doi:10.3969/j.issn.1000-436x.2014.z2.025

基于弹性资源调整的云计算资源分配方法

殷波，张云勇，王志军，房秉毅，冯伟斌
（中国联合网络通信有限公司研究院，北京 100032）

摘 要：针对云计算环境中负载难以预测并且负载量实时变化的应用，设计了一种基于时间协同的资源分配方法，该方法在应用运行过程中，根据负载量以及任务处理时间，制定出多虚拟机协同的资源分配方案。该方法采用时间协同的方式有效实现了资源复用，提高了数据中心资源利用率，并进一步降低了数据中心的运维成本。
关键词：时间协同；云计算；资源分配；数据中心
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2014)Z2-0184-07

Cloud resource allocation method based on time collaboration

YIN Bo, ZHANG Yun-yong,WANG Zhi-jun, FANG Bing-yi, FENG Wei-bin
（Research Institute of China Unicom, Beijing 100032, China）

Abstract：According to the application in the cloud computing environment load and load change is difficult to predict the amount of real time. A resource allocation method based on cooperative time was designed, the application of this method in the process of operation, according to the load and the task processing time, formulate the resource allocation scheme for multi virtual machine cooperation. The method uses time collaborative to effectively reuse resource, improves the utilization rate of data center resources, and further reduces the cost of operation and maintenance of data center. It has realistic significance to the increase of cloud resource provider profit.
Key words: elastic adjustment; cloud computing; resource allocation; data center

1 引言

云计算具有资源按需提供的特征，为用户以弹性伸缩的方式分配计算资源。该方式灵活高效的实现了资源的按需分配，并提高了资源利用率，降低了运营成本。目前，以虚拟化技术为基础的云计算资源分配方法，主要以虚拟机为粒度，对应用进行资源分配。用户与云计算资源提供商签订严格的 SLA，云资源提供商在保证用户 SLA 的前提下，根据用户应用的负载状况，自动弹性调整资源分配量。由于应用负载的到达大多具有突发性的特点，到达时间具有随机性，因此，设计基于时序的虚拟机资源分配方法具有现实意义。虚拟机协同任务分配问题的目标是在服务器集群的性能范围内，将不同物理服务器和不同收益的目标合理分配给各虚拟机，最小化代价，最大化整体效益，并保证多虚拟机任务处理的时间协同性。研究高效协同的云计算任务分配方法，有利于大幅提升云资源提供商数据中心任务处理能力，具有较大的理论价值和现实意义。

针对现有研究的不足，在弹性工作负载的场景下，分析并提出了多虚拟机任务分配问题模型，并对多虚拟机任务分配方法进行了改进。首先，对弹性工作负载的资源需求分析了异构的资源约束；然后，分析并提出了基于空闲时间窗的多虚拟机任务协同处理机制，并在此基础上提出了多虚拟机任务分配问题模型；最后，针对该模型设计了弹性资源调整的资

收稿日期：2014-10-20
基金项目：国家自然科学基金资助项目（71172134）；“新一代宽带无线移动通信网”国家科技重大专项基金资助项目（2012ZX03002001-002, 2013ZX03002004-002, 2013ZX03002003-005）
Foundation Items: The National Natural Science Foundation of China (71172134); “A New Generation of Broadband Wireless Mobile Communication Network” Major National Science and Technology Funded Projects (2012ZX03002001-002, 2013ZX03002004-002, 2013ZX03002003-005)

源分配算法，并通过仿真实验验证了算法的有效性。对比已有研究，本文的主要贡献有：

1) 提出了一种基于空闲时间窗的弹性资源分配问题模型，考虑了多虚拟机任务分配问题在时间约束和处理效率方面的要求；

2) 提出了一种多虚拟机协同的资源分配方法，有效保证了多虚拟机资源分配问题求解过程中的实时性和有效性，提高了数据中心的资源利用率。

2　基于弹性资源调整的资源分配问题描述

针对的场景是实时变化的工作负载，每个任务具有资源属性和时间属性。在该问题中，以时间为坐标轴，假设在以时间为横坐标，资源需求为纵坐标所表示的范围 $Q=[0,s]\times[0,s]$ 内，分布有 M 个目标 $T=\{T_1,T_2,\cdots,T_M\}$，其中每个目标表示一个任务，其纵坐标表示的是其资源需求，其横坐标表示的是其时间属性，即为处理该任务所要占用的时长。处理 T_i 所需的资源类型及数量表示为 $R_i=\left\{R_i^1,R_i^2,\cdots,R_i^m\right\}$，其中 R_i^k 表示目标任务 T_i 对于第 k 种资源的要求。用 N 个类型异构的虚拟机 $T_A=\{A_1,A_2,\cdots,A_N\}$，表示与用户签订 SLA 时，分配给用户的虚拟机，这些虚拟机正在运行，用其来完成对新到达工作负载的处理。对应于目标任务中各属性的需求，不同类型虚拟机所具有的剩余资源量 $R_j=\left\{R_j^1,R_j^2,\cdots,R_j^m\right\}$ 各不相同。

对新到达任务的处理，首先，弹性资源调整的资源分配的目标包括 2 个方面：1) 合理调配机群的资源，实现对范围 Q 内目标的资源分配，同时最大化任务处理收益；2) 对于新增加的弹性资源请求，要及时响应并进行处理，在尽可能短的时间内处理尽可能多的目标。

2.1　资源需求

用户与云资源提供商签订 SLA 后，首先，进行任务初始化，云资源提供商为用户开启若干虚拟机处理用户的任务，多虚拟机根据初始任务队列执行既定任务。期间某些虚拟机上出现资源空闲时，若此时到达了新目标任务 T_j，该虚拟机随即对其进行处理。由于虚拟机受到自身资源上限的约束，因此，存在单个虚拟机资源不足以处理新任务的情况，如式(1)所示

$$x_{i,j}^k r_i^k \leqslant R_i^k, \forall k\in\{1,2,\cdots,m\} \tag{1}$$

其中，$x_{i,j}^k$ 表示虚拟机 i 的第 k 种属性是否满足目标任务 T_j 相应的资源需求，用 r_i^k 表示虚拟机 i 的第 k 种资源的资源量；$x_{i,j}^k=1$ 时表示将任务分配给虚拟机 i 进行处理，否则，$x_{i,j}^k=0$；同时，为了完成对任务 T_j 的处理，还需要与其他具有资源空闲的虚拟机 A_i 与虚拟机 i 一起组成多虚拟机集合 I_A。I_A 总的资源需满足目标 T_j 对各类资源的要求如下

$$\sum_{A_i\in I_A} x_{i,j}^k r_i^k \geqslant R_j^k, \forall k\in\{1,2,\cdots,m\} \tag{2}$$

2.2　执行收益

虚拟机处理弹性资源调整后，产生任务处理收益。首先，应抽象出任务处理收益。由于云计算的工作负载具有突发的特性，因此，假设预测未到达目标任务的资源需求属于各类别的概率均为 $1/N$，进行工作负载预测所产生的信息增量可以用预测前后熵的变化量来衡量。假设虚拟机 A_i 能够准确预测目标任务的概率为 p_i，则其预测目标所产生的信息增益，表示为

$$I_i=\mathrm{lb}(N)+(1-p_i)\mathrm{lb}\left(\frac{1-p_i}{N-1}\right)+p_i\mathrm{lb}(p_i) \tag{3}$$

假设目标 T_j 的价值分别为 V_j，则分配虚拟机 A_i 来处理 T_j 任务所得的净收益表示为

$$C_{i,j}^D=I_iV_{j,D} \tag{4}$$

在预测成功工作负载后，要对任务进行处理，因此，需要建立任务处理的数学模型。当虚拟机对目标任务进行处置时，处理目标任务的不同类型资源需求可以得到不同的收益，则将虚拟机 A_i 分配来处理任务 T_j 时，所得的净收益可以表示为

$$C_{i,j}^S=\sum_{k=1}^{l}G_{i,j}^k-I(x_{i,j}^D=0),(I=1\ iff\ x_{i,j}^D=0) \tag{5}$$

$G_{i,j}^k$ 为 A_i 对 T_j 实施处理时与属性 k 相关的收益，表示为

$$G_{i,j}^k=p_{i,j}^kV_j^k,(p_{i,j}^k=f_i^k\frac{R_i^k}{R_j^k}) \tag{6}$$

其中，$p_{i,j}^k$ 表示将虚拟机 A_i 的第 k 类型属性满足资源需求，将其分配来处理任务 T_j 的完成概率；f_i^k 表示虚拟机 A_i 的第 k 类型属性可以处理目标任务 T_j 的任务完成率。

描述完任务预测收益和任务处理收益后，弹性

资源调整的资源分配总收益，可表示为

$$C_A = \sum_{i=1}^{n}\sum_{k=1}^{m} x_{i,j}^{k} C_{i,j}^{k},\ \ k \in \{D,S\} \tag{7}$$

其中，$x_{k,j}^{l}$ 表示虚拟机 i 上的资源属性 k 是否能处理目标任务 T_j。

2.3 任务处理时间

处理的问题场景是在云计算数据中心内，分配给某个用户的多虚拟机处理任务时，由于工作负载的波动，针对突发任务量，当前运行的多虚拟机在各自资源剩余和时间空闲的情况下，组成多虚拟机集合共同完成弹性资源的任务分配。多虚拟机共同处理任务时，需要各虚拟机之间必须能够实现空闲时间一致，并且满足任务所需的资源量，即各虚拟机在特定的时间段内，均能有合适的资源空闲。由于虚拟机在初始任务分配时任务不同，并且各虚拟机也可能存在异构的情况，剩余资源量和时间属性均不相同。这就需要对各虚拟机进行时间和资源协同，资源量合适的虚拟机在等待后续资源合适的虚拟机时有空闲时间的等待。对于这段等待的时间，后续实验中利用 Dubins 曲线来对其进行表示。

如图 1(a)所示为虚拟机初始任务序列，$(t_{j-1,1},t_{j,0})$ 为虚拟机从前序目标 T_{j-1} 中恢复资源可以处理后序目标 T_j 的最短时间；$(t_{j,1},t_{j,2})$ 表示对目标的处理时间；空闲时间窗即为 $(t_{j,0},t_{j,1})$，表示虚拟机理论上可以处理目标的最早时刻 $t_{j,0}$ 与任务开始时刻 $t_{j,1}$ 之间的同步等待时间。

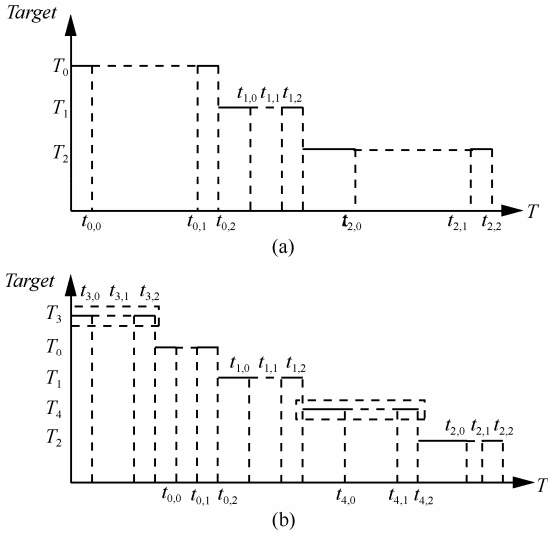


图 1 虚拟机任务处理序列

利用空闲时间窗，对任务的资源分配过程中，虚拟机的资源分配进行调整，缩短虚拟机群对新目标的响应和处理时间。例如，在 T_0 之前插入 T_3，在 T_1 和 T_2 之间插入 T_4，可以得到图 1(b)所示的任务序列，较之图 1(a)，可使虚拟机在相同时间内处理更多的任务。

将虚拟机可以处理新任务 T_j 的时刻 t_k^j ,称为空闲时间窗的起始时刻，针对虚拟机所处的不同情况，将空闲时间窗的起始时刻 t_k^j 分为 3 种类型，如图 2 所示，其中，T_0 为待执行任务，T_j 为新任务，cur 为当前虚拟机资源分配状态。

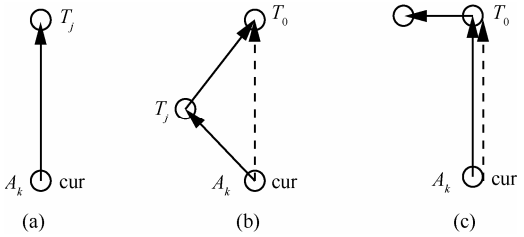


图 2 虚拟机能够处理目标的时间类型

如图 2(a)，在虚拟机当前没有初始任务进行处理时，可以将该虚拟机直接分配给任务 T_j，因此

$$t_k^j = t_k^{\mathrm{cur}-j} + t^{\mathrm{cur}} \tag{8}$$

其中，t^{cur} 表示当前时间，$t_k^{\mathrm{cur}-j}$ 表示 A_k 被分配给任务 T_j 进行任务处理的等待时间。

如图 2(b)，将虚拟机初始等待处理任务表示为 T_0，而当前虚拟机的执行时间允许先执行任务 T_j，因此，可以将 t_k^j 的计算方式同式（2）。在虚拟机执行队列中加入任务 T_j 可能对已分配任务 T_0 的处理时间产生影响，因此，需要重新计算虚拟机可以开始处理 T_0 的时间

$$t_k^{0\prime} = t_k^{j\_\mathrm{wait}} + t_k^{j-0} + t_k^{j} \tag{9}$$

其中，$t_k^{j\_\mathrm{wait}}$ 为处理 T_j 之前的与其他多虚拟机集合中虚拟机同步的等待时间，t_k^{j-0} 为从 T_j 到 T_0 的等待执行时间。

如图 2(c)，将虚拟机初始等待处理任务表示为 T_0，但当前虚拟机的执行时间不允许先执行 T_j，因此，虚拟机能够处理被分配目标的时间为

$$t_k^{j} = t_k^{\mathrm{cur}-0} + t_k^{0\_\mathrm{wait}} + t_k^{0-j} + t^{\mathrm{cur}} \tag{10}$$

将虚拟机处理完新任务 T_j 准备处理新目标的等待时刻，称为时间窗终止时刻。令 $t_k^{0\_\mathrm{process}}$ 为将该

虚拟机分配给任务 T_0 的初始预约时间，则空闲时间窗的结束时刻 $t_k^{j'}$ 可以表示

$$t_k^{j'} = t_k^{0\_\text{process}} - t^{j-0} \tag{11}$$

根据式(8)~式(11)求得处理新任务的上述起止时间后，各虚拟机需要根据如下时序约束，判断是否能将任务加入到各自虚拟机的初始任务处理列表中，以形成空闲时间窗，当第 k 个虚拟机加入新任务 T_j 后，新任务的起始时刻与前序任务 T_{pre} 和后序任务 T_{follow} 之间要满足如下时序约束

$$t_k^{pre} + t_k^{pre-j} \leqslant t_k^j \leqslant t_k^{\text{follow}} - t_k^{j-\text{follow}} \tag{12}$$

当多个虚拟机对任务进行协同处理时，要保证多虚拟机能同时具有时间空闲，并且满足资源需求的约束。如图 3 所示，用横坐标表示虚拟机 A_i 的空闲时间窗，将具有空闲时间窗重叠特征的虚拟机进行组合，形成多虚拟机集合，来共同完成对新任务的处理，已实现弹性资源调整的资源分配。用 W_i 表示虚拟机 A_i 的空闲时间窗，因此，具有如下约束条件

$$\forall A_i, A_j \in I_A, w_i \bigcap w_j \neq \phi \tag{13}$$

2.4　资源分配模型

通过 3 个方面来对协同多虚拟机时间形成集合以处理任务的收益进行评价：1) 最大化多虚拟机集合执行任务的总体收益；2) 实时处理当前新到达的弹性资源请求，符合任务的突发特性，满足任务处理的实时性要求；3) 最小化多虚拟机集合的个数，这样能保证便捷地处理当前任务，并且能保证多虚拟机及时处理更多其他新到达的任务，同时，保证使用较为集中的虚拟机来处理弹性资源请求，能实现工作的虚拟机较为集中的放置，有利于提高资源利用率，进一步节能降耗。

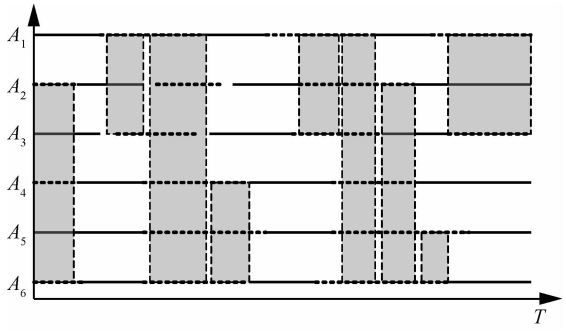


图 3　多虚拟机集合空闲时间窗

由于多虚拟机集合处理任务的开始时刻由最晚具有空闲时间的虚拟机开始时刻 t_A^j 以及多虚拟机集合中成员的总数 N_A 决定，因此，可以用多虚拟机集合中的总等待时间来表示处理任务的及时程度，以及集合中的虚拟机数量。结合任务执行收益，得到以总等待时间最短和任务执行收益最大为目标的资源分配模型，v 表示虚拟机对任务的单位处理能力，如式(14)所示

$$\begin{aligned}
&\max \sum_{i=1}^{n}\sum_{k=1}^{m} x_{i,j}^k C_{i,j}^k \Big/ \sum_{i=1}^{n} x_{i,j}^k v T_A \\
&\text{s.t.} \begin{cases}
x_{i,j}^k r_{i,j}^k \leqslant R_i^k, \forall k \in \{1,2,\cdots,m\}; \\
\sum_{i=1}^{n} x_{i,j}^k r_i^k \geqslant R_j^k, \forall k \in \{1,2,\cdots,m\}; \\
w_A = \bigcap_{i=1}^{n} w_i \neq \phi, \forall A_i \in I_A; \\
T_A = \sup(\min w_A)。
\end{cases}
\end{aligned} \tag{14}$$

由于时间窗约束的限制，目前现有的针对整数规划问题的方法并不适用于求解本文模型，接下来本文设计方法对上述数学模型进行求解。

3　多虚拟机时间协同的资源分配方法

设计了一种多虚拟机时间协同的资源分配方法，对上述基于时间窗的多虚拟机资源分配模型进行求解。首先，当虚拟机 A_i 预测到新到达的任务后，立刻将目标集合 $T_i = \{T_i^1, T_i^2, \cdots, T_i^j, \cdots\}$ 与事先进行分配的令牌 $token_i$ 在当前租户内进行广播，通知到所有的当前用户的虚拟机，并且避免多虚拟机资源竞争，避免资源死锁情况的形成。当虚拟机 A_i 获得请求资格后，对任务 T 以格式 $R_i = \{A_i, T\}$ 进行拍卖。其他虚拟机根据式(8)~式(12)，各自分别计算将任务 T 加入到当前任务处理序列 $E_j = \{E^1, E^2, \cdots, E^m\}$ 的不同位置后，所能产生的新任务处理序列时间窗 $tw_j = \{w^1, w^2, \cdots, w^m\}$，并且根据式(3)~式(6)计算得到任务处理收益，并且将任务处理收益值返回给虚拟机 A_i。虚拟机 A_i 收到各个邻居虚拟机的应答后，根据式(14)计算出最佳的多虚拟机集合。随后将任务与各个虚拟机的任务处理序列进行发布，如果虚拟机 A_i 进入到多虚拟机集合中，则该虚拟机将对其任务处理序列进行相应的调整，将新任务加入到其任务处理序列中，否则，将该虚拟机任务请求序列中的重复任务进行删除。

任务请求发起者采用基于时间窗的多虚拟机集合形成方法（TWVM, time window based virtual

machine set algorithm）组成多虚拟机集合。

为了消除时间窗重叠要求对问题求解的影响，按照各虚拟机空闲时间窗起始时间的升序对虚拟机进行选择，形成候选虚拟机集合。一旦满足处理任务所需要的资源约束后，候选虚拟机集合即形成。因为集合的形成时间由最晚能够处理新任务的虚拟机决定，当候选虚拟机集合形成时，便可以确定多虚拟机集合进行任务处理的最早时间。将式(14)中的时间窗约束去掉，转化为标准的整数规划问题，可以求接出候选虚拟机集合的最优解。引入各虚拟机的资源贡献度 d_i，根据 d_i 从候选集合中采用贪婪策略依次选出对集合资源贡献度最高的成员，形成最终集合组成方式。算法如下。

算法 1 (TWVM, time window based virtual machineset algorithm)

输入：候选虚拟机集合

输出：多虚拟机任务处理集合

1) 初始化候选虚拟机集合。

2) 选择应答中时间窗起始时刻最早的 A_j 加入候选集。若候选集满足目标资源要求，则转阶段 2；否则转步骤 3)。

3) 从与候选虚拟机集合中虚拟机 A_j 相异的剩余虚拟机中，选择时间窗起始时刻最早的虚拟机 A_k 加入候选虚拟机集合，并检查是否与当前候选集中现有虚拟机的时间窗相互重叠。若重叠($\forall A_j \in I_A, W_k \bigcap W_j \neq \phi$)，则删除 A_k。若当前候选虚拟机集合满足目标资源要求，则转到阶段 2，选择最优组合；否则，重复步骤 3)。

4) 若所有的时间窗都被遍历，或已经包含了当前用户的所有虚拟机，仍无结果，则（集合）组成失败，则为用户的弹性工作负载开启新的虚拟机。

5) 初始化任务联盟集合，选择候选虚拟机集合中最后加入的成员加入任务集合：若任务集合满足目标资源要求，则求得最终任务集合，转步骤 8)；否则转步骤 6)。

6) 选择候选虚拟机集合中资源贡献度 d_i 最大的 A_j 加入任务集合。若任务集合满足目标资源要求，则求得最终集合，转步骤 8)；否则转步骤 7)。

7) 从与任务集合中成员 A_j 相异的剩余虚拟机中，选择资源贡献度最大的 A_k 加入任务集合。若当前任务集合满足目标资源要求，则求得最终集合，转步骤 8)；否则，重复步骤 7)。

8) 形成最终任务集合，算法结束。

资源贡献度 d_i 表示虚拟机 i 所能提供的资源对于处理目标所需资源的贡献比例。

$$d_i = \sum_{j=0}^{l} \omega^j \frac{D_i^j}{R_A^j} \tag{15}$$

其中，ω^j 表示第 j 类资源的权重；R_A^j 表示在已加入现有虚拟机的条件下，集合尚缺少的第 j 类资源的数量；D_i^j 表示虚拟机 i 所具有的第 j 类资源对于集合的贡献程度

$$D_i^j = \begin{cases} R_i^j, R_i^j \leqslant R_A^j \\ R_A^j, \text{其他} \end{cases} \tag{16}$$

在进行优化求解时，需要保留最后加入的成员，否则，最终任务集合将无法满足目标所需要的资源。

4 仿真结果

设计了仿真实验，来验证所提方法的效果，评价指标为任务的总体完成率。仿真结果表明，TCVM 算法能够有效提高弹性资源调整的突发任务完成率，并有效缩短总体的任务执行时间，有效保证了在突发任务的情况下，弹性资源调整的资源分配问题，保证了云计算弹性扩展的特性。

实验环境为 8-core 2 GHz AMD Opteron 2350 服务器和 4-core 2.4 GHz Intel Xeon X3220 系统。所有服务器都是用 Xen 3.3 和 Linux 2.6.18（64 bit 内核），开发工具为 Microsoft Visual Studio 2005。实验结果重复 1 000 次平均获得（每次执行时间上限为 1 000 s）。假设虚拟机的单位任务处理速度为 100，各虚拟机和任务突发的资源属性和时间点均随机产生，并且各类资源数量的生成曲线均服从正太分布。

图 4 表示的是在一段时间内，任务的总体完成情况。对比算法是传统的 first-fit profit（FFP）装箱算法，从图 4 中，可以看出，当虚拟机的数量与任务到达数量基本一致时，表明虚拟机具有足够多的资源能力处理突发的弹性任务，此时，FFP 算法的任务完成率与 TWVM 算法的完成率基本一致，二者均能较好地完成工作任务，保持较高的任务总体完成率。当任务所需的资源能力大于当前虚拟机所能提供的资源能力时，此时 FFP 算法中的虚拟机不足以组成多虚拟机集合实时处理弹性资源需求，因

此，FFP 的任务总体完成率相较 TWVM 算法开始明显下降。

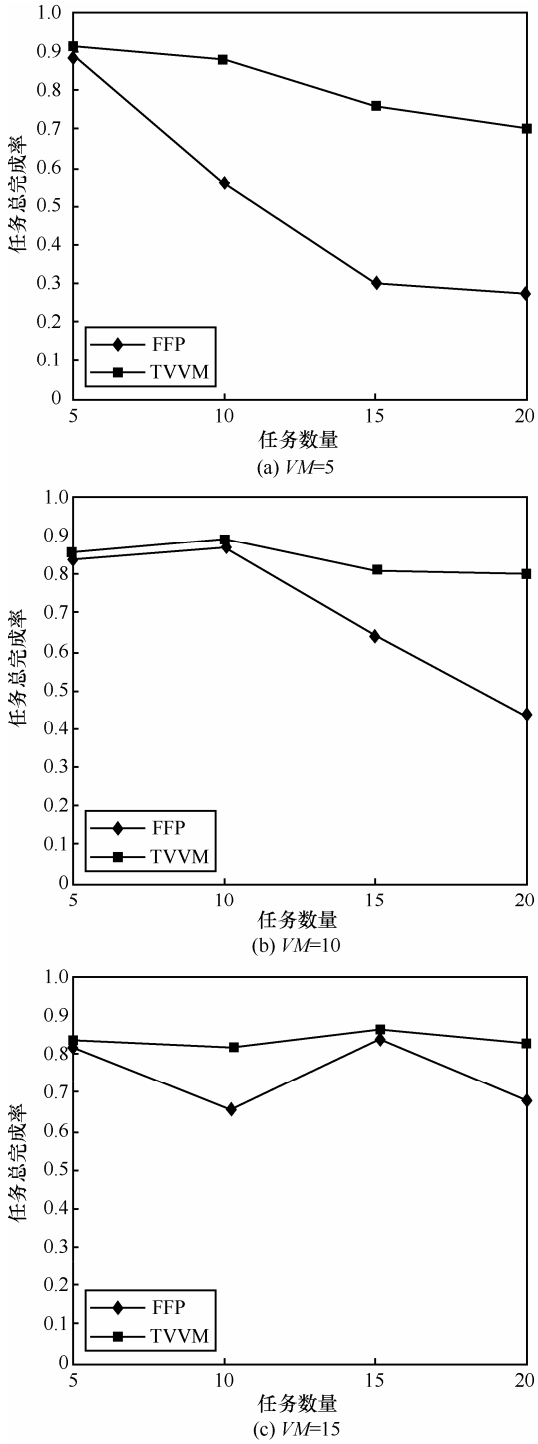


图 4　任务总完成率对比

实验从任务数量为 5、10、15 的 3 种情况尽心讨论。多虚拟机集合的成员个数决定了时间窗的数量，因此，时间窗的数量也由任务数量来决定。仿真实验中的虚拟机数量分别为 5、10、15，实验中任务数量的大小也依次设为 0、5、10。

5　相关工作

目前，基于 SLA 的资源管理策略的研究有很多，文献[1]提出了一种弹性可扩展的 Web 服务提供平台，该平台包括动态负载预测模块和 SLA 驱动的资源管理模块。文献[2]为降低云资源提供商的总体资源成本，提出了一种全局资源优化模型。文献[3]以最大化响应时间并且最大化物理服务器的资源利用率来降低资源成本。以降低云资源提供商成本为目标的研究中，文献[4]提出了一种单层次的虚拟机资源调整方法，实现了 IPs 的收益优化。文献[5]针对动态 Web 应用提出了一种自动弹性伸缩策略。文献[6]提出了一种虚拟机弹性调整机制，用来解决虚拟化环境中的资源分配问题。文献[6]提出了一种细粒度控制虚拟资源的平台 Mesos。针对当前应用架构和调度器之间不具有兼容性的特点，设计了一种两层架构的资源共享平台，用来实现细粒度的资源管理，并提出了资源分配优化策略 DRF，实现了异构平台之间策略的通用性。

目前，基于弹性资源调整的资源分配方法中，大多仅采用虚拟机大小调整、虚拟机数目调整或物理资源调整等某一种资源调整方法，缺乏整合多种资源的调整方式。此外，目前缺乏能有效解决时变的弹性资源调整策略。因此，需要设计一种在保证用户 SLA 的前提下，有效实现资源复用，提高资源利用率的资源调整方法。提出的策略能满足上述需求，以较小的资源开销，达到资源调整的目标。

6　结束语

基于云资源提供商的立场，针对实时变化的负载请求以及应用的多样性，提出了一种弹性资源调整方法。该方法在保证用户 SLA 的前提下，实现了资源复用，有效提高了资源利用率，降低了数据中心的运营成本，对增加云资源提供商的收益具有现实意义。

参考文献：

[1] ARDAGNA D, TRUBIAN M, ZHANG L. SLA based resource allocation policies in autonomic environments[J]. Journal of Parallel and Distributed Computing, 2007, 67(3): 259-270.

[2] VAN H N, TRAN F D, MENAUD J M. SLA-aware virtual resource

management for cloud infrastructures[A]. Proceedings of IEEE 9th International Conference on Computer and Information Technology[C]. Xiamen, 2009.357-362.

[3] IQBAL W, DAILEY M, CARRERA D. SLA-driven Adaptive Resource Management for Web Applications on a Heterogeneous Compute Cloud[M]. Cloud Computing. Springer Berlin Heidelberg, 2009. 243-253.

[4] TSAKALOZOS K, KLLAPI H ANDSITARIDI E, *et al*. Flexible use of cloud resources through profit maximization and price discrimination[C]. Proceedings of 2011 IEEE 27th International Conference on-Data Engineering[C]. Hannover, 2011.75-86.

[5] SAMPAIO L R, LOPES R V. Towards practical auto scaling of user facing applications[A]. Proceedings of 2012 IEEE Latin America Conference on Cloud Computing and Communications[C]. Porto Alegre, 2012.60-65.

[6] DAWOUD W, TAKOUNA I, MEINEL C. Elastic VM for rapid and optimum virtualized resources' allocation[A]. Proceedings of 2011 5th International DMTF Academic Alliance Workshop on Systems and Virtualization Management[C]. Paris, 2011.1-4.

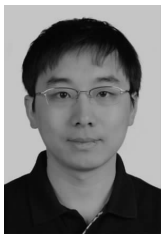
[7] HINDMAN B, KONWINSKI A, ZAHARIA M, *et al*. Mesos: a platform for fine-grained resource sharing in the data center[A]. Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation[C]. USENIX Association, 2011.22.

作者简介：

殷波（1984-），女，山东聊城人，博士，中国联合网络通信有限公司研究院工程师，主要研究方向为软件定义网络、云计算、网络融合等。

张云勇（1978-），男，江苏盐城人，博士后，中国联合网络通信有限公司研究院副院长、高级工程师，主要研究方向为下一代网络、网络融合、云计算、大数据等。

王志军（1976-），男，辽宁丹东人，中国联合网络通信有限公司研究院高级工程师，移动互联网实验室、云计算实验室主任，主要研究方向为业务平台及支撑系统体系架构、面向移动互联网的开放平台、新业务及下一代网络等。



房秉毅（1980-），男，山东泰安人，博士，中国联合网络通信有限公司研究院高级工程师，主要研究方向为云计算、核心网、网络融合等。

冯伟斌（1982-），男，山西运城人，中国联合网络通信有限公司研究院工程师，主要研究方向为云计算、互动媒体、行业信息化等。