

基于动态任务量的高校云计算中心能耗算法分析

杜召彬 李廷锋

(郑州职业技术学院信息工程与大数据学院, 河南 郑州 450121)

摘要: 随着云计算技术的普及, 用户规模的不断扩大, 我国云计算中心的建设速度得到飞速发展, 伴随其巨大的能耗消耗问题引发了广泛关注, 同时云计算中心的资源利用率普遍不高。针对这一问题, 提出了动态任务量的能耗优化算法, 该算法能够根据高校云计算中心动态变化的任务量, 及时调整服务器的运行状态, 从而达到节约能耗的效果。通过仿真实验显示, 该算法能够有效降低云计算机中心的能源消耗, 降低其运营成本, 特别是在低任务量时, 节能效果较为明显。

关键词: 云计算; 计算中心; 能耗

中图分类号: TP391.9

文献标识码: A

文章编号: 1008-6609 (2023) 06-0057-05

DOI:10.15966/j.cnki.dnydx.2023.06.015

1 引言

云计算中心在给我们工作生活带来极大方便的同时, 也造成了严重的能源消耗问题, 因而云计算中心已被列为高耗能产业。据《2022年版中国数据中心市场深度研究与投资可行性报告》数据显示, 2021年中国云计算中心市场规模(含服务器)达3013亿元, 同比增长34.6%, 2021年中国云计算中心机架数量为520万架^[1]; 2021年国内云计算中心总耗电量达到2166亿千瓦时, 占社会用电量的2.6%, 相当于两座三峡水电站一年的发电量^[2]。云计算中心的能耗主要来源于服务器(IT)设备、制冷设备及供电系统三部分, 其中服务器(IT)设备和制冷设备的能源消耗占据了云计算中心能耗占比的绝大部分, 服务器(IT)设备所消耗的能量又占云计算中心总体能耗近50%的份额^[3, 4]。

PUE (Power Usage Effectiveness), 是评价云计算中心能源利用率的主要技术指标, 是云计算中心所有能源消耗与其信息负载设备能源消耗的比值^[5], 计算公式如下:

$$PUE = \frac{P_{Total}}{P_{IT}} \quad (1)$$

式中 P_{Total} 代表云计算中心总的能源消耗, P_{IT} 代表云计算中心信息设备(主要是服务器)的能源消耗, 单位为kWh。目前, 我国已建成的云计算数据中心的PUE值大多处于1.6~2.3内, 其比值较大, 也即意味着能耗消耗较

高。国家为此出台了《数据中心能效限定值及能效等级》(GB 40879—2021) PUE能耗新国标, 该国标由国家市场监督管理总局国家标准化管理委员会于2021年10月11日批准发布, 并于2022年11月1日起在全国范围内强制实施^[6]。在新国标中将数据中心能效等级分为3级, 一级能耗比不高于1.2, 二级能效比不高于1.3, 三级能效比不高于1.5。随后各地相继出台自己的PUE标准, 均高于新国标, 如上海市要求对已有云计算中心进行升级改造(PUE<1.4), 新建云计算中心PUE值不大于1.3^[6]。

针对云计算中心中各种设备的能耗进行监控测量, 发现服务器等信息设备的能耗占云计算中心总能耗的46%左右, 服务器产生高能耗的主要原因有以下四个方面^[7-9]:

(1) 服务器任务负载的波动性

为了提高用户的体验效果, 云计算中心往往会加大硬件设备的投入, 以应对流量高峰的出现, 这就会造成更多的IT信息设备在大多数时间中处于一定低利用率, 从而造成巨大的能耗。

(2) 服务器任务负载的随机性

云计算中心流量高峰会在什么时刻出现, 这是一个随机事件, 当多数服务器处于低任务状态时, 为了维护服务器的基本运行状态, 这些服务器依然消耗了60%左右的能量。

(3) 服务器任务与服务器不匹配

由于不同服务器对不同任务的响应效果不同, 如果

作者简介: 杜召彬(1975—), 男, 河南南阳人, 硕士研究生, 副教授, 研究方向为数字图像处理。

基金项目: 河南省高等学校重点科研项目, 项目编号: 23B520031, 22B520053。

服务器接收到的任务与服务器类型不匹配,会造成不同任务执行时引起额外的能耗,这种能源消耗也能产生极大的电能消耗。

(4) 云计算中心制冷设备的能耗

为了保证云计算中心服务器的正常运行,云计算中心的温度多大控制在26℃左右,制冷温度每降低1℃,就会增加制冷设备的能源消耗量,进而导致云计算中心总的能源消耗上升。

2 相关研究

通过对云计算中心服务器的使用情况进行统计分析,绝大多数的服务器使用率在10%~50%之间,这个使用效率是比较低的,所以很多文献资料的研究都集中在如何提高服务器的使用率上。文献[10]提出了基于能耗感知的虚拟机节能调度算法,构建云计算中心能耗模型,通过虚拟机的分配和迁移,以减少服务器的运行数量;文献[11]提出了基于物理服务器温度感知的虚拟机迁移模型,对服务器的CPU温度进行监测,根据CPU温度来判断该服务器的负载状态,进行虚拟机迁移操作,从而达到物理服务器负载均衡的状况;文献[12]提出了一种根据服务器的温度状态,采用贪婪随机自适应算法搜索可能产生的热点主机,进行虚拟机的迁移合并来实现降低能耗的目的;文献[13]根据服务器的CPU使用率变化,提出了基于Kalman滤波的CPU使用率预测模型,根据CPU的使用情况进行虚拟机的迁移合并;文献[14]提出利用神经网络能耗模型,通过实时动态监测CPU的使用率,来选择虚拟机的分配及迁移,实现节约能耗的目的;文献[15]通过监测CPU的使用率,设置多个监控阈值,避免服务器出现负载过载(热点主机)或过轻的状态,及时进行虚拟机迁移从而达到降低能耗的效果;文献[16]通过监测服务器任务工作状态来调整CPU的电压、运行频率,从而达到降低能耗的效果;以上7篇文献提出的节能算法基本上都是通过一定的方法触发虚拟机迁移机制,来减少服务器的运行数量,从而达到节能的目的,没有考虑频繁地触发虚拟机的迁移在一定程度上又增加了服务器自身的运行负担。文献[13]的研究中,已经明确指出,虚拟机的频繁迁移会增加大约10%的额外能源消耗。同时虚拟机迁移合并机制,可能会导致服务器出现局部热点现象(服务器负载过载),将对整个云计算中心带来极大的可靠运行风险;一旦热点现象出现,将进一步增加制冷系统的工作负载,同样增加了云计算中心的能源消耗。

针对上述观点中出现的可能频繁触发虚拟机的迁移,又引起额外能耗的问题,本文采用对云计算中心的任务量进行动态分析,及时感知任务量的变化,在保证服务器正常运行效率的同时,将到达的任务分配给已运行而未达到满负荷运行的服务器来执行,尽可能减少服

务器的运行数量;当任务量减少时,再进行一定量的服务器运行任务合并操作,不会频繁地触发大量的虚拟机合并与部署操作,最大限度保证已运行服务器达到满负荷运行状态,且运行服务器的数量最小化,进而达到降低能耗的目的,通过实际的仿真实验,该方法在减少虚拟机迁移及能耗节约上具有一定的优势。

3 高校云计算中心动态任务量算法设计

当前,许多高校都在开展实训室建设,以为学生提供充分的实践环境。以我校的信息工程学院为例,已建设有30多个一体化教室,近1200台计算机的实训室规模,全校建设的实训室计算机装机总量将近5000台。传统单机系统一般会安装还原卡、杀毒软件等防护措施来维护计算机的正常运行,但这种防护措施本身可能存在设计缺陷及漏洞,在实际使用中出現很多问题。针对高校不同专业,实训室需要安装不同的软件来满足教学、科研的需求,以及各类网络培训、实践考试的需求,高校实训室较高的使用频率,不同品牌、型号的硬件差异,伴随着高校计算机实训室规模的不断扩大,后期设备的维护对于机房技术人员来说是一项艰巨且繁重的工作。

云计算是通过互联网技术,融合了并行计算、存储及虚拟化技术,实现网络服务的一种软硬件资源平台。其优点是可以根据客户的需求,进行定制式的服务。随着云计算技术的日渐成熟,云计算技术具有部署灵活、易扩展、成本低廉的特点,高校在进行教育信息化建设的同时,加大了云计算平台的建设。一些院校还利用云计算平台实现办公自动化的全校部署,有效解决了校内信息孤岛的问题,加快了信息资源的整合及共享。云计算数据中心的桌面系统应用、存储及网络设备的全面虚拟化,能够根据用户的需要进行定制分配,可以实现统一的数据管理,减少了不同规格硬件的数量,可以降低后期的维护管理成本,提高设备的运行维护效率。

根据高校学生的作息规律性,在实际使用中,高校计算机实训室在某个时间段内,使用频率相对集中,平时又会处于低任务量的状态。低任务量运行期间,云计算中心的多数服务器又在正常运行,造成大量的能源消耗。根据高校云计算中心运行的特点,提出动态任务量计算算法,旨在根据动态任务量变化情况,及时调整运行服务器的数量,以满足云计算平台正常平稳的运行需求,即达到不影响客户体验感的同时,又能节省能耗的目的。

3.1 实时动态任务自动监控模块

云计算中心实时动态任务监控模块结构图如图1所示,包括实时任务监控模块、数据存储模块及服务器集群监控器三部分。通过实时监控模块接收到云计算平台的任务量,将根据服务器集群监控器中监控到的已

运行物理主机的状态情况，把任务分配给已运行但未满负荷运转的物理主机，待一台物理主机处于满负荷运行状态时，再给下一台物理主机分配运行任务。通过设置周期性的监控时间段，由任务监控模块及时监控任务量的变化情况，来分配任务给已运行物理主机来执行任务，完成任务的动态分配。

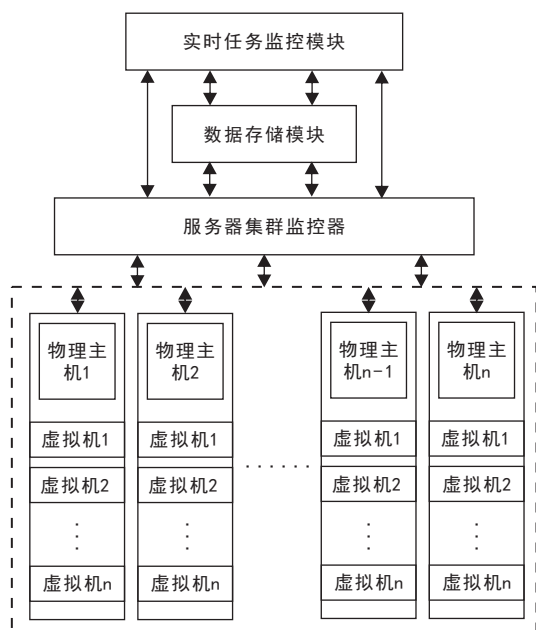


图1 实时任务监控模块架构图

本算法根据高校云计算任务量在正常的学习期间比较集中的特点，基于对任务量的动态监控分析，实时监控云计算任务量（或者通过设置工作时间段的形式）。假设进行动态调配前，云计算中心运行服务器主机的数量已达稳定状态（即已运行主机数量能够满足正常的运算需求），根据实时监控任务量的变化情况（比如设置5分钟的调度周期），计算出需要唤醒或休眠的服务器主机数量，以期云计算环境得以正常平稳运行。

动态任务量算法设计部署分为以下步骤：

- (1) 虚拟机发出启动请求；
- (2) 实时任务监控模块收到虚拟机启动请求后，与服务器集群控制器、数据存储模块交换数据信息；
- (3) 服务器集群控制器根据已运行服务器状态，来判断将该虚拟机分配给哪台物理主机运行；
- (4) 同时更新服务器集群控制器中服务器的运行状态信息；
- (5) 根据步骤2中得到的任务量计算需要唤醒或休眠的服务器主机数量；
- (6) 相应物理主机启动对应的虚拟机服务；
- (7) 唤醒或休眠一定数量的服务器主机。

3.2 动态任务量算法设计

云计算中心能耗主要由运行服务器主机产生，主机运行总能耗的构成为：

$$EE_{total} = \sum_{i=0}^1 \sum_{j=1}^N X_i Ser_j \quad (2)$$

根据公式2可以计算出云计算中心所有服务器主机的能耗和，公式中 X_i 表示当前服务器主机是否处于唤醒或休眠状态（其值1代表唤醒，0代表休眠）， Ser_j 代表第 j 台主机的能耗， EE_{total} 表示云计算中心主机的总能耗。

该算法的主要思想是根据实时监控云计算平台的任务量，及时调整（唤醒或休眠）运行服务器主机的数量，即要使得每台运行主机处于满任务量的状态，又要保证云平台的正常平稳运行，从而实现云计算平台运行主机的数量最小化，进而达到每台运行主机基本处于满任务量的运行状态，无任务的服务器主机则处于休眠状态（处于休眠状态的主机，其云计算消耗很小）。

算法：动态任务量的能耗算法

输入：taskNumber, hostNumber, timeSlot

输出：energy, hostNumber

1. Initialize taskNumber, hostNumber, timeSlot
2. while timeSlot < 5 do
3. for hostNumber = 1 to max
4. taskNumber = getTaskNumber()
5. if taskNumber > LastTaskNumger()
6. if hostNumber < max
7. runHost = getDormancyHost()
8. runHost = true
9. else
10. runHost = getActivityHost()
11. runHost = false
12. endif
13. timeSlot --
14. if timeSlot == 0
15. timeSlot = 5
16. endwhile

在算法中，在第一步对任务量taskNumber, 待机的主机量hostNumber, 自动监控时间间隔timeSlot进行初始化。在每个监控时间段内，步骤3对所能使用的主机进行标识（唤醒或休眠），在这里可以构建一个活动服务器主机的数组和休眠服务器主机的数组，根据任务量的变化，及时调整主机活动标识，并执行主机激活或休眠的动作。

3.3 仿真实验

结合现有云计算中心的服务器实际情况，虚拟出一

个拥有40台物理服务器的云计算中心，由这40台物理服务器模拟出1200台虚拟主机进行性能测试，所有主机的软硬件运行环境相同，用户发出的任务请求随机产生（即1200台虚拟主机随机开机运行）。假设物理服务器在休眠状态下的能耗为30瓦，空闲状态下的能耗为200瓦，最大运行载荷状态下功耗为800瓦，在这里不考虑数据中心制冷等设备的能耗问题。

根据上述假设情况，从1200台虚拟主机中随机选择100台、300台、500台、800台、1100台和1200台虚拟主机启动运行，让启动运行的每台虚拟主机执行300、600、1000次指令数（MIPS），并进行128MB的随机存取存储操作。根据虚拟主机休眠状态、半任务工作状态及满任务工作状态下的能耗消耗情况，每个实验模拟进行10次，求出平均值。

根据上面分配的任务量，比较不同的算法效果：在任务到达时，执行随机算法将虚拟主机启动任务随机分配给服务器运行的能耗消耗；实现文献[10]提出尽可能使得服务器的工作效率达到其工作的下限，从而触发虚拟机迁移策略来实现节能的目的；采用本算法，当任务产生时，将任务优先分配给已运行而未达最大载荷的服务器主机来运行该任务，当所有已运行服务器主机都处于最大载荷状态时，再唤醒下一台处于休眠的服务器主机来运行新的虚拟主机任务。在不同任务量下的服务器主机运行能耗对比如图2所示。

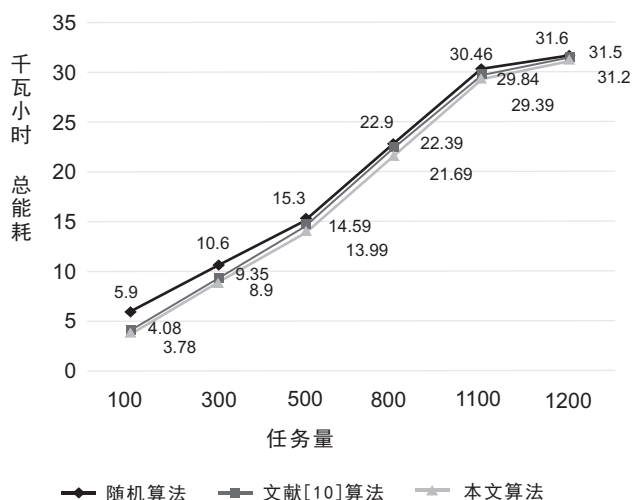


图2 随机分配算法、文献[10]算法与本文算法能耗对比

3.4 仿真结果分析

根据图2的对比结果，任务量随机算法、文献[10]算法和本文算法三种算法的对比效果图，可以得出结论：在任务量较少时，使用本算法的能耗比随机算法和文献[10]算法都要节约能耗，当虚拟主机运行数为

1100和1200台时，三种算法的能耗基本一致，这是因为在满任务量（即虚拟主机接近全部运行）的时候，全部服务器主机基本都处于满负荷运行状态，所以能耗消耗基本相同。但当虚拟主机运行数量较少时，虚拟主机运行数为100台、300台、500台、800台时，本算法的能耗比随机算法分别降低了35.9%、16.04%、8.5%和6.59%，比文献[10]算法能耗下降了7.35%、4.81%、4.11%、3.13%。本算法优于文献[10]算法能耗，经过分析文献[10]算法，当服务器工作效率达到其工作的下限，会触发虚拟机的迁移，通过迁移虽然可以让部分服务器进入休眠状态以达到节能的效果，但频繁的迁移同样会增加额外的能耗。本算法采用对任务量进行动态分析的方法，根据任务量将运行虚拟机尽可能分配给已运行服务器（未达满负荷运行的服务器）来运行，当一台服务器的运行任务达到满任务量的时候，再唤醒一台休眠服务器来运行新的任务，以减少运行服务器的数量，相对文献[10]算法来说会减少虚拟机的迁移次数，所以比文献[10]算法节约能耗。当任务量较少时，本算法的能耗状态明显优于随机算法、文献[10]算法的能耗。

4 结语

本文提出了基于动态任务量的云计算中心能耗算法，该算法综合考虑了云计算中心执行任务量的随机性。当任务量较少时，大量服务器处于低任务状态时，依然会造成较多的能源消耗，本算法根据任务量的动态变化，建立算法模型，可以实现虚拟机部署的动态优化选择，保证尽可能少的服务器主机处于满负荷运行状态，其余服务器主机处于休眠状态来降低能源消耗。同时在一定程度上可以减少虚拟机的迁移，根据模拟实验结果表明，在保证服务品质的同时，该算法能够有效降低云计算中心的能源消耗。本算法在建立能耗模型时，仅仅考虑了物理服务器的CPU和内存能耗，而没有考虑云计算中心其他IT设备（如网络设备等）的能耗。如若综合考虑云计算中心能耗模型，也应将这些因素考虑进去，这将是更深层级的研究内容及方向。

参考文献：

- [1]立鼎产业研究网.2017-2021年我国数据中心市场规模、数据中心机架数及能耗增长分析[EB/OL].(2022-11-23).<http://www.leadingir.com/datacenter/view/8368.html>.
- [2]界面新闻.大模型浪潮推动之下,数据中心的“液冷时代”将给谁带来机会?[EB/OL].(2023-06-13).<https://www.jiemian.com/article/9569312.html>.
- [3]周细佳.云数据中心能耗-性能均衡方案研究[D].长沙:湖南大学,2020.
- [4]吴雷,郭聪,田振武.双碳政策下数据中心节能创新技

术研究[J].产业科技创,2022,4(6):82-85.

[5]苏志.基于PUE分析的某数据中心制冷系统优化研究[J].制冷与空调,2021,35(2):162-168.

[6]搜狐新闻.新国标《数据中心能效限定值及能效等级》11月起强制实施[EB/OL].(2022-11-04).http://society.sohu.com/a/602538995_121469436.

[7]宋达.云数据中心中能耗感知的虚拟机迁移研究[D].南京:南京邮电大学,2020.

[8]王子懿,唐碧华,陈日凡,等.一种基于神经网络的数据中心机房空调节能控制方法[J].广东通信,2022,42(6):76-79.

[9]张诗良.数据中心基础设施系统中的能耗管理分析[J].集成电路应用,2022,39(10):182-183.

[10]郑瑛.云计算数据中心节能调度算法改进研究[J].西南大学学报(自然科学版),2019,41(12):135-142.

[11]周震,袁正道,李俊峰.云数据中心基于温度感知的虚拟机迁移模型[J].计算机应用与软件,2021,38(11):70-76.

[12]张静.云数据中心能量与热量感知的虚拟机合并与部署[J].计算机应用与软件,2021,38(1),63-69.

[13]何丽,汤莉.基于Kalman滤波的云数据中心能耗和性能优化[J].计算机工程与科学,2018,40(7):1165-1172.

[14]智伟威,周新星.神经网络在数据中心能耗模型研究中的应用[J].计算机仿真,2020,37(10):273-277.

[15]王伟.一种低能耗的云计算虚拟机部署方案[J].控制工程,2018,25(7):1291-1296.

[16]黄海莹.面向移动边缘计算的CNN推理的能耗优化调度方法[D].深圳:深圳大学,2020.

Energy Consumption Algorithm Analysis of University Cloud Computing Center Based on Dynamic Task Volume

DU Zhao-bin LI Ting-feng

(Zhengzhou Vocational and Technical College, Zhengzhou 450121, Henan)

Abstract: With the popularization of cloud computing technology and the continuous expansion of user scale, the construction speed of cloud computing centers in China has been rapidly developed. At the same time, its huge energy consumption has aroused widespread concern. At the same time, the resource utilization rate of computing centers is generally low. To solve this problem, this paper proposes an energy consumption optimization algorithm for dynamic task volume, which can adjust the running state of the server in time according to the dynamic task volume of the computing center, thus achieving the effect of energy saving. Simulation experiments show that this algorithm can effectively reduce the energy consumption of cloud computing centers and reduce their operating costs, especially when the task load is low, the energy-saving effect is relatively obvious.

Keywords: cloud computing; compute center; energy consumption