软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(2):179−199 [doi: 10.3724/SP.J.1001.2012.04146]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

面向 Internet 数据中心的资源管理*

张 伟[1,2], 宋 莹[3], 阮 利[1,2,4+], 祝明发[1,2], 肖利民[1,2]

[1](软件开发环境国家重点实验室(北京航空航天大学),北京 100191)

[2](北京航空航天大学 计算机学院,北京 100191)

[3](计算机体系结构国家重点实验室(中国科学院 计算技术研究所),北京 100190)

[4](轨道交通控制与安全国家重点实验室(北京交通大学),北京 100044)

Resource Management in Internet-Oriented Data Centers

ZHANG Wei[1,2], SONG Ying[3], RUAN Li[1,2,4+], ZHU Ming-Fa[1,2], XIAO Li-Min[1,2]

[1](State Key Laboratory of Software Development Environment (BeiHang University), Beijing 100191, China)

[2](School of Computer Science and Technology, BeiHang University, Beijing 100191, China)

[3](State Key Laboratory of Computer Architecture (Institute of Computing Technology, The Chinese Academy of Sciences), Beijing 100190, China)

[4](State Key Laboratory of Rail Traffic Control and Safety (Beijing Jiaotong University), Beijing 100044, China)

+ Corresponding author: E-mail: ruanli@buaa.edu.cn

Zhang W, Song Y, Ruan L, Zhu MF, Xiao LM. Resource management in Internet-oriented data centers. *Journal of Software*, 2012,23(2):179−199. http://www.jos.org.cn/1000-9825/4146.htm

Abstract: Internet data centers are developing towards a diversified, intelligent, automated, large-scaled, and standardized direction. With the increasing of scale and complexity, it brings great challenges in how to effectively manage resources. Currently, resource management has become a major issue in Internet data centers, and its importance and urgency cannot be ignored. This paper analyzes two major challenges of resource management with which the Internet data center is facing: (1) meeting the compatibility of concurrent and multiple application SLAs (service level agreements); (2) improving the energy efficiency of system service. Based on the challenges, this paper thoroughly analyzes and summarizes the related work of resource management guaranteeing the SLA, reducing power, and incorporating the objectives of guaranteeing the SLA and reducing power simultaneously during the last ten years. Finally, the paper summaraizes the research and points out future research directions.

Key words: data center; resource management; SLA (service level agreement); power; virtualization

摘 要: Internet 数据中心向多元化、智能化、自动化、规模化与标准化道路发展,其规模越来越大、越来越复杂,这为如何有效管理资源带来极大的冲击与挑战.当前,资源管理已成为 Internet 数据中心亟待解决的重要问题,其重要性与紧迫性已不容忽视.分析了 Internet 数据中心资源管理面临的两大挑战:(1) 满足并发多应用 SLAs(service

* 基金项目: 国家自然科学基金(60973007, 61003015, 60973008, 60921002); 国家高技术研究发展计划(863)(2011AA01A205); 国家教育部高等学校博士学科点专项科研基金(20101102110018); 核高基重大专项(2010ZX01036-001-001); 2011 专项移动互联网总体架构研究项目(2011ZX03002-001-01); 轨道交通控制与安全国家重点实验室(北京交通大学)开放课题基金(RCS2008K001)

收稿时间: 2011-07-15; 修改时间: 2011-09-07; 定稿时间: 2011-11-14

level agreements)的兼容性;(2) 提高系统服务的能量有效性.以挑战为主线,对近十几年来国内外在满足 SLA、降低功耗、同时满足 SLA 和降低功耗方面所取得的资源管理研究成果进行了全面的概括总结和分析,最后进行总结并对未来的研究发展趋势提出观点.

关键词: 数据中心;资源管理;SLA(service level agreement);功耗;虚拟化

中图法分类号: TP316 **文献标识码**: A

Internet 数据中心(Internet data center,简称 IDC)是当前信息技术领域的研究热点,是产业界、学术界、政府等各界十分关注的焦点.它体现了网络就是计算机的思想,将大量计算资源、存储资源与软件资源链接在一起,形成巨大规模的共享虚拟 IT 资源池,是数据运算、交换、存储的中心,以其便利、经济、高可扩展性等优势吸引了越来越多的企业和用户的目光,承载网络中 80%以上的服务请求和数据存储量,为业务体系的健康运转提供服务和运行平台.IDC 是一个重要 IT 产业增长点,具有巨大的市场增长前景.根据 IDC 咨询公司调查统计,2009 年中国数据中心外包服务市场整体规模达 70 亿人民币,增长率高达 35.2%[1].在目前全业务发展的趋势下,将会催生 IDC 数据中心的快速发展期,引领其向多元化、智能化、自动化、规模化与标准化的道路发展.

但当前,Internet 数据中心面临许多关键性问题,而资源管理问题首当其冲.并且随着 Internet 数据中心的快速发展和不断普及,资源管理的重要性呈现逐步上升趋势.H3C 2009 年提出,IDC 资源管理不仅要综合考虑厂商、设备、应用、用户、技术等各种因素,而且需要考虑与数据中心 IT 部门的运维流程结合,要建立一个开放式、标准化、易扩展、可联动的统一智能管理平台绝非易事.数据中心的规模越来越大、越来越复杂,且应用种类越来越多、越来越复杂,这更加剧了资源管理的难度.

目前,Internet 数据中心资源管理问题得到越来越多的关注.著名的国际会议如 IWQoS,LISA,NSDI,NOMS,ICAC,IM 等将 Internet 数据中心资源管理列为焦点问题,NOMS 专门设置关于 Internet 数据中心资源管理的研讨会.许多企业组织、研究团体都启动了相关研究,各大 IT 厂商也在关注各类资源管理产品.企业组织 HP,IBM 等,研究团体 Berkeley,CMU,Massachusetts,Michigan,Rice,Duke,Illinois,Florida,George Mason,Tennesse 等对此展开研究.本文分析 Internet 数据中心的发展趋势,总结 Internet 数据中心所面临的资源管理挑战,对近十几年针对挑战展开的相关研究进行分类、分析与总结,并对有待解决的问题及未来研究重点给出观点.

1 Internet 数据中心发展趋势

Internet 数据中心为集中式收集、存储、处理和发送数据的设备提供运行维护的设施以及相关的服务体系,最主要的业务是资源租赁.传统 IDC 无法面对突发流量所带来的影响,容易造成瘫痪与资源浪费.根据美国国家标准与技术研究院(NIST)的定义,当前云计算服务可分为 3 个层次[2],分别是:(1) 基础设施即服务(IaaS),将虚拟机等资源作为服务提供给用户,如 Amazon 的弹性计算云;(2) 平台即服务(PaaS),为开发人员提供应用程序开发及部署平台,如微软的 Azure 平台等;(3) 软件即服务(SaaS),将应用作为服务提供给用户,如 Google 的 Gmail 等.其中,IaaS 解决传统 Internet 数据中心弊病,使云计算提供弹性资源,根据每个租用者的需要,在一个超大的资源池里动态分配或释放资源,为用户提供“召之则来,挥之即去”且“能力无限”的 IT 服务.因此,云计算是 Internet 数据中心未来发展的趋势.Gartner2010 年的调查结果显示,云计算是 IT 用户最关注的重要技术之一.目前 IBM,HP,Amazon,Google,Microsoft 等大型 IT 公司和数据中心分别建立并对外提供各种云端资源.

随着当前云计算等的日益澎湃发展,资源管理动态化、弹性化和自动化需求更加突出.实现资源的按需动态伸缩对于云计算数据中心的可用性是至关重要的.例如,Amazon AWS 的服务器由于不能抵抗高并发应用的冲击,出现服务器宕机,造成巨大损失[3].目前,越来越多的服务提供商选择租赁 IDC 的资源对外提供服务,业务和用户的需求使 Internet 数据中心向多元化服务推进.总体来看,资源服务化已经成为 IDC 的重要外部特征,资源已经呈现出了聚集化、并行化日益突出等趋势,相应资源管理需向动态化、弹性化、自动化方向发展.

未来 Internet 数据中心趋向集中式管理和一站式购买方向发展,满足高并发应用的性能、低功耗、自动化、高利用率将成为 IDC 资源管理的关注点.由于 Internet 数据中心代表了未来信息技术领域的核心竞争力,当前世

界各国政府、国内研究界和工业界都十分重视 Internet 数据中心的发展,力争在未来信息技术领域占据一席之地.例如,2010 年美国对 IDC 的开支超过 1 250 亿美元.我国政府对 IDC 的发展也高度重视,2011 年 9 月 15 日来自云计算产业链上的各方代表齐聚一堂,对中国 IDC 发展趋势进行深入探讨.

2 Internet 数据中心资源管理挑战

随着 Internet 应用的飞速发展,由于数据、计算、网上交易、通过 Internet 进行即时交互需求的急剧增长使得对 IDC 的需求呈爆炸式增长.如,社交网站 Facebook 自 2007 年 1 月起用户数量以每周平均 3%的速度增长,目前已经创建一个用户人数超过 6 亿的平台.Twitter 在一年内的客户群从 300 000 增长到了 800 000,Twitter 服务器每秒需要处理几千条即时消息.IDC 承载多种应用,不同应用具有不同的负载特征,资源需求特征也存在差异,如何满足高并发多应用 SLA(service level agreement)的兼容性是 IDC 资源管理者面临的一大挑战.

需求的飞速增长和多样化使得数据中心的规模越来越大,越来越复杂.据统计[4],Google、微软、拍卖网站 eBay、雅虎、Facebook、亚马逊等所拥有的服务器数量均在几十万台.我国 2009 年各类数据中心和机房的总数量达到 519 990 个,超过 70%的服务器安装运行在数据中心.IDC 预计[5],到 2012 年我国数据中心数量将以复合年增长率 1.3%达到约 540 777 个左右.40.6%的数据中心服务器数量以 20%~50%速度增长.随着 IDC 规模的激增,供应系统正常运行的电能和制冷等成本也呈爆炸式增长,由此造成的系统过热现象严重[6].然而据统计,IDC 的资源利用率很低,平均只有 30%左右,服务器在一天中有很大一部分时间处于闲置,即使空闲时也会带来满载时 60%的功耗[7].基于以上的分析,如何提高系统服务的能量有效性是 IDC 资源管理者面临的又一大挑战.

2.1 满足并发多应用SLAs的兼容性

SLA 即服务等级协议,是 QoS 的另一种表达方式.据统计[8],若应用性能超出用户期望的 QoS,不仅存在提升用户失望程度的风险,更有甚者会带来巨大商业损失.用户也有可能转向其他提供商,给公司产品或公司本身带来负面影响.以下原因使得同时满足并发多应用 SLAs 的兼容性难度更大:

(1) IDC 承载多种应用.不同应用具有不同负载特征和资源需求特征.

(2) Internet 应用种类越来越多,越来越复杂.以前仅由简单静态 Web 内容组成,而目前提供更灵活的 Web 内容和安全能力以保护用户信息.基于 Web 计算模式的出现导致(如 Web service)更复杂结构产生.

(3) Internet 应用工作负载具有随时变动性、突发性.

2.2 提高系统服务的能量有效性

能量有效性即单位电能所产生的性能,以 performance/watt 表示.海量的能源消耗和恼人的散热问题也是一直成为数据中心发展的瓶颈问题.如何提高系统服务的能量有效性,控制数据中心功耗、降低电能和制冷成本、降低服务器过热导致的故障是数据中心资源管理者亟待解决的问题.其原因如下:

(1) 环境污染的主要来源之一.数据中心为全球“贡献”30%的 CO_2 排放量[9],美国环境保护机构发起“能源之星计划”,呼吁数据中心采取措施降低 CO_2 排放量.

(2) 电能、制冷成本急剧上升.2010 年美国电能和制冷成本花费大约 74 亿美元,在数据中心总成本中所占的比重越来越大.据预测不加以控制,将来会赶超硬件成本.

(3) 产生过多热量.由于系统过热导致服务器出现故障的概率增加.2011 年 3 月 25 日,欧洲数据中心就由于系统过热导致维基百科出现大范围宕机,严重影响到了欧洲所有维基百科用户[10].

2.3 以挑战为主线的相关研究发展脉络

以资源管理挑战为主线,将近十几年 Internet 数据中心资源管理的相关研究划分为满足应用 SLA、降低功耗、同时满足 SLA 且降低功耗三大类进行分析与总结,相应章节中给出划分的依据,如图 1 所示为整个分类.

第 3 节对以满足 SLA 为目标的资源管理研究进展进行分析与讨论,按资源共享还是独占划分为两大类.其中,资源共享方式按粒度进一步划分为以进程为粒度的资源共享和以虚拟机为粒度的资源共享.独占平台涉及到资源框架的设计和服务器动态分配策略,资源分配策略按所用的方法进行分析与总结.在共享平台中,以进程

为粒度的方式按解决子问题分为两类:应用放置/迁移与同结点应用资源的动态调整,其中同结点应用资源的动态调整涉及到性能隔离和动态资源分配策略两个技术细节.以虚拟机为粒度的共享涉及虚拟机放置/迁移与同结点虚拟机资源的动态调整两大问题.虚拟机放置/迁移主要使用启发式算法、优化理论两种方法.同结点虚拟机资源的动态调整主要使用控制论、模糊逻辑、效用函数和优化方法.

第 4 节对以降低功耗为目标的资源管理研究进展分析与讨论,按功耗控制粒度和手段划分为 DVS/DVFS、开/关机、服务器整合 3 类.DVS/DVFS 按非虚拟化、虚拟化平台进行分析与总结.开/关机按达到目的的手段分为资源分配和负载分发两类.服务器整合按服务器整合评价系统、实际服务器整合系统进行分析与总结.

第 5 节对同时满足 SLA 且降低功耗的资源管理研究按非虚拟化平台与虚拟化平台进行分析与总结.

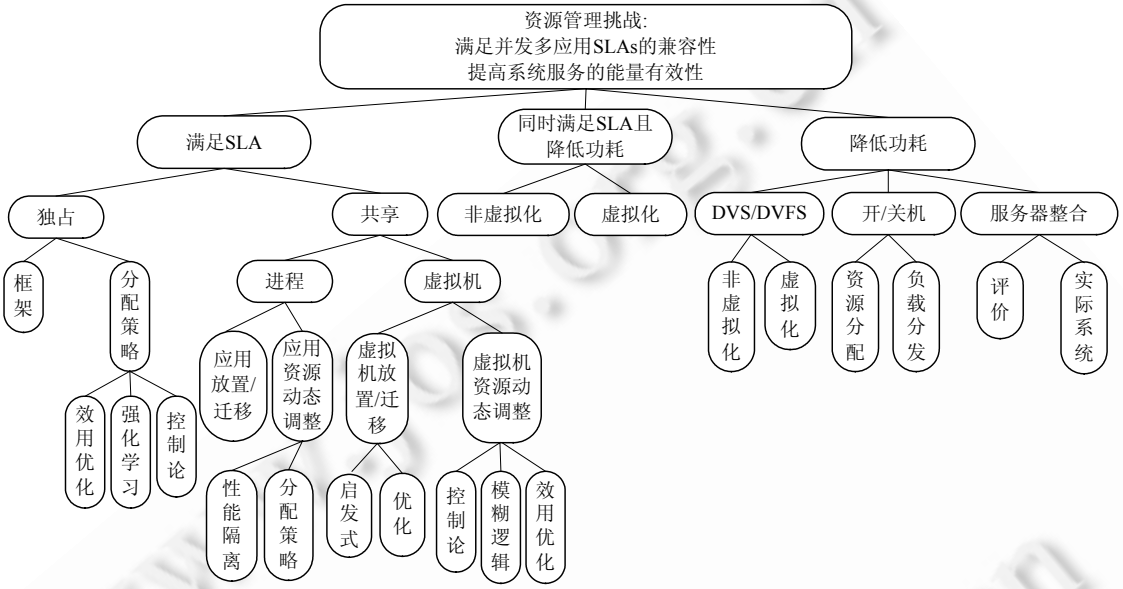

Fig.1 Related research trends of resource management

图 1 资源管理相关研究发展脉络

3 以满足 SLA 为目标的资源管理研究现状

数据中心使用基于虚拟机粒度的共享平台之前,广泛使用基于进程粒度的共享平台和基于物理机粒度的独占平台.早期出于成本考虑,倾向于使用基于进程粒度的共享平台.这种方式下,资源利用率高,但应用之间隔离性差.随着硬件成本的下降,为降低应用之间的干扰,开始采取基于服务器粒度的独占平台,但存在严重资源浪费.随着技术进步和硬件支持,虚拟化技术再一次兴起,集结了基于进程粒度的共享平台和基于服务器粒度独占平台的优势,以虚拟机为粒度实现资源共享,不仅存在资源利用率高的优势并且应用之间隔离性好,但虚拟化本身带来一定的性能开销.独占平台仅涉及以物理机为单位的服务器分配,而在共享平台下,不仅涉及物理机的分配(应用或虚拟机放置/迁移),而且涉及同台物理机上应用/虚拟机资源的动态调整.两者资源分配粒度和涉及的方面有所不同,纵观以满足 SLA 为目标的资源管理相关工作,本文主要以资源共享或独占进行划分,其中共享方式又进一步细分为以进程为粒度的共享和以虚拟机为粒度的共享.

3.1 基于物理机粒度的独占平台资源管理

随着硬件成本的下降,为了减少应用之间的相互干扰,数据中心倾向于把应用或应用的一个组件部署在独立的物理机上,服务器的共享是通过时分复用实现的.每个应用或组件运行在不同的操作系统核心,相互之间不存在干扰.但由于不同应用(CPU 密集型、内存密集型、I/O 密集型)对不同资源具有偏好性,如 CPU 密集型应用

所运行的物理机出现内存、磁盘等资源相对闲置的状态,并且应用负载具有随时变动性,负载处于波谷状态时,对所偏好资源的利用率也是很低的.以上现象造成物理机资源利用率不高、存在严重浪费的后果.

在这种平台下以单台物理机作为资源分配的粒度,通常把资源池划分为多个域,每个域为一类应用服务,根据应用负载转化为对物理服务器数量的需求,从而对域进行动态划分.相关研究包括:(1) 资源分配框架的设计;(2) 服务器在域之间进行动态分配的具体策略.

3.1.1 资源分配框架

IBM 提出集中式动态按需分配框架[11]和两层动态分配架构[12].Appleby 等人[11]首次针对这种平台提出一种集中式动态按需分配框架——Oceano,把整个资源池划分为多个客户域,整个系统存在一个集中资源控制器,通过 SLMonitor 监控每个域各服务器的 SLA Metrics(如每个服务器的活动连接数、响应时间、输出带宽等),与设定的固定阈值比较.若超出固定阈值,则启用域内的空闲服务器;若不存在,则向其他资源使用率低的域暂借使用.Walsh 等人[12]提出了两层动态资源分配架构,每个域存在局部资源控制器,根据负载量提出资源需求.整个资源池存在一个全局资源裁决器,根据各局部资源控制器提交的资源需求申请和资源总量的限制,做出最终决策.由于其相对于集中式架构,具有更好的可扩展性,这种架构被以后很多学者所采用[13−19].

集中式架构实现简单,但可扩展性差,易出现单点故障.两层架构中每个域存在“组长”,由其再与全局控制器交互,可扩展性更好,但全局控制器仍是整个系统单点故障之所在.系统中可考虑用主备模式提高系统可靠性.

3.1.2 服务器动态分配策略

服务器动态资源分配需要根据应用负载量转化为资源需求量,考虑资源池的容量,采取方法为应用分配服务器,从而满足 SLA 目标.

基于 Oceano 框架展开研究的相关工作有文献[20,21].Duke 大学 Moore 等人[20]设计了 COD(cluster on demand)系统,实现了一种能够动态调整虚拟集群大小的协议,使得物理机能在多个虚拟集群之间快速、自动、在线迁移.Rice 大学 Ranjan 等人提出了 QUID-online[21]服务器迁移算法,为每个服务器定义期望的 CPU 利用率,依据 $G/G/N$ 排队模型把平均响应时间表示为所分配服务器个数和请求完成量的函数,以此来计算实际需要的服务器数量.但是他们没有考虑整个资源池服务器数量有限的限制条件.

基于 Walsh 提出的混合式两层架构展开研究的相关工作有文献[13,22−26].所用的方法主要有:(1) 基于效用函数和优化理论;(2) 强化学习;(3) 控制论.基于效用函数和优化理论的代表性工作是 IBM 的 Levy 等人[22]所做的工作.使用 $M/M/1$ 排队模型为每类应用构建响应时间与请求到达率和服务器分配数量间的函数模型,根据实际响应时间和期望时间的差异为每类应用构建局部效用函数,为使整个资源池全局效用最小为目标进行资源分配.针对文献[21]没有考虑资源池服务器数量有限的缺陷,在优化模型限制条件中加入此限制.基于强化学习的代表性工作是 IBM 的 Tesauro 等人的工作[23−25].基于控制论的代表性工作是 Bouchenak 等人[26]设计的 Jade 系统,决策逻辑基于实际资源使用率与期望资源使用率的差异,使用反馈控制理论调整资源.

基于效用函数和优化理论方法的难点在于效用函数的构建,需要专业领域知识,而强化学习和控制论不需要专业领域知识.其中,强化学习不需要建立显式的系统模型,但需要对数据进行训练,过长的训练时间导致可扩展性差.基于控制论的方法不需要准确预测应用的负载参数,将整个系统看成一个“黑盒”,具有稳定性、准确性、稳定时间短和不过度满足指标等特点.

3.1.3 小 结

在基于物理机粒度的独占平台下,每个应用或应用的一个组件运行在一台物理机上.应用之间相互隔离,不会相互干扰.然而以整个物理机作为资源重新分配的单位,粒度很大.由于不同应用对资源的偏好和负载到达波峰、波谷的差异性,导致物理机的资源利用率很低,如 HP 数据中心中平均每台物理机的利用率在 15%~20%之间,存在严重的资源浪费.从空间、功耗、制冷和成本方面考虑,代价都很高.

3.2 基于进程粒度或虚拟机粒度的共享平台资源管理

早期计算机非常昂贵,为了节约购置成本,将多个应用或应用的多个组件运行于一台物理机上,应用以进程的方式共享物理机资源.这种方式下,物理机资源利用率高,但需要操作系统提供动态资源分配机制支持

(CPU,Memory,I/O 资源的动态分配),并且应用之间易受干扰,隔离性差.

随着技术进步和硬件支持,提高硬件资源使用率并在应用之间提供更好隔离性的现实需求,推动虚拟化技术再一次兴起.这种方式下,一个物理机上可运行多台虚拟机,每台虚拟机运行一个应用或应用的一个组件,以虚拟机为粒度共享物理机资源.这种方式下,每个应用在自己的操作系统环境中独立运行,相互间干扰小.虚拟化技术提供细粒度资源动态分配机制支持(如 Xen 的 Credit 调度支持 CPU 资源的动态分配,balloon driver 机制支持内存资源的动态分配),越细粒度的资源分配越接近优化分配.不同应用对资源的偏好和负载的波峰、波谷到来时机不同,可利用分配机制为虚拟机动态分配资源,提高物理机资源利用率.同时,虚拟化技术提供虚拟机迁移机制支持,便于实现整个资源池的优化分配.虚拟化技术通过添加一层中间层-虚拟机监控器,带来以上灵活性的同时,也带来一定的开销,目前最先进的虚拟化技术至少也给系统带来 5%~10%的性能损失[27],这种开销主要来自内存虚拟化和 I/O 虚拟化.内存虚拟化的性能开销主要源于虚拟机监控器截获虚拟机内存申请、释放函数和影子页表导致虚拟机陷入.I/O 虚拟化的性能开销主要源于虚拟机监控器截获 I/O 指令导致虚拟机陷入.

基于以上的分析,我们根据资源共享粒度以进程方式还是虚拟机方式对共享平台进行进一步划分.

3.2.1 基于进程粒度的共享平台资源管理

这种方式下,资源分配的相关工作涉及到两个方面:(1) 应用放置/迁移,即根据应用资源需求进行应用到物理机的放置或重放置[28−30].(2) 同结点应用资源的动态调整.在这种方式下多个应用共享物理机资源,而应用负载又具有随时变动性,静态部署无论从满足 SLA 或资源使用率方面都会受到冲击.若应用负载量增大,则应增加资源分配量,从而避免 SLA 受到影响;若应用负载量减小,则应减少资源分配量,从而避免资源浪费.需要设计策略根据应用变化的负载获得资源需求,动态分配物理机资源,降低 SLA 受到影响或资源浪费概率的发生[14,31−34].

3.2.1.1 应用放置/迁移

基于进程粒度的共享平台运行多个应用,首先面临的问题是考虑物理机资源、硬件或其他因素的限制,解决应用到物理机的映射问题.应用部署首先需要获得其资源需求,综合考虑需求和物理机容量,解决应用到物理机的放置问题.随着负载变化,应用资源需求之和可能超出物理机资源容量限制,发生资源竞争,这时需要进行应用到物理机的重新放置.

较早从事这方面工作的知名学者是 Massachusetts 大学的 Urgaonkar[28,30].在 Urgaonkar 等人的工作中[28],准确获得应用在独占结点上的资源使用率,以此来指导应用在共享平台上的资源需求,满足应用 CPU 和网络资源需求的物理结点作为可选结点.他们把放置问题抽象为图中顶点和边之间的关系.顶点代表应用和物理机,若应用可放置到物理结点上,则它们之间存在一条边,然而这种技术在寻找新放置方式时对以前的放置方式不具有记忆性,因此非常不适合大规模和频繁进行在线放置的场景.为了解决先前工作的可扩展性问题,Urgaonkar 等人设计了 Shark 系统[29].为了解决文献[28]中重新放置时对以前放置的无记忆性,IBM 的 Karve 等人[30]引入放置控制器,使用放置矩阵保存先前的放置方式.他们把问题抽象为多目标优化问题,在满足应用性能的同时,选取使先前放置矩阵和新放置矩阵之间改变最少的方式进行应用放置.

以上研究[28−30]存在以下缺陷:(1) 没有考虑放置改变成本.重新部署涉及到应用在服务器之间的迁移,需要在源物理结点暂停应用的执行,在目的物理结点重新启动运行或者采取进程迁移的方式都会对应用性能产生较大影响,因此部署策略中应该显式考虑迁移成本.(2) 资源池有限.应用资源需求之和超出整个资源池容量限制,应考虑应用优先级,对其提供差异服务,优先保证高优先级应用的服务.(3) 缺乏对异构服务器的考虑.

3.2.1.2 同结点应用资源的动态调整

多个应用运行在同一物理机上,需要操作系统提供性能隔离[31,32],根据应用负载计算获得资源需求[14,33,34],调用操作系统提供的细粒度资源分配接口,为应用动态增加或减少资源.

性能隔离:要实现性能隔离,需要对传统操作系统进行扩展.Rice 大学 Druschel 等人[31]设计了 Resource Container 系统,实现了单台物理机上多个应用间的性能隔离和 CPU 细粒度资源分配机制支持,然而局部隔离并不能保证全局隔离.次年,Druschel 等人设计了 Cluster Container 系统[32],以解决应用在集群范围内的隔离问题.

动态资源分配策略:需要根据应用的负载量决定对资源的需求,考虑各个应用的资源需求和资源总量进行

最终资源分配决定.IBM 设计的 MBRP 系统[33]是较早解决应用对物理机资源动态划分的工作.该系统把服务请求响应时间视为 CPU 处理时间和磁盘处理时间之和,使用基于模型的方法对内存和存储资源进行动态分配,但是这种方法仅适用于负载稳定且规模小的集群范围,可扩展性差.为了应对负载具有高可变的环境,Massachusetts 大学 Chandra 等人[14,34]提出使用在线观测的方法适应负载的改变,控制 CPU 和接收队列资源以满足各类负载 QoS 需求.其中,文献[34]使用时间序列预测分析技术,根据在线监测的值预测未来工作负载,使用时域队列模型决定期望的资源需求,基于受限制优化技术动态决定服务器资源的分配.

3.2.2 基于虚拟机粒度的虚拟化平台资源管理

2003 年 IBM 的 Chandra 等人[35]首次提出使用基于虚拟机粒度的共享资源管理平台.类似于进程粒度的共享平台资源管理,基于虚拟机粒度的虚拟化平台资源管理同样划分为两大类:(1) 虚拟机的放置/迁移.借助虚拟机迁移机制支持,根据应用资源需求,进行虚拟机到物理机的放置或重放置;(2) 同结点虚拟机资源的动态调整.由于虚拟机中应用负载具有随时变动性,同样需要借助虚拟化技术提供的细粒度资源分配机制支持,为虚拟机动态增加或减少资源.

3.2.2.1 虚拟机放置/迁移

虚拟机放置/迁移首先面临的是根据虚拟机的资源需求和物理机容量,把虚拟机放置到物理机上.随着应用负载的变化,可能发生虚拟机资源需求之和超出某些物理机容量的现象,需借助虚拟机迁移机制设计虚拟机迁移策略,决策何时进行迁移、迁移哪些虚拟机、被迁移的虚拟机放置到哪些物理机上等问题,即所谓的 3W 问题.

当前虚拟化平台提供两种虚拟机迁移类型,离线迁移、在线迁移[36].下面从原理、迁移时间、宕机时间、灵活性、对应用性能影响大小几个方面进行对比分析,见表 1.

Table 1 Virtual machine migration comparison

表 1 虚拟机迁移类型对比

迁移类型	原理	迁移时间	宕机时间	灵活性	对应用性能的影响
离线迁移	首先暂停虚拟机的运行,把状态保存到存储设备,然后将保存的虚拟机状态迁移到另一个物理机上,恢复继续运行	比在线迁移短	比在线迁移长	迁移计划灵活性差	比在线迁移影响大
在线迁移	虚拟机在迁移过程中一直处于运行状态,只有在最后一次 post-copy 阶段才需要暂停虚拟机	比离线迁移长	比离线迁移短	迁移计划灵活性好	比离线迁移影响小

虚拟机放置/迁移策略所使用的方法主要有:(1) 启发式算法;(2) 优化理论.

(1) 启发式算法

现有基于启发式的虚拟机放置/迁移研究主要关注 CPU、内存资源[37−41].Purdue 大学 Khanna 等人[37]为每个物理机设定期望 CPU 和内存使用率,通过在线观测以事件触发被动实现虚拟机的重新放置.IBM 的 Bobroff[38]根据监测的 CPU 使用率历史数据,使用时间序列预测虚拟机未来 CPU 需求,使用首次适应启发式算法进行虚拟机的放置/迁移,但没有考虑迁移成本.VMWare 设计的 DRS 系统(dynamic resource scheduler)[39]通过监控 CPU 和内存压力,使用迁移技术实现各个物理机之间的负载均衡,但没有利用应用日志信息进行性能监控.Massachusetts 大学 Wood 等人设计了 Sandpiper 系统[40],使用贪心启发式算法,结合应用性能实现负载均衡.

另一些研究主要关注网络 I/O 资源[42,43].文献[42,43]通过贪心启发式算法解决虚拟机到物理机的放置/迁移.Northwestern 的 Sundararaj 等人[42]设计了 Virtuoso 系统,通过 VTTIF 组件监控应用之间的通信行为,VADAPT 根据应用之间网络通信强度、网络拓扑、网络负载强度信息做出决策.Wisconsin-Madison 的 Shrivastava 设计的 AppAware 系统[43]不仅考虑网络拓扑的影响,而且考虑多层应用中不同组件之间的相互依赖.

总体来看,基于启发式算法的虚拟机放置/迁移通常比较简单且容易实现,但需要大量的专家知识,因此应用领域的专属性较强.此外,启发式算法不保证稳定性,可能导致某些系统性能参数的较大震荡[37,38,40,42−44].

(2) 优化理论

清华大学 Wang 等人[45,46]提出一种面向虚拟服务计算环境的自主管理框架,使用受限制的非线性优化技

术,解决虚拟 CPU 资源到物理资源的映射问题.为简化处理,他们假设所有虚拟机具有相同容量,实用性差.France Orange Labs的Van Nguyen等人[47,48]打破虚拟机容量相同的限制条件,设计自主资源管理器,通过效用函数和受限制的编程方法取得优化放置结果,通过减少虚拟机迁移次数和提高迁移并行度降低迁移成本.IBM 的Meng 等人设计了 TVMPP[49]系统,考虑应用之间的通信强度,把通信量大的虚拟机尽量映射到同一台物理机.

总体来看,基于优化理论的虚拟机放置/迁移方法具有全局优化、可扩展性好的优势,但得到优化解的时间长,比较适用于长时间范围内的操作[45−50].

3.2.2.2 同结点虚拟机资源的动态调整

应用负载具有随时变动性,若根据峰值负载为虚拟机分配资源,则浪费资源;若根据波谷负载分配,则影响性能.理想的资源分配方式是根据应用变化的负载,设计资源分配策略裁决虚拟机应分得的资源量,调用细粒度资源分配接口,为虚拟机动态分配资源.从使用方法上主要分为 3 大类:(1) 控制论;(2) 模糊逻辑;(3) 效用函数和优化理论.

(1) 控制论

控制论是这种平台下使用最多的方法.为达到用户期望的 SLA,目前激发资源动态分配时机的手段分两类:(1) 根据为应用定义的期望资源使用率.资源使用率信息容易获得,类 Unix 系统下提供很多收集工具,如 top,sysstat,memstat,dstat 等.(2) 根据为应用定义的期望 SLA,此处主要指响应时间.类 Unix 系统没有直接提供任何获得响应时间的工具,信息难以获得.随着应用种类越来越多、越来越复杂,以对应用或操作系统透明的方式,实时获得响应时间更具挑战.

① 基于期望的资源使用率

在这种方式下,为虚拟机定义期望的资源使用率,通过比较实际资源使用率与期望资源使用率的差异,决策需要为虚拟机增加或减少的资源量.很多研究根据管理员的经验,采取静态指定方式对期望资源使用率进行定义[15−18,51−55].如果定义得太高,则 SLA 受影响的概率提升;如果定义得太低,则浪费资源.实际中期望资源使用率的定义不仅与应用类型有关,而且依赖于负载强度.为了弥补固定阈值的缺陷,目前有些学者采取收集应用性能和工作负载,周期性修正期望资源使用率[56,57].

固定阈值的方式:早期 HP 针对单层应用 CPU 资源进行动态分配[52,53].随着应用多样化和复杂度的提高,多层应用架构更受青睐,2007 年 Michigan 大学与 HP 合作展开对多层应用资源分配的研究[16,17].Padala[16]为虚拟机设定优先级,使用单输入输出控制器(single-input,single-output,简称 SISO)依据实际 CPU 资源率和期望 CPU 利用率之间的差异进行分配,在发生资源竞争时优先保证高优先级应用的资源需求.类似的考虑虚拟机优先级的工作还有文献[17,19,52,58−60].文献[16]简化了建模过程和控制器的设计,仅使用静态模型捕获输入和输出之间的关系,采取离线方式调整控制器参数.针对以上缺陷,Padala 在文献[17]中对控制器的参数使用最小二乘法进行在线调整,使用多输入多输出控制器(multi-input,multi-output,简称 MIMO)对 CPU 和磁盘资源进行动态分配.HP的Liu等人[15]提出使用自适应多变量控制器,协调多层应用中各层资源分配以满足应用SLA.当多台虚拟机运行于同一个物理机上时,内存更容易成为瓶颈,Illinois大学Heo等人[55]使用反馈控制方法,依据实际内存利用率和期望内存利用率的差异对内存资源进行动态分配.

周期性修正阈值的方式:HP 的 Wang 等人设计了 AppRAISE 系统[56],收集应用的性能和工作负载,周期性修正期望资源使用率的值,集成前向离线预测和被动反馈,控制 CPU 资源的分配,但是前向系统采取离线方式,影响控制系统性能.针对此缺陷,Lund 大学 Kjaer 等人[57]提出使用在线前向预测和反馈控制相结合的方式对 CPU 资源进行分配.

基于期望资源使用率进行资源存在以下缺陷:(a) 在云计算环境下,应用变得越来越复杂,期望的资源使用率如何定义很困难.若定义得太高,则影响应用性能;若定义得太低,则浪费资源.(b) 在虚拟化环境下,准确的资源使用率采样变得更加困难,需要过滤掉 hypervisor 层的开销[35].(c) 当多台虚拟机运行于同一个物理机上时,使用当前的资源使用率预测真实的资源需求效果会很差[61].

② 基于期望响应时间

实时获得响应时间并不是件易事,以期望响应时间作为调整时机的相关研究很少.Illinois 大学 Xue 等人[51]对 httpperf 进行修改,在客户端获得响应时间又传回给服务器端加以利用,存在延迟.并且受网络等因素的干扰,结果不准确.CMU 的 Sangpetch 等人[62]在服务器端使用抓包工具实时收集,分析请求在服务器端的响应时间,开销很大.

(2) 模糊逻辑

Florida 大学 Xu 等人[18]使用这种方法展开研究,他们使用模糊模型和模糊预测方法为每台虚拟机设计局部控制器,自动学习虚拟机运行时的行为.使用模糊模型建立工作负载与 CPU 资源需求之间的关系,使用模糊预测根据当前 CPU 使用率预测未来资源需求,为适应系统改变,用最新信息自动更新模糊模型和模糊预测的相关信息.基于模糊逻辑的方法不需要先验知识和对被管理系统建立数学模型,也不需要长时间训练,具有抗噪声、抗干扰、快速适应改变的能力,尤其适用于实时系统的控制.

(3) 效用函数和优化理论

这类方法大多基于文献[12]所提出的两层架构,每台虚拟机定义一个局部效用函数,提出资源申请,整个物理机定义一个全局效用函数,根据各台虚拟机的资源需求和资源总量,裁决每台虚拟机实际所分得的资源.George Mason 大学 Menasc 等人[19]提供为虚拟机动态设定优先级的接口,为每台虚拟机构建关于与期望响应时间偏离的局部效用函数,采用模拟方式对方法进行验证.IBM 的 Zhang 等人[61]结合预测的思想,预测应用在未来的响应时间,把资源分配问题抽象为非线性优化问题.基于效用函数和优化理论的方法可扩展性好,但难点在于效用函数的合理构建,需要专业领域知识.

3.2.3 小 结

在基于进程粒度的共享平台下,多个应用或应用的多个组件运行在一台物理机上.应用以进程的形式运行于同一个操作系统中,因此需要对操作系统进行扩展,使其具备性能隔离、细粒度资源动态分配功能.由于涉及内核级编程,实现有一定难度,对非开源操作系统更是很大的挑战.由于多个应用共享物理机资源,首先需要解决应用到物理机的放置或重放置.又因为应用负载具有随时变动性,为降低资源不足供给应用 SLA 受影响或过量供给资源浪费现象的发生,需要根据应用负载动态划分物理机资源.此平台下物理机资源利用率高,但应用之间可能会相互干扰、隔离性差.例如,一个应用的运行引起系统崩溃,会导致运行在此物理机上的所有应用都受到影响.

细粒度资源分配和虚拟机迁移是基于虚拟机粒度共享平台下的两种资源分配技术手段,细粒度资源分配开销小、灵活性高,但力度有限.虚拟机迁移资源分配力度大,但迁移过程对网络以及源、目的物理主机的 CPU 带来一定压力.从以上分析可以看出,目前两种技术单独使用,未能有效结合.它们之间的关系可与并行编程中 openMP 与 MPI 之间的关系类比,openMP 在一台物理机上共享内存的方式实现进程间通信,开销小,但频繁的线程创建消耗不少资源,并且容易使物理机内存资源发生竞争.MPI 是不同物理机上以消息传递实现进程间通信的方式,可以更有效地利用多机资源,但对网络带来压力.正如目前发展出现了 openMP+MPI 的混合编程,我们认为将来也需要设计能够有效结合细粒度资源分配和虚拟机迁移的资源分配策略,有效结合两种方式,取长补短.再者,由于虚拟机资源容量在运行过程中动态可变,传统针对物理机环境的负载分发策略也应进行重新设计.

表 2 为平台相关指标的对比.

Table 2 Different platforms comparison

表 2 不同平台对比

平台	特征	操作系统个数	应用隔离性	分配粒度	资源利用率	相关研究
物理机粒度独占	每个应用或组件独享一台物理机	1	好	以物理机为单位	低	文献[11−13,20−26]
进程粒度共享	多个应用或组件共享一台物理机	1	差	细粒度	高	文献[14,28−34]
虚拟机粒度共享	多台虚拟机共享一台物理机	每台虚拟机 1 个	好	细粒度;以虚拟机为单位	高	文献[15−19,51−54,58,60,62]

4 以降低功耗为目标的资源管理研究现状

据统计[6],数据中心功耗成本占 60%,其中处理器占 30%左右,即使在空闲时的功耗也占满载状态下的 60%,是当前学者主要关注的功耗控制部件.本文仅综述了控制 CPU、整机功耗的相关研究.从功耗控制的粒度划分为 3 类:(1) DVS/DVFS,属于组件或部件级控制方式;(2) 开/关机,属于整机级控制方式;(3) 服务器整合,属于资源池级控制方式.我们以功耗控制技术手段为主线,对目前降低功耗的相关研究进行分析和总结.

4.1 DVS/DVFS

这种方式下,在低负载时,通过降低CPU频率或电压进而达到降低功耗的目的,但需要硬件和操作系统的支持.物理机独占环境下采取这种方式降低功耗相对容易,而虚拟化环境下由于一台物理机上同时运行多台虚拟机,硬件状态的改变会影响物理机上运行的所有虚拟机性能,不能片面地仅凭 1 台或某些虚拟机的负载状况就作出 DVS/DVFS 决策.由于不同应用负载的波峰、波谷到来时机存在差异,更加剧了虚拟化环境下采取此种技术进行功耗控制的难度.基于上述分析,我们以非虚拟化平台、虚拟化平台进行划分,对当前采取此技术进行功耗控制的相关研究进行分析与总结.

4.1.1 非虚拟化平台

这种平台下的相关研究,复杂度从单层应用发展到多层应用,空间范围从单层应用运行的单机发展到多层应用运行的多机,进而到整个资源池服务器的协调 DVS/DVFS 操作.早期 IBM 的 Elnozahy 等人[63]提出一种针对 Web 服务器的请求批处理技术,在低负载阶段先把请求保存在网卡的存储器中,处理器处于低功耗状态,当累积的请求超出批处理周期时,把处理器唤醒处理这些请求,但仅针对单层应用.随着多层应用的大量出现,文献[64–66]对多层应用运行的多机环境协调进行 DVS/DVFS 操作.Virginia 大学 Horvath 等人[64,65]对多层 Web 服务器在保证端到端响应时间的限制条件下,使用严格的优化方法基于 DVS 降低总功耗开支.为避免频繁调整,仅当响应时间超出设定的阈值范围时才采取 DVS 操作.中国科学院计算技术研究所 Lin 等人[66]的工作比文献[64,65]更进一步,使用跟踪技术识别多层应用中的主要请求类别,监测主要请求类别在每层的响应时间进行 DVFS 控制.Tennesse 大学 Wang 等人[67]设计了基于多输入多输出控制论(MIMO)的集群级功耗控制器,在整个集群总电能消耗一定的限制条件下,基于服务器性能需求调节 CPU 频率,使电能在各个服务器之间进行转移.

4.1.2 虚拟化平台

在虚拟化环境下,使用 DVS/DVFS 操作降低功耗变得更加困难.这是因为任何硬件组件状态的改变都会影响运行在此物理机上所有虚拟机的性能,仅根据其中 1 台虚拟机的性能就采取 DVS/DVFS 操作,严重影响物理机上运行的其他虚拟机性能.面对这种挑战,Tennessee 大学 Wang 等人[68,69]在这方面展开了研究.文献[68]基于控制论提出一种两层控制架构,主控制循环采取 MIMO 保持各台虚拟机有大致相同的性能,次控制循环控制 CPU 频率以降低功耗.由于 DVS/DVFS 仅允许一定程度的操作,且不能明显降低漏电压,因此在负载极低时功耗控制的效果不明显.基于这些现象,受文献[63]思想的影响,Wang 等人[69]提出一种针对虚拟化平台下的请求批处理技术.当负载低时,动态分配 CPU 资源使各台虚拟机取得大致相同的性能,采取 DVFS 操作降低功耗.当负载极低时,使物理机处于休眠状态,请求在网卡中暂存以降低更多功耗.但文献[68,69]所设计的模型限制其仅适用于虚拟机中运行相同应用的场景,大大简化了设计,实用性差.

4.1.3 小 结

通过以上分析可以看出,目前采取 DVS/DVFS 进行功耗控制的研究大多只针对同构,而目前广为流行的异构多核中如何采取技术是未来值得研究的工作.同时,由于负载的波峰、波谷到来时机存在差异,如何对运行异构应用的多台虚拟机采取 DVS/DVFS 进行功耗控制也是很具挑战性的工作.目前对此问题的研究比较匮乏.

DVS/DVFS 操作通过在负载低时降低 CPU 的频率或电压来降低功耗,仅降低了 CPU 的功耗.很低的电压和频率影响性能,在降低电压的同时,性能也呈线性下降.当处理器的使用率很低时,DVFS 不能明显降低漏电压,并且许多高性能处理器仅允许一定程度上的 DVFS 操作,很低的级别也提供比低负载所要求的更高速度,也即

DVFS 降低功耗的力度是有限的.例如,Intel Xeon 5360 处理器,从最高频率调到最低频率,功耗消耗也仅能从163M 降到 158M[70].

4.2 开/关机

据资料显示[6],服务器在空闲时功耗是其在满载状态下功耗的 60%,因此让更多物理机空闲关机能够在更大程度上降低功耗.通过开/关方式降低功耗的实现手段分两大类:(1) 资源分配.在资源池端,根据负载动态为应用加入或移除服务器[71−73].(2) 负载分发.在负载分发器端通过负载分发尽可能让负载集中,以使某些物理机负载尽快排空,从而关机达到降低功耗的目的[74,75].

4.2.1 资源分配

这种方式下,要求系统具有动态调整服务器分配的能力,根据应用负载匹配资源需求,从而提高能量有效性.早期知名工作是 Duke 大学[71]设计的 Muse 系统,以整个物理机为单位进行动态资源分配.通过比较服务器满足应用性能带来的收益和增加的功耗成本,构建关于利润的效用函数,以此来决定系统需要的活动服务器数量,但这项工作基于整个集群结点同构的假设条件.Rutgers 大学 Heath[72,73]考虑异构集群的场景.

4.2.2 负载分发

数据中心中同一服务存在多个副本,通常前端存在负载分发器,根据设定的负载分发策略,把请求转发到相应结点提供服务.与通过负载分发实现负载均衡的思路相反,把负载尽量集中到少数服务器,尽可能使更多服务器排空负载达到关机条件[74,75].IBM 的 Rajamani 等人[74]设计了 PARD 系统,监测服务器的活动连接数,尽可能地把负载发送到活动连接数多的服务器,直至达到饱和.但所设计的策略仅适应于短时间段的请求响应式服务,而面向 TCP 长连接应用具有独特性,有的连接生命周期可能几天甚至更长,而有的可能就短短几分钟,如 MSN 等服务,单纯通过活动连接数不能进行很好的评判,Microsoft Research 的 Rigas[75]针对长连接展开研究.

4.2.3 小 结

通过关机方式降低功耗是一种很直接、最有效降低功耗的技术手段,降低功耗的幅度也最大.但是服务器在关机后不能服务任何请求,需要花费一段时间重启使其能响应服务,时间开销很大.

4.3 服务器整合

据统计[6],数据中心服务器利用率平均在 30%左右,为服务器整合提供现实需求.不同应用负载的波峰、波谷到来时机存在差异为服务器整合提供现实可能,虚拟机迁移机制为服务器整合提供技术支持.采取有效的虚拟机整合策略可以把负载的波峰、波谷到来时机不同的虚拟机进行整合,更有效地利用物理机资源,尽可能空闲更多的物理机,使其处于休眠或关机状态,达到降低功耗的目的.

从 2007 年开始,利用这种技术进行降低功耗的研究如火如荼地展开[76−85].目前相关研究分两大类:(1) 服务器整合评价系统;(2) 服务器整合系统.

4.3.1 服务器整合评价系统

借助虚拟机迁移技术,进行服务器整合,从而达到降低功耗的目的,是随着虚拟化技术再一次兴起的一种新型功耗控制手段,目前有一些研究对服务器整合进行定量或定性分析[76−78].Michigan 大学 Padala 等人[76]对使用虚拟机迁移技术对服务器整合性能进行定性评价.中国科学院计算技术研究所 Song 等人[77]进行了定量分析,但没有考虑服务器整合过程中的一些限制条件,如可靠性、可用性、安全性等.针对此缺陷,美国 NEC 实验室的 Chen 所设计的 VCAE[78]能够帮助系统管理员有效处理服务器整合计划中的多数限制条件,实用性强.

4.3.2 服务器整合系统

服务器整合技术[79]是抑制数据中心能量消耗快速增长的关键手段之一,设计有效的服务器整合系统,需要对应用功耗消耗特征进行分析[80].服务器整合策略分析工作负载[81],考虑虚拟机迁移成本,提升系统健壮性[82].IBM 的 Verma[83]设计 pMapper 系统,虚拟机迁移策略使用首次适应装箱算法(first fit decreasing,简称 FFD)选择物理服务器以降低功耗,然而系统中所用方法仅适应于较长整合周期,以至于基于应用峰值资源使用率为其分配资源,严重浪费资源.Verma 等人[81]对 pMapper 的缺点进行改进,对工作负载进行分析,按需分配资源,尽可能

减少资源浪费.Georgia Institute 的 Jung 等人[82]设计开销敏感的自适应服务器整合驱动器,以运行时重新配置方案所带来的收益相对于虚拟机迁移所带来的开销进行服务器整合.

4.3.3 小 结

服务器整合通过使虚拟机尽量集中,空闲更多物理机达到降低功耗的目的,是虚拟化平台下的新兴手段[84,85].把负载的波峰、波谷到来存在差异的虚拟机整合到 1 台物理机,使物理机资源从时间和空间都得到更充分的利用,最终还是通过休眠或关机降低功耗,幅度大,但虚拟机迁移会对网络,源、目的物理主机产生一定的压力.

表 3 为功耗控制技术手段的对比.

Table 3 Power control technology comparison

表 3 功耗控制技术对比

控制手段	技术支持	调整范围	控制效果	时间开销	操作灵活性	相关研究
DVS/DVFS	硬件、操作系统	部件级	仅控制 CPU 功耗力度小	比开/关机小	好	文献[63−70]
开/关机	Wake-on-LAN	单机级	降低整机功耗效果显著	大	差	文献[71−75]
服务器整合	虚拟机迁移、Wake-on-LAN	集群级	降低整机功耗效果显著	大	一般	文献[76−85]

5 以同时满足 SLA 和降低功耗为目标的研究现状

降低数据中心功耗可以缩减电能和制冷成本,减少 CO_2 排放量缓解环境污染,降低由于系统过热造成服务器故障的概率,是亟待解决的问题.但单纯为降低功耗而使应用 SLA 受到影响,会使提供商的潜在收益减少.美国 Gomez 网络公司曾做过 2 000 个用户的调查,当一个网站响应速度很慢时,50%的用户会放弃访问而转向竞争对手的网站.在电子商务中,当用户的购买流程失败了 3 次以上时,94%的用户会选择永远放弃这个网站.而更严重的是,近五成的用户会将这种不好的体验告诉 5 个以上朋友或同事.最近有一些研究以同时满足 SLA 和降低功耗为共同目标进行资源管理研究[86−92].由于在虚拟化平台下出现虚拟机迁移新特性,我们把相关研究以平台划分为两大类:(1) 非虚拟化平台;(2) 虚拟化平台.

5.1.1 非虚拟化平台

在非虚拟化平台下,学者们使用动态资源分配结合 DVS/DVFS 或开/关机控制手段同时满足 SLA 和降低功耗目标.Rutgers 大学的 Pinheiro 是较早从事此研究的学者之一[93,94],使用服务器动态分配与负载分发相结合的方式,根据负载量大小,决定服务器的加入或移除,分阶段分别以满足 SLA 和降低功耗为主要目标.当负载量小时,通过负载分发尽量让负载集中以降低功耗.当负载量大时,通过负载分发尽量让负载均衡,以满足 SLA.以上的研究没有考虑操作成本.Penn State 大学 Chen 等人[95]设计的模型中显式考虑操作成本,并且使用服务器动态供应和 DVS 进行功耗控制和性能管理.IBM 的 Tesauro[86,96,97]把同时满足 SLA 和降低功耗的问题抽象为多目标优化问题.针对单纯使用强化学习会随着数据中心规模的扩大存在可扩展性问题[96],文献[97]使用文献[24]中所使用的混合强化学习方法解决此问题.

5.1.2 虚拟化平台

Drexel 大学的 Kusic 是较早从事虚拟化平台下功耗和性能权衡的研究人员之一,尤其是 2008 年的工作[87]设计动态资源供应和功耗管理框架,使用受限前向控制把资源供应问题视为不确定性顺序优化问题,然而并没有对功耗和性能提供显式保证,并且仅用仿真实验对效果进行验证.Tennessee 大学 Wang 等人[88]基于反馈控制论设计功耗控制环和性能控制环,协调性能管理和功耗控制,所设计的性能管理模型限制其仅适用于所有虚拟机中应用具有同类型的场景.为了弥补此缺陷,2010 年 Southern California 大学的 Urgaonkar 等人[89]针对异构应用场景设计复杂模型,使用排队论、优化理论对问题进行求解.

5.1.3 小 结

以上的研究或者只针对非虚拟化平台,或者只针对虚拟化平台,而现实情况是数据中心更倾向于采取物理机和虚拟机共存的混合平台.这是因为采用非虚拟化平台可以避免虚拟化带来的性能开销,在满足应用 SLA 方面具有优势,而虚拟化平台可以利用服务器整合技术更好地控制功耗.在这种混合平台下,需要对物理资源和虚

拟化的资源统一地进行组织.如何超越物理和虚拟化环境,对物理资源和虚拟化共存的环境进行统一管理是数据中心管理的一大难题.同时,虚拟化技术的多元化使得 IT 资源的管理更为棘手.

6 资源管理方法和相关工作总结

根据以上分析,资源管理领域的方法有以下几大类:控制论、优化理论、排队论、启发式算法、强化学习、模糊逻辑、运行时观测.表 4 对各种方法进行对比.

Table 4 Resource management comparison

表 4 资源管理方法对比

方法	优点	缺点	相关研究
控制论	不需要专业领域知识;稳定性、准确性高;适用于短时间范围内的操作	可扩展性差	文献[15−17,26,50−52,57,58,66,68,87,88,95,98]
排队论	请求到达符合排队论系统原理	需假设系统具有稳定,独立性;一般只能得到稳态解或数值;解难以反映资源需求的瞬时特性	文献[21−23,34,45,46,56,95,99]
优化理论	具有全局优化的特点;可扩展性好	得到优化解的时间长	文献[12,18,19,22,23,25,32,34,47,48,61,71,99]
强化学习	不需要专业领域知识;不需要建立显式的系统模型	需要对数据进行训练,可扩展性差	文献[23−25,86,96,97]
启发式算法	比较简单且容易实现	应用领域专属性较强.算法不保证稳定性	文献[38,40,42,44]
模糊逻辑	不需要系统以前运行的经验知识和对被管理系统的数学模型	可扩展性差	文献[56,100]
运行时观测	依赖运行时观测值,实现简单;适用于负载高可变的环境	准确性差,容易出现“事后”效应	文献[11,14,34,39,40,44,50,53,93]

表 5 对目前资源管理工作从所采取的平台、所要达到的目标、使用的方法、是 1 台物理机还是多台物理机之间的资源分配、资源类型、是否考虑迁移成本、重新分配时机这几个方面进行总结归纳.其中,独占表示基于物理机粒度的独占平台;进程代表基于进程粒度的共享平台;虚拟机代表基于虚拟机粒度的共享平台;CT(control theory)代表控制论;QT(queuing theory)代表排队论;OT(optimization theory)代表优化理论;RL(reinforcement learning)代表强化学习;HA(heuristic algorithm)代表启发式算法;FL(fuzzy logic)代表模糊逻辑;OM(observation monitor)代表运行观测;DR(desired resource utilization)代表期望的资源使用率;DT(desired response time)代表期望的响应时间;ON/OFF 代表开/关机;SC(server consolidation)代表服务器整合;SLA 代表以满足 SLA 为目标;Power 代表以降低功耗为目标;SLA&Power 代表以同时满足 SLA 和降低功耗为目标.

Table 5 Resource management summary

表 5 资源管理工作总结

工作	平台	目标	方法	1 台物理机	多台物理机之间	资源类型	是否考虑迁移成本	重新分配时机
2001, IBM, Oceano[11]	独占	SLA	OM		√			DT
2003, IBM, Walsh[12]	独占	SLA	OT		√			
2002, Duke, COD[20]	独占	SLA			√			
2002, Rice, QUID[21]	独占	SLA	QT		√	CPU		DR
2003, IBM, Levy[22]	独占	SLA	QT, OT		√			DT
2005, IBM, Tesauro[23]	独占	SLA	RL		√			
2006, IBM, Tesauro[24]	独占	SLA	QT, RL		√			
2007, IBM, Tesauro[25]	独占	SLA	OT, RL		√			
2006, France, Jade[26]	独占	SLA	CT		√			DR
2002, Massachusetts, Urgaonkar[28]	进程	SLA			√	CPU, Netwrok		
2004, Massachusetts, Urgaonkar[29]	进程	SLA			√	CPU, Network		
2006, IBM, Karve[30]	进程	SLA	OT		√	CPU, Memory		

Table 5 Resource management summary (Continue 1)

表 5 资源管理工作总结(续 1)

工作	平台	目标	方法	1 台物理机	多台物理机之间	资源类型	是否考虑迁移成本	重新分配时机
1999, Rice, Resource Container[31]	进程	SLA		√		CPU		
2000, Rice, Cluster Container[32]	进程	SLA	OT		√	CPU		
2003, IBM, MBRP[33]	进程	SLA	QT	√		Memory, Disk		
2002, Massachusetts, Chandra[14]	进程	SLA	OM	√		CPU		
2003, Massachusetts, Chandra[34]	进程	SLA	QT, OM, OT	√		CPU		
2006, Purdue, Khanna[37]	虚拟机	SLA	HA		√	CPU, Memory	√	DR
2007, IBM, Bobroff[38]	虚拟机	SLA	HA		√	CPU		DR
2008, VMware, DRS[39]	虚拟机	SLA	OM	√	√	CPU, Memory		DR
2007, Massachusetts, Sandpiper[40]	虚拟机	SLA	HA, OM		√			DR
2009, Massachusetts, Memory Buddies[41]	虚拟机	SLA	OM		√	Memory		
2005, Northwestern, Virtuoso[42]	虚拟机	SLA	HA		√			
2007, North Carolina, VIOLIN[44]	虚拟机	SLA	HA		√		√	DR
2007,清华大学,Wang[45]	虚拟机	SLA	QT, OT		√	CPU		
2008,清华大学,Wang[46]	虚拟机	SLA	QT, OT		√	CPU		
2009, France, Van[47]	虚拟机	SLA	OT		√	CPU, Memory	√	
2009, France, Van[48]	虚拟机	SLA	OT		√	CPU, Memory	√	
2008, HP, Zhu[50]	虚拟机	SLA	CT, OT	√	√	CPU		DR
2005, Illinois, Liu[51]	虚拟机	SLA	CT	√		CPU		DT
2005, HP, Wang[52]	虚拟机	SLA	CT	√		CPU		DR, DT
2006, HP, Rolia[53]	虚拟机	SLA	OM	√		CPU		DR
2007, HP, Liu[15]	虚拟机	SLA	CT	√		CPU		
2007, Michigan, Padala[16]	虚拟机	SLA	CT	√		CPU		DR
2009, Michigan, Padala[17]	虚拟机	SLA	CT	√		CPU, Disk		DR
2008,中国科学院计算技术研究所,Song[54]	虚拟机	SLA	OT	√		CPU, Memory		DR
2008, Florida, Xu[18]	虚拟机	SLA	FL	√		CPU		DR
2009, Illinois, Heo[55]	虚拟机	SLA	CT	√		CPU, Memory		DR
2009, HP, AppRAISE[56]	虚拟机	SLA	QT, CT	√		CPU		DR, DT
2009, Lund, Kjaer[57]	虚拟机	SLA	CT	√		CPU		DR, DT
2009,中国科学院计算技术研究所,Song[59]	虚拟机	SLA	OT		√	CPU, Memory		DR
2006, George Mason, Menasc[19]	虚拟机	SLA	OT	√		CPU		DT
2009, IBM, Zhang[61]	虚拟机	SLA	OT	√		CPU		
2003, IBM, Elnozahy[63]	独占	Power	DVFS	√		CPU		
2007, Virginia, Horvath[64]	独占	Power	DVFS, OT		√	CPU		
2007, Virginia, Horvath[65]	独占	Power	DVFS, OT		√	CPU		
2010,中国科学院计算技术研究所,Lin[66]	独占	Power	DVFS, CT		√	CPU		
2008, Tennesse, Wang[67]	独占	Power	DVFS, CT		√	CPU		
2008, Tennesse, Wang[68]	虚拟机	Power	DVFS, CT	√		CPU		
2009, Tennesse, Wang[69]	虚拟机	Power	DVFS, ON/OFF, CT	√		CPU		
2001, Duke, Muse[71]	独占	Power	ON/OFF, OT		√			
2003, Rutgers, Heath[72]	独占	Power	ON/OFF		√			
2005, Rutgers, Heath[73]	独占	Power	ON/OFF		√			
2003, IBM, PARD[74]	独占	Power	ON/OFF		√			
2008, Microsoft, Rigas[75]	独占	Power	ON/OFF		√			

Table 5 Resource management summary (Continue 2)

表 5 资源管理工作总结(续 2)

工作	平台	目标	方法	1 台物理机	多台物理机之间	资源类型	是否考虑迁移成本	重新分配时机
2009, Georgia Institute, Jung[82]	虚拟机	Power	SC		√		√	
2008, Pennsylvania, Choi[80]	虚拟机	Power	SC		√			
2008, IBM, Verma[81]	虚拟机	Power	SC		√			
2008, IBM, Verma[83]	虚拟机	Power	SC		√		√	
2010, IBM, Hanson[85]	虚拟机	Power	DVFS, SC	√	√	CPU	√	
2001, Rutgers, Pinheiro[93]	独占	SLA&Power	OM, ON/OFF		√			
2009, IBM, Korupolu[91]	虚拟机	SLA&Power	SC		√			
2001, Rutgers, Pinheiro[85]	独占	SLA&Power	OM, CT, ON/OFF		√			
2005, Penn State, Chen[95]	独占	SLA&Power	CT, QT, DVFS, ON/OFF	√	√	CPU	√	
2007, IBM, Tesauro[96]	独占	SLA&Power	OT, RL, DVFS	√	√	CPU		
2007, IBM, Tesauro[97]	独占	SLA&Power	CT, OT, RL, DVFS	√	√	CPU		
2008, IBM, Tesauro[86]	独占	SLA&Power	CT, OT, RL, DVFS, ON/OFF	√	√	CPU		
2008, Drexel, Kusic[87]	虚拟机	SLA&Power	OT, SC		√		√	
2010, Southern California, Urgaonkar[89]	虚拟机	SLA&Power	QT, OT, DVFS, SC		√	CPU		
2009, Tennessee, Wang[90]	虚拟机	SLA&Power	OT, DVFS	√	√	CPU		

7 总结与展望

Internet 应用的飞速发展使得越来越多的服务提供商采取租赁数据中心的资源对外提供服务,需求驱动数据中心的规模越来越大,越来越复杂.其中资源管理问题首当其冲.如何同时满足高并发应用 SLAs 的兼容性和提高系统服务的能量有效性是当前数据中心资源管理者面临的两大挑战,而 Internet 应用负载的随时变动性更加剧了以上挑战的解决.针对以上挑战,学者们展开如火如荼的研究.本文对近十几年解决以上挑战的相关资源管理研究发展脉络进行梳理,按照要实现的目标分 3 大类对研究现状进行综述:(1) 以满足 SLA 为目标的研究工作分析与总结;(2) 以降低功耗为目标的研究工作分析与总结;(3) 以同时满足 SLA 和降低功耗为目标的研究工作分析与总结.

综合以上面向 Internet 数据中心资源管理工作的研究进展,我们认为,未来的研究趋势主要有:

(1) 虚拟化平台下细粒度资源分配与虚拟机迁移间的权衡及结合

细粒度资源分配[37,40−49,76,99,101,102]和虚拟机迁移[15−19,51−58,61,62,98,100]是虚拟化平台下两种资源分配技术,然而目前两种技术大多单独使用.细粒度资源分配具有灵活性高、开销小的特点,但资源调整力度有限.单纯使用细粒度资源分配无法解决资源发生竞争,从而造成 SLA 受影响的场景,即使存在物理机资源空闲也无法充分利用这些非本地的空闲资源.可见,局部调整并不能保证全局优化.虚拟机迁移技术资源调整力度大,可进行全局资源优化分配,但迁移过程会对网络带宽与源物理主机、目的主机的 CPU 产生一定压力,迁移过程中对虚拟机上所运行应用的 SLA 也会产生一定的影响.而且,当前的虚拟机迁移技术还存在各种限制,如要求源物理主机与目的物理主机需要具有一致的 CPU 架构.为尽量发挥两种技术的优势,并弥补各自的不足,我们认为,细粒度资源分配与虚拟机迁移相结合是数据中心资源管理研究的必然趋势,细粒度资源分配属于局部轻量级调整范畴,而虚拟机迁移则属于全局重量级调整范畴.若能在应用负载波动剧烈时采用轻量级的细粒度资源分配来保证应用的 SLA,避免重量级的虚拟机迁移带来的开销和延迟,在局部资源无法满足 SLA 时结合虚拟机迁移,则能实现数据中心全局资源优化的目的.因此,两种技术之间如何权衡及结合从而更好地满足应用 SLAs 和高资源利用率是未来面临的挑战之一.

(2) 虚拟化环境下负载分发与细粒度资源分配的有效匹配

数据中心中对同一应用往往存在多个副本同时提供服务,对于到达的请求负载需要根据位于前端负载分发器上的负载分发策略选择合适的结点提供服务.在传统物理机环境下,物理机的容量除非手工进行 scale up 或 scale down,否则资源容量在运行过程中是不变的.而在虚拟化环境下,由于具有细粒度资源动态分配机制支持,虚拟机的资源容量在运行过程中可动态变化,传统针对物理机环境下的负载分发策略在虚拟化环境下已不能很好地适用,原因在于,传统针对物理机环境下的负载分发策略均没有考虑结点容量动态改变的因素.虚拟化技术的引入,为负载分发策略提出新的需求,也为这一传统技术提供了新的发展机遇和生机.如何使负载分发策略感知虚拟机结点容量的变化,从而设计新的负载分发策略,适应虚拟结点容量动态变化的场景?如何将负载分发与细粒度资源分配进行有效匹配,避免出现诸如某虚拟结点容量增加,其分得的负载相对减小;或某些虚拟结点由于容量不断增加,其分得负载相应不断增加,而另外的虚拟结点容量不断减小,负载也不断减小,最终形成虚拟节点容量和负载极端不平衡的现象.两者之间的策略如何匹配才能更好地满足并发多应用 SLAs 的兼容性,减小 SLAs 受影响的概率是未来需要解决的重要问题之一.

(3) 异构多核架构下的功耗控制

目前 DVS/DVFS 功耗控制工作大多针对同构平台,现实中由于 FPGA,GPU 的发展,很多厂商采用异构多核技术.在这种异构多核平台下,根据应用负载调整各部件的频率和电压以提高系统的能量有效性是很困难的.因为不同部件硬件制造工艺存在差别,具有不同的频率,即使 CPU 之间也可能具有不同频率,加剧了功耗控制的难度.如目前大多数工作[68,69]采用同步调整 CPU 频率的方法,虽然已有部分工作开始考虑单独调整每个 CPU 的频率,但它们没有考虑在 FPGA、GPU 同时存在的异构多核平台下的功耗控制问题.并且 GPU 的低功耗研究还处于起步阶段,现代高性能 GPU 中除了显示芯片外,往往还包含一个大容量的存储器,这意味着必须将存储器的功耗控制纳入 GPU 的功耗优化范围才能取得较好的优化效果,单是 GPU 的功耗控制就很复杂,部分或同时考虑 GPU,FPGA,CPU 的功耗控制将变得更加困难.如任务的完成需要 CPU 和 GPU 协作完成,若根据 GPU 当前任务量降低其频率,而 CPU 需要等待 GPU 计算的结果才能进行下一步工作,则会导致整个任务执行时间变长,有可能得不偿失.因此,异构多核架构如何提高系统能量有效性是未来的重要挑战之一.

(4) 混和异构平台下同时满足 SLA 和降低功耗的资源管理

从以上的分析可以看出,目前的资源管理工作或者单独针对非虚拟化平台、或者单独针对虚拟化平台展开研究,而现实中由于性能和功耗的双重需求[103,104],有些数据中心采取物理机和虚拟机共存的混合平台.采用非虚拟化平台可以避免虚拟化带来的性能开销,在满足应用 SLA 方面具有优势,而虚拟化平台可以利用服务器整合技术更好地控制功耗.在这种大规模异构数据中心中,如何对非虚拟化资源、虚拟化资源以统一的标准进行衡量?如何解决应用与资源的映射?不同应用之间共享资源,存在干扰.如两个 I/O 密集型的应用整合到一台物理机上时,其性能下降可超过 10 倍[105].两类应用对于资源的需求总和往往不能进行简单的相加,上下文切换、存储介质访问规律的打乱都会使应用性能受到显著影响.在混合异构平台下,这种干扰的判定更加困难.为此,在混合异构平台下分析应用共存特性,并基于此设计实现同时满足应用 SLA 和降低功耗目标的应用到资源映射更加困难,这是当前大规模数据中心亟需解决的关键问题,也是当前资源管理面临的满足并发多应用 SLAs 的兼容性和提高系统服务的能量有效性两大挑战的具体落实点之一.

References:

[1] http://datacenter.ctocio.com.cn/133/8477133.shtml

[2] http://book.51cto.com/art/201104/253795.html

[3] http://cloud.yesky.com/cloud2011/

[4] http://www.datacenterknowledge.com/

[5] http://www.idcun.com/news/2010080614356.html

[6] McKinsey & Company. Revolutionizing data center efficiency. 2008. http://uptimeinstitute.org

[7] http://tech.watchstor.com/Data-Center-132873_1.htm

[8] Bhatti N, Bouch A, Kuchinsky A. Integrating user-perceived quality into Web server design. In: Proc. of the 9th Int'l World Wide Web Conf. on Computer Networks. Amsterdam, 2000. 1−16. [doi: 10.1016/S1389-1286(00)00087-6]

[9] http://www.greenit-pc.jp/activity/symposium/110223/pdf/bk1_taketani.pdf

[10] http://www.searchdatacenter.com.cn/showcontent_33068.htm.

[11] Appleby K, Fakhouri S, Fong L, Goldszmidt G, Kalantar M, Krishnakumar S, Pazel DP, Pershing J, Rochwerger B. Oceano-SLA based management of a computing utility. In: Proc. of the 7th IFIP/IEEE Int'l Symp. on Integrated Network Management. 2001. [doi: 10.1109/INM.2001.918085]

[12] Walsh WE, Tesauro G, Kephart JO, Das R. Utility functions in autonomic systems. In: Proc. of the 1st Int'l Conf. on Autonomic Computing. IEEE Computer Society, 2004. 70−77. [doi: 10.1109/ICAC.2004.1301349]

[13] Bennani MN, Menasce DA. Resource allocation for autonomic data centers using analytic performance models. In: Proc. of the 2nd Int'l Conf. on Autonomic Computing (ICAC 2005). Washington, 2005. 229−240. [doi: 10.1109/ICAC.2005.50]

[14] Pradhan P, Tewari R, Sahu S, Chandra A, Shenoy P. An observation-based approach towards self-managing Web servers. In: Proc. of the Int'l Workshop on Quality of Service. Miami Beach, 2002. [doi: 10.1109/IWQoS.2002.1006570]

[15] Liu X, Zhu XY, Padala P, Wang ZK, Singhal S. Optimal multivariate control for differentiated services on a shared hosting platform. In: Proc. of the IEEE Conf. on Decision and Control (CDC). 2007. [doi: 10.1109/CDC.2007.4434560]

[16] Padala P, Shin KG, Zhu XY, Uysal M, Wang ZK, Singhal S, Merchant A, Salem K. Adaptive control of virutalized resources in utility computing environments. In: Proc. of the EuroSys. 2007. [doi: 10.1145/1272996.1273026]

[17] Padala P, Hou KY, Shin KG, Zhu XY, Uysal M, Wang ZK, Singhal S, Merchant A. Automated control of multiple virtualized resources. In: Proc. of the EuroSys 2009. Nuremberg, 2009. [doi: 10.1145/1519065.1519068]

[18] Xu J, Zhao M, Fortes J, Carpenter R, Yousif M. Autonomic resource management in virtualized data centers using fuzzy logic-based approaches. Cluster Computing, 2008,11(3):213−227. [doi: 10.1007/s10586-008-0060-0]

[19] Menasce DA, Bennani MN. Autonomic virtualized environments. In: Proc. of the IEEE ICAS 2006. 2006. [doi: 10.1109/ICAS.2006.13]

[20] Moore J, Irwin D, Grit L, Sprenkle S, Chase J. Managing mixed-use clusters with cluster-on-demand, cluster-on-demand draft. Internet Systems and Storage Group, Duke University, 2002.

[21] Ranjan S, Rolia J, Fu H, Knightly E. QoS-Driven server migration for Internet data centers. In: Proc. of the IEEE IWQoS 2002. [doi: 10.1109/IWQoS.2002.1006569]

[22] Levy R, Nagarajarao J, Pacici G, Spreitzer A, Tantawi A, Youssef A. Performance management for cluster-based Web services. In: Proc. of the 8th Int'l Symp. on Integrated Network Management (IM 2003). 2003. 247−261. [doi: 10.1109/INM.2003.1194184]

[23] Tesauro G. Online resource allocation using decompositional reinforcement learning. In: Proc. of the AAAI-2005. 2005 .886−891.

[24] Tesauro G, Jong NK, Das R, Bennani MN. A hybrid reinforcement learning approach to autonomic resource allocation. In: Proc. of the ICAC-2006. 2006. 65−73. [doi: 10.1109/ICAC.2006.1662383]

[25] Tesauro G, Jong NK, Das R, Bennani MN. On the use of hybrid reinforcement learning for autonomic resource allocation. Clustet Computing, 2007,10(3):287−299. [doi: 10.1007/s10586-007-0035-6]

[26] Bouchenak S, de Palma N, Hagimont D, Taton C. Autonomic management of clustered applications. In: Proc. of the IEEE Int'l Conf. on Cluster Computing (Cluster 2006). Barcelona, 2006. [doi: 10.1109/CLUSTR.2006.311842]

[27] Menon A, Santos JR, Turner Y Janakiraman G, Zwaenepoel W. Diagnosing performance overheads in the Xen virtual machine environment. In: Proc. of the 1st ACM/USENIX Int'l Conf. on Virtual Execution Environments (VEE). 2005. 13−23. [doi: 10.1145/1064979.1064984]

[28] Urgaonkar B, Shenoy P, Roscoe T. Resource overbooking and application profiling in shared hosting platforms. In: Proc. of the 5th Symp. on Operating Systems Design and Implementation (OSDI). 2002. 239−254. [doi: 10.1145/844128.844151]

[29] Urgaonkar B, Shenoy P. Share: Managing CPU and network bandwidth in shared clusters. IEEE Trans. on Parallel and Distributed Ststems, 2004,15(1):2−17. [doi: 10.1109/TPDS.2004.1264781]

[30] Karve A, Kimbrel T, Pacifici G, Spreitzer M, Steinder M, Sviridenko M, Tantawi A. Dynamic placement for clustered Web applications. In: Proc. of the 15th Int'l World Wide Web Conf. (WWW). 2006. [doi: 10.1145/1135777.1135865]

[31] Banga G, Druschel P, Mogul JC. Resource containers: A new facility for resource management in server systems. In: Proc. of the 3rd Symp. on Operating Systems Design and Implementation (OSDI'99). New Orleans, 1999. 45−58. [doi: 10.1145/224057.225831]

[32] Aron M, Druschel P, Zwaenepoel W. Cluster reserves: A mechanism for resource management in cluster-based network servers. In: Proc. of the Joint Int'l Conf. on Measurement and Modeling of Computer Systems (ACM SIGMETRICS 2000). 2000. 90−101. [doi: 10.1145/339331.339383]

[33] Doyle RP, Chase JS, Asad OM, Jin W, Vahdat A. Model-Based resource provisioning in a Web service utility. In: Proc. of the USITS 2003. 2003.

[34] Chandra A, Gong WB, Shenoy PJ. Dynamic resource allocation for shared data centers using online measurements. In: Proc. of the IWQoS 2003. 2003. 381−400. [doi: 10.1145/781027.781067]

[35] Chandra A, Goyal P, Shenoy P. Quantifying the benefits of resource multiplexing in on-demand data centers. In: Proc. of the 1st Workshop on Algorithms and Architectures for Self-Managing Systems. 2003.

[36] Clark C, Fraser K, Hand S, Hansen JG, Jul E, Limpach C, Pratt I, Warfield A. Live migration of virtual machines. In: Proc. of the 2nd Conf. on Symp. on Networked Systems Design & Implementation. Berkeley: USENIX Association, 2005.

[37] Khanna G, Beaty K, Kar G, Kochut A. Application performance management in virtualized server environments. In: Proc. of the IEEE Network Opertaions and Management Symp. 2006. 373−381. [doi: 10.1109/NOMS.2006.1687567]

[38] Bobroff N, Kochut A, Beaty K. Dynamic placement of virtual machines for managing SLA violations. In: Proc. of the 10th IFIP/IEEE Int'l Symp. on Integrated Network Management. 2007. 119−128. [doi: 10.1109/INM.2007.374776]

[39] Vmware. VMware dynamic resource scheduler. 2008. http://www.vmware.com/products/vi/vc/drs.html

[40] Wood T, Shenoy P, Venkataramani A, Yousif M. Black-Box and gray-box strategies for virtual machine migration. In: Proc. of the 4th USENIX Symp. on Networked Systems Design and Implementation. Cambridge, 2007. 229−242.

[41] Wood T, Tarasuk-Levin G, Shenoy P, Peter D, Cecchet E, Corner MD. Memory buddies: Exploiting page sharing for smart colocation in virtualized data centers. In: Proc. of the ACM SIGPLAN/SIGOPS Int'l Conf. on Virtual Execution Environments. 2009. 27−36. [doi: 10.1145/1508293.1508299]

[42] Sundararaj AI, Gupta A, Dinda PA. Increasing application performance in virtual environments through run-time inference and adaptation. In: Proc. of the 14th IEEE Int'l Symp. on High Performance Distributed Computing (HPDC 2005). Research Triangle Park, 2005. 47−58. [doi: 10.1109/HPDC.2005.1520935]

[43] Shrivastava V, Zerfos P, Lee KW, Jamjoom H, Liu YH, Banerjee S. Application-Aware virtual machine migration in data centers. In: Proc. of the IEEE INFOCOM 2011. 2001. 66−70. [doi: 10.1109/INFCOM.2011.5935247]

[44] Ruth P, Rhee J, Xu DY, Kennell R, Goasguen S. Autonomic live adaptation of virtual computational environments in a multi-domain infrastructure. In: Proc. of the IEEE ICAC. 2007. [doi: 10.1109/ICAC.2006.1662376]

[45] Wang XY, Lan DJ, Wang J, Fang X, Ye M, Chen Y, Wang QB. Appliance-Based autonomic provisioning framework for virtualized outsourcing data center. In: Proc. of the ICAC 2007. 2007. [doi: 10.1109/ICAC.2007.6]

[46] Wang XY, Du ZH, Chen YN, Li SL. Virtualization-Based autonomic resource management for multi-tier Web applications in shared data center. The Journal of Systems and Software, 2008,81(9):1591−1608. [doi: 10.1016/j.jss.2007.11.719]

[47] Van HN, Tran FD, Menaud JM. Autonomic virtual resource management for service hosting platforms. In: Proc. of the Int'l Conf. on Software Enginnering Workshop on Software Engineering Challenges of Cloud Computing. 2009. 1−8. [doi: 10.1109/CLOUD.2009.5071526]

[48] Van HN, Tran FD, Menaud JM. SLA-Aware virtual resource management for cloud infrastructures. In: Proc. of the IEEE 9th Int'l Conf. on Computer and Information Technology. 2009. [doi: 10.1109/CIT.2009.109]

[49] Meng XQ, Pappas V, Zhang L. Improving the scalability of data center networks with traffic-aware virtual machine placement. In: Proc. of the INFOCOM. 2010. [doi: 10.1109/INFCOM.2010.5461930]

[50] Zhu XY, Young D, Watson BJ, Wang ZK, Rolia J, Singhal S, McKee B, Hyser C, Gmach D, Gardner R, Christian T, Cherkasova L. 1000 islands: Integrated capacity and workload management for the next generation data center. In: Proc. of the 5th Int'l Conf. on Autonomic Computing. Chicago, 2008. [doi: 10.1109/ICAC.2008.32]

[51] Liu X, Zhu X, Singhal S, Arlitt M. Adaptive entitlement control of resource containers on shared servers. In: Proc. of the IFIP/IEEE Int'l Symp. on Integrated Network Management. 2005. [doi: 10.1109/INM.2005.1440783]

[52] Wang ZK, Zhu XY, Singhal S. Utilization and slo-based control for dynamic sizing of resource partitions. Technical Report, HPL-2005-126R1, Hewlett Packard Laboratories, 2005. [doi: 10.1007/11568285_12]

[53] Rolia J, Cherkasova L, Clifford MC. Configuring workload manager control parameters for resource pools. In: Proc. of the 10th IEEE/IFIP Network Operations and Management Symp. 2006. 127−137. [doi: 10.1109/NOMS.2006.1687545]

[54] Song Y, Li YQ, Wang H, Zhang YF, Feng BQ, Zang HY, Sun YZ. A service-oriented priority-based resource scheduling scheme for virtualized utility computing. In: Proc. of the HiPC 2008. 2008. 220−231. [doi: 10.1007/978-3-540-89894-8_22]

[55] Heo J, Zhu XY, Padala P, Wang ZK. Memory overbooking and dynamic control of Xen virtual machines in consolidated environments. In: Proc. of the 11th IFIP/IEEE Int'l Conf. on Symp. on Integrated Network Management (IM 2009). Piscataway: IEEE Press, 2009. 630−637. [doi: 10.1109/INM.2009.5188871]

[56] Wang ZK, Chen Y, Gmach D, Singhal S, Watson BJ, Rivera W, Zhu XY, Hyser CD. Appraise: Application-Level performance management in virtualized server environments. IEEE Trans. on Networking and Service Management, 2009,6(4):240−254. [doi: 10.1109/TNSM.2009.04.090404]

[57] Kjaer MA, Kihl M, Robertsson A. Resource allocation and disturbance rejection in Web servers using slas and virtualized servers. IEEE Trans. on Network and Service Management, 2009,6(4):226−239. [doi: 10.1109/TNSM.2009.04.090403]

[58] Kim D, Kim H, Jeon M, Seo E, Lee J. Guest-Aware priority-based virtual machine scheduling for highly consolidated server. In: Proc. of the Euro-Par. 2008. 285−294. [doi: 10.1007/978-3-540-85451-7_31]

[59] Song Y, Wang H, Li YQ, Feng BQ, Sun YZ: Multi-Tiered on-demand resource scheduling for vm-based data center. In: Proc. of the CCGRID. 2009. 148−155. [doi: 10.1109/CCGRID.2009.11]

[60] Rodríguez F, Freitag F, Navarro L. Towards intelligent management in VM-based resource providers. In: Proc. of the 1st Int'l DMTF Academic Alliance Workshop on Systems and Virtualization Management: Standards and New Technologies. Toulouse, 2007.

[61] Ardagna D, Mirandola R, Trubian M, Zhang L. Run-Time resource management in SOA virtualized environments. In: Proc. of the QUASOSS 2009. 2009. [doi: 10.1145/1596473.1596484]

[62] Sangpetch A, Turner A, Kim H. How to tame your vms: An automated control system for virtualized services. In: Proc. of the 24th Int'l Conf. on Large Installation System Administration, LISA 2010. Berkeley: USENIX Association, 2010. 1−16.

[63] Elnozahy M, Kistler M, Rajamony R. Energy conservation policies for Web servers. In: Proc. of the 4th USENIX Symp. on Internet Technologies and Systems. 2003.

[64] Horvath T, Abdelzaher T, Skadron K, Liu X. Dynamic voltage scaling in multitier Web servers with end-to-end delay control. IEEE Trans. on Computers, 2007,56(4):444−458. [doi: 10.1109/TC.2007.1003]

[65] Horvath T, Skadron K, Abdelzaher T. Enhancing energy efficiency in multi-tier Web server clusters via prioritization. In: Proc. of the 2007 IEEE Parallel and Distributed Processing Symp. (IPDPS 2007). 2007. 1−6. [doi: 10.1109/IPDPS.2007.370509]

[66] Yuan L. PowerTracer, tracing requests in multi-tier services to save cluster power consumption. Technical Report, 2010.

[67] Wang XR, Chen M. Cluster-Level feedback power control for performance optimization. In: Proc. of the HPCA. 2008. [doi: 10.1109/HPCA.2008.4658631]

[68] Wang YF, Wang XR, Chen M, Zhu XY. Power-Efficient response time guarantees for virtualized enterprise servers. In: Proc. of the RTSS. 2008. [doi: 10.1109/RTSS.2008.20]

[69] Wang YF, Deaver R, Wang XR. Virtual batching: Request batching for energy conservation in virtualized servers. In: Proc. of the IWQoS. 2010. [doi: 10.1109/IWQoS.2010.5542736]

[70] Flautner K, Reinhardt S, Mudge T. Automatic performance setting for dynamic voltage scaling. In: Proc. of the MOBICOM. 2001. [doi: 10.1023/A:1016546330128]

[71] Chase JS, Anderson DC, Thackar PN, Vahdat AM, Doyle RP. Managing energy and server resources in hosting centers. In: Proc. of the 18th Symp. on Operating Systems Principles. 2001. [doi: 10.1145/502034.502045]

[72] Heath T, Diniz B, Carrera EV, Meira W Jr, Bianchini R. Self-Configuring heterogeneous server clusters. In: Proc. of the Workshop on Compilers and Operating Systems for Low Power. 2003.

[73] Heath T, Diniz B, Carrera EV, Jr Meira W, Bianchini R. Energy conservation in heterogeneous server clusters. In: Proc. of the ACM SIGPLAN Symp. on Principles and Practice of Parallel Programming (PPoPP). 2005. [doi: 10.1145/1065944.1065969]

[74] Rajamani K, Lefurgy C. On evaluating request-distribution schemes for saving energy in server clusters. In: Proc. of the IEEE Int'l Symp. on Performance Analysis of Systems and Software. 2003. [doi: 10.1109/ISPASS.2003.1190238]

[75] Chen G, He WB, Liu J, Nath S, Rigas L, Xiao L, Zhao F. Energy-Aware server provisioning and load dispatching for connection-intensive Internet services. In: Proc. of the NSDI. 2008.

[76] Padala P, Zhu XY, Wang ZK, Singhal S, Shin KG. Performance evaluation of virtualization technologies for server consolidation. Technical Report, HPL-2007-59, HP Labs, 2007.

[77] Song Y, Zhang YW, Sun YZ, Shi WS. Utility analysis for Internet-oriented server consolidation in VM-based data centers. In: Proc. of the CLUSTER 2009. 2009. 1−10. [doi: 10.1109/CLUSTR.2009.5289190]

[78] Chen HF, Huang H, Jiang GF, Yoshihira K, Saxena A. VCAE: A virtualization and consolidation analysis engine for large scale data centers. In: Proc. of the 4th IEEE Int'l Conf. on Self-Adaptive and Self-Organizing Systems (SASO 2010). Budapest, 2010. [doi: 10.1109/SASO.2010.25]

[79] Srikantaiah S, Kansal A, Zhao F. Energy aware consolidation for cloud computing. In: Proc. of the USENIX Workshop on Power Aware Computing and Systems. 2008.

[80] Choi J, Govindan S, Urgaonkar B, Sivasubramaniam A. Profiling, prediction, and capping of power consumption in consolidated environments. In: Miller EL, Williamson CL, eds. Proc. of the MASCOTS. IEEE Computer Society, 2008. 3−12. [doi: 10.1109/MASCOT.2008.4770558]

[81] Verma A, Dasgupta G, Nayak TK, De P, Kothari R. Server workload analysis for power minimization using consolidation. In: Proc. of the 2009 Usenix ATC. 2009.

[82] Jung GY, Joshi KR, Hiltunen MA, Schlichting RD, Pu C. A cost-sensitive adaptation engine for server consolidation of multitier applications. In: Middleware 2009: Proc. of the 10th ACM/IFIP/USENIX Int'l Conf. on Middleware. 2009. 163−183. [doi: 10.1007/978-3-642-10445-9_9]

[83] Verma A, Ahuja P, Neogi A. pMapper: Power and migration cost aware application placement in virtualized systems. In: Proc. of the ACM/IFIP/Usenix Middleware. 2008. 243−264. [doi: 10.1007/978-3-540-89856-6_13]

[84] Kumar S, Talwar V, Kumar V, Ranganathan P, Schwan K. vManage: Loosely coupled platform and virtualization management in data centers. In: Proc. of the IEEE Intl. Conf. on Autonomic Computing (ICAC 2009). 2009. [doi: 10.1145/1555228.1555262]

[85] Hanson JE, Whalley I, Steinder M, Kephart JO. Multi-Aspect hardware management in enterprise server consolidation. In: Proc. of the Network Operations and Management Symp. (NOMS). 2010. 543−550. [doi: 10.1109/NOMS.2010.5488460]

[86] Das R, Kephart JO, Lefurgy C, Tesauro G, Levine DW, Chan H. Autonomic multi-agent management of power and performance in data centers. 2008.

[87] Kusic D, Kephart JO, Hanson JE, Kandasamy N, Jiang GF. Power and performance management of virtualized computing environments via lookahead control. In: Proc. of the ICAC. 2008. 1−15. [doi: 10.1007/s10586-008-0070-y]

[88] Wang XR, Wang YF. Co-Con: Coordinated control of power and application performance for virtualized server clusters. In: Proc. of the 17th IEEE Int'l Workshop on Quality of Service (IWQoS). Charleston, 2009. [doi: 10.1109/IWQoS.2009.5201388]

[89] Urgaonkar R, Kozat UC, Igarashi K, Neely MJ. Dynamic resource allocation and power management in virtualized data centers. In: Proc. of the IEEE/IFIP NOMS. 2010. [doi: 10.1109/NOMS.2010.5488484]

[90] Vassala CP, Tanelli M, Lovera M. Dynamic trade-off analysis of QoS and energy saving in admission control for Web service Systems. In: Proc. of the 4th Int'l ICST Conf. on Performance Evaluation Methodologies and Tools 2009. 2009. [doi: 10.4108/ICST.VALUETOOLS2009.7941]

[91] Cardosa M, Korupolu MR, Singh A. Shares and utilities based power consolidation in virtualized server environments. In: Proc. of the IFIP/IEEE Integrated Network Management (IM). 2009. 327−334. [doi: 10.1109/INM.2009.5188832]

[92] Steinder M, Whalley I, Hanson JE, Kephart JO. Coordinated management of power usage and runtime performance. In: Proc. of the Network Operations and Management Symp. (NOMS 2008). 2008. 387−394. [doi: 10.1109/NOMS.2008.4575159]

[93] Pinheiro E, Bianchini R, Carrera EV, Heath T. Load balancing and unbalancing for power and performance in cluster-based systems. In: Proc. of the Workshop on Compilers and Operating Systems for Low Power. 2001.

[94] Pinheiro E, Bianchini R, Carrera EV, Heath T. Dynamic cluster reconfiguration for power and performance. In: Compilers and Operating Systems for Low Power. Kluwer Academic Publishers, 2003.

[95] Chen YY, Das A, Qin WB, Sivasubramaniam A, Wang Q, Gautam N. Managing server energy and operational costs in hosting centers. In: Proc. of the Sigmetrics. 2005. 303−314. [doi: 10.1145/1064212.1064253]

[96] Kephart JO, Chan H, Das R, Levine DW, Tesauro G, Rawson F, Lefurgy C. Coordinating multiple autonomic managers to achieve specified power-performance tradeoffs. In: Proc. of the ICAC. 2007. [doi: 10.1109/ICAC.2007.12]

[97] Tesauro G, Das R, Chan H, Kephart JO, Lefurgy C, Levine DW, Rawson F. Managing power consumption and performance of computing systems using reinforcement learning. In: Advances in Neural Information Processing Systems. 2007.

[98] Kalyvianaki E, Charalambous T, Hand S. Self-Adaptive and self-configured cpu resource provisioning for virtualized servers using Kalman filters. In: Proc. of the 6th Int'l Conf. on Autonomic Computing (ICAC). New York: ACM, 2009. 117−126. [doi: 10.1145/1555228.1555261]

[99] Abrahao B, Almeida V, Almeida J, Zhang A, Beyer D, Safai F. Self-Adaptive SLA-driven capacity management for Internet services. In: Proc. of the 10th IEEE/IFIP NOMS. Vancouver, 2006. [doi: 10.1109/NOMS.2006.1687584]

[100] Xu J, Zhao M, Fortes J, Carpenter R, Yousif M. On the use of fuzzy modeling in virtualized data center management. In: Proc. of the ICAC. 2007. [doi: 10.1109/ICAC.2007.28]

[101] Kallahalla M, Uysal M, Swaminathan R, Lowell DE, Wray M, Christian T, Edwards N, Dalton CI, Gittler F. SoftUDC: A software-based data center for utility computing. Computer, 2004,37(11):38−46. [doi: 10.1109/MC.2004.221]

[102] Hyser C, Mckee B, Gardner R, Watson BJ. Autonomic virtual machine placement in the data center. Technical Report, HPL-2007-189, HP Labs, 2007.

[103] Steinder M, Whalley I, Carrera D, Gaweda I, Chess D. Server virtualization in autonomic management of heterogeneous workloads. In: Proc. of the IEEE Symp. on Integrated Network Management. 2007. 139−148. [doi: 10.1109/INM.2007.374778]

[104] Raghavendra R, Ranganathan P, Talwar V, Wang ZK, Zhu XY. No "power" struggles: Coordinated multi-level power management for the data center. In: ASPLOS XIII: Proc. of the 13th Int'l Conf. on Architectural Support for Programming Languages and Operating Systems. 2008. 48−59. [doi: 10.1145/1346281.1346289]

[105] Chiang RC, Huang HH. TRACON: Interference-Aware scheduling for data-intensive applications in virtualized environments. In: Proc. of the Int'l Conf. for High Performance Computing, Networking, Storage and Analysis. 2011. http://dl.acm.org/citation.cfm?id=2063447&dl=ACM&coll=DL&CFID=69602907&CFTOKEN=43862475 [doi: 10.1145/2063384.2063447]

张伟(1984－),女,河北石家庄人,博士生,CCF 学生会员,主要研究领域为云计算,虚拟化,资源管理.

宋莹(1979－),女,博士,助理研究员,CCF 会员,主要研究领域为云计算,虚拟化,资源管理.

阮利(1978－),女,博士,讲师,CCF 会员,主要研究领域为计算机系统结构,计算机系统软件,高性能计算,虚拟化,云计算.

祝明发(1945－),男,博士,教授,博士生导师,主要研究领域为计算机系统结构,计算机系统软件,高性能计算,虚拟化,云计算.

肖利民(1970－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机系统结构,计算机系统软件,高性能计算,虚拟化,云计算.