}$/分	$Cost^m_{re}$/分
550(1260 s)	1.655 8	1.989 2	1.782 9
650(1240 s)	1.680 3	1.912 3	1.754 6

通过对比发现,采用内存资源复合预留方案租借资源的总开销低于其他两种方案。如果以采用预留计划租借资源的开销为基准,其他两种方案与其对比结果如图 3 所示。

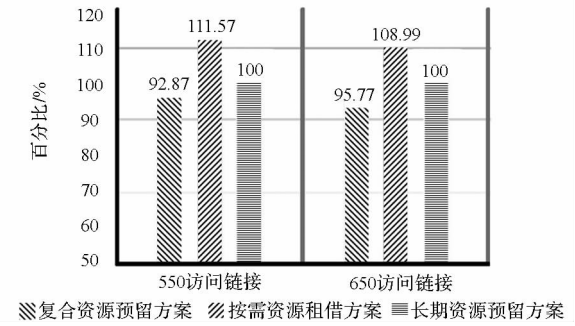

图 3　采用各方案租借内存开销对比图

Fig. 3　Cost of renting memory resources by different plans

另外表 5 中数据是租借内存约 20 min 的开销,多数消费者租借虚拟机时间一般在 1 年期及以上,假设虚拟机承载的任务负载较为稳定,那么可预计内存开销如表 6 所示。

表 6　不同负载下用各方案租借内存 1 年预期开销

Tab. 6　Expected cost of renting memory resources by different plans under different load in one year

Requests(峰值)	$Cost^m_{mix}$/(元/a)	$Cost^m_{on}$/(元/a)	$Cost^m_{re}$/(元/a)
550	435.14	522.77	468.55
650	441.59	502.55	461.11

4.3 与相关研究方案对比分析

Chaisiri 等[10]提出了云资源优化配置算法，以虚拟机为对象为消费者预留资源。在当前规模实验中采用该算法预留方案的内存开销 $Cost^m_{\mathrm{OCRP}}$ 与采用内存复合预留方案租借内存的开销 $Cost^m_{\mathrm{mix}}$ 对比如表 7 所示。

表 7 不同负载下用 2 种方案租借内存 20 min 开销

Tab. 7 Cost of renting memory resources by the two plans under different load in 20 min

Requests（峰值）	$Cost^m_{\mathrm{mix}}$/分	$Cost^m_{\mathrm{OCRP}}$/分
550	1.655 8	1.782 9
650	1.680 3	1.754 6

在当前实验中，以虚拟机为研究对象的算法预留的内存量必然是云虚拟机额定内存量，而内存复合预留方案以虚拟机内存资源为对象细粒度地研究资源预留量，其值小于标准虚拟机额定内存资源量、小于 OCRP 算法的资源预留量，所以租借内存的开销较低。

4.4 监测时间间隔对预留方案的影响分析

以上实验以 VMware 云计算系统为实验平台，但不同云计算系统监测资源消耗量的时间间隔并不相同，如果监测时间周期较短，那么不同系统的内存消耗量 ε_{t_i} 数据集有一定差异，导致分布概率 $P(\varepsilon)$ 不同，从而影响 x^r_n 的求解精度。

但可以看出 t_i 的数量决定 $P(\varepsilon)$ 是否能真实反映内存消耗情况，随着 i 增大分布概率 $P(\varepsilon)$ 趋于真实反映内存消耗情况，当 i 足够大时就会屏蔽由于云计算系统监测时间间隔不同而造成的数据集差异以及对 x^r_n 求解精度的影响。

而且在商业云环境中，具有持续不间断性任务负载的消费者租借 VM 的时间较长，如：1 月、1 年。为制订合理的预留方案，可以在实验及应用中选取较长时间段（如 1 星期）的内存消耗量作为求解数据集，可在一定程度上屏蔽由于云计算系统监测时间间隔不同而造成的数据集差异，进而保障最优预留量 x^r_n 的准确性，提高内存优化预留方法的适用性。

5 结论

近年来，云计算系统资源管理配置已成为云计算研究领域的热点问题，降低消费者租借云计算资源成本也逐渐得到重视，本文提出了一种虚拟机内存资源优化预留方法，不以虚拟机为对象而以虚拟机各项资源为对象细粒度地研究资源预留方案，通过与标准云计算系统资源租借方案以及相关的研究成果对比试验发现该方案能够降低用户的总开销，同时降低资源闲置率，有利于云提供商动态配置资源。

内存资源优化预留方法以降低消费者成本为目标，为提高方案的适应性应进一步结合云计算系统服务器负载均衡、能耗[21]等指标研究综合性资源配置及预留方案。

参考文献(References)

[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] Püschel T, Schryen G, Hristova D, et al. Revenue management for cloud computing providers: decision models for service admission control under non-probabilistic uncertainty[J]. European Journal of Operational Research, 2015, 244(2): 637 - 647.

[3] Tang S, Yuan J, Wang C, et al. A framework for Amazon EC2 bidding strategy under SLA constraints [J]. IEEE Transactions on Parallel & Distributed Systems, 2014, 25(1): 2 - 11.

[4] Juve G, Deelman E, Berriman G B, et al. An evaluation of the cost and performance of scientific workflows on Amazon EC2[J]. Journal of Grid Computing, 2012, 10(1): 5 - 21.

[5] Barrett E, Howley E, Duggan J. Applying reinforcement learning towards automating resource allocation and application scalability in the cloud [J]. Concurrency & Computation Practice & Experience, 2013, 25(12): 1656 - 1674.

[6] Goiri í, Guitart J, Torres J. Economic model of a cloud provider operating in a federated cloud [J]. Information Systems Frontiers, 2012, 14(4): 827 - 843.

[7] Wu L, Garg S K, Versteeg S, et al. SLA-based resource provisioning for hosted software-as-a-service applications in cloud computing environments [J]. IEEE Transactions on Services Computing, 2014, 7(3): 465 - 485.

[8] Al-Ayyoub M, Jararweh Y, Daraghmeh M, et al. Multi-agent based dynamic resource provisioning and monitoring for cloud computing systems infrastructure [J]. Cluster Computing, 2015, 18(2): 919 - 932.

[9] Chaisiri S, Lee B S, Niyato D. Optimal virtual machine placement across multiple cloud providers [C]// Services Computing Conference, IEEE Asia-Pacific, 2009: 103 - 110.

[10] Chaisiri S, Lee B S, Niyato D. Optimization of resource provisioning cost in cloud computing[J]. IEEE Transactions on Services Computing, 2012, 5(2): 164 - 177.

[11] Di S, Wang C L. Error-tolerant resource allocation and payment minimization for cloud system [J]. IEEE Transactions on Parallel & Distributed Systems, 2015, 24(6): 1097 - 1106.

[12] Chard K, Bubendorfer K. High performance resource allocation strategies for computational economies[J]. IEEE Transactions on Parallel and Distributed Systems, 2013,

24(1): 72 – 84.

[13] Di S, Wang C L. Dynamic optimization of multiattribute resource allocation in self-organizing clouds [J]. IEEE Transactions on Parallel & Distributed Systems, 2013, 24(3): 464 – 478.

[14] Min C, Kim I, Kim T, et al. VMMB: virtual machine memory balancing for unmodified operating systems [J]. Journal of Grid Computing, 2012, 10(1): 69 – 84.

[15] Rathor V S, Pateriya R K, Gupta R K. An efficient virtual machine scheduling technique in cloud computing environment[J]. International Journal of Modern Education & Computer Science, 2015, 7(3): 39 – 46.

[16] 刘婷婷, 王文彬. 云计算中基于公平的安全判定相等协议的身份认证方案[J]. 国防科技大学学报, 2013, 35(5): 120 – 123,139.
LIU Tingting, WANG Wenbin. An authentication scheme based on fair equality-determination protocol in cloud computing[J]. Journal of National University of Defense Technology, 2013, 35(5): 120 – 123, 139. (in Chinese)

[17] 付松龄, 廖湘科, 黄辰林, 等. FlatLFS:一种面向海量小文件处理优化的轻量级文件系统[J]. 国防科技大学学报, 2013, 35(2): 120 – 126.
FU Songling, LIAO Xiangke, HUANG Chenlin, et al. FlatLFS: a lightweight file system for optimizing the performance of accessing massive small files[J]. Journal of National University of Defense Technology, 2013, 35(2): 120 – 126. (in Chinese)

[18] Macías M, Guitart J. SLA negotiation and enforcement policies for revenue maximization and client classification in cloud providers [J]. Future Generation Computer Systems, 2014, 41(C): 19 – 31.

[19] Guan X, Cheng B, Song A, et al. Modeling users′ behavior for testing the performance of a web map tile service [J]. Transactions in GIS, 2014, 18(S1): 109 – 125.

[20] Kılıç Y E, Tuzkaya U R. A two-stage stochastic mixed-integer programming approach to physical distribution network design[J]. International Journal of Production Research, 2015, 53(4): 1291 – 1306.

[21] Mastelic T, Oleksiak A, Claussen H, et al. Cloud computing: survey on energy efficiency[J]. ACM Computing Surveys, 2015, 47(2): 1 – 36.