

云数据中心虚拟资源管理研究综述*

钱琼芬^{1,2}, 李春林¹, 张小庆¹, 李腊元¹

(1. 武汉理工大学 计算机科学与技术学院, 武汉 430063; 2. 空军雷达学院 四系, 武汉 430019)

摘要: 云计算数据中心是数据中心架构和发展的方向,云数据中心的虚拟资源管理问题是当前的研究热点。总结了数据中心的发展历程,综述了云数据中心的虚拟化技术、虚拟资源提供与部署、资源调度模型与算法以及基于能量优化和负载均衡的虚拟机迁移技术的研究现状。最后展望了云数据中心虚拟资源管理领域有待进一步研究的方向。

关键词: 云计算; 虚拟化; 虚拟资源; 资源管理; 资源提供; 虚拟机迁移

中图分类号: TP393 **文献标志码:** A **文章编号:** 1001-3695(2012)07-2411-05

doi:10.3969/j.issn.1001-3695.2012.07.003

Survey of virtual resource management in cloud data center

QIAN Qiong-fen^{1,2}, LI Chun-lin¹, ZHANG Xiao-qing¹, LI La-yuan¹

(1. School of Computer Science & Technology, Wuhan University of Technology, Wuhan 430063, China; 2. No. 4 Department, Air Force Radar Academy, Wuhan 430019, China)

Abstract: Cloud data center is the future of data center and the virtual resource management issues is the current research focus in cloud data center. This paper summarized the development of data center. Then summarized the overviews of the cloud data center on virtualization, virtual resource provisioning, resource scheduling and virtual machine migration technology based on energy-efficient and load-balance. Finally, this paper concluded with some future prospects of the research directions in this area.

Key words: cloud computing; virtualization; virtual resource; resource management; resource provisioning; virtual machine migration

云计算是一种基于网络的计算模式,云计算提供的服务类型可以分为三类,即将基础设施作为服务(IaaS)、将平台作为服务(PaaS)以及将软件作为服务(SaaS)。Amazon、Google、Salesforce等在世界各地建立庞大的数据中心,实现海量的计算任务和海量信息的存储,为用户提供云计算服务,在商业应用上取得了成功。云计算技术的迅速发展带动了传统数据中心的变革,产生了新一代数据中心——云数据中心。

云数据中心包含大量计算机,运作成本很高。有效整合资源、提高资源利用率、节约能源、降低运行成本是云数据中心关注的热点。云数据中心通过虚拟化技术将计算资源、存储资源和网络资源构建成动态的虚拟资源池;使用虚拟资源管理技术实现云计算资源自动部署、动态扩展、按需分配;用户采用按需和即付即用的方式获取资源。因此,虚拟资源管理成为当前云计算研究的热点和难点。

1 数据中心的发展

数据中心是信息系统的核心,主要功能是通过网络向用户提供信息服务。数据中心的演变经历了四个阶段:

a) 数据存储中心阶段。数据中心最早出现在20世纪60年代,采用的是以主机为核心的计算方式,一台大型主机就是

数据中心,如IBM360系列计算机,其主要业务是数据的集中存储和管理。

b) 数据处理中心阶段。20世纪70年代以后,随着计算需求的不断增加、计算机价格的下降以及广域网和局域网的普及、应用,数据中心的规模不断增大,数据中心开始承担核心的计算任务。

c) 信息中心阶段。20世纪90年代,互联网的迅速发展使网络应用多样化,客户端/服务器的计算模式得到广泛应用。数据中心具备了核心计算和核心业务运营支撑功能。

d) 云数据中心阶段。进入21世纪,数据中心规模进一步扩大,服务器数量迅速增长。虚拟化技术的成熟应用和云计算技术的迅速发展使数据中心进入了新的发展阶段。数据中心承担着核心运营支持、信息资源服务、核心计算、数据存储和备份等功能。

由于数据中心规模的扩大和功能的多样性,随之出现的问题是数据中心的可靠性以及维护管理的巨大成本。传统数据中心主要关注应用的稳定性、数据的安全性和运行的可靠性,而对资源的利用率、节能高效等问题考虑较少。云数据中心通过虚拟化技术,一方面可以在硬件服务器上实现多个虚拟服务器,通过应用和故障隔离,提高数据中心的计算能力和可靠性;

收稿日期: 2012-02-02; **修回日期:** 2012-03-03 **基金项目:** 国家自然科学基金资助项目(60970064); 国家教育部新世纪优秀人才计划资助项目(NCET-08-0806); 武汉市科技攻关资助项目(201010621207); 霍英东教育基金资助项目(121067); 中央高校基本科研业务费专项资金资助项目(2010-YB-19); 软件开发环境国家重点实验室开放课题资助项目(SKLSDE-2011KF-01)

作者简介: 钱琼芬(1970-),女,博士研究生,主要研究方向为云计算、网格计算(qqfld@hotmail.com); 李春林(1974-),女,教授,博导,主要研究方向为云计算、网格计算。

另一方面能够整合数据中心的计算资源、网络资源和存储资源,并将它们动态地分配给虚拟机,实现数据中心资源的动态部署,提高资源利用率,减少能耗。与传统的数据中心相比,云数据中心具有资源共享、资源动态调整、绿色环保、高自动化和高可用性的特点。

2007年,全球首个虚拟化数据中心——Sun公司的黑盒子面世,该数据中心可以容纳200多台Sun服务器。IBM2008年推出了便携式模块化数据中心PMDC;惠普公司2010年推出了性能优化数据中心POD;思科公司2010年推出了统一计算系统UCS,集中统一管理计算、网络、存储等虚拟化资源。Google数据中心采用标准的集装箱设计,每个集装箱可以容纳1000多个服务器,并配备了冷却系统。芝加哥数据中心是微软最大的数据中心,占地面积70万平方英尺,集装箱里放置着微软云计算产品的重要组件,每个集装箱都存放了上千台服务器,为微软的云计算提供服务。

云计算技术是在网络技术和Web应用技术的推动下产生的网络计算模式。2006年,Amazon发布了简单存储服务S3;2007年,Salesforce.com公司推出了SaaS服务,客户可以根据需要订购软件应用服务,按服务多少和时间长短支付费用;2007年,Google推出了Google Docs在线办公服务,2008年推出Google AppEngine程序开发平台,将平台作为一种服务提供给用户。IBM于2007年推出了蓝云计算平台,为客户带来即买即用的云计算平台,它包括一系列虚拟化软件,使来自全球的用户可以访问云计算的大型服务器资源池。微软2008年推出了Windows Azure蓝天操作系统,基于互联网架构,打造新的云计算平台,将微软所拥有的数以亿计的Windows用户和桌面接到云中。目前,成功的云计算应用实例还有GoGrid、Joyent、MapReduce、Aneka以及Animoto等。

虚拟化概念及技术起源于20世纪50年代。20世纪60年代,IBM360系统首次采用了虚拟化平台Hypervisor。IBM、VMware、Xen/Citrix和Microsoft是虚拟化业界的四大厂商,他们提供了成熟、多方位、广泛的虚拟化产品,实现了从服务器、存储、网络到桌面、应用、甚至数据中心的虚拟化。

数据中心的发展与云计算技术的演进、虚拟化技术的成熟应用是密切相关、相互促进的;数据中心的发展离不开计算机技术、网络技术、云计算技术的发展,云计算技术是在网络技术的广泛发展、虚拟化技术的日益成熟以及网络应用需求的推动下产生的,虚拟化技术一直伴随着计算机技术、网络技术和数据中心的发展,实现了不同层面的虚拟化。上述相关领域技术的发展如图1所示。

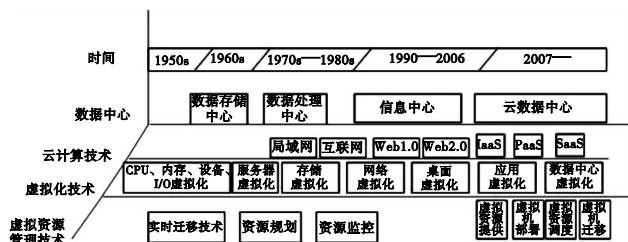


图1 数据中心及相关技术的发展

云数据中心是未来数据中心建设和发展的方向,实现云数据中心的关键技术——虚拟资源管理已成为目前研究的热点和难点。深入研究虚拟资源的管理有利于设计与实现高效的云数据中心系统,提高云数据中心资源利用效率、系统可靠性,降低运行管理成本。本文从虚拟资源提供与部署、虚拟资源调

度、虚拟机迁移等角度,针对当前相关研究缺少系统总结的情况,较全面地概括、分析了虚拟资源管理技术的研究现状。

2 虚拟资源提供与部署

云数据中心资源管理的特点是统一部署,集中管理,分布使用。在资源虚拟化技术的基础上,采用各种技术实现资源的自动部署、集中监控、动态优化、节能降耗等功能。

2.1 资源虚拟化

数据中心硬件设备之间的差异导致兼容性很差,为统一的资源管理提出了挑战。资源的抽象和虚拟化是屏蔽资源差异、建立虚拟计算环境的基础。iVCE虚拟计算环境针对资源多样性、自主性带来的访问方式不一致问题,深入研究了资源虚拟化模型与方法,提出了具备环境动态感知和自主行为决策特征的自主元素资源抽象模型^[1]。文献[2]通过虚拟化实现虚拟集群系统,消除系统软件与硬件之间的耦合,实现集群的快速部署和快速切换,并获得了明显的性能优势。云数据中心通过虚拟化技术构建虚拟资源池来实现对大规模基础资源的有效、统一管理,然后将计算任务分布在大量动态、可扩展的虚拟资源池上,用户按需获取计算能力、存储空间和信息服务。

虚拟化技术屏蔽了低层物理设备的差异,对云数据中心的低层架构进行抽象,从而实现资源的统一管理。IBM的TPM(tivoli provisioning manager)是数据中心实现资源虚拟化管理的核心产品,可以自动完成服务器、存储器、网络设备、操作系统、中间件、应用程序的部署和配置。通过实现从低层硬件、系统的虚拟化到存储、网络的虚拟化,使数据中心的IT资源成为一个虚拟的资源池,可以按照一定的粒度来实现资源分配。VMware的数据中心产品VMware Infrastructure虚拟化软件套件能提供虚拟化基础架构、应用程序和管理等多种服务。其中,ESX Server能整合数据中心的计算、存储和网络资源,构建动态虚拟资源池。微软公司的虚拟化集成管理软件是System Center。其中,Virtual Machine Manager能管理数据中心的基础架构,包括服务器以及虚拟机,实现平台中虚拟化资源的动态优化。浪潮云海·云计算操作系统架构于服务器、存储、网络等基础硬件资源之上,通过基于Xen平台的虚拟化技术,构建计算节点、存储节点,形成动态虚拟计算资源池、存储资源池和网络资源池,实现了资源的有效监控、动态流转与伸缩,以IaaS的模式为用户提供云计算服务。

2.2 资源提供策略

云数据中心的功能涉及到很多方面,如数据中心内部的网络结构、存储架构、资源模型、资源发现、资源提供、虚拟机部署、调度策略、资源状态监测评估、能耗管理、数据安全以及QoS支持等。资源管理系统的基本功能是接受来自云计算用户的资源请求,并且把特定的资源分配给资源请求者,合理地调度相应的资源,使用户请求资源的任务得以运行。云数据中心资源管理系统提供四种基本的服务,即资源发现、资源监控、资源存储和资源调度。资源发现在数据中心发现适合用户应用的虚拟资源,能自动发现和应用虚拟资源与物理资源之间的关联;资源监控能自动监测基础硬件资源的状态、性能;资源调度采用不同策略,把所需资源分配到相应的用户任务上。虚拟资源的提供和虚拟机资源的部署是资源管理的基础^[3-7]。

典型的云计算资源提供策略包括基于租借理论和动态多

级资源池的策略、基于经济学原理的资源提供策略、基于一般优化算法的资源提供策略以及基于随机整数规划的最优资源提供策略。

a) 基于租借理论和多级资源池的云计算资源提供 文献[8]基于云计算的虚拟化技术,提出了一种租借理论和动态多级资源池相结合的资源调度策略,将资源虚拟化成多个槽(slot),根据资源的某个共同的特性将资源归类,形成资源池,并建立多级资源池。其中有一个资源池充当服务器,提供与云外面的交互、维护资源池负载平衡以及分配任务等服务,同时,结合共享策略、私有策略、借还策略和重声明策略完成对资源的调度。文献[9]提出一种面向虚拟资源的云计算资源管理机制,通过资源划分和资源预留策略来实现对虚拟资源的分配,确保用户对虚拟资源使用的有效性,并提出一种借入/借出调度策略来实现虚拟资源利用率的最大化。

b) 基于经济学的云计算资源提供策略 文献[10,11]从经济学原理的角度出发,提出了云计算经济学架构,设计了基于QoS的云资源管理经济模型。依据经济学的理念,云计算环境被视为一个云市场,资源被视为商品,多个计算云、存储云抽象为资源提供者,云计算用户被视为资源的消费者。该策略为云消费者和供应商提供有关经济激励的反馈,提高资源利用率,有助于实现云计算环境下资源的高效管理、优化配置,同时最大限度地满足用户服务质量需求。但是此策略是从经济学角度出发,并没有为节省客户花费提供相关的策略,也没有考虑资源和价格的动态性。

c) 基于随机整数规划的最优资源提供策略 文献[12~14]使用随机整数规划优化资源提供方法。在云计算中,云提供商可以有两种方式给用户提供服务^[15],即预订方式和即付即用方式。预订方式可以有效减少用户的花费,但是因为用户需求 and 资源价格的不确定性,完全通过预订方式获取资源是很难达到用户的要求。使用启发式方法^[16]或K-邻近算法^[17]可以对用户所需资源进行预测。该策略提供一个动态的资源提供方案来满足客户的需求,考虑了资源提供各个阶段的资源花费,能达到用户资源花费最小化的目的。

2.3 虚拟机部署

资源快速部署是数据中心的一项重要功能需求。云计算数据中心采用虚拟化技术后,资源管理的一个重要功能就是构建虚拟的资源池,将虚拟机部署在不同的物理机上,实现对大规模基础资源有效、统一的管理。云计算环境对虚拟机部署的要求较高,目标是高效快速、节能低耗、均衡负载,充分提高计算资源、存储资源以及网络资源的利用率。虚拟机部署是一个复杂的问题,一方面,在云环境中资源和应用不仅变化范围大而且动态性高,用户所需服务主要采用按需部署方式;另一方面,不同云数据中心、不同层次云计算环境中服务的部署模式是不一样的,部署过程所支持的软件系统形式多样,系统结构各不相同,部署策略具有多样性。

a) 基于网络的虚拟机部署策略 云数据中心的各个物理机上的虚拟机要进行频繁的数据通信,因此在考虑应用进程性能时,要加入网络带宽、延迟等因素。而传统的虚拟机部署主要是考虑效率和计算资源的利用率,没有从网络方面考虑。文献[18]提出一种基于网络的虚拟机部署策略,能达到虚拟机之间数据传输时间的最小化目的,从而实现整个应用性能的优化。

b) 基于约束满足问题的虚拟机部署策略 文献[19]使用约束满足问题对虚拟机部署进行建模,约束条件为用户的服务等级协议,目的是最大化节省能量。其实现思想是最大化空闲物理机数,通过关闭空闲物理机来节省能量。文献[20,21]采用动态虚拟机部署方案,约束条件为满足用户的SLA需求。但使用约束满足问题很容易引入额外的应用约束,如搭配约束和同类约束等。

c) 基于容错机制的虚拟机部署策略 采用虚拟化技术之后,主机服务器系统^[22]的失败是一个需要考虑的问题。由于虚拟机是基于主机服务器的物理设备和虚拟化平台,所以当主机服务器失败时,所有运行在其上的虚拟机也不能幸免,因此需要制定一个针对此问题的对策。文献[23]提出使用冗余配置虚拟机的方法,该方法根据应用程序所需的性能、评估所需的最少虚拟机数量,然后决策出一个最优虚拟机部署,将任意 k 个主机服务器失败所造成的影响最小化。

3 虚拟资源调度

资源调度是根据一定的资源使用规则,在不同资源使用者之间进行资源调整的过程。不同资源使用者对应着不同的计算任务,每个计算任务在操作系统中对应于一个或多个进程。资源调度的目的是将用户任务分配到合适的资源上,使得在满足用户需求的前提下,任务完成时间尽量小,且资源利用率尽量高。资源调度最终要实现时间跨度、服务质量、负载均衡、经济原则最优的目标。由于不同厂商架构的云基础设施不同,资源的管理和调度没有统一的国际标准,基于各种调度基础设施和调度模型的调度算法很多^[24]。

3.1 资源调度模型

文献[25]将任务调度模型分为应用模型、计算平台模型和性能目标模型。应用模型涉及如何将应用划分为任务、如何考虑任务的属性特征等,典型的任务模型有依赖任务模型DAG、独立任务模型IND以及可分任务模型DLM;计算平台模型是对系统中资源的抽象,其中最重要的资源是处理机资源和网络资源;性能指标模型可以分为基于系统的目标和基于用户的目标两类,基于系统的性能模型关注整个系统的吞吐量、资源利用率、效率和公平性,基于用户的性能指标包括应用的最短完成时间、周转时间、平均延迟和带权完成时间等。

资源管理调度模型按照调度实体之间的关系可以分为统一资源代理调度模型和多资源代理调度模型^[24]。按照资源的组织调度形式可分为集中调度模型、层次调度模型和非集中式调度模型^[7]。在集中式的环境中,所有的资源由一个中央调度程序调度,所有可用系统的有关信息被聚集在中心机上。在层次式的调度模型中,有一个集中式的调度程序,作业被提交到集中的调度程序,而每一个资源使用一个独立的调度程序用于本地的调度。此结构的主要优点是采用不同的策略用于本地和全局的作业调度。在非集中式系统中,分布式的调度程序交互作用并且将作业提交到远程系统中,单个组件的失效不会影响整个云计算系统,容错和可靠性更高。但由于一个并行程序的所有部分可能被分配在不同域的资源上,不同的调度程序必须同步作业并且保证同时运行,这使得调度系统的优化相当困难。

资源管理调度模型按照体系结构不同,可分为层次模型、抽象所有者模型(AO)以及市场经济模型。层次模型将管理系

统分为若干功能层,有利于对具有站点自治性和底层异构性资源进行管理,能在一定程度上实现资源的联合分配,具有较强的适用性。抽象所有者模型使用作为资源所有者的资源经纪人与用户进行交互和协商。资源共享过程中遵循类似于快餐店的订购与交货模式。计算经济模型综合了层次模型和抽象所有者模型的核心特征,既可以利用层次模型中相对成熟的技术,又明确强调了以经济为基础的资源管理和调度。基于供求原则的投资回报机制,促进了计算服务质量的提高和资源的升级,经济是调节供求关系的最重要的机制,为访问网格资源的用户提供公平的价格机制,并允许对一切资源进行交易。建立以用户为中心而不是系统为中心的调度政策,提供了资源分配和管理的有效机制^[25]。

3.2 资源调度算法

针对不同的资源调度模型,许多学者都提出了各自不同的算法,根据算法的目标函数,较典型的有时间最优算法、代价最优算法和时间代价最优算法^[26,27]。时间最优算法的出发点是尽量快地在预算范围内完成任务,针对每个资源,考虑到以往分配的任务和完成率,估算一个任务的完成时间,再依据完成时间对资源按升序排序,再从队列中依次取出资源。如果该任务的成本小于或等于该任务的预算,则分配该任务给这个资源。代价最优算法尽量以最小的花费在完成期限内完成任务,基本思想是首先给资源按价格升序排序,对队列中的每个资源,在不超过完成期限的范围内分配尽可能多的任务。时间代价最优算法结合了以上两个算法的优点,在不增加额外处理成本的情况下,最优化处理时间。

考虑算法的不同调度策略和目标函数,按照不同标准可以对资源调度算法进行多种分类:a)传统的调度算法,如轮循调度、最小连接调度、目标地址散列调度、源地地址散列调度等,但算法简单,性能不佳;b)启发式调度算法,由于资源调度因素复杂,通常采用启发式方法,根据调度算法的运行时间,启发式映射算法可以分为静态映射和动态映射,动态调度算法分为在线模式和批模式,典型的在线模式启发是算法有 MCT、MET、SA、KPB、OLB 等,典型的批模式启发算法有 min-min 算法、max-min 算法、快速贪心算法、贪吃算法、忍耐算法和老化算法等;c)基于经济学模型的调度算法,以经济学中的商品市场模型、标价模型、议价模型、招标/合同网模型、拍卖模型等为基础,分别采用代价最优、时间最优或时间代价等目标函数,实现最优资源调度;d)基于 agent 的调度算法的资源调度将每个资源节点封装成一个 agent,资源管理系统成为一种多层次 Agent 系统的集合,调度问题被简化成如何在各 Agent 之间匹配计算任务并随时根据 Agent 的变化情况进行调整,以及在 agent 内如何进行子任务的继续分配的问题;e)基于任务的性质及任务之间的相关性,调度算法可以分为独立任务调度算法、可分任务调度算法、依赖任务调度算法以及多维 QoS 要求及负载均衡的任务调度算法^[25];f)基于博弈论的资源调度算法,博弈论是经济学中的重要理论方法,由于资源分配与社会经济活动的相似性,博弈论也广泛地应用到资源分配研究中;g)其他的调度算法,除以上介绍的资源调度算法之外,还有一些改进和综合的调度算法如基于信任模型的可信资源调度、任务依赖调度算法、多维 QoS 要求调度算法、负载均衡的调度算法、基于能耗的资源调度算法等^[24,25]。

云资源调度算法可以借鉴网格计算和分布式计算中的研

究成果,并关注云计算资源调度的特点。云数据中心资源调度的特点是:资源虚拟化和面向用户的调度性能优化。虚拟机的出现使得所有的计算任务都被封装在一个虚拟机内部。由于虚拟机具有隔离性,可以采用虚拟机的动态迁移技术来完成计算任务的迁移,实现资源优化。在传统的分布式计算环境中,资源均为无偿免费的,系统整体性能最优往往是调度的最优化目标。云计算环境中,云服务提供商提供资源和服务,用户按需付费,只需要为使用的资源或服务付费。因此,云环境中的调度问题必须考虑任务执行的成本约束。此外,与成本相关的如任务完成时间、赔偿率以及用户付费等因素,也是云计算调度问题中需要考虑的重要约束条件。传统分布式环境中,调度的优化目标均是以系统为中心,主要面向系统性能,如系统吞吐量、CPU 利用率等,而对用户的 QoS 需求考虑较少。云计算环境中不仅注重资源利用率及系统性能的提高,而且重视保证用户的 QoS 需求,以实现资源供给与资源消费的双赢局面^[28]。

4 虚拟机迁移

系统虚拟化技术是一种重要的虚拟化技术,具有跨平台、隔离性强、可移动等特点。云数据中心的资源以虚拟机的形式提供给用户,实现资源的动态切割和分配,用户通过虚拟机执行任务。由于云数据中心的虚拟机数量众多,而且虚拟机数量和虚拟机的负载会随用户和应用的需求而经常变化,静态的资源分配往往会使虚拟机产生资源浪费或资源不足的情况,而人工的动态资源调整会有明显的滞后性,因此虚拟机资源需要进行动态调整。虚拟机实时迁移一般运用在虚拟环境中。所谓虚拟机的动态迁移,是指虚拟机在不关机、且能持续提供服务的前提下,从一台虚拟平台服务器迁移到其他的虚拟平台服务器运行,目前 Citrix、VMware 以及微软等几家主要的虚拟平台厂商都提出了自己的虚拟机动态迁移技术。

数据中心通常包含大量计算节点,并且经常会出现节点失败的情况,因此,虚拟机的失效是一种常态。通过虚拟机实时迁移,无缝地将虚拟机从失效的物理机转移到稳定的物理机上,快速部署新节点来代替失效的节点,实现故障隔离,确保计算环境的可用性非常重要。此外,虚拟机实时迁移技术可以从以下两个方面考虑。

4.1 能量优化

在云数据中心,大量时间里物理服务器处于空闲状态。为节省能量,采用实时动态迁移技术将任务从空闲物理机迁出,关闭空闲的服务器或将服务器置于节能休眠状态,实现节省能耗的目的。

文献[26]提出了一种虚拟机在线迁移方法,该方法能最小化迁移过程中的虚拟机停机时间,保证数据的一致性和完整性。Sandpiper 系统^[29]提出了基于黑盒方法和灰盒方法的云数据中心资源监控方法,它能自动监测和探测热点,并通过虚拟机迁移完成新的物理资源到虚拟资源的映射,消除系统热点。

4.2 负载均衡

空闲物理机处于低负载状态或过载时,可以通过负载重分配和迁移,动态关闭和重启物理机,实现负载均衡,获得更高的系统效率。

文献[27]针对虚拟机集群资源负载不平衡的问题,基于虚拟机迁移技术,提出了一种虚拟机集群资源调度策略,为虚

拟机中的应用提供透明的资源调度,并使虚拟机集群达到负载均衡和节能的双重目标。文献[30]实现了一种在虚拟机集群模式下基于动态资源分配原则的一种负载均衡方案,实时监测虚拟机和物理机的资源使用情况,然后对运行在同台物理机上的虚拟机进行资源的重新分配,达到本地虚拟机的负载均衡;而在物理机之间进行虚拟机的热迁移,达到全局的负载均衡。它优化了虚拟机资源的分配,实现了虚拟机集群的全局负载均衡。文献[18]提出一种虚拟机动态迁移框架,实现云计算环境中资源的实时监控,自主发起动态迁移;而且能够在不同类型的虚拟机监控器间进行虚拟机的动态迁移,提高动态迁移的灵活性,同时保证迁移过程中虚拟机的 SLA。

上述云数据中心所涉及的虚拟资源管理技术如图2所示,分为资源提供、虚拟化机部署、资源调度和虚拟机迁移四个部分。

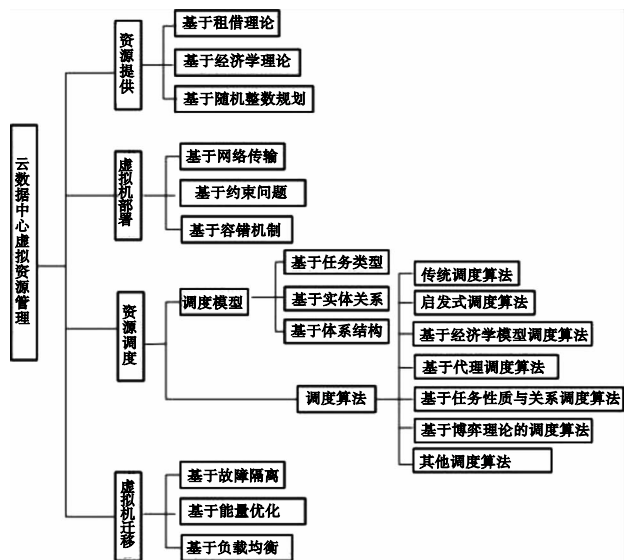


图2 云数据中心虚拟资源管理技术

5 结束语

云计算数据中心通过虚拟资源管理技术,可以实现资源的自动部署、动态优化、节能降耗,提供用户满意的云计算服务。随着云计算技术的产业化,云数据中心如何降低系统运行成本,以经济的价格为用户提供优质服务,是增强市场竞争力的关键。

本文主要研究了云数据中心虚拟资源管理的关键技术:资源虚拟化、资源提供、虚拟机部署、资源调度以及虚拟机迁移。在虚拟资源管理方面,今后有待进一步研究的内容主要包括:a)多阶段的动态资源分配。由于用户使用云资源可以采用资源预约、即用即付等多种方式,数据中心资源提供的实施将分为不同阶段。由于资源的需求具有不确定性,需要考虑在用户需求的动态不确定性情况下,如何提高云数据中心的收益和资源利用率。b)QoS约束问题。云应用呈现的多样性使得用户对服务质量的需求不同。如何为用户提供不同SLA的云应用服务,建立SLA度量、监测、惩罚机制,确保满足用户的QoS需求。c)多数据中心的资源分配问题。通过将多个不同厂商、企业的数据中心联合起来,建立更大范围的动态伸缩资源池,扩展资源提供能力是有待进一步研究的问题。

参考文献:

- [1] 卢锡城. 虚拟计算环境聚合与协同机理研究[R]. 2010.
- [2] 辛军,陈康,郑伟民. 虚拟化的集群资源管理技术研究[J]. 计算机科学与探索, 2010, 4(4): 324-329.
- [3] YOUNG A J, Von LASZEWSKI G, WANG Li-zhe, et al. Efficient resource management for cloud computing environments[C]//Proc of International Conference on Green Computing. 2010:357-364.
- [4] XIONG Peng-cheng, CHI Yun, ZHU Sheng-huo, et al. Intelligent management of virtualized resources for database systems in cloud environment[C]//Proc of the 27th IEEE International Conference on Data Engineering. 2011:87-98.
- [5] ZHENG Zhong-ni, WANG Rui, ZHONG Hai, et al. An approach for cloud resource scheduling based on parallel genetic algorithm[C]//Proc of the 3rd International Conference on Computer Research and Development. 2011:444-447.
- [6] YE Yun-qi, JAIN N, XIA Long-sheng, et al. A framework for QoS and power management in a service cloud environment with mobile devices[C]//Proc of the 5th IEEE International Symposium Service Oriented System Engineering. 2010:236-243.
- [7] HAMSCHER V, SCHWIEGELSHOHN U, STREIT A, et al. Evaluation of job_scheduling strategies for grid computing[C]//Proc of the 1st IEEE/ACM International Workshop on Grid Computing. 2000: 191-202.
- [8] 孙瑞峰,赵政文. 基于云计算的资源调度策略[J]. 航空计算技术, 2010, 3(4): 103-105.
- [9] 袁文成,朱怡安,陆伟. 面向虚拟资源的云计算资源管理机制[J]. 西北工业大学学报, 2010, 5(28): 704-708.
- [10] 高宏卿,那颖. 基于经济学的云资源管理模型研究[J]. 计算机工程与设计, 2010, 31(19): 4139-4146.
- [11] NGUYEN H V, TRAN F D, MENAUD J M. SLA-aware virtual resource management for cloud infrastructures[C]//Proc of the 9th IEEE International Conference on Computer and Information Technology. 2009:357-364.
- [12] CHAISIRI S, LEE B S, NIYATO D. Optimal virtual machine placement across multiple cloud providers[C]//Proc of IEEE Asia-Pacific Services Computing Conference. 2009:103-110.
- [13] CHAISIRI S, LEE B S, NIYATO D. Robust cloud resource provisioning for cloud computing environments[C]//Proc of IEEE International Conference on Service-Oriented Computing and Applications. 2010:1-8.
- [14] CHAISIRI S, LEE B, NIYATO D. Optimization of resource provisioning cost in cloud computing[J]. IEEE Trans on Volume Services Computing, 2011, 99(7): 1-32.
- [15] BOBROFF N, KOCHUT A, BEATY K. Dynamic placement of virtual machines for managing SLA violations[C]//Proc of the 10th IFIP/IEEE International Symposium on Integrated Network Management. 2007:119-128.
- [16] MIYASHITA K, MASUDA K, HIGASHITANI F. Coordinating service allocation through flexible reservation[J]. IEEE Trans on Services Computing, 2008, 2(1): 117-128.
- [17] CHEN Jin, SOUDARARAJAN G, AMAZ C. Autonomic provisioning of backend databases in dynamic content Web servers[C]//Proc of IEEE International Conference on Autonomic Computing. 2010:231-242.
- [18] LIU Shuo, QUAN Gang, REN Shang-ping. On-line preemptive scheduling of real-time services with profit and penalty[C]//Proc of IEEE Southeast Conference. 2011:287-292.

(上接第2415页)

- [19] NGUYEN H V, TRAN F D, MENAUD J M. Performance and power management for cloud infrastructures[C]//Proc of the 3rd IEEE International Conference on Cloud Computing. [S. l.]: IEEE Press, 2010:329-336.
- [20] HYSER C, MCKEE B, GARDNER R, *et al.* HP Labs technical report, Autonomic virtual machine placement in the data center[EB/OL]. (2007). <http://www.hpl.hp.com/techreports/2007/HPL-2007-189.html>.
- [21] BOBROFF N, KOCHUT A, BEATY K. Dynamic placement of virtual machines for managing SLA violations[C]//Proc of the 10th IEEE Symposium on Integrated Management. Wahington DC:IEEE Computer Society,2007:119-128.
- [22] MACHIDA F, KIM D S, PARK J S, *et al.* Toward optimal virtual machine placement and rejuvenation scheduling in a virtualized data center[C]//Proc of IEEE International Conference on Software Reliability Engineering Workshops. 2008:1-3.
- [23] MACHIDA F, KAWATO M, MAENO Y. Redundant virtual machine placement for fault-tolerant consolidated server clusters[C]//Proc of Network Operations and Management Symposium. 2010:32-39.
- [24] 黄昌勤, 李翠菊, 宋广华, 等. 计算网格中的任务管理研究及示范应用[M]. 北京:科学出版社,2009.
- [25] 刘宴兵, 尚明生, 肖云鹏. 网格高性能调度及资源管理技术[M]. 北京:科学出版社,2010.
- [26] BUYYA R, MURSHED M. GridSim: A toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing[J]. *Concurrency and Computation: Practice and Experience*, 2002, 14(13-15):1175-1220.
- [27] BUYYA R, MURSHED M, ABRAMSON D. A deadline and budget constrained cost-time optimization algorithm for scheduling task farming applications on global grids[C]//Proc of International Conference on Parallel and Distributed Processing Techniques and Applications. 2002:540-552.
- [28] 刘晓茜. 云计算数据中心结构及其调度机制研究[D]. 合肥:中国科学技术大学,2011.
- [29] WOOD T, SHENOY P, VENKATARAMANI A, *et al.* Black-box and gray-box strategies for virtual machine migration[C]//Proc of the 4th USENIX Symposium on Networked Systems Design & Implementation. 2007:229-242.
- [30] ZHOU Wen-yu, YANG Shou-bao, FANG Jun, *et al.* VMCTune: a load balancing scheme for virtual machine cluster based on dynamic resource allocation[C]//Proc of the 9th International Conference on Grid and Cloud Computing. 2010: 81-86.
- [31] 刘鹏程, 陈榕. 面向云计算的虚拟机动态迁移框架[J]. 计算机工程, 2010, 36(5): 37-39.
- [32] 张彬彬, 罗英伟, 汪小林, 等. 虚拟机全系统在线迁移[J]. 电子学报, 2009, 37(4): 894-899.
- [33] 周文煜, 陈华平, 杨寿保, 等. 基于虚拟机迁移的虚拟机集群资源调度[J]. 华中科技大学学报, 2011, 39(sup1): 130-133.