云计算资源调度研究综述

林伟伟 齐德昱
(华南理工大学计算机科学与工程学院 广州 510640)

摘 要 资源调度是云计算的一个主要研究方向。首先对云计算资源调度的相关研究现状进行深入调查和分析;然后重点讨论以降低云计算数据中心能耗为目标的资源调度方法、以提高系统资源利用率为目标的资源管理方法、基于经济学的云资源管理模型,给出最小能耗的云计算资源调度模型和最小服务器数量的云计算资源调度模型,并深入分析和比较现有的云资源调度方法;最后指出云计算资源管理的未来重要研究方向:基于预测的资源调度、能耗与性能折衷的调度、面向不同应用负载的资源管理策略与机制、面向计算能力(CPU、内存)和网络带宽的综合资源分配、多目标优化的资源调度,以便为云计算研究提供有益的参考。

关键词 云计算,资源调度,能耗
中图法分类号 TP393 **文献标识码** A

Survey of Resource Scheduling in Cloud Computing

LIN Wei-wei QI De-yu
(School of Computer Engineering and Science, South China University of Technology, Guangzhou 510640, China)

Abstract Resource scheduling is a fundamental issue in cloud computing. The resource scheduling methods of reducing energy consumption and improving resource utilization in cloud computing data center, and economics-based cloud resource management models were discussed. Then cloud computing resource scheduling model of minimizing energy consumption and minimizing number of servers was proposed. Finally, important directions for future research in resource scheduling of cloud computing, which include prediction-based resource scheduling, power and performance trade-off scheduling, resource management policies and mechanisms for different application workload types, comprehensive multi-resource allocation with computing power (CPU, memory) and network bandwidth, multi-objective optimization of resource scheduling, were presented.

Keywords Cloud computing, Resource scheduling, Energy consumption

1 引言

近年来,随着互联网网络规模的不断扩大,互联网所需要处理的业务量也随之快速增长。如何处理海量的数据与服务,以有效地为用户提供方便、快捷的网络服务,已成为互联网当前发展面临的一个问题。在这种背景下,基于分布式计算特别是网格技术的发展,产生了一种新型服务计算模型:云计算[1,2]。中国工程院李德毅院士认为,云计算就是将整个互联网的资源汇聚整合起来,研究云计算模型可以有效地解决互联网云进化、云控制、云推理和软计算等复杂问题。李院士指出云计算将会给信息产业带来巨大的影响,将使信息技术整体结构发生改变,今后更多的软件会逐步转移到云计算环境中,更多的用户也将受益于云计算服务。随着云计算的研究不断深入及其应用不断发展,它必将成为未来主流的应用模式[3]。

云计算的概念从提出到现在已经好几年了,但到目前为止没有统一的定义。美国国家标准技术研究院(NIST)给出了目前最权威的云计算定义:(1)云计算是一种能够通过网络以便利的、按需的方式访问一个可配置的计算资源共享池(包括网络、服务器、存储、应用和服务等)的模式,这个资源共享池能以最少的管理开销及最少的与供应商的交互,迅速配置、提供或释放资源;(2)云计算模式具有 5 个基本特征:按需自助服务、广泛的网络访问、共享的资源池、快速弹性能力、可度量的服务,还包括 3 种服务模式:软件即服务(SaaS)、平台即服务(PaaS)、基础设施即服务(IaaS),以及 4 种部署方式:私有云、社区云、公有云、混合云。

从云计算的这个权威定义可以看出,云计算的核心问题是资源管理。为此,本文对云计算资源管理的相关研究现状进行了深入调查和分析,重点给出了以降低云计算数据中心能耗为目标的云资源管理、以提高系统资源利用率为目标的云资源管理和基于经济学的云资源管理的相关理论和方法,并指明下一步研究的几个重要方向:面向不同类型应用负载

到稿日期:2011-12-28 返修日期:2012-03-10 本文受国家自然科学基金项目(61070015),广东省自然科学基金项目(10451064101005155,S2011010001754),广东省战略性新兴产业核心技术攻关项目(2011A010801002),广州市海珠区科技计划项目(x2jsB2120750)资助。
林伟伟(1980-),男,博士,副教授,主要研究方向为分布式系统,E-mail:linww@scut.edu.cn;齐德昱(1959-),男,博士,教授,主要研究方向为分布式系统、计算机体系结构、网络安全。

的云资源管理策略与机制、能耗与性能折衷的云资源调度、基于预测的云资源调度、面向计算能力(CPU、内存)和网络带宽的综合云资源分配、多目标优化的云资源管理与调度。

2 当前研究热点

云计算具有十分广阔的应用前景,然而,云计算应用的快速发展依赖于云计算关键技术的研究。其中,最核心技术是资源管理,包括异构资源统一管理、资源合理调度与分配等。近几年学者们在云计算资源管理方面进行了较多研究,其中当前主要的几个研究方向有:以降低云计算数据中心能耗为目标的资源分配和调度研究、以提高系统资源利用率为目标的资源管理与调度研究、基于经济学的云资源管理模型研究、其他相关研究。

2.1 以降低云计算数据中心能耗为目标的资源分配和调度研究

目前降低云计算能耗的方法有两类:(1)通过动态调整服务器 CPU 的电压或频率来节省电能;(2)关闭不需要的服务器资源实现节能。

2.1.1 第一类方法:通过动态调整服务器 CPU 的电压或频率来节省电能

文献[4]针对虚拟集群环境的能耗问题进行了研究,给出了一个通过降低处理器速度来降低能耗的调度机制及能耗公式 $E=E_{dynamic}+E_{static}$ 及 $P_{dynamic}=ACv^2s$[5,6],其中 A 为能耗系数,C 为总电容负载,v 为处理器的电压,s 为处理器的频率。由上述的能耗公式可知,降低虚拟机的处理器电压,可以降低系统能耗。为此,该文献给出了根据虚拟机的负载来动态调节处理器电压的调度机制,即提出的调度机制的主要思想是:监视虚拟机的状态,当虚拟机负载减少时调低处理器的速度来降低能耗。但提出的方法没有建立能耗减少和应用性能的关系,降低能耗后可能会对应用性能造成不良影响。

文献[7]研究了云计算环境下优先顺序受限的(precedence-constrained)并行应用的调度问题,分别对优先顺序受限并行应用的完成时间和能耗进行了建模,其中能耗模型为:$P_{dynamic}=ACv^2s$;然后提出了一种平行双目标混合方法,即把该调度问题建模为最小完成时间和能耗双多目标优化问题,并利用 Pareto 遗传算法求解多目标优化问题。提出的并行双目标混合遗传算法兼顾到了能耗的最少和应用完成时间的最小。实验结果证明,提出的算法相对文献[8]的方法在能耗和完成时间上都有明显改进。

2.1.2 第二类方法:关闭不需要的服务器资源实现节能

云计算数据中心的能耗问题也是当前资源管理的一个主要研究方向。文献[9,10]研究云计算数据中心的能耗问题,提出了一个绿色云计算体系结构(见图 1),并重点给出了能耗感知的虚拟机优化放置和选择算法,其通过能耗感知的分配方法来降低数据中心的能耗。下面重点讨论绿色云计算体系结构,它由消费者/用户代理、节能服务分配器、虚拟机、物理主机 4 层组成。(1)消费者/代理人:向云计算中心提交服务请求的用户代理。(2)节能服务分配器:该分配器是消费者和云基础设施之间的接口。节能服务分配器是整个体系结构的核心,它与节能协调器、服务分析器、消费者分析器、定价器、能耗监视器、服务调度器、虚拟机管理器、计费等组件进行交互,实现节能调度的核心功能。(3)虚拟机:云计算分配给应用的虚拟资源。一个物理机器上可以运行多个虚拟机,这些虚拟机有不同的资源配置,运行着不同的操作系统,并能同时工作。而且,这些虚拟机可以被动态迁移,以使整个云的工作负载得到整合,让用不到的资源进入低功耗状态、关闭或将某些物理结点设置成低性能状态以节约能源。(4)物理服务器:是云计算的硬件基础设施,可以根据要求提供虚拟机。在该绿色云计算体系结构基础上给出的能耗模型为:

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

式中,$P_{\max}$ 表示服务器满负载时的最大功耗值,k 表示空闲服务器的能耗所占(满负载)的比例,u 表示 CPU 的利用率。作者在提出该能耗模型的基础上,给出了能耗感知的云计算数据中心资源分配方法。云计算数据中心在接收到虚拟机资源请求时,资源分配方法根据能耗模型计算数据中心的物理服务器的总能耗,获得能耗最小的虚拟机分配方式。

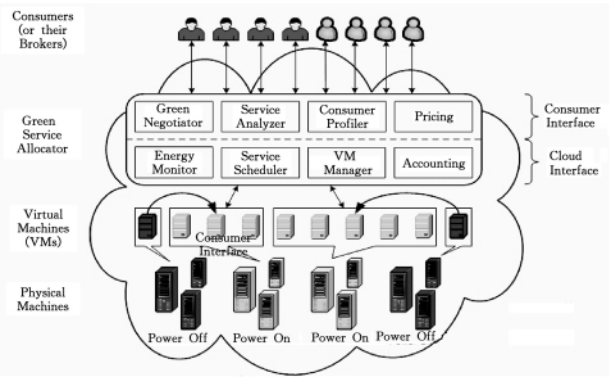

图 1 绿色云计算体系结构

为了实现数据中心的节能,文献[11]提出通过工作负载整合来关闭不必要的服务器,从而减少活动服务器的数量和数据中心的能耗;在工作负载整合过程中,由于调度信息的动态变化,可能出现错误的负载迁移,从而影响系统性能。为了处理这种不确定信息和提高系统性能,作者提出机器学习方法来预测负载迁移后应用和机器的关系模型,从而实现更加智能的负载联合和资源调度。然而,该文献并没有考虑不同类型负载的能耗与性能关系的不同;而且他们假设数据中心的服务器是同构的,而实际上大部分数据中心的服务器都是异构的。

文献[12]通过给虚拟机分配最少数量的服务器来减少服务器的数量和降低数据中心的能耗。与其他方法不同的是,作者利用神经网络方法预测服务器的工作负载,减少关闭和启动服务器的频率,从而减少因此产生的性能下降。作者提出了 Green Scheduling 算法,该算法根据服务器的历史工作负载预测将来的工作负载,从而根据预测结果来决定关闭或打开一个服务器。但是该文献只考虑 HTTP 请求一种负载类型,而在实际的云计算数据中心往往同时存在多种需要调度的负载;而且提出的预测方法只预测工作负载的到达率,不预测工作负载的资源需求和执行时间等。

2.2 以提高系统资源利用率为目标的资源分配与调度研究

在云计算资源调度方面,当前主要的方法是为虚拟资源动态优化分配物理资源,以减少云计算所需的物理资源和提

高系统资源率。Fabien Hermenier 等人针对如何分配和迁移虚拟机到物理主机的问题进行了研究[15]，并在考虑重配置计算时间和虚拟机迁移时间两个因素情况下，给出一种优化总的动态调度时间的资源管理方法 Entropy。为了实现虚拟机的优化分配和放置，Entropy 将 CPU、内存资源约束情况下的虚拟机放置问题建模为约束满足问题模型。在 Entropy 资源管理方法中，假设云计算环境中的物理资源向量为 $P=(p_1,\cdots,p_k)$，需要放置的虚拟资源（虚拟机）向量为 $V=(v_1,\cdots,v_m)$，物理资源 p_i 上的虚拟资源放置位向量为 $H_i=(h_{i1},\cdots,h_{ik})$，$h_{ij}=1$ 表示虚拟资源 v_j 放置在物理资源 p_i 上。R_P 表示每个虚拟资源的 CPU 需求，C_P 表示每个物理资源的 CPU 大小；R_m 表示每个虚拟资源的内存需求，C_m 表示每个物理资源的内存大小。由物理资源与虚拟资源的约束关系可得：

$$R_P \cdot H_i < C_p(p_i), \forall\, p_i \in P$$

$$R_m \cdot H_i < C_m(p_i), \forall\, p_i \in P$$

为了获得最优的虚拟机资源放置（放置虚拟机资源所需要的最少物理资源），给出了如下目标函数：

$$Minimize \sum_{i\in P} u_i, u_i = \begin{cases} 1, & \exists\, v_j \in V \mid h_{ij}=1 \\ 0, & \text{其它} \end{cases}$$

在建立的虚拟机资源放置的约束满足模型基础上，通过求解约束满足模型获得优化的虚拟机资源放置方案，然后基于优化的虚拟机资源放置方案开发和实现从物理资源到虚拟资源的分配算法，进而实现云计算资源的优化分配。

法国的 Jean-Marc Menaud 和 Hien Nguyen Van 等人针对云计算中虚拟资源的管理提出一些动态调度方法[13,14]，其主要是讨论如何为应用选择合适的虚拟机和为虚拟机选择合适的物理计算机的问题，并把这些调度问题转化为约束满足问题，以获得优化调度结果。为了提高云计算系统的资源利用率，文献[18]为云计算环境提出了两级调度器：元调度器和虚拟机调度器。元调度器负责为用户任务选择合适的资源，系统级虚拟机调度器负责动态删除和创建虚拟机，优化各虚拟机的负载，并通过扩展 Cloudsim 类库实现了提出的启发式任务调度算法。通过实验验证了提出的算法的有效性。

此外，也有些方法通过虚拟机的动态迁移和重新分配方法来实现云计算系统的负载均衡，从而达到云计算资源优化分配。魏贵义教授等人利用博弈论的方法来解决云计算资源分配问题[16]。他们设计了一个基于博弈论的资源分配算法，该算法首先利用整数规划方法处理单个参与者的独立优化问题，然后利用进化算法处理多个参与者的综合优化问题。提出的进化算法同时考虑了优化和公平两个方面，能给出一种较好的折衷资源分配方法；这种基于博弈论的资源分配方法重点是针对一些可分任务调度问题的优化，主要适用于处理一些非常复杂和动态的、应用能分成多个协作任务的资源调度问题。华夏渝等人提出一种基于蚁群优化的计算资源分配算法[17]，该算法在分配计算资源时，首先预测潜在可用节点的计算质量，然后根据云计算环境的特点，通过分析诸如带宽占用、线路质量和响应时间等因素对分配的影响，利用蚁群优化算法得到一组最优的计算资源。通过在 Gridsim 环境下的仿真分析和比较表明，这种算法能够在满足云计算环境要求的前提下，获得比其他一些针对网格的分配算法更短的响应时间和更好的运行质量，因而更加适合于云环境。提出的蚁群资源分配算法能够针对云环境的大规模性、共享性和动态性等特点，动态地对用户的作业进行分片搜寻并分配计算资源，该算法能够有效地在云计算环境中完成计算资源搜索与分配的工作。文献[44]为负载动态变化的云应用提出了一种动态资源分配模式：它以单个虚拟资源作为资源动态重配置（再分配）的基本单位，根据云应用的负载变化，为云应用动态配置虚拟资源；并为了避免资源动态调度造成的颠簸（振荡）现象，提出基于门限的调度决策方法。

2.3 基于经济学的云资源管理模型研究

澳大利亚 Rajkumar Buyya 等学者提出的基于经济模型资源调度方法[1,19]是当前的主要调度方法之一。他们提出了面向市场的云计算体系结构和面向市场的资源分配和调度方法，该体系结构通过 SLA 资源分配器来实现资源使用者与资源提供者之间的协商，实现资源优化分配，但该体系结构中很多具体实现问题仍然在研究之中。在此基础上，文献[20]给出了一种基于市场机制的云计算资源分配策略，并设计一个基于遗传基因的价格调节算法来处理市场的需求和供给的平衡问题，但提出的方法只是针对底层资源调度问题，即如何给虚拟资源（虚拟机）分配物理资源（CPU、内存、存储器），而且提出的方法目前仅仅考虑 CPU 资源，无法处理其它类型的物理资源。虽然，使用经济学模型进行资源调度和协同分配可以实现资源的高效调度和提高资源利用率，但目前只研究了底层资源的调度问题，且没有成熟的实现。文献[34]从经济学原理的角度提出了云计算经济学架构，设计了基于 SLA 的云资源管理经济模型。该策略为云消费者和供应商提供了有关经济激励的反馈，提高了资源利用率，避免了信息系统的重复建设，有助于实现云环境下资源的高效管理、优化配置，可最大限度地满足用户服务质量需求。为了处理云计算的应用负载的动态变化，文献[37]以经济模型为基础，提出了一个云资源提供者的联邦体系结构，该体系结构由资源构造层、资源管理层、资源调度层组成。提出的联邦云的核心思想是联邦资源调度，即云资源提供者的资源不够时，通过外包虚拟机提供给其他资源提供者，从而获得利润；当云资源提供者的资源过剩时，提供空闲资源给其他资源提供者，从而提供资源利用率；基于经济模型的联邦云通过联邦资源调度来提高资源利用率，节省云计算运营成本，增加云计算盈利能力。

2.4 其他相关研究

为了提高大规模虚拟化集群环境的资源利用率，文献[35]提出了一种面向数据中心的集群资源按需动态配置方法，该方法基于布尔二次指数平滑法预测用户请求，能够根据不断变化的需求高效实时地调整集群中节点运行的数量，实现资源快速动态配置，从而提高集群利用率，降低能耗。

文献[36]的专利提出了一种基于动态重配置虚拟资源的云计算资源调度方法，该方法以云应用监视器收集的云应用负载信息为依据，基于运行云应用的虚拟资源的负载能力和云应用当前的负载进行动态决策，然后根据决策的结果为云应用动态重配置虚拟资源。通过为云应用重配置虚拟资源的方法实现资源的动态调整，不需要动态重新分配物理资源和

停止云应用的执行。该方法能根据云应用负载变化动态重配置虚拟资源，优化云计算资源分配，实现云计算资源的高效使用和满足云应用动态可伸缩性的需要；而且该方法可以避免云计算资源的浪费，节省云应用客户的资源使用成本。

2.5 云资源调度方法比较

表1给出对现有主要的云资源调度方法的比较。

表1 各种云资源调度方法的比较

文献	资源调度目标	资源调度方法	资源调度的实现关键技术
4	降低能耗	针对计算集群中的虚拟机调度进行建模，根据虚拟机的负载来动态调节处理器电压	基于OpenNebula[38]搭建了Xen虚拟机调度环境，实现了提出的虚拟机调度算法，使用Xen虚拟机的xenpm命令[39]调节CPU电压
7	降低能耗	把云计算环境下优先顺序受限并行应用的调度问题建模为最小完成时间和能耗双多目标优化模型，动态调节处理器电压	采用Pareto遗传算法求解多目标优化模型，并基于软件包ParadisEO实现求解算法
9,10	降低能耗	通过动态分配云计算数据中心的虚拟机，减少服务器的个数	使用CloudSim模拟云计算数据中心，实现提出的虚拟机动态调度和迁移算法
11	降低能耗	利用机器学习方法预测负载迁移后应用和所需要服务器的关系模型，以达到虚拟机负载联合的智能调度	基于OMNeT++[40]实现一个模拟器，对各种应用负载与能耗的关系进行了模拟；并对实际的网格平台Grid5000的工作负载[41]进行调度模拟实验
12	降低能耗	利用神经网络方法预测服务器的工作负载，关闭不必要的服务	使用CloudSim模拟实现了提出算法，并对NASA和ClarkNet两个Web服务器工作负载数据进行模拟测试
13,14	提高资源利用率	将应用调度虚拟机和为虚拟机分配物理计算机资源的问题建模为约束满足问题模型，优化虚拟机资源分配，减少所需的物理服务器数量	在由4个服务器（每个服务器的物理资源为4000MHz，4000MB）组成的集群中模拟了提出的虚拟机资源调度框架和算法，并采用Choco求解约束满足模型
15	提高资源利用率	将CPU、内存资源约束情况下的虚拟机放置问题建模为约束满足问题模型，优化虚拟机资源动态分配，减少所需的物理服务器数量	在Grid5000网格计算平台上测试了提出模型和虚拟机迁移算法，采用Choco求解约束满足模型
16	提高资源利用率	提出了基于博弈论的云计算资源分配算法	先利用二进制整数规划方法对模型进行初始优化，然后给出进化优化算法进行求解
17	提高资源利用率	提出了基于蚁群优化的计算资源分配算法	用Gridsim模拟云计算环境，实现了提出的算法，并与退火算法和遗传算法进行了比较
44	提高资源利用率	给出基于门限的云计算资源动态调度方法	用CloudSim实现了提出的算法，模拟了正态分布的和振荡的工作负载情况下的资源调度情况
1,19	面向经济模型的调度	提出了面向市场的云计算体系结构和面向市场的资源分配和调度方法，该体系结构通过SLA资源分配器来实现资源使用者与资源提供者之间的协商，实现资源优化分配	用基于.NET实现一个面向市场的服务资源管理云平台Aneka[43]，实现了7个价格机制，并对高性能计算应用的工作负载和Internet服务的工作负载进行资源分配调度实验
20	面向经济模型的调度	提出了一种基于市场机制的云计算资源分配策略，设计一个基于遗传基因的价格调节算法处理市场的需求和供给的平衡问题	基于Xen虚拟机模拟实现了提出的资源分配算法
37	面向经济模型的调度	提出了一个云资源提供者的联邦体系结构，通过联邦资源调度来提高资源利用率	在亚马逊弹性计算云EC2环境中评估了提出的联邦云的有效性，并使用实际云工作模拟测试提出的联邦调度策略

3 未来研究方向分析与展望

虽然，近几年在云计算资源管理与调度上学者们已经开展了许多研究工作，给出了不少方法和取得了一些研究成果，但仍然存在不少问题，需要进一步深入研究和给出解决方法。在云计算资源管理与调度方面，未来主要研究方向与机会有：基于预测的资源调度、能耗与性能折衷的调度、面向不同应用负载的资源管理策略与机制、面向计算能力（CPU、内存）和网络带宽的综合资源分配、多目标优化的资源管理与调度。

3.1 面向不同类型应用负载的云资源管理策略与机制

当前的方法主要是通过减少应用服务器数量来降低能耗和提高资源利用率，其核心技术是使用虚拟化技术来联合多个应用负载并部署到单个物理服务器上，从而减少物理主机数量。然而，已有的方法都没有考虑到不同应用负载类型对资源的不同需求（如科学计算应用需要服务器有强大的计算能力、多媒体应用需要服务器有强大的数据处理能力）[21]，也未考虑到工作负载类型的应用联合可能会影响应用的性能和增加服务器的能耗。

未来需要研究不同类型负载的应用共享物理主机的资源方法和机制，利用多种应用资源需求的互补特性，联合不同类型的应用负载并部署到单个物理主机上，以更好地共享物理主机资源。

3.2 能耗与性能折衷的云资源调度

当前云计算数据中心的资源分配方法都考虑同构物理主机，假设数据中心的物理主机有同样物理资源和同样的能耗。然而，现实的数据中心的服务器往往是异构的，比如文献[22,23]描述的数据中心。而且，提出的方法都是尽力去联合最大数量的应用负载并将其部署在单个物理服务器上，这种方法能减少物理服务器的数量，但是服务器的过载可能导致应用的性能下降，增加服务器产生的热量，导致数据中心制冷的能耗[22]。

为此，需要研究实现能耗和性能折衷的资源分配机制，以避免因过度分配资源而造成应用性能的下降；研究分析不同硬件特性适合不同应用负载的特性，利用物理主机的异构性来实现更好的资源分配。

3.3 基于预测的云资源调度

为了提高云计算资源利用率，目前提出的很多方法[13-17,24]都采用动态资源分配来实现负载均衡或弹性调度。

这些方法需要根据应用的负载变化来动态配置资源，而且为了减少物理服务器的数量，往往需要进行虚拟机迁移。然而，基于动态监测的工作负载进行调度和迁移时，可能出现频繁迁移的波动情况，反而影响了系统调度性能。因此，未来需要研究各种类型应用工作负载变化的预测方法和云应用资源需求的预测机制，通过预测资源的需求进行更准确的资源分配，避免不必要的虚拟机迁移开销和满足应用的动态资源需求，实现更有效的资源配置和提高资源利用率。

3.4 面向计算能力和网络带宽的综合云资源分配

在云计算环境网络资源管理方面，当前一些商业的虚拟机资源管理工具（如 VMware Capacity Planner[25] 和 IBM WebSphere CloudBurst[26]）仅仅考虑虚拟机的 CPU、内存等主机资源分配，没有考虑网络资源的分配问题；虽然学者们提出了一些模型和方法优化虚拟机的网络资源分配和放置[27]，但是，这些模型和方法只是近似最优，而且没有同时考虑物理资源分配和网络资源分配[28]。总之，关于虚拟机的主机资源（CPU、内存等）和网络带宽资源的综合优化分配问题仍然是一个开放的问题。

3.5 多目标优化的云资源调度

早期的资源调度研究[29-31]主要集中在保证应用性能的前提下改进资源利用率。而当前的研究工作主要是从数据中心的能耗方面给出改进方法，通过使用虚拟化技术来联合各工作负载，并将其部署到单个物理服务器上，关闭不必要的服务器来降低系统能耗。此外，文献[32,33]也提出了温度感知的应用负载迁移方法，即通过服务器的温度控制进行资源分配。因此，对于云计算中心，一方面为了最大化资源效益需要降低能耗，另一方面需要保证云应用的服务质量和提高应用性能。然而，多个目标可能存在相互冲突，例如最大化资源利用率往往会带来服务质量的下降。因此，为了获得多方面的综合优化，需要研究数据中心资源分配的多目标优化模型与算法。

此外，对于云应用客户来说，希望在应用调度时消耗最小的资源，同时又希望能尽快完成应用。而这两个目标往往是矛盾的，为此，需要研究实现最小完成时间和最小花费的多目标的资源调度算法。

结束语 许多云计算系统已经从工业和学术届走向实用，而且，最近两年在云计算方面有了很大进展。然而，仍然存在不少开放的问题，如安全性、可用性、可扩展性、互操作性、服务水平协议、数据迁移和管理、资源管理、信任问题、以用户为中心的隐私、透明性、政治和法律问题、业务服务管理等。其中云计算资源管理是当前云计算的一个主要研究方向。本文重点研究与分析了以降低云计算数据中心能耗为目标的资源管理方法、以提高系统资源利用率为目标的资源管理方法、基于经济学的云资源管理模型等，深入分析和比较了现有的云资源调度方法，并指出云计算资源管理的未来几个重要研究方向。

参 考 文 献

[1] Buyya R, Yeo C S, Venugopal S, et al. Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility[J]. Future Generation Computer Systems, 2009, 25(6): 599-616

[2] Armbrust M, Fox A, Griffith R, et al. Above the Clouds: A Berkeley View of Cloud Computing [EB/OL]. http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html, February 2009

[3] Lin Wei-wei, Qi De-yu. Research on Resource Self-Organizing Model for Cloud Computing[C]// 2010 International Conference on Internet Technology and Applications. 2010: 1-5

[4] Von L G, Wang L, Younge A J, et al. Power-Aware Scheduling of Virtual Machines in DVFS-enabled Clusters[C]// Proc. of IEEE International Conference on Cluster Computing 2009. New Orleans, LA, USA, 2009: 1-10

[5] Ge R, Feng X, Cameron K. Performance-constrained distributed dvs scheduling for scientific applications on power-aware clusters[C]// Proceedings of the 2005 ACM/IEEE Conference on Supercomputing. IEEE Computer Society, Washington DC, USA, 2005: 34

[6] Venkatachalam V, Franz M. Power reduction techniques for microprocessor systems[J]. ACM Computing Surveys (CSUR), 2005, 37(3): 195-237

[7] Mezmaz M, Melab N, Kessaci Y, et al. A parallel bi-objective hybrid metaheuristic for energy-aware scheduling for cloud computing systems[J]. Journal of Parallel and Distributed Computing (JPDC), 2011, 71(11): 1497-1508

[8] Lee Y C, Zomaya A Y. A novel state transitionmethod formeta-heuristic-based scheduling in heterogeneous computing systems [J]. IEEE Transactions on Parallel and Distributed Systems, 2008, 19(9): 1215-1223

[9] 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

[10] Buyya R, Beloglazov A, Abawajy J. Energy-Efficient Management of Data Center Resources for Cloud Computing: A Vision, Architectural Elements, and Open Challenges[C]// Proceedings of the 2010 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA2010). Las Vegas, USA, July 2010

[11] Berral J L, Goiri I, Nou R, et al. Towards energy-aware scheduling in data centers using machine learning[C]// Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking. Passau, German, 2010: 215-224

[12] Duy T V T, Sato Y, Inoguchi Y. Performance evaluation of a Green Scheduling Algorithm for energy savings in Cloud computing[C]// Proceedings of the IEEE International Symposium on Parallel & Distributed Processing Workshops. 2010: 1-8

[13] Van H N, Tran F D, Menaud J-M. SLA-Aware Virtual Resource Management for Cloud Infrastructures[C]// Ninth IEEE International Conference on Computer and Information Technology. vol. 1, 2009: 357-362

[14] Van H N, Tran F D, Menaud J-M. Autonomic virtual resource

management for service hosting platforms[C]// ICSE Workshop on Software Engineering Challenges of Cloud Computing. 2009: 1-8

[15] Hermenier F, Lorca X, Menaud J-M, et al. Entropy: a Consolidation Manager for Cluster[C]// Proc. of the 2009 International Conferenceon Virtual Execution Environments (VEE'09). Mar. 2009: 41-50

[16] Wei Gui-yi, Vasilakos A, Zheng Yao, et al. A game-theoretic method of fair resource allocation for cloud computing services [J]. The Journal of Supercomputing, 2010, 54(2): 252-269

[17] **华夏渝, 郑骏, 胡文心. 基于云计算环境的蚁群优化计算资源分配算法**[J]. **华东师范大学学报: 自然科学版**, 2010(1): 127-134

[18] Sadhasivam S, Nagaveni N, Jayarani R, et al. Design and Implementation of an efficient Two-level Scheduler for Cloud Computing Environment[C]// 2009 International Conference on Advances in Recent Technologies in Communication and Computing. 2009: 884-886

[19] Buyya R, Yeo C S, Venugopal S. Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities[C]// Keynote Paper, Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications (HPCC 2008. IEEE CS Press, Los Alamitos, CA, USA). Dalian, China, Sept. 2008

[20] You Xin-dong, Xu Xiang-hua, Wan Jian, et al. RAS-M: Resource Allocation Strategy Based on Market Mechanism in Cloud Computing[C]// Fourth ChinaGrid Annual Conference. 2009: 256-263

[21] Srikantaiah S, Kansal A, Zhao Feng. Energy aware consolidation for cloud computing[C]// Proceedings of the 2008 Conference on Power Aware Computing and Systems. 2008: 10

[22] Liu Liang, Wang Hao, Liu Xue, et al. GreenCloud: a new architecture for green data center[C]// Proceedings of the 6th International Conference Industry Session on Autonomic Computing and Communications Industry Session. 2009: 29-38

[23] Ferreira A M. An energy-aware approach for service performance evaluation[C]// the International Conference on Energy-Efficient Computing and Networking. 2010

[24] Lee Y C, Zomaya A Y. Energy efficient utilization of resources in cloud computing systems[J]. The Journal of Supercomputing, 2010(53): 1-13

[25] VMware Capacity Planner [EB/OL]. http://www.vmware.com/products/capacity-planner/, 2011

[26] IBM WebSphere CloudBurst [EB/OL]. http://www-01.ibm.com/software/webservers/cloudburst/, 2011

[27] **胡冷非. 虚拟机** Xen **网络带宽分配的研究和改进**[D]. **上海: 上海交通大学**, 2009

[28] Wang Cong, Wang Cui-rong, Yuan Ying. Dynamic Bandwidth Allocation for Preventing Congestion in Datacenter Networks [J]. Lecture Notes in Computer Science, Advances in Neural Networks, 2011, 6677: 160-167

[29] Urgaonkar B, Chandra A. Dynamic provisioning of multi-tier internet applications[C]// Proceedings of the Second International Conference on Automatic Computing. IEEE Computer Society, Washington DC, USA, 2005: 217-228

[30] Padala P, Shin K G, Zhu X, et al. Adaptive control of virtualized resources in utility computing environments[C]// EuroSys '07: Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007. ACM, New York, NY, USA, 2007: 289-302

[31] Gmach D, Rolia J, Cherkasova L. Satisfying service level objectices in a self-managing resource pool[C]// Proceedings of the 2009 Third IEEE International Conference on Self-Adaptive and Self-Organizing Systems, SASO '09. IEEE Computer Society, Washington DC, USA, 2009: 243-253

[32] Tang Q, Gupta S K S, Varsamopoulos G. Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach[J]. IEEE Trans. Parallel Distrib. Syst., 2008, 19(11): 1458-1472

[33] Moore J, Chase J, Ranganathan P, et al. Making scheduling "cool": temperature-aware workload placement in data centers [C]// ATEC '05: Proceedings of the annual conference on USENIX Annual Technical Conference. USENIX Association, Berkeley, CA, USA, 2005: 5

[34] **高宏卿, 邢颖. 基于经济学的** U **资源管理模型研究**[J]. **计算机工程与设计**, 2010, 31(19): 4139-4212

[35] **米海波, 王怀民, 尹刚, 等. 一种面向虚拟化数字中心资源按需重配置方法**[J]. **软件学报**, 2011, 22(9): 2193-2205

[36] **林伟伟, 齐德昱. 一种基于动态重配置虚拟资源的云计算资源调度方法**[P]. **中国, 申请号**: 201010268105.7. 2011.01.05

[37] 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

[38] Fontan J, Vazquez T, Gonzalez L, et al. OpenNEbula: The Open Source Virtual Machine Manager for Cluster Computing[J]. Open Source Grid and Cluster Software Conference. San Francisco, CA, USA, May 2008

[39] Wei G, Liu J, Xu J, et al. The On-going Evolutions of Power Management in Xen [EB/OL]. www.xen.org/files/xensummit_oracle09/xensummit_intel.pdf, 2012

[40] Omnet[OL]. http://www.omnet.org, 2009

[41] The Grid Workloads Archive [EB/OL]. http://gwa.ewi.tudelft.nl, 2012

[42] Traces in the Internet Traffic Archive[EB/OL]. http://ita.ee.lbl.gov/html/traces.html, 2012

[43] Chu X, Nadiminti K, Jin C, et al. Aneka: Next-Generation Enterprise Grid Platform for e-Science and e-Business Applications [C]// Proceedings of the 3th IEEE International Conference on e-Science and Grid Computing (e-Science 2007). Bangalore, India, Dec. 2007

[44] Lin Wei-wei, Wang J Z, Liang Chen, et al. A Threshold-based Dynamic Resource Allocation Scheme for Cloud Computing[J]. Procedia Engineering, 2011, 23: 695-703

[45] **李乔, 郑啸. 云计算研究现状综述**[J]. **计算机科学**, 2011, 38(4): 32-37