

云环境下基于功耗感知的虚拟机博弈迁移算法^{*}

曾 凯, 余 堃, 敬思远

(电子科技大学 计算机科学与工程学院, 成都 611731)

摘 要: 为了达到降低云环境下数据中心过高能耗的目的, 提出一种基于功耗感知的虚拟机非合作博弈迁移算法 PANMA。该算法主要分为基于中值绝对偏差的物理机选择、基于负载最大相关的虚拟机选择和基于非合作博弈的虚拟机迁移三个步骤。实验结果表明, 该算法极大地减少了数据中心能量消耗。

关键词: 云环境; 非合作博弈; 能耗; 虚拟机迁移

中图分类号: TP301.6

文献标志码: A

文章编号: 1001-3695(2013)06-1668-04

doi:10.3969/j.issn.1001-3695.2013.06.016

Game theoretic migration algorithm based on power-aware for virtual machines in cloud

ZENG Kai, SHE Kun, JING Si-yuan

(School of Computer Science & Engineering, University of Electronic Science & Technology of China, Chengdu 611731, China)

Abstract: For reducing energy consumption in data center, this paper proposed a new non-cooperative migration algorithm PANMA based on power-aware. The power saving was achieved by three steps, including host selection based on median absolute deviation, virtual machine selection based on maximum correlation policy, virtual machine migration based on non-cooperative game. The experimental results show that the proposed algorithm greatly saves the energy consumption in data center.

Key words: cloud environment; non-cooperative game; power consumption; virtual machine migration

0 引言

随着信息技术的高速发展, 云计算逐渐成为学术界研究的热点。其中数据中心作为云计算服务载体, 扮演了至关重要的角色。传统的研究热点基本集中于通过对数据中心资源的合理管理, 满足上层应用高性能计算的需求, 却极少关注云环境下数据中心过高能源消耗问题。麦肯锡研究报告^[1]指出, 一个普通数据中心的平均能源消耗量相当于 25 000 个普通家庭能耗总和; 2010 年全球数据中心能耗花费约为 115 亿美金, 并估算一个普通的数据中心能耗花费平均每 5 年翻一倍。伴随着能源价格的不断增长和可利用能源的不断减少, 在满足数据中心高性能计算的同时兼顾能耗的合理利用这一需求被逐渐提出, 并得到业界的极大关注。

动态电压缩放技术 (dynamic voltage scaling, DVS) 是一种有效的传统节能技术。文献[2]中提出, 在程序运行过程中动态调节处理器电压使得处理器不总以最高电压工作, 以此达到节能的目的, 但其没有考虑到电压状态频繁切换所带来的时耗, 无法满足云环境下大规模计算机集群的节能需求。数据中心通常布设大量的物理主机, 对这些设备制冷降温往往需要消耗更大的电能。针对这一问题, 文献[3]提出了基于空气热力学原理, 通过合理布局物理主机位置空间来达到降低数据中心制冷能耗的目的。但此类方法需要改变机房的物理布局, 部分机房可能需要迁移到高原地区, 实施起来具有一定的局限性。文献[4]提出了基于温度感知的负载迁移节能算法, 通过合理

分配过热主机上的负载来降低能耗。其主要思想是通过数据中心内部负载的合理整合 (如虚拟机的迁移) 来实现节能的目的。这种软件实现方式具有操作简单、方便移植等优点, 逐渐成为数据中心节能的主流解决方案。

虚拟机迁移^[5]、系统负载均衡等组合优化问题的传统解决办法中, 往往集中关注最终结果的整体最优, 却对参数的自私性行为缺乏合理的解释。例如在以节能效果为最终目标的云环境下, 一方面系统要降低整体能耗, 为不均匀的虚拟机选择合适的物理主机; 另一方面每一个虚拟机都不愿意受系统的束缚, 希望在当前条件下寻找自身对能耗影响最小的物理主机, 以达到更好的节能表现, 这对于以往的虚拟机迁移策略研究提出了具有挑战性的新要求。因此, 有必要引入新的研究方法和理论, 博弈论为该研究提供了坚实的数学基础和合理的解释。

本文针对以上问题, 提出了基于功耗感知的虚拟机非合作博弈迁移算法 (power-aware non-cooperative migration algorithm, PANMA)。一方面, 通过虚拟机的高效迁移, 释放出低负载的物理主机并将其切入休眠状态, 达到降低能耗的目的; 另一方面, 在虚拟机迁移过程中, 充分考虑到对迁入目的主机的能耗影响, 通过虚拟机之间的非合作博弈, 以近似纳什均衡状态作为最终的迁移策略, 并在实验平台上对该算法进行了测试。

1 系统分析

1.1 功耗模型

数据中心的能量消耗主要来源于计算节点的 CPU、内存、

收稿日期: 2012-09-11; 修回日期: 2012-11-08 基金项目: 国家“863”计划资助项目 (2008AA04A107); 粤港关键领域重点突破项目 (2009A98B21)

作者简介: 曾凯 (1985-), 男, 辽宁营口人, 博士研究生, 主要研究方向为云计算、绿色计算 (zengkailink@sina.com); 余堃 (1967-), 男, 教授, 博导, 主要研究方向为云计算; 敬思远 (1981-), 男, 博士研究生, 主要研究方向为云计算、绿色计算。

硬盘等物理设备。由于 CPU 消耗了其中绝大部分能量,本文的研究重点集中于 CPU 的能耗影响。近期的研究结果^[6,7]表明,CPU 的当前功率和其效用呈线性关系。同时指出,空闲状态的 CPU 的能耗相当于满载状态 CPU 的 70%。本文充分利用这一结论,利用虚拟机的迁移来释放低负载的物理主机,并将其切换到休眠状态,以达到降低能耗的目的。其中能耗与负载的关系如式(1)所示。

$$P(u) = k \times P_{\max} + (1 - k) \times P_{\max} \times u \quad (1)$$

其中: $P_{\max}$ 代表 CPU 满载时的最大功率; k 代表空闲状态 CPU 的功耗率(通常取 70%); u 代表 CPU 当前效用。云环境下,由于负载的动态变化,CPU 效用是一个关于时间 t 的函数 $u(t)$,因此在一段时间内,某计算节点的能耗如式(2)所示。

$$E = \int (P(u(t))) \quad (2)$$

1.2 绿色数据中心架构

本文的研究内容是通过合理的虚拟资源整合,在满足计算服务需求的同时减少能源消耗,实现绿色数据中心。绿色数据中心的架构如图 1 所示。

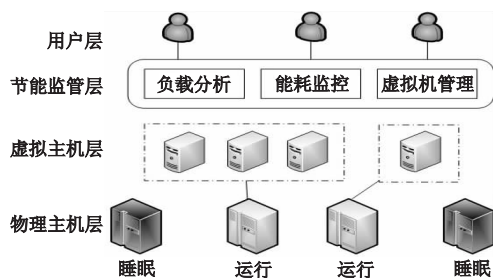


图1 绿色数据中心架构

其内部按层次划分由以下四部分组成:

- 用户层。云服务的使用者,如终端用户、Web 服务开发者等。
- 节能监管层。负责负载分析、能耗监控、虚拟机管理。
- 虚拟主机层。可以启动、停止、挂起以满足不同服务需求;通过虚拟机主动迁移,释放低利用率物理主机来达到降低能耗的目的。
- 物理主机层。向上层虚拟机提供物理运行环境。

节能监管层负责监控系统负载及主机能耗,同时向虚拟主机层、物理主机层反馈这些信息,并响应下层虚拟机的迁移请求,这一过程对上层用户是完全透明的;虚拟主机层计算节点获得当前系统整体负载和能耗状况,向上层提交迁移策略,通过与其他虚拟主机的非合作博弈寻找使整体能耗最低的物理主机;最下层物理主机负载为空时,及时切换到休眠状态,避免空负载下的能耗浪费。

1.3 非合作博弈迁移模型

博弈论是经济学、运筹学、计算机科学等领域的重要分析工具之一。本文中非合作博弈的引入为虚拟机主动的迁移行为进行了更加合理的科学解释。

定义 1 博弈系统表示为 $G = \{P, S, F\}$ 。其中: P 代表博弈的参与者集合 $P = \{P_1, P_2, \dots, P_n\}$; S 代表策略矩阵,其中 $S^i = \{S_1^i, S_2^i, \dots, S_n^i\}$ 表示第 i 个博弈参与者的策略集合; F 代表收益函数集合, $F = \{F_1, F_2, \dots, F_n\}$ 。本文中,博弈参与者为虚拟机;选择策略为要迁移的目标主机 ID;根据式(1),收益函数为 $\min \Delta(i) = P(u^i) - P(u^{-i})$,其中 $\Delta(i)$ 表示某主机在迁入第 i 台虚拟机后与不迁入这台虚拟机的功耗差。换句话说,虚拟机

每轮博弈都选择使得 $\Delta(i)$ 最小的物理主机。

本文系统选择了博弈理论中的非合作博弈模型。非合作博弈是指一种参与者不可能达成具有约束力的协议的博弈类型,这是一种具有互不相容的情形。非合作博弈研究人们在利益相互影响的局势中如何选决策使自己的收益最大,即策略选择问题。在非合作博弈中,参与者独立做决策,相互之间没有合作。博弈最终达到一种均衡状态,称之为纳什均衡。纳什均衡指的是这样一种策略组合,这种策略组合由所有参与人最优策略组成。即在给定别人策略的情况下,没有人能够在其他人不改变策略的情况下,通过修改自己的策略获得更大的收益。在本文系统环境中,虚拟机不满足系统分配,希望通过主动迁移而获得更好的节能效果,并最终所有虚拟机都选择了这种最满意的物理主机,达到虚拟机博弈的纳什均衡。

定义 2 博弈参与者连续两次博弈收益变化百分比小于 ξ ,则不修改其策略,称之为最优反映,博弈的最终结果称为 ξ -均衡。

由于严格的纳什均衡收敛时间的不确定性,本文采用这种近似纳什均衡的解决办法,更加符合实际系统的要求。

2 算法设计

虚拟机可以动态迁移的特性,充分满足了不同级别计算需求。当物理主机上的负载较小时,可以将在其上运行的虚拟机迁移到其他适合物理主机上,并将这些空闲出来的物理主机切换到睡眠状态,以此减少空闲主机的能量消耗。在这一过程中需要解决三个问题:a)哪些物理机上的虚拟机需要调整;b)选择哪些虚拟机进行迁移;c)虚拟机具体迁入到哪台物理机。下面针对这三个问题进行详细讨论并给出相应算法。

2.1 物理机选择策略

物理机选择就是要确定负载过高和过低的物理主机。最简单的做法是设定物理 CPU 效用的上、下界限作为判断标准。然而在很多情况下,固定的阈值很难适应复杂的云环境,系统需要一个可以根据负载情况自适应调整的阈值。中值绝对偏差(median absolute deviation, MAD)比样本方差或标准差等统计参数更加具有鲁棒性,因此本文使用中值绝对偏差策略来动态调整阈值。其主要思想是以物理主机生命周期内的历史 CPU 效用的中值绝对偏差作为物理机负载阈值的计算参数。这里的 CPU 效用是指假设某台物理主机的 CPU 计算能力为 1 000 MIPS,上层虚拟机带来的计算负载为 750 MIPS,则 CPU 效用为 75%。设递增数列 $X_1, X_2, \dots, X_n$ 为某物理主机的历史 CPU 效用,MAD 的值如式(3)所示。

$$MAD = \text{median}_i (|X_i - \text{median}_j (X_j)|) \quad (3)$$

阈值上界 T_u 、下界 T_l 定义如式(4)所示。

$$\begin{cases} T_u = 1 - s \times MAD, T_l = s \times MAD & \text{当 } s \times MAD < 0.5 \\ T_u = 0.9, T_l = 1 - s \times MAD & \text{当 } s \times MAD \geq 0.5 \end{cases} \quad (4)$$

根据式(4),依次找到 CPU 效用超过阈值范围的主机,剩下的中间物理主机作为适合迁入的物理主机。其中, s 代表对虚拟机迁移程度的需求,通常 s 取值在 2.5 左右较为合理。

2.2 虚拟机选择策略

云计算的本质是计算服务,如果仅为了节约能耗而将过多的负载集中,将大大降低云的服务质量。因此,虚拟机选择包

含两个部分:a)将虚拟机从过低负载的物理主机上全部迁出,并休眠该物理主机以降低能耗;b)将过高负载物理机上的部分虚拟机迁出,通过降低该主机上的负载,确保云计算服务质量。确定高负载的物理主机上哪些虚拟机需要迁移是本节讨论的问题。对此,本文采取负载最大相关虚拟机选择策略。文献[8]提出在物理主机上不同应用负载之间的相关性越大,则该主机越可能处于过高的负载。换句话说,在某个过高负载的物理主机上,应首先选择那些CPU负载相关性较高的虚拟机。本文采用重相关系数平方值来评估这些虚拟机CPU负载的相关程度。这里的CPU负载等同于2.1节中提到的负载。假设某台物理主机有 n 台虚拟机,在连续 w 个时间采样点内,第 i 台虚拟机的CPU历史负载为 $X_i = \{X_1^i, X_2^i, \dots, X_w^i\}$,为了判断第 k 台虚拟机是否需要迁移,用 Y 表示该台虚拟机的负载, X 表示剩余虚拟机的负载增广矩阵,如式(5)所示。

$$X = \begin{bmatrix} 1 & x_1^1 & \dots & x_w^1 \\ \vdots & \vdots & & \vdots \\ 1 & x_1^{k-1} & \dots & x_w^{k-1} \\ 1 & x_1^{k+1} & \dots & x_w^{k+1} \\ \vdots & \vdots & & \vdots \\ 1 & x_1^n & \dots & x_w^n \end{bmatrix}, Y = \begin{bmatrix} x_1^k \\ \vdots \\ x_w^k \end{bmatrix} \quad (5)$$

进行重回归分析,因变量的预测值用 $\bar{Y}$ 表示,具体计算为

$$\bar{Y} = Xb; b = (X^T X)^{-1} X^T Y \quad (6)$$

根据预测值,重相关系数平方 $R_{X^k, X^1, \dots, X^{k-1}, X^{k+1}, \dots, X^w}^2$,如式(7)所示:

$$R_{X^k, X^1, \dots, X^{k-1}, X^{k+1}, \dots, X^w}^2 = \frac{\sum_{i=1}^w (Y_i - m_Y)^2 (\bar{Y}_i - m_{\bar{Y}})^2}{\sum_{i=1}^w (Y_i - m_Y)^2 \sum_{i=1}^w (\bar{Y}_i - m_{\bar{Y}})^2} \quad (7)$$

其中: Y_i 和 $\bar{Y}_i$ 分别表示真实值和预测值向量第 i 个元素的值; m_Y 和 $m_{\bar{Y}}$ 分别表示真实值和预测值的均值。综上,当该物理主机的CPU效用在阈值上界与下界之间时,不进行迁移;当小于下界时,全部迁移;当大于上界时,依次选择 $R_{X^k, X^1, \dots, X^{k-1}, X^{k+1}, \dots, X^w}^2$ 的值最大的第 k 台虚拟机,直至该物理主机的CPU效用在上界与下界之间。具体算法如下:

输入:主机列表 hostList;负载阈值为 T_u 、 T_l 。

输出:迁移虚拟机列表 migrationList;

for hostList 中的每个 host

vmList←当前物理主机上的虚拟机列表

util←当前物理主机负载

while util > 负载上界

vm←非合作博弈找到能耗最小物理主机

vm←加入迁移队列并从当前队列中移除

util←计算移除 vm 后的物理主机负载

end while

if util < 负载下界

所有虚拟机全部加入迁移队列

休眠该物理主机

end if

end for

return 迁移虚拟机队列

2.3 非合作博弈迁移策略

根据定义1、2,虚拟机通过寻找使 $\Delta(i)$ 最小的物理主机来给出自己的最优反映策略。为了快速收敛,本文采用虚拟机序贯出价的博弈方式,虚拟机依次给出自己的策略。当第 i 台虚

拟机连续两次博弈的 $\Delta(i)$ 变化百分比小于 ξ ,则退出博弈;另一方面,当有超过90%的虚拟机退出博弈,则认为达到了近似的纳什均衡状态。具体算法如下:

输入:迁移虚拟机队列 migrationList;可迁入虚拟机的主机列表 adapList。

输出:退出博弈虚拟机队列 exitGameList。

while(1)

for 在 migrationList 中的每一个 vm

host←在 adapList 中找到最小能耗影响主机

t ←连续两次迁移策略带来的能耗变化差

if(能耗变化比 < ξ)

迁移该 vm,并加入退出博弈列表

else

该 vm 继续参加博弈

end if

end for

if(退出博弈虚拟机百分比 > 0.9)

博弈结束

end if

end while

return exitGameList

算法整体执行过程中,系统节能管理层组件不断轮询各物理主机的当前负载并分析历史记录,确定 T_u 和 T_l 。当出现过高或过低负载物理主机时,则启动非合作博弈迁移算法,直至达到虚拟机迁移的纳什均衡状态,在确保服务质量的同时降低数据中心的能量消耗。

3 实验分析

本文实验有三个目的:a)通过对比实验,评估本文算法的节能效果;b)通过非合作博弈实验分析,阐述虚拟机博弈过程,以此科学解释虚拟机希望得到更好节能表现的自私性行为;c)分别测量每次通过PANMA算法休眠物理主机数量和切换时耗,验证实验设计与仿真的合理性。

本文算法若首先实验于实际的数据中心,需要搭建大规模的物理主机和虚拟机,并在此之上做大量重复实验,这将大大降低实验的效率,影响算法初步的评估。因而本文采用仿真实验的方法。本文使用CloudSim^[9]作为仿真工具,相比其他的仿真工具如Sim++、GangSim等,CloudSim支持按需自定义物理主机、虚拟机以及上层应用负载,其内核支持能耗测量扩展,这为本文实验提供了便利。本实验采用仿真HP ProLiant ML110 G5系列和IBM server x3550系列服务器组合的方式作为物理主机群;仿真CPU计算能力分别为500、750和1000 MIPS,1 GB内存,350 GB硬盘作的虚拟主机群;上层应用负载数据来源于CoMon^[10]项目,本文随机选择2011年3~4月,由CoMon监控获得的全球500多个不同地区的1000多台虚拟主机历史负载,为本实验中虚拟机随机分配。

3.1 节能效果实验分析

本实验将分别仿真800、1400、2000台物理主机作为测试环境。以非能量感知算法(non-power aware algorithm, NPAA)和DVS算法作为本文PANMA算法对比实验。其中NPAA算法在系统运行时只考虑计算效率而不考虑任何节能因素,DVS方法则不通过虚拟机迁移来节能。实验结果如图2所示。

实验表明,PANMA 算法比 NPAA 算法在节能方面有了显著提高。当负载一定、物理主机数量提高时,由于 NPAA 算法只考虑计算性能,使尽可能多的物理主机参与计算服务,导致能耗过高;而 PANMA 算法在保证满足服务质量的同时,使虚拟机运行在相对集中的物理主机内,使更多的其他物理主机转入睡眠状态,从而使能耗维持在较低的稳定状态。PANMA 算法与传统的 DVS 算法相比也有较好的节能表现,但两者并不冲突,在实际系统中可以配合使用,使整体达到最佳节能效果。

3.2 非合作博弈迁移策略

本实验将分别仿真 800 台物理主机和 2 000 台虚拟机作为测试环境,其 ξ 分别取 0.1、0.01 和 0.001。实验记录系统虚拟机第一次博弈迁移时,轮博弈后退出博弈的虚拟机总数量。实验结果如图 3 所示。

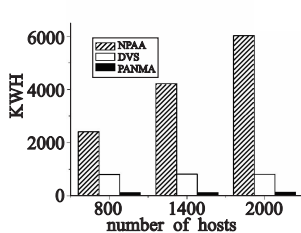


图2 节能效果对比

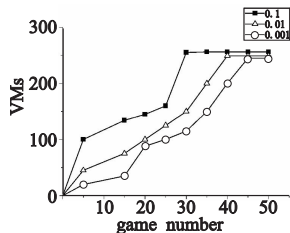


图3 虚拟机非合作博弈迁移

图 3 中横坐标表示博弈轮次序号,纵坐标表示在博弈中达到均衡状态退出博弈的虚拟机台数。实验结果表明,随着博弈次数的增加,越来越多的虚拟机无法通过修改迁移策略来获得更好的节能表现。当这些虚拟机的数量达到所有需要迁移虚拟机总数的 90% 时,达到近似纳什均衡状态,博弈过程结束。这一过程合理解释了虚拟机不满足系统分配,希望自身获得更好的节能表现的自私性行为,解决了传统的资源分配优化等问题中,采用一次性决策,参与者服务需求满意率较低的问题,提高了系统资源分配的科学性和公平性。另一方面,当 ξ 取值较大时,博弈收敛速度较快,但与严格的纳什均衡差距增大,因为严格的纳什均衡要求所有个体都达到最优反映,而 ξ 恰好反映了博弈参与者对近似最优与严格最优差异的容忍度。然而虚拟机负载有一定的随机性与复杂性,因而在实际的系统中,需要动态选择合适的 ξ 来满足实际需要。

3.3 物理主机切换睡眠状态效果分析

本文主要是通过对低负载物理主机的休眠状态切换达到节能的目的,所以对这一情景的测量与分析是必不可少的。本实验将仿真 800 台物理主机和 2 000 台虚拟机,分别测量每次通过 PANMA 算法休眠物理主机数量和切换时耗。具体数据如表 1、2 所示。

表 1 主机休眠台数测量

平均	中值	标准差
199.1	204	18.2

表 2 休眠切换时耗

平均	中值	标准差
221.1	160	35.1

从表 1 数据可得,在该实验规模下,PANMA 算法平均每次可以将近 25% 的物理主机转换至休眠状态,以降低总体能耗;从表 2 数据可得,每台主机状态切换平均耗时为 221.1 ms。文献[11]指出,普通的刀锋服务器从满负载时的 450 W 功耗切换到睡眠状态的 10.4 W 功耗,需要耗时 300 ms 左右。这些数字表明本文实验结果与实际环境是吻合的,更进一步说明了本文实验参数的设计和对负载、虚拟机、物理机的仿真是科学合理的。

4 结束语

本文提出了一种基于能量感知的虚拟机非合作博弈算法,在实验条件下取得了较好的节能表现,并验证了实验设计和仿真的合理性。在下一步的研究中将着重于将算法移植到真实的云环境下^[12,13],如 OpenStack^[14]等。另外,本文系统没有考虑虚拟机迁移性能开销^[15],因此在后续的研究中,将引入此问题进一步讨论。

参考文献:

- [1] KAPLAN J, FORREST W, KINDLER N. Revolutionizing data center energy efficiency[R]. California: McKinsey & Company, 2009.
- [2] CHASE J S, ANDERSON D C, THAKAR P N, et al. Managing energy and server resources in hosting centers[C]//Proc of the 18th ACM Symposium on Operating Systems Principles. New York: ACM Press, 2001: 103-116.
- [3] TANG Qing-hui, GUPTA S K S, VARSAMPOPOULOS G. Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: a cyber-physical approach[J]. IEEE Trans on Parallel and Distributed Systems, 2008, 19(11): 1458-1472.
- [4] MOORE J, CHASE J, RANGANATHAN P, et al. Making scheduling "cool": temperature-aware workload placement in data center[C]//Proc of USENIX Annual Technical Conference. San Diego: USENIX, 2005: 61-75.
- [5] 刘永, 王新华, 王朕, 等. 节能及信任驱动的虚拟机资源调度[J]. 计算机应用研究, 2012, 29(7): 2479-2483.
- [6] RAGHAVENDRA R, RANGANATHAN P, TALWAR V, et al. No "power" struggles: coordinated multi-level power management for the data center[C]//Proc of the 13th International Conference on Architectural Support for Programming Languages and Operating Systems. New York: ACM Press, 2008: 48-59.
- [7] KUSIC D, KEPHART J O, HANSON J E, et al. Power and performance management of virtualized computing environments via look ahead control[J]. Cluster Computing, 2009, 12(1): 1-15.
- [8] VERMA A, DASGUPTA G, NAYAK T K, et al. Server workload analysis for power minimization using consolidation[C]//Proc of USENIX Annual Technical Conference. San Diego: USENIX, 2009.
- [9] BUYYA R, RANJAN R, CALHEROS R N. Modeling and simulation of scalable cloud computing environments and the cloudsims toolkit: challenges and opportunities[C]//Proc of the 7th High Performance Computing and Simulation Conference. [S. l.]: IEEE Press, 2009: 1-11.
- [10] PARK K S, PAI V S. CoMon: a mostly-scalable monitoring system for PlanetLab[J]. ACM SIGOPS Operating Systems Review, 2006, 40(1): 65-74.
- [11] MEISNER D, GOLD B Y, WENISCH T F. Powernap: eliminating server idle power[J]. ACM SIGPLAN Notices, 2009, 44(3): 205-216.
- [12] 钱琼芬, 李春林, 张小庆, 等. 云数据中心虚拟资源管理研究综述[J]. 计算机应用研究, 2012, 29(7): 2411-2415, 2421.
- [13] 敬思远, 余望, 钟毅. 用于多核嵌入式环境的硬实时任务感功调度算法[J]. 计算机应用, 2011, 31(11): 2936-2939.
- [14] BELOGLAZOV A, ABAWAJY J, BUYYA R. Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing[J]. Future Generation Computer Systems, 2012, 28(5): 755-768.
- [15] 敬思远, 余望. 基于混合粒子群算法的数据中心能耗优化[J]. 计算机工程, 2012, 38(15): 276-278.